통계적 예측모형

서울시립대 통계학과 이용희

2025-04-25

Table of contents

Preface **1**

1. 선형 회귀모형의 소개 **3**
1.1. 예제-단순 회귀모형 . . . 3
1.2. 선형 회귀모형 . . . 6
1.3. 최소제곱법 . . . 7
1.4. 결정계수 . . . 15
1.5. 중회귀 모형 . . . 18
1.5.1. 최소제곱추정 . . . 20
1.6. 예제-중회귀모형 . . . 21
1.7. 최소제곱 추정량의 분포 . . . 22
1.8. 가우스-마코브 정리 . . . 23
1.9. 최대가능도 추정 . . . 25

2. 선형회귀에서의 추론 **29**
2.1. 제곱합의 분포 . . . 29
2.1.1. 잔차제곱합의 분포 . . . 29
2.1.2. 회귀제곱합의 분포 . . . 30
2.1.3. 잔차제곱합과 회귀제곱합의 독립 . . . 32
2.1.4. 총제곱합의 분포 . . . 32
2.2. 모분산의 추정 . . . 32
2.3. 최소제곱 추정량의 성질 . . . 33
2.4. 모형의 적합도 검정과 분산분석 . . . 33

3. 모형의 비교 **35**
3.1. 직교하는 설명 변수 . . . 35
3.2. 설명변수의 추가 . . . 39
3.3. 부분 F-검정과 가능도비 검정 . . . 45
3.3.1. 부분 F-검정 통계량의 분포 . . . 48

4. 모형의 진단 **50**
4.1. 등분산성 가정의 위반 . . . 50
4.1.1. 가중 최소제곱법 . . . 51
4.2. 변수변환 . . . 53
4.2.1. 지수모형과 멱함수: 로그변환 . . . 53
4.2.2. 쌍곡선과 역변환 . . . 55

4.2.3. Box-Cox 변환 . . . 57
4.3. 다중공선성 . . . 58
4.3.1. 독립변수간의 상관계수 . . . 58
4.3.2. $\boldsymbol{X}^t\boldsymbol{X}$ 의 고유값(Eigenvalues) . . . 58
4.3.3. 조건지수 (condition number) . . . 60
4.3.4. 분산팽창계수 (Variance Inflation Factor ;VIF) . . . 60

5. 관측값에 대한 진단 61
5.1. 서론 . . . 61
5.2. 이상점의 유형 . . . 61
5.3. 지렛값 . . . 63
5.4. 내 표준화 잔차 . . . 64
5.5. 관측값의 영향: 계수 추정 . . . 64
5.6. 외 표준화 잔차와 PRESS 잔차 . . . 65
5.7. 관측값의 영향: 분산 추정 . . . 66
5.8. 영향력의 측도 . . . 67

6. 모형의 선택 69
6.1. 서론 . . . 69
6.2. 모형선택의 측도 . . . 70
6.2.1. 결정계수 . . . 70
6.2.2. 결정계수의 수정 . . . 71
6.2.3. Mallow's C_p . . . 71
6.2.4. PRESS . . . 72
6.3. AIC 와 BIC . . . 73
6.4. 변수 선택법 . . . 74

7. 분산분석 모형 75
7.1. 서론 . . . 75
7.2. 일원배치법 . . . 75
7.2.1. set-to-zero condition . . . 76
7.2.2. sum-to-zero condition . . . 76
7.3. 선형모형과 제약 조건 . . . 77
7.3.1. Set-to-zero 조건에서의 모형과 최소제곱 추정량 . . . 78
7.3.2. Sum-to-zero 조건에서의 모형과 최소제곱 추정량 . . . 79
7.3.3. 예제 7.3 . . . 80
7.4. 불완전 계수행렬에서의 추정 . . . 86
7.4.1. 모수의 재조정 (reparameterization) . . . 86
7.4.2. 부가 조건의 이용 . . . 86
7.4.3. 일반화 역행렬의 이용 . . . 87
7.5. 추정 가능한 함수 . . . 87
7.5.1. 일원배치법에 추정가능한 모수 . . . 87
7.5.2. 추정가능한 모수의 함수 . . . 88

7.5.3. 예제 . . . 89
7.6. 가설 검정 . . . 91
7.6.1. 변동의 분해 . . . 91
7.6.2. 제곱합과 F-통계량 . . . 92
7.7. 다중비교 . . . 94
7.7.1. 일원배치에서 평균의 비교 . . . 94
7.7.2. 두 개 이상의 가설 . . . 94
7.7.3. 실험단위 오류 . . . 95
7.7.4. 예제: 2개의 가설을 가진 임상실험 . . . 96
7.7.5. 다중비교 . . . 96

8. 공분산분석 **99**
8.1. 공분산분석 개요 . . . 99
8.2. 공분산분석의 모형 . . . 101
8.3. 모수의 추정 . . . 101
8.4. 가설검정 . . . 102
8.5. 예제 1: 영어교육 방법 . . . 103
8.6. 예제 2: 저혈당 실험 . . . 106

9. 최소제곱 추정량의 계산 **111**
9.1. 촐레스키 분해의 이용 . . . 111
9.2. QR 분해의 이용 . . . 112
9.3. SVD 을 이용 . . . 114

10. 회귀모형의 확장 **116**
10.1. 개요 . . . 116
10.2. 자료 만들기 . . . 116
10.3. 로버스트 회귀 . . . 117
10.3.1. M-추정량 . . . 118
10.3.2. 가중치 함수 . . . 119
10.3.3. 기타 방법 . . . 120
10.3.4. 예제 . . . 121
10.4. 비선형 회귀분석 . . . 127
10.4.1. 예제: 철도 여객 운송 . . . 127
10.4.2. 선형모형의 한계 . . . 129
10.4.3. 비선형 모형 . . . 131
10.4.4. 비선형모형의 적합 . . . 133
10.4.5. 비선형 회귀의 추론 . . . 135
10.4.6. 유의할 사항 . . . 135
10.5. 비모수 회귀모형 . . . 136
10.5.1. 편이-분산의 관계 . . . 137
10.5.2. Regressogram . . . 137
10.5.3. 커널추정 . . . 140

10.6. 스플라인 회귀모형 145
10.6.1. 함수의 기저 145
10.6.2. 직선 스플라인 147
10.6.3. 스플라인 회귀 150
10.6.4. 평활 스플라인 155

11. 일반화 선형모형 **161**
11.1. 일반화 선형모형 161
11.1.1. 지수군 분포와 일반화 선형모형 162
11.2. 일반화 선형모형의 가능도함수 163
11.3. 최대가능도추정 164
11.3.1. 가능도 방정식의 유도: 첫 번째 방법 166
11.3.2. 가능도 방정식의 유도: 두 번째 방법 166
11.4. 최대가능도추정량의 계산 169
11.5. 편차 171
11.6. 이항변수에 대한 회귀모형 175
11.6.1. 단순 로지스틱 회귀모형 175
11.6.2. 오즈비 176
11.6.3. 로지스틱 회귀식 추정 및 예측 178
11.7. 발생 횟수에 대한 모형 179
11.7.1. 포아송 회귀모형 179
11.7.2. 비율 모형 181
11.7.3. 음이항 분포 182
11.7.4. 영과잉모형 183

References **185**

Appendices **186**

A. 행렬의 기초 **186**
A.1. 벡터와 행렬 186
A.2. 두 행렬의 덧셈 186
A.3. 스칼라곱 187
A.4. 벡터와 행렬의 곱셈 187
A.4.1. 행과 열의 내적 187
A.4.2. 열벡터의 선형조합 188
A.5. 행렬의 전치 188
A.6. 행렬의 곱셈 188
A.7. 단위벡터와 항등행렬 190
A.8. 대각합 191
A.9. 행렬식 192
A.10.직교행렬 192
A.11.우드베리 공식 193

B. 벡터공간 **194**
B.1. 벡터공간의 정의와 의미 . . . 194
B.2. 벡터의 선형독립 . . . 196
B.3. 역행렬 . . . 197
B.4. 행렬의 계수 . . . 197
B.5. 생성집합과 기저 . . . 198
B.6. 벡터공간의 차원 . . . 199
B.7. 행렬의 열공간과 영공간 . . . 199
B.8. 두 벡터의 사영 . . . 200
B.9. 최소제곱법과 사영 . . . 201

C. 고유값과 고유벡터 **205**
C.1. 특성다항식 . . . 205
C.2. 고유값과 고유벡터 . . . 205
C.2.1. 정의 . . . 205
C.2.2. 계산 . . . 206
C.2.3. 중복도와 고유공간 . . . 207
C.3. 대칭행렬의 대각화 . . . 211

D. 행렬의 분해 **213**
D.0.1. Gram–Schmidt 방법 . . . 213
D.0.2. LU 분해 . . . 215
D.0.3. QR 분해 . . . 216
D.0.4. SVD 분해 . . . 219

E. 벡터 미분 **222**
E.1. 용어 . . . 222
E.2. 벡터 미분의 표기법 . . . 222
E.3. 함성함수의 미분법 . . . 224
E.4. 두 벡터 내적의 미분 . . . 225
E.4.1. 상수벡터와 변수벡터의 내적 . . . 225
E.4.2. 상수벡터와 함수벡터의 내적 . . . 225
E.4.3. 함수벡터와 함수벡터의 내적 . . . 225
E.5. 벡터 미분의 응용 . . . 226
E.5.1. 선형사상의 미분상 . . . 226

F. 다변량 확률변수의 성질 **228**
F.1. 일변량분포 . . . 228
F.2. 확률벡터와 분포 . . . 228
F.3. 다변량 정규분포 . . . 230
F.4. 표준정규분포로의 변환 . . . 232
F.5. 예제 . . . 233

G. 이차형식과 제곱합의 분포 **234**
G.1. 이차형식 . . . 234
G.2. 대칭행렬의 대각화 . . . 235
G.3. 멱등행렬 . . . 236
G.4. 이차형식의 분포 . . . 236
G.4.1. 카이제곱 분포 . . . 236
G.4.2. 비중심 카이제곱 분포 . . . 237
G.4.3. 이차형식의 분포 . . . 237
G.4.4. 이차형식의 독립 . . . 238
G.4.5. 이차형식의 차이 . . . 239
G.5. 코크란의 정리 . . . 239

H. 지수군 분포 **241**
H.1. 가능도함수와 그 성질 . . . 241
H.2. 독립표본 . . . 243
H.3. 지수군 분포 . . . 243
H.3.1. 평균, 분산과 기본형 모수의 관계 . . . 244
H.3.2. 지수군 분포의 예제 . . . 245
H.4. 최대가능도추정법 . . . 247

I. 모형선택의 정보 기준 **248**
I.1. Kullback-Leibler 정보 . . . 248
I.2. 가능도 함수 . . . 249
I.3. AIC . . . 252
I.4. BIC . . . 254

J. R-실습: 중회귀 모형 적합 **256**
J.1. 예제 3.3 자료 . . . 256
J.1.1. 산점도 행렬 . . . 257
J.2. 중회귀 모형의 적합 . . . 257
J.2.1. 회귀계수의 추정과 결정계수 . . . 261
J.2.2. 예측값 . . . 263
J.3. 제곱합의 종류 . . . 264
J.3.1. 순차제곱합 . . . 265
J.4. 부분 F 검정 . . . 268
J.5. 모형의 진단 . . . 269

K. R-실습: 중회귀 모형 진단 **271**
K.1. 변수변환 . . . 271
K.1.1. 예제 4.8 . . . 271
K.2. Box-Cox 변환 . . . 274
K.2.1. 예제 4.10 . . . 274
K.2.2. 예제 4.11 . . . 275

K.3. 다중공선성 280
K.3.1. 고유값과 고유벡터에 대한 이론 280
K.3.2. 고유값과 고유벡터에 대한 예제: 두 개의 독립변수 282
K.3.3. 예제 4.13 285
K.3.4. 예제 4.14 285

L. R-실습: 관측값에 대한 진단 **290**
L.1. 교과서 예제 3.8 290
L.1.1. 잔차그림 290
L.1.2. 잔차 293
L.1.3. 영향점 측도 294
L.2. 교과서 연습문제 5.14 297
L.2.1. 회귀모형 적합 298
L.2.2. 잔차그림 299
L.2.3. 잔차와 지렛값 분석 301
L.2.4. 영향점 측도 305
L.2.5. 결론 310

M. R-실습: 모형의 선택 **313**
M.1. AIC 와 BIC 계산 313
M.2. 예제 6.1 314
M.2.1. 변수선택의 통계량 314
M.2.2. 이상치를 제거한 경우 319
M.3. 변수선택 방법 322
M.3.1. 단계별 선택의 적용 323
M.3.2. 함수 `step()` 327
M.3.3. 패키지 `olsrr` 의 이용 330
M.4. 연습문제 6.10 334
M.4.1. 완전모형 334
M.4.2. 모든 가능한 회귀 335
M.4.3. 전진선택법 336
M.4.4. 후방제거 338
M.4.5. 단계별 선택 339
M.5. 연습문제 6.12 348

N. R-실습: 분산분석 모형 **353**
N.1. 일원배치: 예제와 R 프로그램 353
N.1.1. 예제: 실험계획 353
N.1.2. 자료의 생성 353
N.1.3. 선형모형의 적합(set-to-zero) 354
N.1.4. 선형모형의 적합 (sum-to-zero) 356
N.1.5. 분산분석 358
N.1.6. 다중비교 예제 359

N.2. 예제 7.1 - 일원배치 362
N.3. 예제 7.2 - 교호작용 364
N.4. 예제 7.3 - 일원배치 2 366

List of Figures

1.1. 오차제곱합의 함수 형태 8
1.2. 표준화 시 오차제곱합의 함수 형태 10

4.1. 오차항의 등분산성이 위반된 경우 53
4.2. 지수모형과 로그변환 54
4.3. 역지수모형과 로그/역변환 55
4.4. 쌍곡선과 역변환 56
4.5. Additive error model 과 로그변환 57
4.6. 회귀분석에서의 다중공선성의 정도(Fox 2008, p310) 59

5.1. 여러가지 종류의 이상점과 그 영향 62

8.1. 실험과 관측연구에서의 전체 변동의 설명 99
8.2. 공분산분석에 의한 전체 변동의 설명 100
8.3. 공분산분석의 효과 100

10.1. M-추정량의 이용되는 함수 120
10.2. 10일전 예약 인원수와 실제 탑승객수의 관계 122
10.3. 탑승객수 추정 모형 123
10.4. 10일전 예약 인원수와 실제 탑승객수의 관계 - 이상점이 포함된 경우 124
10.5. 탑승객수 추정 모형 - 이상점이 포함된 경우 -단순회귀 126
10.6. 탑승객수 추정 모형 - 이상점이 포함된 경우 - MM 추정 127
10.7. 예약 20일 전에서 출발일까지의 열차 예약 현황 128
10.8. 로지스틱 함수와 모수 133

B.1. 벡터의 사영 201
B.2. 최소제곱법을 설명한 그림 202
B.3. 포함관계에 있는 열공가에 대한 사영의 관계 204

D.1. Gram–Schmidt 방법(출처:Introduction to Applied Linear Algebra by Boyd and Vandenberghe, 2019) 215
D.2. LU 분해 216
D.3. SVD 분해의 기하학적 의미 221

List of Tables

2.1. 적합도 검정을 위한 분산분석표 34

8.1. 공분산분석 모형의 분산분석표 103

10.1. 미캘리스-멘텐(Michaelis-Menten) 비선형 함수의 모수와 의미 132
10.2. 3개의 모수로 이루어진 로지스틱 비선형 함수의 모수와 의미 132

11.1. 나이와 만성심장질환의 관계 177
11.2. 나이와 만성심장질환의 관계 177

N.2. LSD, Bonferoni, HSD 방법의 p-값 비교 361

Preface

이 책은 통계적 예측모형에 대한 교재이며 일반 선형모형을 포함하여 예측에 사용되는 기본적인 통계 모형에 대한 이론을 최대 가능도 추정법의 관점에서 설명합니다. 또한 실제 예제를 통한 실습, 모형을 적합하는 계산방법과 연관된 행렬이론에 대하여 다루고자 합니다.

> **Note**
>
> 이 책에서 사용된 기호, 표기법, 프로그램의 규칙과 쓰임은 다음과 같습니다.
>
> - 스칼라(scalar)와 일변량 확률변수는 일반적으로 보통 글씨체의 소문자로 표기한다. 특별한 이유가 있는 경우 대문자로 표시할 것이다.
> - 벡터, 행렬, 다변량 확률벡터는 굵은 글씨체로 표기한다.
> - 통계 프로그램은 R을 이용하였다. 각 예제에 사용된 R 프로그램은 코드 상자를 열면 나타난다.

강의의 부교재는 강근석 and 유형조 (2016) 을 사용한다.

이 교과서에서 이용하는 R 패키지는 다음과 같다.

```
library(here)           # file pathways
library(tidyverse)      # data management, summary, and visualization
library(MASS)
library(knitr)
library(kableExtra)

library(agricolae)
library(emmeans)
library(car)

library(olsrr)
library(leaps)

library(plotly)
library(plot3D)

library(mvtnorm)
library(nlstools)
```

```
library(faraway)


# 라이브러리 `ElemStatLearn`는 더 이상 R 라이브러리에 없으므로 다음 사이트에서 다운로드 받아서 직접 화일로 설치해야 한다**
# 다운로드 사이트: https://cran.r-project.org/src/contrib/Archive/ElemStatLearn/
# install.packages("~/Downloads/ElemStatLearn_2015.6.26.2.tar.gz", repos = NULL, type = "sourc

library(ElemStatLearn)

# 참고문헌 "R을 활용한 선형회귀분석. 1st ed. 교우사" 에 나오는 데이터는 regbook 이라는
# 패키지를 설치해야 합니다. 다음 사이트에 안내된 방법으로 설치합니다.
# https://github.com/regbook/regbook

library(regbook)

# 아래 3 문장은 한글을 포함한 ggplot 그림이 포함된 HTML, PDF로 만드는 경우 사용
library(showtext)
font_add_google("Nanum Pen Script", "gl")
showtext_auto()

#강의 부교재 자료를 포함한 패키지 설치
install.packages("remotes")
remotes::install_github("regbook/regbook")
library(regbook)
```

1. 선형 회귀모형의 소개

1.1. 예제-단순 회귀모형

Example 1.1 (자동차의 제동거리). 자동차가 달리는 속도(speed,단위는 mph; mile per hour)와 제동거리(dist, 단위는 ft;feet)의 관계를 알아보기 위하여 50대의 자동차로 실험한 결과의 자료 cars 는 다음과 같다(처음 10개의 자료만 보여준다). 자료는 R 의 data.frame 형식으로 저장되어 있다.

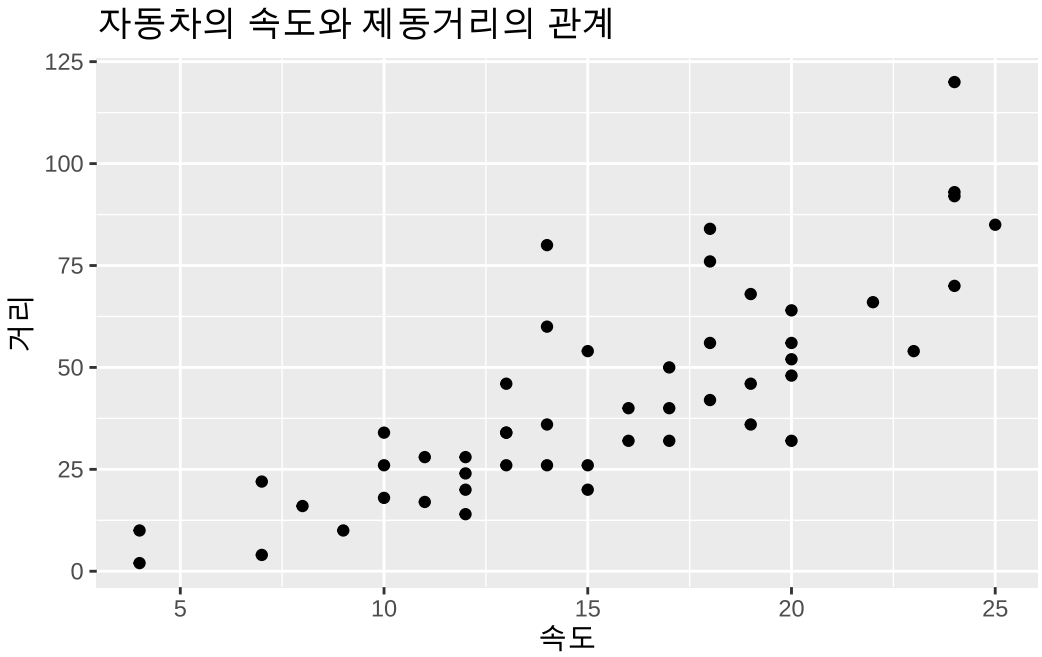
아래 자료를 보면 실험에서 2대의 자동차는 7 mph 로 달리다가 브레이크를 밟고 정지하는 경우 각각 4, 22 feet 의 제동거리가 필요한 것으로 나타났다. 또한 3대의 는 10 mph 로 달리다가 각각 18, 26, 34 feet 의 제동거리가 필요한 것으로 나타났다.

```
cars %>% head(n=10)
```

```
   speed dist
1      4    2
2      4   10
3      7    4
4      7   22
5      8   16
6      9   10
7     10   18
8     10   26
9     10   34
10    11   17
```

자동차의 속도와 제동거리에 대한 산포도는 아래와 같다.

```
ggplot(cars, aes(x=speed, y=dist)) + geom_point() + labs(x = "속도", y = "거리") +
  labs(title="자동차의 속도와 제동거리의 관계")
```



위와 같은 자료를 이용하여 자동차의 속도가 주어졌을 경우 제동거리를 예측하려고 한다면 어떤 방법을 사용해야 할까?

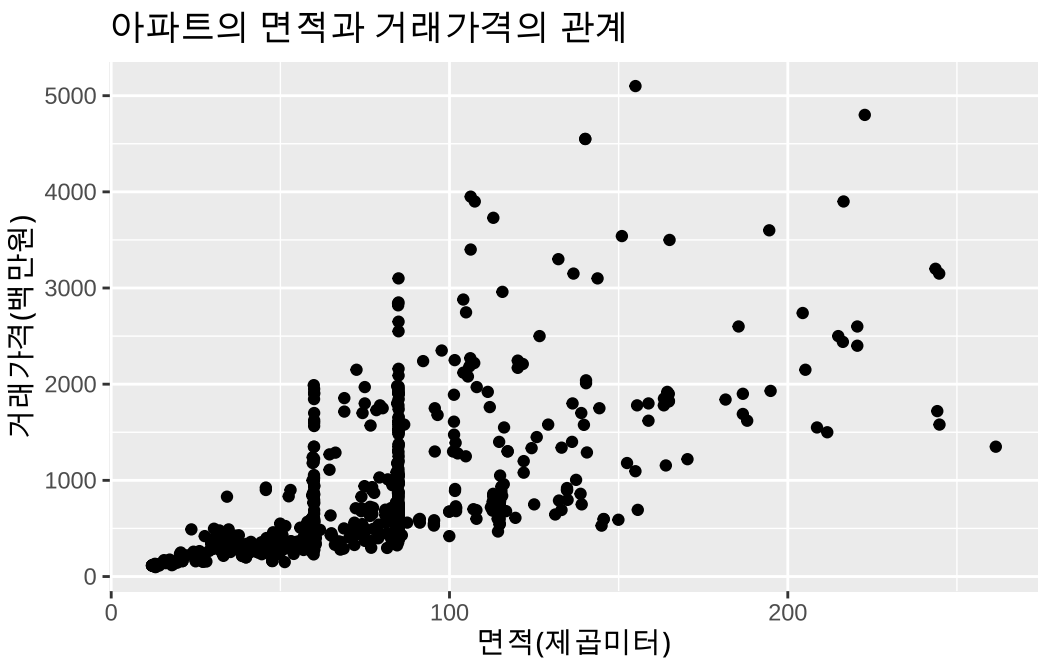
Example 1.2 (아파트 판매가격)**.** 다음 살펴볼 자료는 2019년 거래된 서울 아파트의 실거래 데이터 중 4개의 구(동대문구, 서초구, 관악구, 노원구)에서 거래된 아파트 중 1000개의 아파트를 임의로 추출한 자료이다.

```
apart_2019 <- read.csv(here("data", "seoul_apartment_2019_sample.csv"), header = T)
head(apart_2019,10)
```

```
       gu year  area price
1   관악구 1974 65.09   450
2   관악구 1978 56.86   276
3   관악구 1982 91.24   599
4   관악구 1982 91.24   560
5   관악구 1984 60.27   425
6   관악구 1984 60.27   420
7   관악구 1985 38.92   217
8   관악구 1988 61.74   485
9   관악구 1991 71.90   328
10  관악구 1991 84.44   438
```

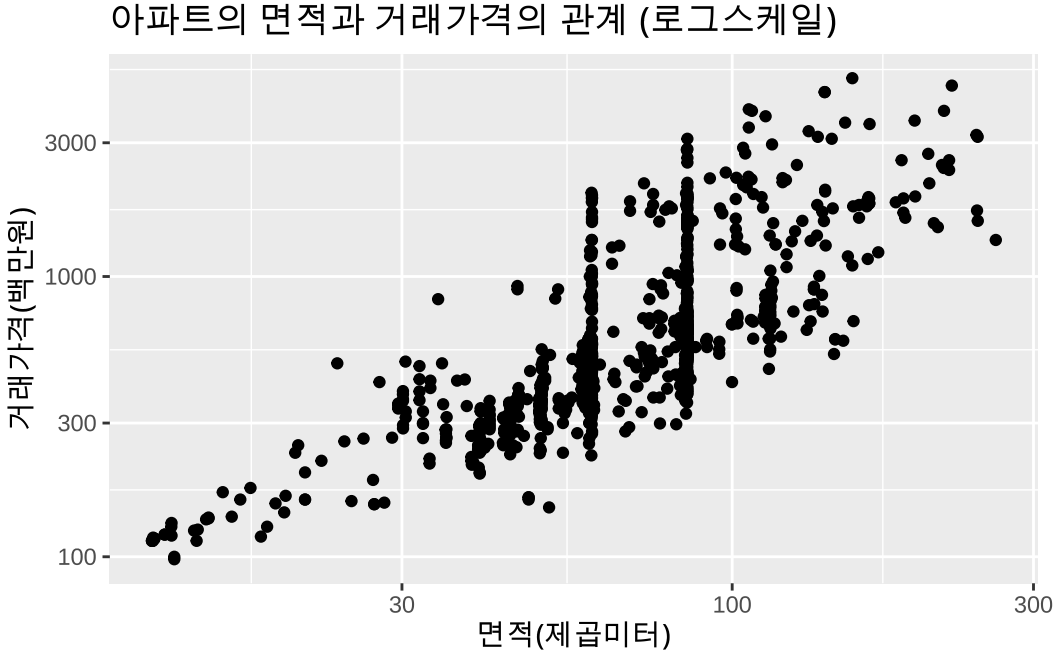
아파트의 면적(area;제곱미터)에 따른 거래가격(price;백만원)의 변화는 다음 그림과 같다.

```
ggplot(apart_2019, aes(x=area, y=price)) + geom_point() + labs(x = "면적(제곱미터)", y = "거래가격(
   labs(title = "아파트의 면적과 거래가격의 관계")
```



만약 아파트의 면적(x)과 거래가격(y) 대신 각각의 로그값(log(x), log(y))을 시용하면 다음과 같은 산포도가 나타난다.

```
# log scale for y
ggplot(apart_2019, aes(x=area, y=price)) + geom_point() + labs(x = "면적(제곱미터)", y = "거래가격(
  scale_y_log10() +
  scale_x_log10() +
  labs(title = "아파트의 면적과 거래가격의 관계 (로그스케일)")
```



1.2. 선형 회귀모형

회귀모형(regression model)는 변수들의 함수적 관계를 분석하는 통계적 방법이다. 일반적으로 한 개 또는 여러 개의 설명변수들(explanatory variables, x)이 관심있는 반응변수(response variable, y)에 어떤 형태로 영향을 미치는지에 파악하고 설명변수와 반응변수의 함수 관계를 통계적으로 추론하는 것이 회귀분석의 목적이다.

위에서 살펴본 두 예제에서 자동차의 속도(x)가 증가하면 제동거리(y)가 증가하는 경향이 있다는 것을 알 수 있으며, 아파트의 면적(x)과 거래가격(y)도 유사한 관계임을 알 수 있다.

이러한 두 변수의 관계를 다음과 같은 반응변수 y 와 설명변수의 선형 예측식(linear predictor)으로 나타내어 보자. 이러한 관계는 반응변수의 값의 변화를 근사적으로 설명변수의 선형식으로 예측할 수 있다는 의미이다.

$$y \approx \beta_0 + \beta_1 x$$

위와 같은 근사적인 관계를 더 구체화하여 다음과 같이 반응변수의 평균값이 설명변수의 선형식으로 나타나는 것을 가정할 수 있으며 이를 선형 회귀모형(linear regression model)이라고 한다.이

$$E(y|x) = \beta_0 + \beta_1 x \tag{1.1}$$

1.1 은 반응변수 y의 평균이 설명변수 x 의 선형 예측식으로 나타나는 관계를 가정한 것이며 절편 β_0 와 기울기 β_1 는 회귀계수(regression coefficient)라고 부르는 모수(parameter)로서 추정해야 한다.

특별히 하나의 설명변수를 사용하는 회귀 모형을 단순선형 회귀모형(simple linear regression model)이라고 한다.

위에서 본 두 예제와 같이 n 개의 자료 $(x_1, y_1), (x_2, y_2), .., (x_n, y_n)$을 독립적으로 추출하였다면 자료의 생성과정을 다음과 같은 단순선형 회귀모형으로 나타낼 수 있다. 반응변수 y_i는 설명변수 x_i의 선형함수로 표현된 선형 예측식 1.1 과 임의의 오차항 (random error) e_i 의 합으로 나타내어진다고 가정하자.

$$y_i = \beta_0 + \beta_1 x_i + e_i, \quad i = 1, 2, \dots, n \tag{1.2}$$

여기서 오차항 e_i는 평균이 0이고 분산이 σ^2 인 임의의 확률분포를 따르며 서로 독립이라고 가정한다.

$$E(e_i) = 0, \quad V(e_i) = \sigma^2 \quad i = 1, 2, \dots, n$$

오차항의 분산 σ^2도 추정해야할 모수(parameter)이다.

1.3. 최소제곱법

앞에서 언급한 것과 같이 선형회귀모형 1.2 에서 모수 β_0와 β_1를 회귀계수라고 하며 자료를 이용하여 추정해야 한다. n개의 자료를 이용하여 회귀계수 β_0와 β_1를 추정하려고 할 때 사용할 수 있는 방법들 중에서 가장 쉽고 유용한 방법은 최소제곱법(least square method)이다.

회귀모형 1.2 에서 β_0와 β_1의 값이 주어졌다면 설명변수 x_i 에서 반응변수의 관측값 y_i에 가장 합리적인 예측값은 무었일까? 가장 합리적인 예측값은 주어진 x_i에서 반응변수의 평균값인 $E(y_i|x_i) = \beta_0 + \beta_1 x_i$이다. 여기서 실제 관측하여 얻어진 값 y_i와 예측값 $\beta_0 + \beta_1 x_i$ 사이에는 오차에 의해서 차이가 발생할 수 있다. 그 차이를 잔차(residual)라고 하며 r_i 라고 표기한다.

$$r_i = y_i - E(y_i|x_i) = y_i - (\beta_0 + \beta_1 x_i)$$

잔차는 위에 식에서 알 수 있듯이 관측값과 회귀식을 통한 예측값의 차이를 나타낸 것이다. 그러면 자료를 가장 잘 설명할 수 있는 회귀직선을 얻기 위해서는 잔차 r_i를 가장 작게하는 회귀모형을 세워야 한다. 잔차들을 최소로 하는 방법들 중 하나인 최소제곱법은 잔차들의 제곱합을 최소로 하는 회귀계수 β_0와 β_1를 추정하는 방법이다. 잔차들의 제곱합은 다음과 같이 표현된다.

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} r_i^2 = \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2 \tag{1.3}$$

> **Note**
>
> 1.3 를 잔차제곱합(residusl sum of square)이라고 부른다. 일반적으로 회귀계수의 값이 특정지어져서 실제로 잔차를 계산할 수 있는 경우 잔차제곱합이라고 부른다. 뒤에 분산분석에서는 잔차제곱합을 SSE(sum of square error)라고 부른다.
> 잔차제곱합을 최소로 하는 회귀계수의 값을 찾는 최적화의 목표로 잔차제곱합이 제시될 때 이를 오차제곱합(error sum of square)이라고 부른다.

위의 오차제곱합 $S(\beta_0, \beta_1)$ 을 최소화하는 β_0와 β_1의 값을 구하는 방법은 오차제곱합이 β_0와 β_1의 미분 가능한 2차 함수이고 아래로 볼록한 함수(convex function)임을 이용한다.

```
gridnum <- 60
sizing <- 5
extrascale <- 10
extrascale2 <- 0.7
b0 <- seq(-17.6-sizing*extrascale,  -17.6+sizing*extrascale, length=gridnum )
b1 <- seq(4-sizing*extrascale2, 4+sizing*extrascale2, length=gridnum )

SSE <- matrix(0, gridnum, gridnum )
for (i in 1:gridnum ) {
  for (j in 1:gridnum ){
    r <- cars$dist- b0[i] -b1[j]*cars$speed
```

```
    SSE[i,j]  <- (sum(r^2))/1000
  }
}

persp3D(b0, b1, SSE, theta =10, phi = 20, expand = 1)

## Interactive 3d graph
#fig <- plot_ly(z = ~SSE)
#fig <- fig %>% add_surface()
#fig
```

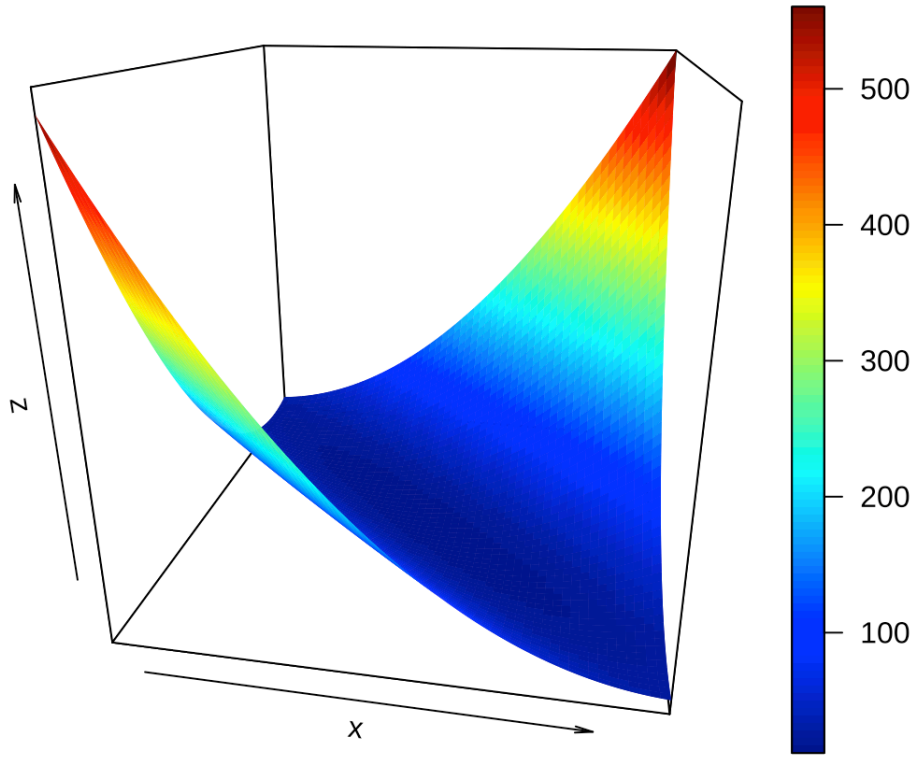


그림 1.1.: 오차제곱합의 함수 형태

위의 그림을 보면 볼록한 모양이 너무 평평하여 오차제곱합이 최소가 되는 β_0와 β_1의 위치가 명확하지 않다.

이제 모든 변수들을 표준화하고 표준화된 변수들에 단순회귀모형에 대한 오차제곱합을 β_0와 β_1의 함수로서 그림을 그리면 아래와 같다.

$$v_i = \beta_0 + \beta_1 w_i + e_i, \quad i = 1, 2, \dots, n \tag{1.4}$$

여기서

$$v_i = \frac{y_i - \bar{y}}{s_y}, \quad w_i = \frac{x_i - \bar{x}}{s_x}$$

```
# 변수들을 표준화!
std_cars <- as.data.frame(scale(cars))
gridnum <- 60
sizing <- 1
b0 <- seq(0-sizing,  0+sizing, length=gridnum )
b1 <- seq(1-sizing, 1+sizing, length=gridnum )

SSE <- matrix(0, gridnum, gridnum )
for (i in 1:gridnum ) {
  for (j in 1:gridnum ){
    r <- std_cars$dist- b0[i] -b1[j]*std_cars$speed
    SSE[i,j]  <- sum(r^2)
  }
}

persp3D(b0, b1, SSE, theta =40, phi = 15, expand = 1)

## Interactive 3d graph
#fig <- plot_ly(z = ~SSE)
#fig <- fig %>% add_surface()
#fig
```

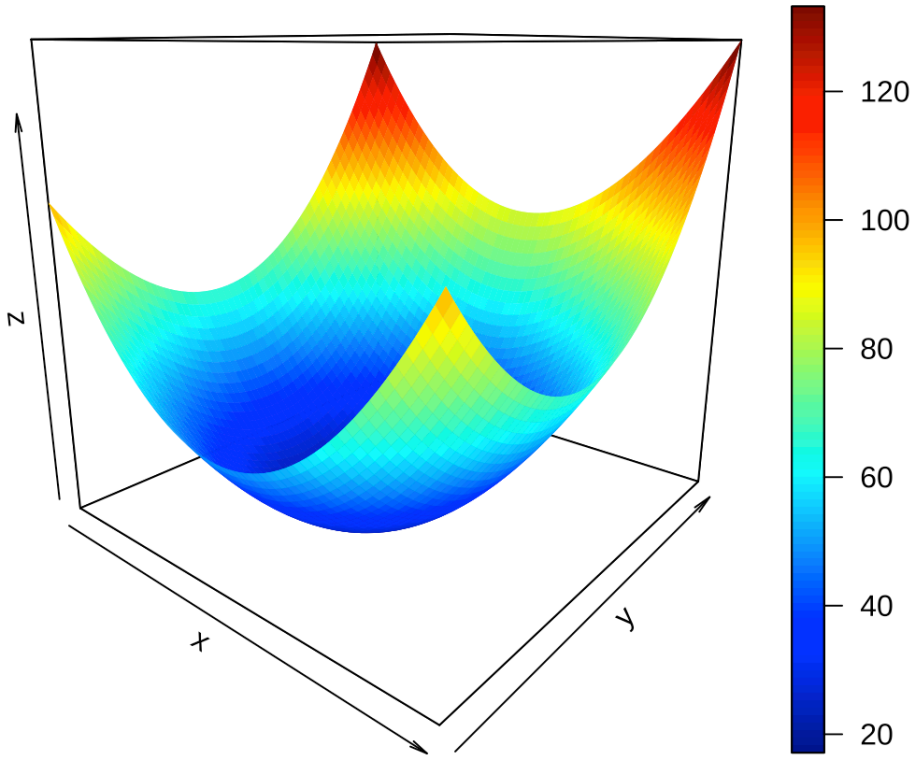


그림 1.2.: 표준화 시 오차제곱합의 함수 형태

위위 같이 변수들을 표준화하면 오차제곱합 함수의 볼록한 정도가 덜 평평하게 변하여 최적값을 더 확실하게 보인다. 기계학습이나 인공지능 모형에서 적합하기 전에 모든 변수를 표준화하는 이유가 위의 그림에서 나타난다.

1.3 의 오차제곱합을 각 회귀계수에 대해서 편미분을 하고 0으로 놓으면 아래와 같이 두 방정식이 얻어진다.

$$\frac{\partial S(\beta_0, \beta_1)}{\partial \beta_0} = \sum_{i=1}^{n} (-2)[y_i - (\beta_0 + \beta_1 x_i)] = 0$$
$$\frac{\partial S(\beta_0, \beta_1)}{\partial \beta_1} = \sum_{i=1}^{n} (-2x_i)[y_i - (\beta_0 + \beta_1 x_i)] = 0$$

위의 연립방정식을 행렬식으로 표시하면 다음과 같이 나타낼 수 있다.

$$\begin{bmatrix} n & \sum_i x_i \\ \sum_i x_i & \sum_i x_i^2 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \begin{bmatrix} \sum_i y_i \\ \sum_i x_i y_i \end{bmatrix}$$

위의 방정식을 풀어서 구한 회귀계수의 추정치를 $\hat{\beta}_0$, $\hat{\beta}_1$ 이라고 하면 다음과 같이 주어진다.

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$
$$\hat{\beta}_1 = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}$$

최소제곱법에서 얻어진 회귀계수의 추정량 $\hat{\beta}_0$과 $\hat{\beta}_1$ 을 이용한 반응변수 y_i 에 대한 예측값 $\hat{y}_i$는 다음과 같이 정의되고

$$\hat{y}_i = \hat{E}(y_i|x_i) = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

> 표준화 전과 후
>
> 두 개의 회귀방정식 1.2 과 1.4 에서 각각 최소제곱법으로 구한 기울기의 추정치 $\hat{\beta}_1$ 이 동일하게 나타나는 경우는 어떤 경우일까 생각해보자.

잔차 r_i는 다음과 같이 계산한다.

$$r_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) = y_i - \hat{y}_i \tag{1.5}$$

잔차 r_i는 다음과 같은 성질을 가진다.

$$\sum_{i=1}^{n} r_i = 0$$

$$\sum_{i=1}^{n} x_i r_i = 0$$

이제 위에서 본 `cars` 자료를 가지고 선형회귀모형 1.2 에 나타난 회귀계수를 추정해보자. 아래는 `R` 프로그램에서 함수 `lm`을 이용한 추정결과이다.

```
lm_car <- lm(dist~speed, data=cars)
summary(lm_car)
```

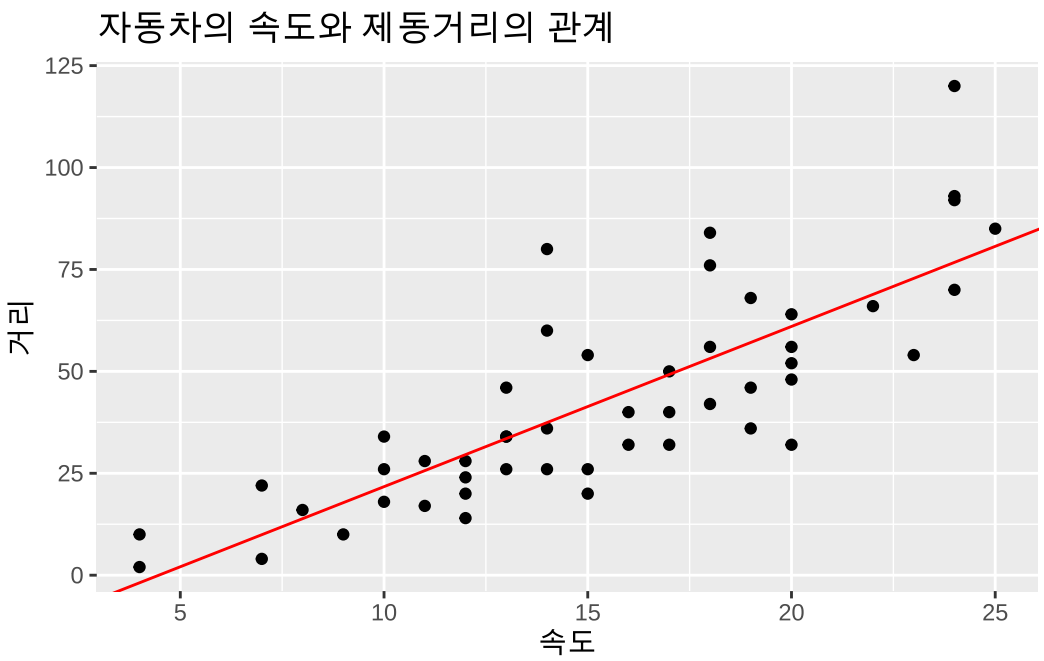
```
Call:
lm(formula = dist ~ speed, data = cars)

Residuals:
    Min      1Q  Median      3Q     Max
-29.069  -9.525  -2.272   9.215  43.201

Coefficients:
```

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -17.5791     6.7584  -2.601    0.0123 *
speed         3.9324     0.4155   9.464 1.49e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.38 on 48 degrees of freedom
Multiple R-squared:  0.6511,    Adjusted R-squared:  0.6438
F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```

위에서 주어진 선형회귀모형 1.2 에 대한 추정 결과를 이용하면 자동차의 속도($x =$ speed)와 제동거리($y =$ dist)의 관계는 다음과 같은 회귀식으로 나타낼 수 있다.

$$\hat{E}(y|x) = -17.58 + 3.93x$$

```
ggplot(cars, aes(x=speed, y=dist)) + geom_point() + labs(x = "속도", y = "거리") +
  labs(title="자동차의 속도와 제동거리의 관계") +
  geom_abline(intercept = -17.58, slope = 3.93, color = "red")
```



위의 추정식을 이용하면 주어진 자동차의 속도에서 제동거리를 예측할 수 있다. 예를 들어 자동차의 속도가 25 mph인 경우에는 제동거리의 평균이 80.73 mph 임을 알 수 있다.

$$E(y|x = 25) = -17.58 + 3.93(25) = 80.73$$

```
newcars <- data.frame(speed = c(25))
predict(lm_car, newdata=newcars)
```

```
       1
80.73112
```

기울기의 추정값 $\hat{\beta}_1 = 3.93$ 은 자동차의 속도 (x)가 1 mph 증가할 때 평균 제동거리 ($E(y|x)$)가 3.93 ft 증가한다는 의미이다.

이제 아파트 거애 가격에 대한 단순선형회귀모형을 적합해보자. 이 경우 면적과 가격대신 각각의 로그값을 사용하여 회귀모형을 적합해 보자. 아래는 아파트의 면적과 거래가격에 대한 단순선형 회귀모형을 적합한 결과이다.

```
apart_2019_log <- apart_2019 %>%
  mutate(log_area = log10(area), log_price = log10(price)) %>%
  dplyr::select(log_area, log_price)

head(apart_2019_log,10)
```

```
   log_area log_price
1  1.813514  2.653213
2  1.754807  2.440909
3  1.960185  2.777427
4  1.960185  2.748188
5  1.780101  2.628389
6  1.780101  2.623249
7  1.590173  2.336460
8  1.790567  2.685742
9  1.856729  2.515874
10 1.926548  2.641474
```

```
lm_apart <- lm( log_price~ log_area, data=apart_2019_log)
summary(lm_apart)
```

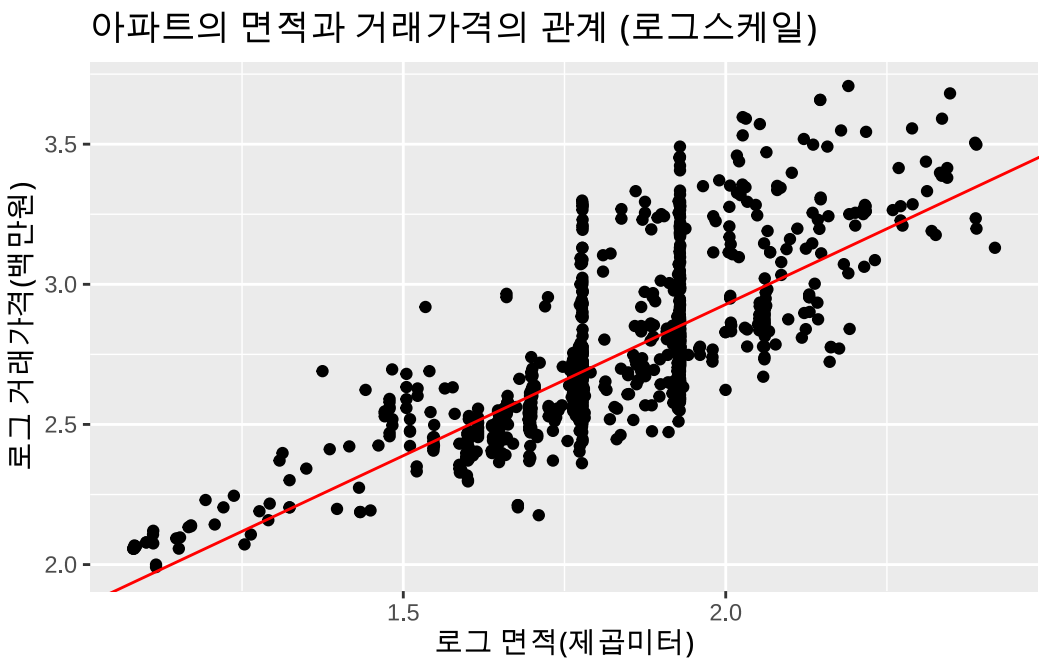
```
Call:
lm(formula = log_price ~ log_area, data = apart_2019_log)

Residuals:
     Min       1Q   Median       3Q      Max
-0.45348 -0.12132 -0.04075  0.07531  0.62358
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.76902    0.05217   14.74   <2e-16 ***
log_area     1.08797    0.02843   38.27   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1894 on 998 degrees of freedom
Multiple R-squared:  0.5948,    Adjusted R-squared:  0.5944
F-statistic:  1465 on 1 and 998 DF,  p-value: < 2.2e-16
```

위의 결과는 다음과 같이 나타낼 수 있다.

$$\hat{E}(\log 10y|x) = 0.769 + 1.0797 \log 10(x)$$

```
ggplot(apart_2019_log, aes(x=log_area, y=log_price)) + geom_point() + labs(x = "로그 면적(제곱미터
  labs(title = "아파트의 면적과 거래가격의 관계 (로그스케일)") +
  geom_abline(intercept = 0.769, slope = 1.0797, color = "red")
```



이제 위의 결과를 응용하면 아파트의 면적이 100 제곱미터인 경우의 아파트의 평균 거래가격을 880(백만원)으로 예측할 수 있다.

```
newapart <- data.frame(log_area = c(log10(100)))
pred_y <- 10^predict(lm_apart, newdata=newapart)
pred_y
```

```
      1
880.9605
```

1.4. 결정계수

고려한 설명변수와 반응변수에 대하여 제시된 회귀식을 적합한 후 회귀모형이 두 변수의 관계를 얼마나 잘 설명하는지에 대한 기준이 필요하다. 회귀식의 적합에 대한 기준으로서 결정계수(coefficient of determination; R^2)가 있다. 결정계수는 적합의 정도(degree of fitting)를 측정한다. 즉 "설명변수는 반응변수를 얼마나 잘 예측하느냐"에 대한 정도를 수치로 표현한 것이다.

회귀분석에서 설명변수와 반응변수 간에 전혀 관계가 없다면 당연히 반응변수의 값은 설명변수 값의 변동 여하에 전혀 영향을 받지 않아야 한다. 단순회귀모형에서 설명변수 x의 값의 변화를 반응변수 y로 값으로 표현하는것이 바로 기울기 β_1이다. 이렇게 고려한 설명변수 x가 반응변수 y를 예측하는데 전혀 소용이 없다면 이는 기울기에 대한 회귀계수가 0 $\beta_1 = 0$ 이라는 것을 의미이다. 이러한 경우에 대하여 다음과 같은 모형을 생각할 수 있다.

$$y_i = \beta_0 + e_i, \quad e_i \sim (0, \sigma^2) \tag{1.6}$$

기울기에 대한 회귀계수가 0인 경우에 대한 모형을 1.6 과 같이 표현할 수 있으며 평균 모형(mean model)이라고 부른다. 평균 모형은 우리가 생각할 수 있는 모형 중에서 가장 간단한, 하지만 별로 쓸모없는 모형이라고 할 수 있다.

이러한 평균 모형에 대한 최소제곱법을 적용하여 β_0의 추정량을 구하면 추정량 $\hat{\beta}_0$는 $\bar{y}$가 된다. 그 이유는 위의 모형에 오차제곱합을 구해보면 다음과 같은 형식이 된다

$$S(\beta_0) = \sum_{i=1}^{n} [y_i - \beta_0]^2$$

여기서 β_0에 대하여 최고로 하는 지점을 찾아보면 다음과 같은 방정식을 얻을 수 있다.

$$\frac{\partial S(\beta_0)}{\partial \beta_0} = 0 \Rightarrow \sum_{i=1}^{n} [y_i - \beta_0] = 0$$

이 방정식을 풀면 $\hat{\beta}_0 = \bar{y}$가 됨을 알 수 있다. 결국 설명변수가 반응변수에 아무른 영향을 주지 못하게 되면 y의 예측값은 평균 $\bar{y}$ 임을 알 수 있다. 참고로 평균모형 1.6 경우 $\bar{y}$는 β_0 의 최소제곱추정량이다.

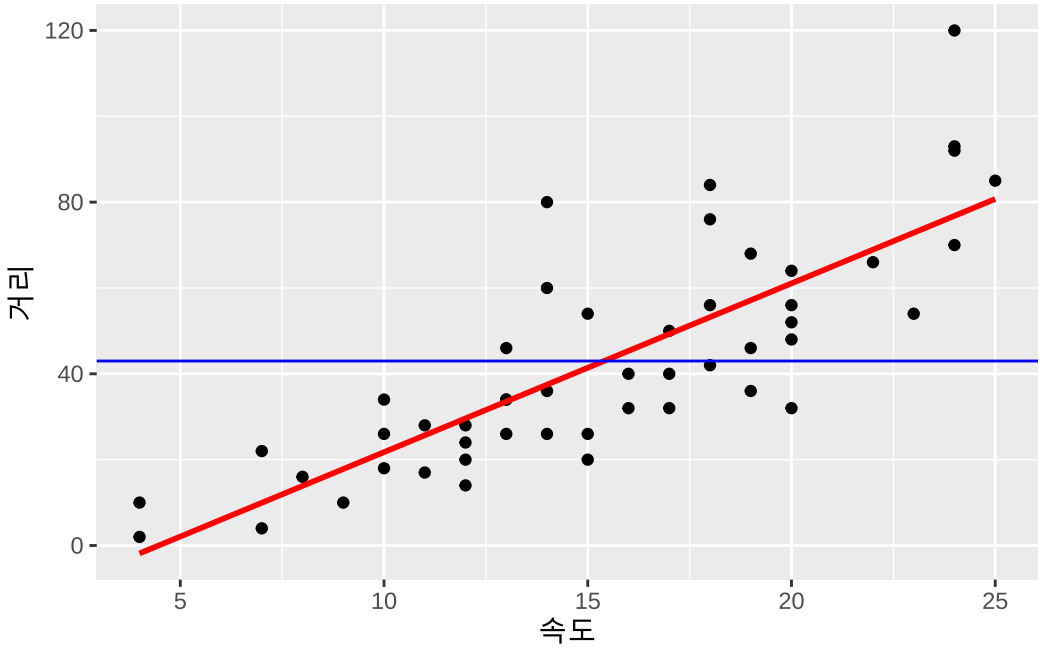
여기서 주목해야할 점은 평균 모형 1.6 에서의 잔차 r_{0i}는 다음과 같이 정의된다.

$$r_{0i} = y_i - \hat{\beta}_0 = y_i - \bar{y}$$

주어진 회귀식이 유의한 경우, 즉 회귀식의 기울기가 0이 아닌 경우 ($\beta_1 \neq 0$) 적합된 회귀식에 대한 잔차는 1.5 과 같이 나타난다. 만약 회귀식이 유의하다면 1.5 으로 구해진 잔차 $r_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$ 와 평균 모형에서 구해지는 잔차 $r_{0i} = y_i - \bar{y}$ 간의 어떤 차이가 있을까?

아래의 그림은 앞의 예제 cars 자료에 대하여 설명변수가 없는 평균 모형(파란 선)과 설명변수가 있는 회귀모형(빨간 선)을 나타낸 그림이다. 잔차는 적합된 직선과 반응 변수 간의 차이를 의미하며 차이의 절대값이 작을 수록 좋은 모형이다.

```
ggplot(cars, aes(x=speed, y=dist)) +
     geom_point() +
     labs(x = "속도", y = "거리") +
     geom_smooth(method = lm, color='red', se = FALSE) +
     geom_hline(yintercept = mean(cars$dist), color='blue')
```



잔차의 절대값보다 제곱한 양이 다루기가 쉬우므로(why?) 평균 모형과 회귀 모형의 적합도를 비교하는 양으로서 다음과 같은 각각의 모형에서 나온 두개의 잔차제곱합을 생각할 수 있다.

먼저 평균 모형은 예측에 사용할 변수가 없는 경우로서 이때의 잔차는 각 관측값에 대한 예측값이 관측값의 평균이다. 이러한 경우 잔차는 관측값 자체가 가지고 있는 변동으로 생각할 수 있다. 이러한 평균모형에서의 잔차 또는 관측값이 가지고 있는 변동을 총제곱합(Total Sum of Squares ; SST)이다

$$\begin{aligned}\sum_{i=1}^{n} r_{0i}^2 &= \sum_{i=1}^{n}(y_i - \bar{y})^2 \\ &= \text{Residual Sum of Squares from mean model} \\ &= \text{Variation of response variables} \\ &= \text{Total Sum of Squares} \\ &= SST\end{aligned}$$

이제 설명변수가 있는 회귀모형에서 예측치 $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$를 고려하면 이 경우의 잔차들의 제곱합은 회귀식의 잔차제곱합(Residual Sum of Squares; SSE)이라고 부르며 아래와 같이 정의한다.

$$\begin{aligned}\sum_{i=1}^{n} r_i^2 &= \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) \\ &= \text{Residual Sum of Squares from linear regression model} \\ &= \text{ Residual Sum of Squares} \\ &= SSE\end{aligned}$$

만약 회귀식에서 고려한 설명변수가 반응변수를 예측하는데 매우 적합하다면 회귀모형에서 구한 잔차들의 제곱합이 평균모형에서 구한 잔차들의 제곱합보다 작을 것이다. 이러한 차이를 비교하려면 두 제곱합 SST 와 SSE의 관계를 이해하는 것이 중요하다.

두 제곱합 SST 와 SSE의 관계를 보기 위하여 먼저 두 잔차 r_i^0 와 r_i의 차이를 비교해 보자

$$r_i^0 - r_i = (y_i - \bar{y}) - (y_i - \hat{y}_i) = \hat{y}_i - \bar{y}$$

위의 식에서 두 잔차의 차이 $\hat{y}_i - \bar{y}$는 예측값과 평균 간 차이로서 그 절대값이 크면 회귀직선이 반응변수를 설명할 수 있는 능력이 크다는 것을 의미한다.

위의 식을 다시 쓰면 다음과 같다.

$$(y_i - \bar{y}) = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})$$

즉, (평균모형의 잔차)=(회귀모형의 잔차))+(회귀모형의 설명부분)으로 분해되는 것으로 이해할 수 있다. 이 분해에서 회귀모형의 잔차가 작을수록 회귀 모형의 예측 능력, 즉 적합도가 커지는 것을 알 수 있다.

이제 총제곱합은 다음과 같이 분해할 수 있다.

$$\begin{aligned}\sum_{i=1}^{n}(y_i - \bar{y})^2 &= \sum_{i=1}^{n}[(y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})]^2 \\ &= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + 2\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) \\ &= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + 0 \quad \text{(why?)}\end{aligned}$$

따라서 다음과 같은 제곱합의 분해를 얻게 된다.

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$

여기서 모형제곱합(regression sum of square; SSR)를 다음과 같이 정의하면

$$SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y}_i)^2$$

총제곱합은 잔차제곱합 과 모형제곱합으로 분해된다.

$$SST = SSE + SSR \tag{1.7}$$

관측값들이 보여주는 총 변동인 총제곱합(SST)에서 회귀모형으로 설명할 수 있는 변동, 즉 모형제곱합(SSR)이 차지하는 비율을 결정계수(coefficient of determination)라 하며 R^2 으로 표현한다.

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

위에서 정의된 R^2 는 평균 모형의 잔차제곱합 SST과 회귀모형의 잔차제곱합 SSE의 비율로 정의되는 것으로 해석할 수 있다. 즉,

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{\text{Residual SS from regression model}}{\text{Residual SS from mean model}}$$

결정계수의 정의를 보면 회귀모형의 잔차제곱합(SSE)가 평균 모형의 잔차제곱합(SST)에 대하여 **상대적으로** 작아질수록 결정계수가 커진다. 결정계수 R^2는 언제나 0 이상 1 이하의 값을 갖는다. 회귀모형이 데이터에 아주 잘 적합되면 결정계수의 값은 1 에 가깝게 된다.

1.5. 중회귀 모형

일반적으로 회귀모형에서 반응변수의 수는 하나인 경우가 많지만 설명변수의 수는 여러 개인 경우가 많다. 이런 경우 중회귀 모형(multiple linear regression)은 다음과 같이 표현할 수 있고, $p-$개의 설명변수가 있다고 가정하고 $(x_1, x_2, \cdots, x_{p-})$ 표본의 크기 n인 자료가 얻어지면 선형회귀식을 행렬로 다음과 같이 표현할 수 있다.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{i,p} + e_i = \boldsymbol{x}_i^t \boldsymbol{\beta} + e_i \tag{1.8}$$

위의 식을 다시 표현하면 다음과 같이 쓸 수 있다.

$$y_i = \boldsymbol{x}_i^t \boldsymbol{\beta} + e_i = \begin{bmatrix} 1 & x_{i1} & x_{i2} & \cdots & x_{i,p} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix} + e_i$$

이제 n개의 관측치 $y_1, y_2, \ldots, y_n$ 으로 이루어진 관측값 벡터 $\boldsymbol{y}$를 고려하면 n개의 관측치에 대한 회귀식을 행렬식으로 다음과 같이 표현할 수 있다.

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1,p} \\ 1 & x_{21} & \cdots & x_{2,p} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \cdots & x_{n,p} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

위의 식을 간단히 행렬식으로 표시하면 다음과 같다.

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e} \tag{1.9}$$

위의 행렬식에서 각 벡터와 행렬의 차원은 다음과 같다.

- $\boldsymbol{y}$: $n \times 1$
- $\boldsymbol{X}$: $n \times (p+1)$
- $\boldsymbol{\beta}$: $(p+1) \times 1$
- $\boldsymbol{e}$: $n \times 1$

여기서 회귀분석의 오차항 e_i은 서로 설명이고 동일한 분산을 갖는다. 즉, 오차항은 다음의 분포를 따른다.

$$\boldsymbol{e} \sim (\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n)$$

따라서 관측값 벡터 $\boldsymbol{y}$의 평균은 다음과 같고

$$E(\boldsymbol{y}|\boldsymbol{X}) = E(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}) = \boldsymbol{X}\boldsymbol{\beta} + E(\boldsymbol{e}) = \boldsymbol{X}\boldsymbol{\beta} \tag{1.10}$$

$\boldsymbol{y}$의 분산은 아래와 같이 주어진다.

$$V(\boldsymbol{y}|\boldsymbol{X}) = E[(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^t] = E(\boldsymbol{e}\boldsymbol{e}^t) = \sigma^2 \boldsymbol{I}_n \tag{1.11}$$

여기서 오차항이 정규분포를 따른다면

$$\boldsymbol{e} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n)$$

관측값 벡터 $\boldsymbol{y}$ 또한 정규분포를 따른다

$$\boldsymbol{y} \sim N(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_n)$$

1.5.1. 최소제곱추정

이제 중회귀모형 1.9 에서 회귀계수벡터 $\boldsymbol{\beta}$의 추정량을 구하기 위하여 최소제곱법을 적용해보자.

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} (y_i - \boldsymbol{x}_i^t \boldsymbol{\beta})^2 = \min_{\boldsymbol{\beta}} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^t (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \tag{1.12}$$

1.5.1.1. 방법 1

$\hat{\boldsymbol{\beta}}$는 잔차의 제곱합 7.8 을 최소로 하는 최소제곱 추정량이다. 잔차의 제곱합을 $S(\boldsymbol{\beta})$이라고 하면

$$\begin{aligned} S(\boldsymbol{\beta}) &= (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^t (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \\ &= \boldsymbol{y}^t\boldsymbol{y} - \boldsymbol{y}^t\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{\beta}^t\boldsymbol{X}^t\boldsymbol{y} + \boldsymbol{\beta}^t\boldsymbol{X}^t\boldsymbol{X}\boldsymbol{\beta} \\ &= \boldsymbol{y}^t\boldsymbol{y} - 2\boldsymbol{\beta}^t\boldsymbol{X}^t\boldsymbol{y} + \boldsymbol{\beta}^t\boldsymbol{X}^t\boldsymbol{X}\boldsymbol{\beta} \end{aligned}$$

여기서 $S(\boldsymbol{\beta})$를 최소로 하는 회귀계수벡터의 값을 구하기 위하여 $S(\boldsymbol{\beta})$를 회귀계수벡터 $\boldsymbol{\beta}$로 미분한후 $\mathbf{0}$ 으로 놓고 선형 방정식을 풀어야 한다.

벡터미분을 이용하면

$$\begin{aligned} \frac{\partial S(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} &= \frac{\partial}{\partial \boldsymbol{\beta}} (\boldsymbol{y}^t\boldsymbol{y} - 2\boldsymbol{\beta}^t\boldsymbol{X}^t\boldsymbol{y} + \boldsymbol{\beta}^t\boldsymbol{X}^t\boldsymbol{X}\boldsymbol{\beta}) \\ &= \mathbf{0} - 2\boldsymbol{X}^t\boldsymbol{y} + 2\boldsymbol{X}^t\boldsymbol{X}\boldsymbol{\beta} \\ &= \mathbf{0} \end{aligned}$$

최소제곱 추정량을 구하기 위한 정규방정식은 다음과 같이 쓸 수 있다.

$$\boldsymbol{X}^t\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}^t\boldsymbol{y} \tag{1.13}$$

방정식 1.13 를 정규방정식(normal equation)이라고 한다. 만약 $\boldsymbol{X}^t\boldsymbol{X}$가 정칙행렬일 경우 최소제곱법에 의한 회귀계수 추정량 $\hat{\boldsymbol{\beta}}$ 다음과 같다.

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{y} \tag{1.14}$$

예측값 벡터 $\hat{\boldsymbol{y}}$ 는 $E(\boldsymbol{y}|\boldsymbol{X})$의 추정치로서 다음과 같다.

$$\hat{E}(\boldsymbol{y}|\boldsymbol{X}) = \hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t y$$

만약 $\boldsymbol{X}^t\boldsymbol{X}$가 정칙행렬이 아닐 경우 최소제곱법에 의한 회귀계수 추정량 $\hat{\boldsymbol{\beta}}$은 $\boldsymbol{X}^t\boldsymbol{X}$의 일반화 역행렬 $(\boldsymbol{X}^t\boldsymbol{X})^-$를 이용하여 다음과 같이 구한다. 이 경우 일반화 역행렬이 유일하지 않기 때문에 회귀계수 추정량도 유일하지 않다.

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^t\boldsymbol{X})^{-}\boldsymbol{X}^t\boldsymbol{y}$$

1.5.1.2. 방법 2

7.8 에서 나오는 오차벡터를 정의하고 $\boldsymbol{e} = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$ 오차벡터를 모수벡터 $\boldsymbol{\beta}$로 미분하면 다음과 같은 결과를 얻는다.

$$\frac{\partial \boldsymbol{e}}{\partial \boldsymbol{\beta}} = \frac{\partial(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -\frac{\partial \boldsymbol{X}\boldsymbol{\beta}}{\partial \boldsymbol{\beta}} \equiv -\frac{\partial \boldsymbol{\beta}^t\boldsymbol{X}^t}{\partial \boldsymbol{\beta}} = -\boldsymbol{X}^t$$

이제 오차제곱합 $S(\boldsymbol{\beta}) = \boldsymbol{e}^t\boldsymbol{e}$ 를 모수벡터로 미분하면 이차형식의 미분공식과 합성함수 미분공식을 차례로 적용하면 된다.

$$\frac{\partial S(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \frac{\partial \boldsymbol{e}^t\boldsymbol{e}}{\partial \boldsymbol{\beta}} = \frac{\partial \boldsymbol{e}}{\partial \boldsymbol{\beta}}\frac{\partial \boldsymbol{e}^t\boldsymbol{e}}{\partial \boldsymbol{e}} = -\boldsymbol{X}^t\,(2\boldsymbol{e}) = -2\boldsymbol{X}^t(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$$

위의 방정식을 $\mathbf{0}$으로 놓으면 최소제곱 추정량 (열)벡터를 구한다.

$$\boldsymbol{X}^t\boldsymbol{y} - \boldsymbol{X}^t\boldsymbol{X}\boldsymbol{\beta} = \mathbf{0} \quad \rightarrow \quad \hat{\boldsymbol{\beta}} = (\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{y}$$

1.6. 예제-중회귀모형

Example 1.3 (중고차 가격자료)**.** 강의 부교재의 `usedcars` 자료를 이용하여 중회귀모형을 적합해보자. 자료를 구성하는 변수는 다음과 같다.

- `price` : 자동차 가격
- `yesr` :연식
- `mileage` : 주행거리
- `cc` : 엔진 크기
- `automatic` : 자동 변속기 여부

```
usedcars %>% head(n=10)
```

```
   price year mileage   cc automatic
1    790   78  133462 1998         1
2   1380   39   33000 2000         1
3    270  109  120000 1800         0
4   1190   20   69727 1999         1
5    590   70  112000 2000         0
6   1120   58   39106 1998         1
```

```
7    815   53   95935 1800         1
8    450   68  120000 1800         0
9   1290   15   20215 1798         1
10   420   96  140000 1800         0
```

자동차의 가격을 반응변수로 한고 나머지 변수를 설명변수로 설정한 중회귀 모형에 대한 모수의 추정 결과는 다음과 같다.

```
usedcars_lm <- lm(price ~ year + mileage + cc + automatic, data=usedcars)
summary(usedcars_lm)
```

```

Call:
lm(formula = price ~ year + mileage + cc + automatic, data = usedcars)

Residuals:
    Min      1Q  Median      3Q     Max
-177.35  -63.91   -0.99   70.34  212.69

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.253e+02  3.998e+02   1.314 0.200823
year        -5.800e+00  9.283e-01  -6.247 1.55e-06 ***
mileage     -2.263e-03  7.211e-04  -3.138 0.004324 **
cc           3.888e-01  2.022e-01   1.923 0.065958 .
automatic    1.653e+02  3.986e+01   4.147 0.000339 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 101.1 on 25 degrees of freedom
Multiple R-squared:  0.9045,    Adjusted R-squared:  0.8892
F-statistic: 59.21 on 4 and 25 DF,  p-value: 2.184e-12
```

1.7. 최소제곱 추정량의 분포

회귀식을 추정하기 위한 회귀계수 추정값인 $\hat{\boldsymbol{\beta}}$의 분포를 알아보기 위해서 우선 선형추정량을 보면 다음과 같다.

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} \equiv \boldsymbol{M}\boldsymbol{y}$$

따라서 최소제곱 추정량은 관측값들의 선형 변환이다. 회귀계수 추정값 $\hat{\boldsymbol{\beta}}$의 기대값은

$$\begin{aligned} E(\hat{\boldsymbol{\beta}}) &= E(\boldsymbol{M}\boldsymbol{y}) = E((\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}) \\ &= (\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}'E(\boldsymbol{y}) \\ &= (\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{X}\boldsymbol{\beta} \\ &= \boldsymbol{\beta} \end{aligned}$$

따라서 최소제곱 추정량 $\hat{\boldsymbol{\beta}}$는 $\boldsymbol{\beta}$의 불편추정량이다. 최소제곱 추정량 $\hat{\boldsymbol{\beta}}$의 공분산 행렬을 전개해보면

$$\begin{aligned} Var(\hat{\boldsymbol{\beta}}) &= Var((\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{y}) \\ &= (\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\ Var(\boldsymbol{y})\ \boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{X})^{-1} \\ &= (\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t[\sigma^2\boldsymbol{I}_n]\boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{X})^{-1} \\ &= \sigma^2(\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{X})^{-1} \\ &= \sigma^2(\boldsymbol{X}^t\boldsymbol{X})^{-1} \end{aligned}$$

위에서 최소제곱 추정량의 평균과 공분산을 구할 때에는 정규성 가정이 필요하지않다. 만일 $\boldsymbol{y}$가 정규분포를 따른다면 $\boldsymbol{y}$의 선형변환으로부터 얻어진 $\hat{\boldsymbol{\beta}}$의 분포는 정규분포이며 다음과 같다.

$$\hat{\boldsymbol{\beta}} \sim N\left(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^t\boldsymbol{X})^{-1}\right)$$

1.8. 가우스-마코브 정리

Theorem 1.1 (가우스-마코브 정리)**.** 선형회귀모형 $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$에서 $E(\boldsymbol{e}) = 0, Var(\boldsymbol{e}) = \sigma^2\boldsymbol{I}$이 성립하면 최소제곱 추정량

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{y}$$

는 $\boldsymbol{\beta}$의 최소분산 선형 불편추정량이다.

위의 정리를 가우스-마코브 정리 (Gauss-Markov Theorem)라고 하며 이는 회귀계수 $\boldsymbol{\beta}$의 모든 선형 불편 추정량들 중에 최소제곱 추정량 $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{y}$이 가장 작은 분산을 가짐을 뜻한다 (Best Linear Unbiased Estimator; BLUE).

가우스-마코브 정리를 정확하게 표현하면 $E(\boldsymbol{L}\boldsymbol{y}) = \boldsymbol{\beta}$를 만족하는 모든 $n \times n$ 차원의 행렬 $\boldsymbol{L}$과 임의의 벡터 $\boldsymbol{c}$에 대하여 다음이 성립한다.

$$V(\boldsymbol{c}^t\hat{\boldsymbol{\beta}}) \leq V(\boldsymbol{c}^t\boldsymbol{L}\boldsymbol{y})$$

이제 가우스-마코브 정리를 증명해보자. 관측벡터 $\boldsymbol{y}$에 대한 임의의 선형 추정량 $\boldsymbol{\beta}^* = \boldsymbol{L}\boldsymbol{y}$를 생각해보면 다시 다음의 형태로 표시할 수 있다.

$$\boldsymbol{\beta}^* = \boldsymbol{L}\boldsymbol{y} = (\boldsymbol{M} + \boldsymbol{L} - \boldsymbol{M})\boldsymbol{y} = (\boldsymbol{M} + \boldsymbol{A})\boldsymbol{y}$$

여기서 $\boldsymbol{M} = (\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t$ 이고 $\boldsymbol{A} = \boldsymbol{L} - \boldsymbol{M}$ 이다. 임의의 선형 추정량 $\boldsymbol{\beta}^*$가 불편 추정량일 조건을 구해보자

$$\begin{aligned} E(\boldsymbol{\beta}^*) &= E[(\boldsymbol{M} + \boldsymbol{A})\boldsymbol{y}] \\ &= (\boldsymbol{M} + \boldsymbol{A})E(\boldsymbol{y}) \\ &= (\boldsymbol{M} + \boldsymbol{A})X\boldsymbol{\beta} \\ &= (\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{A}\boldsymbol{X}\boldsymbol{\beta} \\ &= \boldsymbol{\beta} + AX\boldsymbol{\beta} \end{aligned}$$

여기서 불편추정량이 되기 위해서는 $E(\boldsymbol{\beta}^*) = \boldsymbol{\beta}$ 조건을 만족 해야되며 따라서 $\boldsymbol{A}\boldsymbol{X} = 0$이되어야한다 (이 조건은 $\boldsymbol{A} = 0$를 의미하는 것은 아니다).

이제 최소분산을 가지기 위해서 $\boldsymbol{A}\boldsymbol{X} = 0$을 만족하는 행렬 $\boldsymbol{A}$중에서 $Var(\boldsymbol{\beta}^*)$을 최소로하는 행렬 $\boldsymbol{A}$를 구해야 한다. $\boldsymbol{\beta}^*$의 공분산 행렬은 $AX = 0$이므로

$$\begin{aligned} V(\boldsymbol{\beta}^*) &= (\boldsymbol{M} + \boldsymbol{A})V(\boldsymbol{y})(\boldsymbol{M} + \boldsymbol{A})^t \\ &= (\boldsymbol{M} + \boldsymbol{A})\sigma^2 I_n(\boldsymbol{M} + \boldsymbol{A})^t \\ &= \sigma^2(\boldsymbol{M}\boldsymbol{M}^t + \boldsymbol{A}\boldsymbol{M}^t + \boldsymbol{M}\boldsymbol{A}^t + \boldsymbol{A}\boldsymbol{A}^t) \\ &= \sigma^2[(\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{X})^{-1} + \boldsymbol{A}\boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{X})^{-1} + (\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{A}^t + \boldsymbol{A}\boldsymbol{A}^t] \\ &= \sigma^2[(\boldsymbol{X}^t\boldsymbol{X})^{-1} + \boldsymbol{A}\boldsymbol{A}^t] \\ &= V(\hat{\boldsymbol{\beta}}) + \sigma^2\boldsymbol{A}\boldsymbol{A}^t \end{aligned}$$

이제 임의의 벡터 $\boldsymbol{c}$에 대하여

$$\begin{aligned} V(\boldsymbol{c}^t\boldsymbol{\beta}^*) &= \boldsymbol{c}^t V(\boldsymbol{\beta}^*)\boldsymbol{c} \\ &= \boldsymbol{c}^t V(\hat{\boldsymbol{\beta}})\boldsymbol{c} + \sigma^2\boldsymbol{c}^t\boldsymbol{A}\boldsymbol{A}^t\boldsymbol{c} \\ &= V(\boldsymbol{c}^t\hat{\boldsymbol{\beta}}) + \sigma^2\boldsymbol{c}^t\boldsymbol{A}\boldsymbol{A}^t\boldsymbol{c} \end{aligned}$$

다음이 성립하므로

$$\boldsymbol{c}^t\boldsymbol{A}\boldsymbol{A}^t\boldsymbol{c} = \boldsymbol{u}^t\boldsymbol{u} = \sum_{i=1}^{n} u_i^2 \geq 0$$

임의의 벡터 $\boldsymbol{c}$에 대하여

$$V(\boldsymbol{c}^t\boldsymbol{\beta}^*) \geq V(\boldsymbol{c}^t\hat{\boldsymbol{\beta}})$$

이제 $V(\boldsymbol{c}^t\boldsymbol{\beta}^*)$ 이 $V(\boldsymbol{c}^t\hat{\boldsymbol{\beta}})$과 같으려면 다음 조건이 성립해야 하며

$$\boldsymbol{u} = \boldsymbol{c}^t \boldsymbol{A} = \boldsymbol{0}$$

임의의 모든 벡터 $\boldsymbol{c}$에 대해서 위의 조건 성립해야 하므로 이는 $\boldsymbol{A} = \boldsymbol{0}$ 이 성립해야 한다. 또한 이조건은 $\boldsymbol{AX} = 0$ 도 만족 시켜준다. 따라서 $\boldsymbol{\beta}$의 최소분산 선형 불편추정량은 최소제곱법으로 구한 추정량이다.

여기서 주의할 점은 가우스-마코브 정리에서 관측값 $\boldsymbol{y}$에 대한 가정은 평균과 공분산의 가정만 주어졌으며 $\boldsymbol{y}$의 분포에 대한 가정이 없다. 참고로 만약에 $\boldsymbol{y}$가 정규분포를 따른다면 최소제곱 추정량은 최소분산 불편추정량이다.

1.9. 최대가능도 추정

관측값 벡터 $\boldsymbol{y}$ 가 다음과 같이 선형모형이며 정규분포를 따른다고 가정하자.

$$\boldsymbol{y} \sim N(\boldsymbol{X\beta}, \sigma^2 \boldsymbol{I}_n) \tag{1.15}$$

선형모형 1.15 에 대한 가능도 함수는 다음과 같이 주어진다.

$$\begin{aligned} L_n(\boldsymbol{\theta}; \boldsymbol{y}) &= L(\boldsymbol{\beta}, \sigma^2 | \boldsymbol{y}) \\ &= \prod_{i=1}^n f(y_i) \\ &= \prod_{i=1}^n (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left[-\frac{1}{2\sigma^2}(y_i - \boldsymbol{x}_i^t \boldsymbol{\beta})^2\right] \\ &= (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left[-\frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X\beta})^t(\boldsymbol{y} - \boldsymbol{X\beta})\right] \end{aligned}$$

또한 분산에 대한 모수를 $\tau = \sigma^2$ 과 같이 쓰면 로그 가능도함수는 다음과 같다.

$$\begin{aligned} \ell_n(\boldsymbol{\theta}; \boldsymbol{y}) &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{(\boldsymbol{y} - \boldsymbol{X\beta})^t(\boldsymbol{y} - \boldsymbol{X\beta})}{2\sigma^2} \\ &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\tau - \frac{(\boldsymbol{y} - \boldsymbol{X\beta})^t(\boldsymbol{y} - \boldsymbol{X\beta})}{2\tau} \end{aligned}$$

이제 로그가능도함수로부터 구할 수 있는 스코어함수 $s(\boldsymbol{\theta}; \boldsymbol{y})$ 와 그에 대한 관측 피셔정보 $J_n(\boldsymbol{\theta}; \boldsymbol{y})$ 은 다음과 같이 주어진다.

$$\begin{aligned} s(\boldsymbol{\theta}; \boldsymbol{y}) &= \frac{\partial}{\partial \boldsymbol{\theta}} \ell_n(\boldsymbol{\theta}; \boldsymbol{y}) \\ &= \begin{bmatrix} \frac{\partial}{\partial \boldsymbol{\beta}} \ell_n(\boldsymbol{\theta}; \boldsymbol{y}) \\ \frac{\partial}{\partial \tau} \ell_n(\boldsymbol{\theta}; \boldsymbol{y}) \end{bmatrix} \\ &= \begin{bmatrix} \boldsymbol{X}^t(\boldsymbol{y} - \boldsymbol{X\beta})/\tau \\ -\frac{n}{2\tau} + \frac{(\boldsymbol{y} - \boldsymbol{X\beta})^t(\boldsymbol{y} - \boldsymbol{X\beta})}{2\tau^2} \end{bmatrix} \end{aligned}$$

$$\begin{aligned} J_n(\boldsymbol{\theta};\boldsymbol{y}) &= -\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}}\ell_n(\boldsymbol{\theta};\boldsymbol{y}) \\ &= -\begin{bmatrix} \frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}}\ell_n(\boldsymbol{\theta};\boldsymbol{y}) & \frac{\partial^2}{\partial\boldsymbol{\beta}\partial\tau}\ell_n(\boldsymbol{\theta};\boldsymbol{y}) \\ \frac{\partial^2}{\partial\tau\partial\boldsymbol{\beta}}\ell_n(\boldsymbol{\theta};\boldsymbol{y}) & \frac{\partial^2}{\partial\tau\partial\tau}\ell_n(\boldsymbol{\theta};\boldsymbol{y}) \end{bmatrix} \\ &= \begin{bmatrix} \boldsymbol{X}^t\boldsymbol{X}/\tau & -\boldsymbol{X}^t(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})/\tau^2 \\ -(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^t\boldsymbol{X}/\tau^2 & -\frac{n}{2\tau^2}+\frac{(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^t(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})}{\tau^3} \end{bmatrix} \end{aligned}$$

이제 중회귀모형에서 회귀계수 $\boldsymbol{\beta}$에 대한 최대가능도 추정량은 스코어함수로 부터 얻어진 방정식 $s(\boldsymbol{\theta};y)=0$ 으로부터 얻어지며 다음과 같은 형태를 가진다.

$$\hat{\beta} = (\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{y}$$

$$\hat{\sigma}^2 = \hat{\tau} = (\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})^t(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})/n = \frac{SSE(\hat{\boldsymbol{\beta}})}{n}$$

여기서 유의할 점은 회귀계수 $\boldsymbol{\beta}$ 의 최대가능도 추정량은 최소제곱법으로 구한 추정량과 동일하다. 따라서 $\hat{\boldsymbol{\beta}}$은 최소분산 불편 추정량이다. 하지만 오차항의 분산 σ^2 에 대한 최대가능도 추정량은 불편추정량이 아니다.

$$E(\hat{\sigma}^2) = E\left[(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})^t(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})/n\right] = E\left[\frac{SSE}{n}\right] \neq \sigma^2$$

참고로 오차항의 분산 σ^2에 대한 불편추정량은 $SSE/(n-p)$이다. 오차항의 분산에 대한 불편추정량은 다음 장에서 논의할 것이다.

최대가능도 추정량의 점근적 분포를 이용하면 다음과 같이 말할 수 있다. 오차항이 정규분포인 선형모형인 경우 아래의 분포는 점근분포가 아닌 정확한 분포이다.

$$\hat{\boldsymbol{\theta}} \sim N(\boldsymbol{\theta}, I_n^{-1}(\boldsymbol{\theta}))$$

여기서

$$I_n(\boldsymbol{\theta}) = E[J(\boldsymbol{\theta};y)] = \begin{bmatrix} \boldsymbol{X}^t\boldsymbol{X}/\tau & \mathbf{0} \\ \mathbf{0}^t & \frac{n}{2\tau^2} \end{bmatrix}$$

그리고

$$I_n^{-1}(\boldsymbol{\theta}) = \begin{bmatrix} \tau(\boldsymbol{X}^t\boldsymbol{X})^{-1} & \mathbf{0} \\ \mathbf{0}^t & \frac{2\tau^2}{n} \end{bmatrix} = \begin{bmatrix} \sigma^2(\boldsymbol{X}^t\boldsymbol{X})^{-1} & \mathbf{0} \\ \mathbf{0}^t & \frac{2\sigma^4}{n} \end{bmatrix}$$

따라서 회귀계수 추정량 $\hat{\beta}$의 분포는 평균이 $\boldsymbol{\beta}$ 이고 공분산이 $\sigma^2(\boldsymbol{X}^t-\boldsymbol{X})^{-1}$ 인 정규분포를 따른다.

여기거 주목할 점은 가능도함수에 최대가능도추정량을 대입하면 그 값이 $SSE(\hat{\beta})$의 함수로 나타난다.

$$
\begin{aligned}
L_n(\hat{\boldsymbol{\theta}}) &= L_n(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2) \\
&= (2\pi\hat{\sigma}^2)^{-\frac{n}{2}} \exp\left[-\frac{1}{2\hat{\sigma}^2}(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})^t(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})\right] \\
&= (2\pi\hat{\sigma}^2)^{-\frac{n}{2}} \exp\left[-\frac{n}{2}\right] \\
&= \left(2\pi\frac{SSE(\hat{\boldsymbol{\beta}})}{n}\right)^{-\frac{n}{2}} \exp\left[-\frac{n}{2}\right]
\end{aligned}
$$

또한 가능도함수의 값은 다음과 같다.

$$
l_n(\hat{\boldsymbol{\theta}}) = l_n(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2) = \text{constant} - \frac{n}{2}\log\frac{SSE(\hat{\boldsymbol{\beta}})}{n} \tag{1.16}
$$

따라서 잔차제곱함 $SSE(\hat{\boldsymbol{\beta}})$ 작아지면 가능도함수는 커진다.

앞 절에서 언급한 평균 모형 1.6 에서 최대가능도 추정을 알아보자. 관측값 벡터는 다음과 같은 분포를 따른다.

$$
\boldsymbol{y} \sim N(\beta_0\mathbf{1}, \sigma^2\boldsymbol{I}_n) \tag{1.17}
$$

선형모형 1.15 에 대한 로그 가능도 함수는 다음과 같이 주어진다. 분산에 대한 모수를 $\tau = \sigma^2$ 로 바꾸어 사용하면 모수 벡터는 $\boldsymbol{\theta} = (\beta_0, \tau)^t$ 이다.

$$
\ell_n(\boldsymbol{\theta}; \boldsymbol{y}) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\tau - \frac{(\boldsymbol{y}-\beta_0\mathbf{1})^t(\boldsymbol{y}-\beta_0\mathbf{1})}{2\tau}
$$

두 개의 모수 β_0와 τ에 대하여 미분하여 가능도 방정식을 구하면 다음과 같다.

$$
\begin{aligned}
s(\boldsymbol{\theta}; \boldsymbol{y}) &= \frac{\partial}{\partial\boldsymbol{\theta}}\ell_n(\boldsymbol{\theta}; \boldsymbol{y}) \\
&= \begin{bmatrix} \frac{\partial}{\partial\beta_0}\ell_n(\boldsymbol{\theta}; \boldsymbol{y}) \\ \frac{\partial}{\partial\tau}\ell_n(\boldsymbol{\theta}; \boldsymbol{y}) \end{bmatrix} \\
&= \begin{bmatrix} \mathbf{1}^t(\boldsymbol{y}-\beta_0\mathbf{1})/\tau \\ -\frac{n}{2\tau} + \frac{(\boldsymbol{y}-\beta_0\mathbf{1}))^t(\boldsymbol{y}-\beta_0\mathbf{1})}{2\tau^2} \end{bmatrix} \\
&= \mathbf{0}
\end{aligned}
$$

위의 방정식을 풀면 다음과 같은 최대가능도 추정량을 구할 수 있다.

$$
\hat{\beta}_0 = \bar{y}, \quad \hat{\sigma}^2 = \frac{\sum_{i=1}^n (y_i-\bar{y})^2}{n} = \frac{SST}{n}
$$

그리고 가능도함수에 최대가능도추정량을 대입하면 그 값이 다음과 같다.

$$
L_n(\hat{\boldsymbol{\theta}}) = L_n(\hat{\beta}_0, \hat{\sigma}^2) = \left(2\pi\frac{SST}{n}\right)^{-\frac{n}{2}} \exp\left[-\frac{n}{2}\right] \tag{1.18}
$$

두 개의 모형, 즉 선형회귀모형 1.15 과 평균모형 1.6 의 가능도 함수의 비, 즉 1.16 과 1.18 의 비율을 구해보면 결정계수 R^2 외의 관계를 볼 수 있다.

$$\frac{L_n(\hat{\beta}_0, \hat{\sigma}^2)}{L_n(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2)} = \left(2\pi\frac{SST}{n}\right)^{-\frac{n}{2}} / \left(2\pi\frac{SSE}{n}\right)^{-\frac{n}{2}} \propto \left[\frac{SSE}{SST}\right]^{\frac{n}{2}} = \left[1 - R^2\right]^{\frac{n}{2}}$$

2. 선형회귀에서의 추론

2.1. 제곱합의 분포

앞 장의 중회귀 모형 1.9 에서 관측값 벡터 $\boldsymbol{y}$가 다변량 정규분포 $N(\boldsymbol{X\beta}, \sigma^2\boldsymbol{I})$를 따를 때 회귀계수의 추정량 $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{y}$ 은 다음과 같은 분포를 따르는 것을 보였다.

$$\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^t\boldsymbol{X})^{-1})$$

반응변수의 추정값을 구하는 식에서 다음과 같은 모자행렬(hat matrix) $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t$ 을 정의하자. 여기서 중요한 점은 모자행렬은 대칭인 멱등행렬 ($\boldsymbol{HH} = \boldsymbol{H}$)이며 이는 모자행렬이 사영행렬임을 의미한다.

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{y} = \boldsymbol{Hy} \tag{2.1}$$

2.1.1. 잔차제곱합의 분포

이제 제곱합들의 분포를 알아보기로 하자. 먼저 잔차제곱합 SSE를 이차 형식으로 표시해보자.

$$\begin{aligned} SSE &= \sum_{i=1}^{n}(y_i - \hat{y}_i) \\ &= (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^t(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) \\ &= (\boldsymbol{y} - \boldsymbol{Hy})^t(\boldsymbol{y} - \boldsymbol{Hy}) \\ &= \boldsymbol{y}^t(\boldsymbol{I} - \boldsymbol{H})^t(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y} \\ &= \boldsymbol{y}^t(\boldsymbol{I} - \boldsymbol{H})(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y} \\ &= \boldsymbol{y}^t(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y} \end{aligned}$$

위의 식에서 $\boldsymbol{I} - \boldsymbol{H}$는 멱등행렬이고 다음이 성립한다.

$$(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{X} = \boldsymbol{X} - \boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{X} = \boldsymbol{0}$$

따라서

$$\boldsymbol{\mu}^t(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{\mu} = \boldsymbol{\beta}^t\boldsymbol{X}^t(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{X}\beta = 0$$

이므로 비중심 모수는 0이다.

또한

$$\begin{aligned} r(\boldsymbol{I}-\boldsymbol{H}) &= tr(\boldsymbol{I}-\boldsymbol{H}) \\ &= tr(\boldsymbol{I}_n) - tr\left[\boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\right] \\ &= n - tr\left[(\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{X}\right] \\ &= n - tr(\boldsymbol{I}_p) \\ &= n-p \end{aligned}$$

이므로 부록의 정리에 의하여 SSE는 다음과 같이 중심 카이제곱 분포를 따른다.

$$\frac{SSE}{\sigma^2} \sim \chi^2(n-p) \tag{2.2}$$

2.1.2. 회귀제곱합의 분포

다음으로 회귀제곱합 SSR의 분포를 유도해보자.

$$\begin{aligned} SSR &= \sum_{i=1}^{n}(\hat{y}_i - \bar{y}) \\ &= (\boldsymbol{X}\hat{\boldsymbol{\beta}} - \bar{y}\mathbf{1})^t(\boldsymbol{X}\hat{\boldsymbol{\beta}} - \bar{y}\mathbf{1}) \\ &= \left(\boldsymbol{X}\hat{\boldsymbol{\beta}} - \mathbf{1}(\mathbf{1}^t\boldsymbol{y})/n\right)^t\left(\boldsymbol{X}\hat{\boldsymbol{\beta}} - \mathbf{1}(\mathbf{1}^t\boldsymbol{y})/n\right) \\ &= \left(\boldsymbol{H}\boldsymbol{y} - \tfrac{1}{n}\mathbf{1}\mathbf{1}^t\boldsymbol{y}\right)^t\left(\boldsymbol{H}\boldsymbol{y} - \tfrac{1}{n}\mathbf{1}\mathbf{1}^t\boldsymbol{y}\right) \\ &= \boldsymbol{y}^t\left(\boldsymbol{H} - \tfrac{1}{n}\boldsymbol{J}\right)^t\left(\boldsymbol{H} - \tfrac{1}{n}\boldsymbol{J}\right)\boldsymbol{y} \\ &= \boldsymbol{y}^t\left(\boldsymbol{H} - \tfrac{1}{n}\boldsymbol{J}\right)\left(\boldsymbol{H} - \tfrac{1}{n}\boldsymbol{J}\right)\boldsymbol{y} \\ &= \boldsymbol{y}^t\left(\boldsymbol{H} - \tfrac{1}{n}\boldsymbol{J}\right)\boldsymbol{y} \end{aligned}$$

위의 유도식에서 다음 두 가지 성질을 이용하였다. 첫 번째 성질은 모자행렬이 사영행렬이며 모자행렬이 투영하는 공간은 일벡터 $\mathbf{1}$을 포함한 공간이다. 이는 계획 행렬 $\boldsymbol{X}$의 첫 번째 열이 절편에 대한 값으로 모두 1인 것 때문이다. 따라서

$$\begin{aligned} \boldsymbol{H}\boldsymbol{J} &= \boldsymbol{H}\mathbf{1}\mathbf{1}^t \\ &= [\boldsymbol{H}\mathbf{1}]\mathbf{1}^t \\ &= \left[\boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\mathbf{1}\right]\mathbf{1}^t \\ &= \mathbf{1}\mathbf{1}^t \\ &= \boldsymbol{J} \end{aligned}$$

두 번째로 다음과 같이 $\boldsymbol{J}\boldsymbol{J} = n\boldsymbol{J}$이므로 $\frac{1}{n}\boldsymbol{J}$는 멱등행렬이다.

$$\begin{aligned}\boldsymbol{J}\boldsymbol{J} &= \mathbf{1}\mathbf{1}^t\mathbf{1}\mathbf{1}^t \\ &= \mathbf{1}[\mathbf{1}^t\mathbf{1}]\mathbf{1}^t \\ &= \mathbf{1}[n]\mathbf{1}^t \\ &= n\mathbf{1}\mathbf{1}^t \\ &= n\boldsymbol{J}\end{aligned}$$

> **Note**
>
> 참고로 평균모형 1.6 에서 $\boldsymbol{X} = \mathbf{1}$ 으므로 이 경우 모자행렬이 다음과 같다.
>
> $$H_0 = \mathbf{1}(\mathbf{1}^t\mathbf{1})^{-1}\mathbf{1}^t = \tfrac{1}{n}\boldsymbol{J}$$

다음으로 비중심 모수를 유도하자.

$$\begin{aligned}\boldsymbol{\mu}^t\left(\boldsymbol{H} - \tfrac{1}{n}\boldsymbol{J}\right)\boldsymbol{\mu} &= \boldsymbol{\beta}^t\boldsymbol{X}^t\left(\boldsymbol{H} - \tfrac{1}{n}\boldsymbol{J}\right)\boldsymbol{X}\boldsymbol{\beta} \\ &= \boldsymbol{\beta}^t\left(\boldsymbol{X}^t\boldsymbol{H}\boldsymbol{X} - \tfrac{1}{n}\boldsymbol{X}^t\boldsymbol{J}\boldsymbol{X}\right)\boldsymbol{\beta} \\ &= \boldsymbol{\beta}^t\left(\boldsymbol{X}^t\boldsymbol{X} - \tfrac{1}{n}\boldsymbol{X}^t\boldsymbol{J}\boldsymbol{X}\right)\boldsymbol{\beta} \\ &= \boldsymbol{\beta}^t\boldsymbol{X}^t\left(\boldsymbol{I} - \tfrac{1}{n}\boldsymbol{J}\right)\boldsymbol{X}\boldsymbol{\beta} \\ &\equiv \delta(\boldsymbol{\beta})\end{aligned}$$

또한

$$\begin{aligned}r\left(\boldsymbol{H} - \tfrac{1}{n}\boldsymbol{J}\right) &= tr(\boldsymbol{H}) - tr\left[\tfrac{1}{n}\boldsymbol{J}\right] \\ &= p - \tfrac{1}{n}tr(\mathbf{1}\mathbf{1}^t) \\ &= p - \tfrac{1}{n}tr(\mathbf{1}^t\mathbf{1}) \\ &= p - \tfrac{1}{n}n \\ &= p - 1 \\ &= p - 1\end{aligned}$$

위의 결과를 종합하면 회귀제곱합 SSR은 다음과 같은 분포를 따른다.

$$\frac{SSR}{\sigma^2} \sim \chi^2(p-1, \lambda^2), \tag{2.3}$$

위에서 비중심 모수는 다음과 같다.

$$\lambda^2 = \tfrac{1}{\sigma^2}\delta(\boldsymbol{\beta}) = \tfrac{1}{\sigma^2}\boldsymbol{\beta}^t\boldsymbol{X}^t\left(\boldsymbol{I} - \tfrac{1}{n}\boldsymbol{J}\right)\boldsymbol{X}\boldsymbol{\beta} \tag{2.4}$$

2.1.3. 잔차제곱합과 회귀제곱합의 독립

잔차제곱합과 회귀제곱합에서 나타난 이차형식의 두 멱등행렬의 곱은 $\mathbf{0}$이다.

$$\begin{aligned}(\boldsymbol{I}-\boldsymbol{H})\left(\boldsymbol{H}-\tfrac{1}{n}\boldsymbol{J}\right) &= \boldsymbol{H}-\tfrac{1}{n}\boldsymbol{J}-\boldsymbol{H}\boldsymbol{H}+\tfrac{1}{n}\boldsymbol{H}\boldsymbol{J} \\ &= \boldsymbol{H}-\tfrac{1}{n}\boldsymbol{J}-\boldsymbol{H}+\tfrac{1}{n}\boldsymbol{J} \\ &= \mathbf{0}\end{aligned}$$

따라서 부록의 정리에 의하여 잔차제곱합(SSE)과 회귀제곱합(SSR)은 서로 독립이다.

2.1.4. 총제곱합의 분포

총제곱합 SST의 분포는 위의 결과들을 이용하면 쉽게 구할 수 있다.

$$SST=\sum_{i=1}^{n}(y_i-\bar{y})^2=\boldsymbol{y}^t\left(\boldsymbol{I}-\tfrac{1}{n}\boldsymbol{J}\right)\boldsymbol{y}$$

위의 결과를 종합하면 회귀제곱합 SST은 다음과 같은 분포를 따른다.

$$\frac{SST}{\sigma^2}\sim\chi^2(n-1,\lambda^2), \tag{2.5}$$

위에서 비중심 모수 λ^2은 식 2.4 과 같다.

2.2. 모분산의 추정

최소제곱법을 통해서 회귀분석을 실시하였을때 우리는 적합된 회귀선이 얼마나 실제 관측값들을 잘 설명하고 있는지를 파악하는 것이 모형의 유용성을 판단하는데 중요한 작업이다. 즉, 적합된 회귀선이 관측값을 예측할 때의 변동성을 측정하는 것이 중요하다. 그 변동의 정도를 나타내는 것이 모분산 σ^2의 추정이다.

식 2.2 에 나타난 잔차제곱합의 분포를 이용하면 다음과 같은 결과를 얻는다.

$$E\left[\frac{SSE}{\sigma^2}\right]=n-p$$

위의 방정식에 적률법(Method of Moments)를 적용하면 모분산 σ^2에 대한 불편추정량을 얻을 수 있다. 평균 잔차 제곱합(mean residual sum of square; S^2 또는 MSE)를 다음과 같이 정의하자.

$$MSE=\frac{SSE}{n-p}=\frac{\sum r_i^2}{n-p}=\frac{\sum(y_i-\hat{y}_i)^2}{n-p}\equiv s^2 \tag{2.6}$$

$S^2=MSE$은 모분산의 불편 추정량이다.

$$E(s^2) = E(MSE) = \sigma^2$$

모분산의 추정량이 작을수록 관측값 y의 변동 중 회귀식이 설명할 수 변동이 크다는 것을 나타낸다. 관측값들이 회귀식으로부터 멀리 떨어져 있으면 MSE 는 커진다.

회귀계수들의 공분산을 추정하는 경우에도 s^2이 사용된다.

$$\hat{V}(\hat{\boldsymbol{\beta}}) = \hat{\sigma}^2(\boldsymbol{X}^t\boldsymbol{X})^{-1} = s^2(\boldsymbol{X}^t\boldsymbol{X})^{-1}$$

2.3. 최소제곱 추정량의 성질

최소제곱 추정량의 분포에 대한 성질은 다음과 같다.

- $\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^t\boldsymbol{X})^{-1})$
- $\hat{\boldsymbol{\beta}}$와 SSE는 독립이다.
- 잔차제곱합(SSE)과 회귀제곱합(SSR)은 서로 독립이다.
- SSE/σ^2는 자유도가 $n-p$인 카이제곱 분포를 따른다.
- $(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^t(\boldsymbol{X}^t\boldsymbol{X})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})/\sigma^2$ 는 자유도가 p인 카이제곱분포를 따른다.

2.4. 모형의 적합도 검정과 분산분석

회귀식을 적합하고 가장 먼저 고려해야할 사항은 적합된 회귀식이 유의한 의미를 가지는지 알아보는 것이다. 회귀식이 가지고 있는 의미는 설명변수의 변화에 따라서 반응변수가 변한다는 것이다. 따라서 회귀 모형이 유의하다는 것은 최소한 하나 이상의 설명변수가 반응변수의 변화를 예측하는데 의미가 있다는 것을 뜻한다. 모든 회귀계수의 값이 0이면 반응변수를 예측하는데 모든 설명변수가 필요가 없다는 것을 의미한다. 이러한 무의미한 모형은 앞장에서 나온 평균모형 1.6 이다.

이제 제시된 회귀식이 유의한 지에 대한 검정은 다음과 같은 두 가설 중 하나를 선택하는 것이다.

$$H_0 : \text{mean model} \quad vs. \quad H_1 : \text{ not } H_0$$

위의 가설을 바꾸어 쓰면 선형 회귀모형의 유의성 또는 적합도을 검정하는 가설이 된다.

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_{p-1} = 0 \quad vs. \quad H_1 : \text{ At least one of } \beta_i \text{ is not equal to } 0 \tag{2.7}$$

위의 가설 2.7 를 검정하는 방법이 분산분석표를 이용한 F-검정이다.

가설 2.7 에서 귀무가설 H_0가 참인 경우는

$$\boldsymbol{X\beta} = [\mathbf{1}\ \boldsymbol{x}_1\ \ldots\ \boldsymbol{x}_{p-1}] \begin{bmatrix} \beta_0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \beta_0 \mathbf{1}$$

이 성립하여 식 2.4 에 나타난 비중심 모수가 0이 된다.

$$\lambda^2 = \frac{1}{\sigma^2}\boldsymbol{\beta}^t \boldsymbol{X}^t \left(\boldsymbol{I} - \frac{1}{n}\boldsymbol{J}\right)\boldsymbol{X\beta} = \frac{\beta_0^2}{\sigma^2}\mathbf{1}^t \left(\boldsymbol{I} - \frac{1}{n}\boldsymbol{J}\right)\mathbf{1} = \mathbf{0}$$

따라서 귀무가설에서는 회귀제곱합이 자유도가 $p-1$인 중신ㅁ 카이제곱 분포를 따르게 되고 잔차제곱합과 독립이므로 다음의 통계량 F_0가 자유도가 $p-1$가ㄴ $n-p$를 가지는 F-분포를 따른다.

$$F_0 = \frac{SSR/(p-1)}{SSE/(n-p)} = \frac{MSR}{MSE} \sim F(p-1, n-p) \quad \text{under } H_0 \tag{2.8}$$

따라서 위의 검정 통계량의 p-값이 유의수준보다 크면 적합성 검정에 대한 가설 2.7 의 귀무가설을 기각한다. 귀무가설의 기각은 회귀모형의 계수 중 적어도 하나는 0이 아니므로 회귀 모형이 유의하다는 의미이다.

위에서 안급한 F-검정을 위한 통계량들은 다음과 같은 분산분석(Analysis of Variance; ANOVA) 표를 사용하면 쉽게 계산할 수 있다.

표 2.1.: 적합도 검정을 위한 분산분석표

요인	제곱합	자유도	평균제곱합	F-통계량	p-값
회귀	SSR	$p-1$	MSR	$F_0 = \frac{MSR}{MSE}$	$P(F > F_0)$
오차	SSE	$n-p$	MSE		
전체	SST	$n-1$			

3. 모형의 비교

3.1. 직교하는 설명 변수

다음과 같은 2개의 설명변수(x_1, x_2) 와 반응변수 y 를 가진 자료(데이터프레임)이 있다고 하자.

```
x1 <- c(1,  1, 1,  1, -1, -1, -1, -1)
x2 <- c(1, -1, 1, -1, 1, -1, 1, -1)
y <- c( 2, 5, 3, 4, 6, 9, 5, 10)
df <- data.frame(x1, x2, y)
df
```

```
  x1 x2  y
1  1  1  2
2  1 -1  5
3  1  1  3
4  1 -1  4
5 -1  1  6
6 -1 -1  9
7 -1  1  5
8 -1 -1 10
```

이제 위의 자료로 선형회귀모형을 적합해 보자.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + e_i \tag{3.1}$$

```
fm1 <- lm(y ~ x1 + x2, data=df)
summary(fm1)$coefficients
```

```
            Estimate Std. Error   t value     Pr(>|t|)
(Intercept)      5.5  0.3162278 17.392527 0.0000115141
x1              -2.0  0.3162278 -6.324555 0.0014565818
x2              -1.5  0.3162278 -4.743416 0.0051344617
```

이제 모형 3.1 에서 각각 x_1과 x_2를 제거한 축소된 모형을 적합해보자

$$y_i = \beta_0 + \beta_1 x_{i1} + e_i, \qquad y_i = \beta_0 + \beta_2 x_{i2} + e_i \tag{3.2}$$

```
fm21 <- lm(y ~ x1 , data=df)
summary(fm21)$coefficients
```

```
            Estimate Std. Error   t value     Pr(>|t|)
(Intercept)      5.5  0.6770032  8.124038 0.0001867963
x1              -2.0  0.6770032 -2.954196 0.0254739283
```

```
fm22 <- lm(y ~ x2 , data=df)
summary(fm22)$coefficients
```

```
            Estimate Std. Error   t value     Pr(>|t|)
(Intercept)      5.5  0.8660254  6.350853 0.0007143845
x2              -1.5  0.8660254 -1.732051 0.1339745962
```

두 개의 독립변수가 있는 모형 3.1 에서 하나의 독립변수를 제거해도 남아 있는 독립 변수의 회귀계수 추정량은 모형 3.1 과 같은 것을 알 수 있다. 이렇게 여러 개의 독립변수가 있는 모형에서 하나의 변수를 제거해도 다른 복립변수의 추정에 영향을 미치지 않는 경우는 어떤 경우일까?

이제 모형 3.1 의 설계행렬($\boldsymbol{X}$, design matrix)를 구해서 $\boldsymbol{X}^t\boldsymbol{X}$를 구해보자.

```
X <- model.matrix(fm1)
X
```

```
  (Intercept) x1 x2
1           1  1  1
2           1  1 -1
3           1  1  1
4           1  1 -1
5           1 -1  1
6           1 -1 -1
7           1 -1  1
8           1 -1 -1
attr(,"assign")
[1] 0 1 2
```

```
t(X) %*% X
```

```
            (Intercept) x1 x2
(Intercept)           8  0  0
x1                    0  8  0
x2                    0  0  8
```

모형 3.1 의 설계행렬 $\boldsymbol{X}$의 각 열들은 서로 직교하는것을 알 수 있다.

만약 여러 개의 독립변수를 가진 선형모형에서 모든 설명 변수들의 열들이 모두 서로 직교한다면(절편에 대한 열도 포함해서) 회귀계수의 추정값은 독립변수가 줄어든 축소된 모형에서도 원래의 모형과 같은 것을 알 수 있다.

선형 모형에서 설계행렬 $\boldsymbol{X}$의 각 열벡터를 각각 $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p$라고 하자. 만약 모든 열들이 서로 직교한다면 (즉 $\boldsymbol{x}_i^t \boldsymbol{x}_j = 0$ for $i \neq j$) 선형회귀 모형에서 회귀계수의 추정치는 설명 변수의 유무에 관계없이 일정하게 나타난다.

이러한 상황을 모형식으로 다시 써보자. 만약 다음이 성립하면

$$\boldsymbol{X}^t\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1^t \\ \boldsymbol{x}_2^t \\ \vdots \\ \boldsymbol{x}_p^t \end{bmatrix} \begin{bmatrix} \boldsymbol{x}_1 & \boldsymbol{x}_2 & \ldots & \boldsymbol{x}_p \end{bmatrix} = \begin{bmatrix} \boldsymbol{x}_1^t\boldsymbol{x}_1 & 0 & 0 & \cdots & 0 \\ 0 & \boldsymbol{x}_2^t\boldsymbol{x}_2 & 0 & \cdots & 0 \\ 0 & 0 & \boldsymbol{x}_3^t\boldsymbol{x}_3 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & 0 \\ 0 & 0 & \cdots & 0 & \boldsymbol{x}_p^t\boldsymbol{x}_p \end{bmatrix}$$

회귀계수의 추정량은 다음과 같이 나타난다.

$$\left(\boldsymbol{X}^t\boldsymbol{X}\right)^{-1}\boldsymbol{X}^t\boldsymbol{y} = \begin{bmatrix} \frac{1}{\boldsymbol{x}_1^t\boldsymbol{x}_1} & 0 & 0 & \cdots & 0 \\ 0 & \frac{1}{\boldsymbol{x}_2^t\boldsymbol{x}_2} & 0 & \cdots & 0 \\ 0 & 0 & \frac{1}{\boldsymbol{x}_3^t\boldsymbol{x}_3} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & 0 \\ 0 & 0 & \cdots & 0 & \frac{1}{\boldsymbol{x}_p^t\boldsymbol{x}_p} \end{bmatrix} \begin{bmatrix} \boldsymbol{x}_1^t\boldsymbol{y} \\ \boldsymbol{x}_2^t\boldsymbol{y} \\ \boldsymbol{x}_3^t\boldsymbol{y} \\ \vdots \\ \boldsymbol{x}_p^t\boldsymbol{y} \end{bmatrix} = \begin{bmatrix} \frac{\boldsymbol{x}_1^t\boldsymbol{y}}{\boldsymbol{x}_1^t\boldsymbol{x}_1} \\ \frac{\boldsymbol{x}_2^t\boldsymbol{y}}{\boldsymbol{x}_2^t\boldsymbol{x}_2} \\ \vdots \\ \frac{\boldsymbol{x}_p^t\boldsymbol{y}}{\boldsymbol{x}_p^t\boldsymbol{x}_p} \end{bmatrix}$$

위의 결과를 조금 더 일반화해보자. 만약 계획행렬 $\boldsymbol{X}$를 다음과 같은 p개의 부분 계획행렬 $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_p$로 나누고

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e} = \sum_{k=1}^{p} \boldsymbol{X}_k\boldsymbol{\beta}_k + \boldsymbol{e}$$

부분 계획행렬들이 다음과 같은 성질을 가지고 있다고 하자.

$$\boldsymbol{X} = [\boldsymbol{X}_1\ \boldsymbol{X}_2\ \ldots\ \boldsymbol{X}_p] \quad \text{and} \quad \boldsymbol{X}_i^t\boldsymbol{X}_j = \boldsymbol{0}, i \neq j \tag{3.3}$$

이러한 조건 하에서는 회귀계수의 추정량이 다음과 같이 나타난다.

$$
\begin{aligned}
\hat{\boldsymbol{\beta}} &= \left(\boldsymbol{X}^t\boldsymbol{X}\right)^{-1}\boldsymbol{X}^t\boldsymbol{y} \\
&= \begin{bmatrix}
(\boldsymbol{X}_1^t\boldsymbol{X}_1)^{-1} & 0 & 0 & \cdots & 0 \\
0 & (\boldsymbol{X}_2^t\boldsymbol{X}_2)^{-1} & 0 & \cdots & 0 \\
0 & 0 & (\boldsymbol{X}_3^t\boldsymbol{X}_3)^{-1} & \cdots & 0 \\
\vdots & \vdots & \vdots & \vdots & 0 \\
0 & 0 & \cdots & 0 & (\boldsymbol{X}_p^t\boldsymbol{X}_p)^{-1}
\end{bmatrix}
\begin{bmatrix}
\boldsymbol{X}_1^t\boldsymbol{y} \\ \boldsymbol{X}_2^t\boldsymbol{y} \\ \boldsymbol{X}_3^t\boldsymbol{y} \\ \vdots \\ \boldsymbol{X}_p^t\boldsymbol{y}
\end{bmatrix} \\
&= \begin{bmatrix}
(\boldsymbol{X}_1^t\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^t\boldsymbol{y} \\
(\boldsymbol{X}_2^t\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2^t\boldsymbol{y} \\
\vdots \\
(\boldsymbol{X}_p^t\boldsymbol{X}_p)^{-1}\boldsymbol{X}_p^t\boldsymbol{y}
\end{bmatrix} \\
&= \begin{bmatrix}
\hat{\boldsymbol{\beta}}_1 \\ \hat{\boldsymbol{\beta}}_2 \\ \hat{\boldsymbol{\beta}}_3 \\ \vdots \\ \hat{\boldsymbol{\beta}}_p
\end{bmatrix}
\end{aligned}
$$

위의 결과는 전체 모형에서의 추정량 $\hat{\boldsymbol{\beta}}$의 j 번째 부분 $\hat{\boldsymbol{\beta}}_j$이 $(\boldsymbol{X}_j^t\boldsymbol{X}_j)^{-1}\boldsymbol{X}_j^t\boldsymbol{y}$로 구성되며 다른 $\boldsymbol{X}_i$와는 관계가 없다.

이러한 결과를 이용하면 설명변수들이 서로 직교하는 조건 3.3 를 만족하면 축소모형

$$\boldsymbol{y} = \boldsymbol{X}_j\boldsymbol{\beta}_j + \boldsymbol{e}$$

에서 계수 추정치 $\hat{\boldsymbol{\beta}}_j = (\boldsymbol{X}_j^t\boldsymbol{X}_j)^{-1}\boldsymbol{X}_j^t\boldsymbol{y}$ 는 모든 설명 변수를 고려한 완전 모형에서의 추정치와 같은 것을 알 수 있다. 설명변수들이 서로 직교하는 조건 3.3 을 만족하면 하나의 축소모형에 대한 추정량는 직교하는 다른 설명변수들의 영향을 받지 않는다.

더 나아가서 직교하는 설명변수들이 회귀제곱합에 미치는 영향은 각 축소모형들의 기여도를 단순하게 더한 결과와 같다.

$$
\begin{aligned}
SSR &= \boldsymbol{y}^t\left(\boldsymbol{H} - \tfrac{1}{n}\boldsymbol{J}\right)\boldsymbol{y} \\
&= \boldsymbol{y}^t\boldsymbol{H}\boldsymbol{y} - \tfrac{1}{n}\boldsymbol{y}^t\boldsymbol{J}\boldsymbol{y} \\
&= \boldsymbol{y}^t\boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{y} - n(\bar{y})^2 \\
&= \boldsymbol{y}^t\boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{X})^{-1}(\boldsymbol{X}^t\boldsymbol{X})(\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{y} - n(\bar{y})^2 \\
&= \hat{\boldsymbol{\beta}}^t(\boldsymbol{X}^t\boldsymbol{X})\hat{\boldsymbol{\beta}} - n(\bar{y})^2 \\
&= \sum_{k=1}^{p}\hat{\boldsymbol{\beta}}_k^t(\boldsymbol{X}_k^t\boldsymbol{X}_k)\hat{\boldsymbol{\beta}}_k - n(\bar{y})^2
\end{aligned}
$$

또한 회귀계수 추정량의 분산을 보면 부분 회귀계수 추정량 $\hat{\boldsymbol{\beta}}_j$ 들은 서로 독립이며 축소 모형에서의 분산과 동일함을 알 수 있다.

$$
\begin{aligned}
Cov(\hat{\boldsymbol{\beta}}) &= \sigma^2(\boldsymbol{X}^t\boldsymbol{X})^{-1} \\
&= \begin{bmatrix}
\sigma^2(\boldsymbol{X}_1^t\boldsymbol{X}_1)^{-1} & 0 & 0 & \cdots & 0 \\
0 & \sigma^2(\boldsymbol{X}_2^t\boldsymbol{X}_2)^{-1} & 0 & \cdots & 0 \\
0 & 0 & \sigma^2(\boldsymbol{X}_3^t\boldsymbol{X}_3)^{-1} & \cdots & 0 \\
\vdots & \vdots & \vdots & \vdots & 0 \\
0 & 0 & \cdots & 0 & \sigma^2(\boldsymbol{X}_p^t\boldsymbol{X}_p)^{-1}
\end{bmatrix} \\
&= \begin{bmatrix}
Cov(\hat{\boldsymbol{\beta}}_1) & 0 & 0 & \cdots & 0 \\
0 & Cov(\hat{\boldsymbol{\beta}}_2) & 0 & \cdots & 0 \\
0 & 0 & Cov(\hat{\boldsymbol{\beta}}_3) & \cdots & 0 \\
\vdots & \vdots & \vdots & \vdots & 0 \\
0 & 0 & \cdots & 0 & Cov(\hat{\boldsymbol{\beta}}_p)
\end{bmatrix}
\end{aligned}
$$

3.2. 설명변수의 추가

먼저 계획행렬 $\boldsymbol{X}_1$을 고려한 선형모형을 고려하자.

$$\boldsymbol{y} = \boldsymbol{X}_1\beta_{1*} + \boldsymbol{e} \tag{3.4}$$

이 경우 회귀계수의 최소제곱추정량은 $\hat{\boldsymbol{\beta}}_{1*} = (\boldsymbol{X}_1^t\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^t\boldsymbol{y}$ 이다.

이제 위의 모형 3.4 에 설명변수를 추가한 모형을 생각해 보자. 추가된 설명변수로 이루어진 계획행렬을 $\boldsymbol{X}_2$라고 하면 다음과 같이 쓸수 있다.

$$
\begin{aligned}
\boldsymbol{y} &= \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{e} \\
&= [\boldsymbol{X}_1\ \boldsymbol{X}_2]\begin{bmatrix}\boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2\end{bmatrix} + \boldsymbol{e} \\
&= \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}
\end{aligned} \tag{3.5}
$$

위의 식에서

$$\boldsymbol{X} = [\boldsymbol{X}_1\ \boldsymbol{X}_2] \quad \text{and} \quad \boldsymbol{\beta} = \begin{bmatrix}\boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2\end{bmatrix}$$

설명변수를 추가한 확대 모형 3.5 에서 회귀계수의 최소제곱추정량은 $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{y}$이다.

여기서 주의할 점은 3.4 의 회귀 계수 $\boldsymbol{\beta}_{1*}$ 의 추정량과 3.5 의 회귀 계수 $\boldsymbol{\beta}_1$ 의 추정량은 일반적으로 같지 않다.

확대모형 3.5 의 회귀계수 추정량을 구하려면 최소제곱법을 다시 확장모형에 적용해야 하지만 원래의 모형 3.4 에서 구해진 추정량 $\hat{\boldsymbol{\beta}}_{1*} = (\boldsymbol{X}_1^t\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^t\boldsymbol{y}$ 을 이용하여 유도할 수 있다.

이제 원래의 모형 3.4 에서 모자행렬을

$$\boldsymbol{H}_1 = \boldsymbol{X}_1(\boldsymbol{X}_1^t\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^t$$

라고 하고 확대 모형 3.5 에 대하여 다음과 같이 모형을 다시 표현해보자.

$$\begin{aligned}
\boldsymbol{y} &= \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{e} \\
&= \boldsymbol{X}_1\boldsymbol{\beta}_1 + (\boldsymbol{H}_1 + \boldsymbol{I} - \boldsymbol{H}_1)\boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{e} \\
&= \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{H}_1\boldsymbol{X}_2\boldsymbol{\beta}_2 + (\boldsymbol{I} - \boldsymbol{H}_1)\boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{e} \\
&= [\boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{H}_1\boldsymbol{X}_2\boldsymbol{\beta}_2] + (\boldsymbol{I} - \boldsymbol{H}_1)\boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{e} \\
&= \boldsymbol{X}_1[\boldsymbol{\beta}_1 + (\boldsymbol{X}_1^t\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^t\boldsymbol{X}_2\boldsymbol{\beta}_2] + \tilde{\boldsymbol{X}}_2\boldsymbol{\beta}_2 + \boldsymbol{e} \\
&= \boldsymbol{X}_1\tilde{\boldsymbol{\beta}}_1 + \tilde{\boldsymbol{X}}_2\boldsymbol{\beta}_2 + \boldsymbol{e}
\end{aligned}$$

이제 다음과 같은 변환된 모형을 고려하자.

$$\boldsymbol{y} = \boldsymbol{X}_1\tilde{\boldsymbol{\beta}}_1 + \tilde{\boldsymbol{X}}_2\boldsymbol{\beta}_2 + \boldsymbol{e} \tag{3.6}$$

위의 식에서 다음과 같이 새로운 계수벡터 $\tilde{\boldsymbol{\beta}}_1$ 와 변환된 계획행렬 $\tilde{\boldsymbol{X}}_2$를 정의하였다.

$$\tilde{\boldsymbol{\beta}}_1 = \boldsymbol{\beta}_1 + (\boldsymbol{X}_1^t\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^t\boldsymbol{X}_2\boldsymbol{\beta}_2, \quad \tilde{\boldsymbol{X}}_2 = (\boldsymbol{I} - \boldsymbol{H}_1)\boldsymbol{X}_2 \tag{3.7}$$

이제 변환된 모형 3.6 에서 두 계획행렬 $\boldsymbol{X}_1$과 $\tilde{\boldsymbol{X}}_2$가 서로 직교하는 것을 알 수 있다.

$$\boldsymbol{X}_1{}^t\tilde{\boldsymbol{X}}_2 = \boldsymbol{X}_1^t(\boldsymbol{I} - \boldsymbol{H}_1)\boldsymbol{X}_2 = \boldsymbol{X}_1^t(\boldsymbol{I} - \boldsymbol{X}_1(\boldsymbol{X}_1^t\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^t)\boldsymbol{X}_2 = \boldsymbol{0}$$

이제 확대된 모형 3.6 의 두 계획행렬 $\boldsymbol{X}_1$과 $\tilde{\boldsymbol{X}}_2$가 서로 직교하므로 앞에서 나온 직교하는 계획행렬에 대한 회귀계수에 대한 결과를 이용하면 다음과 같이 회귀계수 추정량을 얻을 수 있다.

이제 모형 3.6 의 회귀계수 $\tilde{\boldsymbol{\beta}_1}$ 과 $\boldsymbol{\beta}_2$의 추정량을 구해보면 다음과 같다.

$$\hat{\tilde{\boldsymbol{\beta}}}_1 = (\boldsymbol{X}_1^t\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^t\boldsymbol{y}, \quad \hat{\boldsymbol{\beta}}_2 = (\tilde{\boldsymbol{X}}_2^t\tilde{\boldsymbol{X}}_2)^{-1}\tilde{\boldsymbol{X}}_2^t\boldsymbol{y} \tag{3.8}$$

먼저 3.6 에서 $\boldsymbol{\beta}_2$ 에 대한 추정량 $\hat{\boldsymbol{\beta}}_2$는 반응변수 $\boldsymbol{y}$ 를 변환된 계획 행렬 $\tilde{\boldsymbol{X}}_2 = (\boldsymbol{I} - \boldsymbol{H}_1)\boldsymbol{X}_2$로 적합할 때의 회귀계수이다.

$$y = [(\boldsymbol{I} - \boldsymbol{H}_1)\boldsymbol{X}_2]\boldsymbol{\beta}_2 + \boldsymbol{e}$$

3.8 에 주어진 추정량 $\hat{\boldsymbol{\beta}}_2$ 를 다시 다음과 같이 유도할 수 있다.

$$
\begin{aligned}
\hat{\boldsymbol{\beta}}_2 &= (\tilde{\boldsymbol{X}}_2^t \tilde{\boldsymbol{X}}_2)^{-1} \tilde{\boldsymbol{X}}_2^t \boldsymbol{y} \\
&= [\boldsymbol{X}_2^t (\boldsymbol{I} - \boldsymbol{H}_1)(\boldsymbol{I} - \boldsymbol{H}_1) \boldsymbol{X}_2]^{-1} \boldsymbol{X}_2^t (\boldsymbol{I} - \boldsymbol{H}_1) \boldsymbol{y} \\
&= [\boldsymbol{X}_2^t (\boldsymbol{I} - \boldsymbol{H}_1)(\boldsymbol{I} - \boldsymbol{H}_1) \boldsymbol{X}_2]^{-1} \boldsymbol{X}_2^t (\boldsymbol{I} - \boldsymbol{H}_1)(\boldsymbol{I} - \boldsymbol{H}_1) \boldsymbol{y} \\
&= (\tilde{\boldsymbol{X}}_2^t \tilde{\boldsymbol{X}}_2)^{-1} \tilde{\boldsymbol{X}}_2^t [(\boldsymbol{I} - \boldsymbol{H}_1) \boldsymbol{y}] \\
&= (\tilde{\boldsymbol{X}}_2^t \tilde{\boldsymbol{X}}_2)^{-1} \tilde{\boldsymbol{X}}_2^t \tilde{\boldsymbol{y}}
\end{aligned}
$$

위의 유도를 보면 반응변수 $\boldsymbol{y}$ 에서 $\boldsymbol{X}_1$ 로 적합한 후에 구한 잔차벡터 $\tilde{\boldsymbol{y}} = (\boldsymbol{I} - \boldsymbol{H}_1)\boldsymbol{y}$ 를 새로운 반응변수로 고려한 후, 추가된 변수에 대한 계획행렬 $\boldsymbol{X}_2$ 에서 먼저 고려한 변수의 계획행렬 $\boldsymbol{X}_1$ 의 효과를 제거한 $(\boldsymbol{I} - \boldsymbol{H}_1)\boldsymbol{X}_2$ 로 적합한 경우의 회귀계수로 나타난다.

$$
\tilde{\boldsymbol{y}} = \tilde{\boldsymbol{X}}_2 \boldsymbol{\beta}_2 + \boldsymbol{e} \quad \rightarrow \quad (\boldsymbol{I} - \boldsymbol{H}_1) y = [(\boldsymbol{I} - \boldsymbol{H}_1)\boldsymbol{X}_2] \boldsymbol{\beta}_2 + \boldsymbol{e}
$$

또한 3.6 의 회귀계수 $\tilde{\boldsymbol{\beta}}_1$ 의 추정량은 직교성에 의하여 $(\boldsymbol{X}_1^t \boldsymbol{X}_1)^{-1} \boldsymbol{X}_1^t \boldsymbol{y}$ 으로 주어지며 이는 모형 3.4 에서 구한 회귀계수의 추정량과 같다.

$$
\hat{\tilde{\boldsymbol{\beta}}}_1 = \hat{\boldsymbol{\beta}}_{1*} = (\boldsymbol{X}_1^t \boldsymbol{X}_1)^{-1} \boldsymbol{X}_1^t \boldsymbol{y}
$$

이제 3.7 의 관계를 이용하면 모형 3.5 에서 나타난 회귀계수 $\boldsymbol{\beta}_1$ 의 추정량을 다음과 같이 표현할 수 있다.

$$
\begin{aligned}
\hat{\boldsymbol{\beta}}_1 &= \hat{\tilde{\boldsymbol{\beta}}}_1 - (\boldsymbol{X}_1^t \boldsymbol{X}_1)^{-1} \boldsymbol{X}_1^t \boldsymbol{X}_2 \hat{\boldsymbol{\beta}}_2 \\
&= \hat{\boldsymbol{\beta}}_{1*} - (\boldsymbol{X}_1^t \boldsymbol{X}_1)^{-1} \boldsymbol{X}_1^t \boldsymbol{X}_2 \hat{\boldsymbol{\beta}}_2 \\
&= (\boldsymbol{X}_1^t \boldsymbol{X}_1)^{-1} \boldsymbol{X}_1^t (\boldsymbol{y} - \boldsymbol{X}_2 \hat{\boldsymbol{\beta}}_2)
\end{aligned}
\tag{3.9}
$$

3.9 에 주어진 회귀계수 $\boldsymbol{\beta}_1$ 의 추정량은 새로운 변수를 추기하기 전의 모형에서 구한 추정량 $\hat{\boldsymbol{\beta}}_{1*}$ 을 새로운 변수를 추가한 후의 모형에서 추가된 변수에 대한 회귀계수의 추정량 $\hat{\boldsymbol{\beta}}_2$ 으로 보정힌 형태이다.

이제 간단한 예제를 통하여 위에서 유도한 공식을 적용해 보자.

먼저 3 개의 설명변수 x_1, x_2, x_3 를 가진 10 개의 자료를 임의로 만들어 보자.

```
set.seed(23123)
x1 <- c(1,2,3,4,5,6,7,8,9,9)
x2 <- c(1,4,2,5,3,2,4,3,1,2)
x3 <- c(6,3,2,3,1,4,5,3,2,1)
y <- 2 + 3*x1 + 4*x2 + 5*x3 + rnorm(10)

df <- data.frame(x1,x2,x3,y)
df
```

```
   x1 x2 x3        y
1   1  1  6 38.99584
2   2  4  3 38.17609
3   3  2  2 27.13590
4   4  5  3 49.86576
5   5  3  1 32.88329
6   6  2  4 48.43808
7   7  4  5 63.30271
8   8  3  3 53.54604
9   9  1  2 41.86736
10  9  2  1 43.30445
```

먼저 2개의 독립변수 x_1 과 x_2 가 있는 모형을 적합해 보자.

$$y = \beta_{0*} + \beta_{1*}x_1 + \beta_{2*}x_2 + e \tag{3.10}$$

```
lm1s <- lm(y ~ x1 + x2, data=df)
lm1s$coefficients
```

```
(Intercept)          x1          x2
  22.964557    1.948430    3.802026
```

또한 모형 3.10 에서 사용한 계획행렬 $\boldsymbol{X}_1$ 을 구해보자

```
X1 <- model.matrix(lm1s)
X1
```

```
   (Intercept) x1 x2
1            1  1  1
2            1  2  4
3            1  3  2
4            1  4  5
5            1  5  3
6            1  6  2
7            1  7  4
8            1  8  3
9            1  9  1
10           1  9  2
attr(,"assign")
[1] 0 1 2
```

다음으로 모든 독립변수가 있는 모형을 적합해 보자.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + e \tag{3.11}$$

```
lmAll <- lm(y ~ x1 + x2 + x3, data=df)
lmAll$coefficients
```

```
(Intercept)          x1          x2          x3
-0.05735613  3.15491311  4.16172306  5.17857500
```

또한 모형 3.11 에서 모든 독립변수를 사용한 경우 계획행렬 $\boldsymbol{X}$ 와 추가된 변수 x_3 에 대한 열만 가지는 계획행렬 $\boldsymbol{X}_2$ 를 구해보자

```
X <- model.matrix(lmAll)
X
```

```
   (Intercept) x1 x2 x3
1            1  1  1  6
2            1  2  4  3
3            1  3  2  2
4            1  4  5  3
5            1  5  3  1
6            1  6  2  4
7            1  7  4  5
8            1  8  3  3
9            1  9  1  2
10           1  9  2  1
attr(,"assign")
[1] 0 1 2 3
```

```
X2 <- X[,4]
X2
```

```
 1  2  3  4  5  6  7  8  9 10
 6  3  2  3  1  4  5  3  2  1
```

이제 3.7 에 주어진 $\tilde{\boldsymbol{X}}_2$ 를 계산하고 이를 이용하여 $\hat{\boldsymbol{\beta}}_2$ 를 구해보자.

```
H1 <- X1 %*% solve(t(X1) %*% X1) %*% t(X1)
X2t <- (diag(10) - H1) %*% X2
X2t
```

```
          [,1]
1    1.8568267
2   -0.7018216
3   -1.6077630
4   -0.1664113
5   -2.0723527
6    1.0911645
7    2.4630575
8    0.6265747
9   -0.2793667
10 -1.2099081
```

```
beta2 <- solve(t(X2t) %*% X2t) %*% t(X2t) %*% y
beta2
```

```
          [,1]
[1,] 5.178575
```

위에서 구한 회귀계수 beta2 는 모든 독립 변수가 있는 모형에서의 x_3 에 대한 회귀계수 추정량과 같다.

이제 절편, x_1, x_2 만 있는 모형에서 구한 회귀계수를 위에서 구한 beta2를 이용하여 보정해 보자. 아래 보정된 회귀계수 추정량은 모든 독립변수를 고려한 모형에서의 회귀계수 추정량과 같다.

```
beta1 <- lm1s$coefficients - solve(t(X1) %*% X1) %*% t(X1) %*% X2 %*% beta2
beta1
```

```
                   [,1]
(Intercept) -0.05735613
x1           3.15491311
x2           4.16172306
```

이제 앞에서 본 예제와 같이 특별하게 1개의 설명변수를 추가하는 경우를 알아보자. 이 경우는 추가된 변수에 대한 계획행렬 $\boldsymbol{X}_2 = \boldsymbol{x}_p$는 하나의 벡터이다. 따라서

$$\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 = \boldsymbol{X}_1\boldsymbol{\beta} + \boldsymbol{x}_p\beta_p + \boldsymbol{e} \tag{3.12}$$

위의 3.8 에서

$$\tilde{\boldsymbol{X}}_2 = (\boldsymbol{I} - \boldsymbol{H}_1)\boldsymbol{x}_p \equiv \tilde{\boldsymbol{x}}_p$$

로 정의하면 회귀식 3.12 에서 하나 추가된 설명변수에 대한 회귀계수의 추정량은 다음과 같다.

$$\begin{aligned}
\hat{\beta}_p &= (\tilde{\boldsymbol{X}}_2^t \tilde{\boldsymbol{X}}_2)^{-1} \tilde{\boldsymbol{X}}_2^t \boldsymbol{y} \\
&= [\tilde{\boldsymbol{x}}_p^t \boldsymbol{x}_p]^{-1} \tilde{\boldsymbol{x}}_p^t \boldsymbol{y} \\
&= [\tilde{\boldsymbol{x}}_p^t \boldsymbol{x}_p]^{-1} \boldsymbol{x}_p^t (\boldsymbol{I} - \boldsymbol{H}_1) \boldsymbol{y} \\
&= [\tilde{\boldsymbol{x}}_p^t \boldsymbol{x}_p]^{-1} \boldsymbol{x}_p^t (\boldsymbol{I} - \boldsymbol{H}_1)(\boldsymbol{I} - \boldsymbol{H}_1) \boldsymbol{y} \\
&= [\tilde{\boldsymbol{x}}_p^t \boldsymbol{x}_p]^{-1} \tilde{\boldsymbol{x}}_p^t (\boldsymbol{I} - \boldsymbol{H}_1) \boldsymbol{y} \\
&= [\tilde{\boldsymbol{x}}_p^t \boldsymbol{x}_p]^{-1} \tilde{\boldsymbol{x}}_p^t \tilde{\boldsymbol{y}} \\
&= \frac{\tilde{\boldsymbol{x}}_p^t \tilde{\boldsymbol{y}}}{\tilde{\boldsymbol{x}}_p^t \tilde{\boldsymbol{x}}_p}
\end{aligned}$$

위의 식에서 $\tilde{\boldsymbol{y}} = (\boldsymbol{I} - \boldsymbol{H}_1)\boldsymbol{y}$ 이다.

또한 3.12 과 같이 새로운 변수 x_p 가 추가되면 새로운 변수가 추가된 후에는 회귀계수 추정량이 다음과 같아 보정된다.

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_1 &= (\boldsymbol{X}_1^t \boldsymbol{X}_1)^{-1} \boldsymbol{X}_1{}^t [\boldsymbol{y} - \boldsymbol{X}_2 \hat{\boldsymbol{\beta}}_2] \\
&= (\boldsymbol{X}_1{}^t \boldsymbol{X}_1)^{-1} \boldsymbol{X}_1{}^t \boldsymbol{y} - (\boldsymbol{X}_1{}^t \boldsymbol{X})_1^{-1} \boldsymbol{X}_1^t \boldsymbol{x}_p \frac{\tilde{\boldsymbol{x}}_p^t \tilde{\boldsymbol{y}}}{\tilde{\boldsymbol{x}}_p^t \tilde{\boldsymbol{x}}_p} \\
&= (\boldsymbol{X}_1^t \boldsymbol{X})_1^{-1} \boldsymbol{X}_1^t \boldsymbol{y} - (\boldsymbol{X}_1^t \boldsymbol{X})_1^{-1} \boldsymbol{X}_1^t [\boldsymbol{x}_p (\tilde{\boldsymbol{x}}_p^t \tilde{\boldsymbol{x}}_p)^{-1} \tilde{\boldsymbol{x}}_p^t] \tilde{\boldsymbol{y}} \\
&= \hat{\boldsymbol{\beta}}_{1*} - (\boldsymbol{X}_1^t \boldsymbol{X})_1^{-1} \boldsymbol{X}_1^t [\boldsymbol{x}_p (\tilde{\boldsymbol{x}}_p^t \tilde{\boldsymbol{x}}_p)^{-1} \tilde{\boldsymbol{x}}_p^t] \tilde{\boldsymbol{y}}
\end{aligned}$$

3.3. 부분 F-검정과 가능도비 검정

앞 절에서 회귀모형에 새로운 독립변수를 1개 이상 추가할 경우 회귀계수 추정량과 제곱합의 변화를 살펴보있다.

실제 자료를 분석하여 회귀 모형식을 만드는 경우 일반적으로 중요한 몇 개의 설명변수부터 모형에 포함시키고 다른 변수들을 추가한다. 반대로 중요한 변수에 대한 사전 정보가 없다면 가능한 모든 변수를 모두 포함시킨 후에 중요하지 않은 변수들을 제거하기도 한다. 이런 두 가지 방법 모두 축차적으로 변수를 추가 또는 제거하는 방법으로 고려하는 모형들이 포함 관계를 가진다.

이렇게 포함관계를 가지는 두 모형을 고려해 보자. 먼저 설명변수의 수가 많은 모형을 최대 모형(full model)이라고 부르자.

$$y_i = \beta_0 + \beta_1 x_{i1} + \dots \beta_{p-1} x_{i,p} + \beta_p x_{i,p} + \cdots + \beta_{p+q} + e_i$$

위의 식은 모두 절편을 제외하면 모두 $p+q$ 개의 설명변수를 가진 선형 모형이다. 위의 최대모형을 다음과 같은 행렬식으로 써보자
아래에 정의된 최대

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}, , \quad \boldsymbol{e} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n) \quad \text{Full model} \tag{3.13}$$

이제 축소된 모형으로 최대모형에서 마자막 q개의 설명변수가 모형에 포함되지 않은 경우를 생각하자.

$$y_i = \beta_0 + \beta_1 x_{i1} + ... \beta_{p-1} x_{i,p} + \beta_p x_{i,p} + e_i$$

축소모형은 다음과 같은 행렬식으로 표시한다.

$$\boldsymbol{y} = \boldsymbol{X}_1 \boldsymbol{\beta}_1 + \boldsymbol{e}, , \quad \boldsymbol{e} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n) \quad \text{Reduced Model} \tag{3.14}$$

참고로 최대모형 3.13 의 계획행렬 $\boldsymbol{X} = [\boldsymbol{X}_1, \boldsymbol{X}_2]$ 의 차원은 $n \times (p+q+1)$이고 축소 모형 3.14 의 계획행렬 $\boldsymbol{X}_1$ 의 차원은 $n \times (p+1)$ 이다.

여기서 유의할 점은 최대 모형은 축소모형을 포함한다는 것이다. 만약 최대모형에서 마지막 q개의 설명변수들에 대한 회귀계수들이 모두 0이면, 즉 $\beta_{p+1} = \cdots \beta_{p+q} = 0$ 이면 축소모형이 된다.

교재에서는 추가제곱합을 이용한 부분 F-검정을 설명한다. 즉, 다음과 같은 가설을 검정하고자 한다.

$$H_0 : \beta_{p+1} = \beta_{p+2} = \cdots = \beta_{p+q} = 0 \text{ (Reduced)} \quad \text{vs.} \quad H_1 : \text{ not } H_0 \text{ (Full)} \tag{3.15}$$

위와 같은 가설을 검정하기 위한 부분 F-검정의 검정 통계량 F_0 는 다음과 같이 주어진다 (교과서 식 4.4).

$$F_0 = \frac{[SSE(R) - SSE(F)]/(df_R - df_F)}{SSE(F)/df_F} \tag{3.16}$$

만약 H_0 가 참이면 검정 통계량 F_0 는 자유도가 각각 $df_R - df_F$와 df_F 를 가지는 F-분포를 따르므로 이를 이용하여 모형에 새로운 변수를 추가하는 검정을 수행하는 부분 F-검정을 실시할 수 있다.

참고로 위의 식들에서 자유도 df_R 과 df_F는 다음과 같이 주어진다.

$$df_R = n - (p+q+1), \quad df_F = n - (p+1)$$

선형모형에 대한 최대 가능도 추정법은 Chapter 1 에서 설명하였다. 위의 최대 모형과 축소모형에 대한 최대가능도 추정법을 설명하기 위하여 다음과 같은 식을 사용할 것이다.

임의의 벡터 $\boldsymbol{v}$에 대하여 노름 $\|\boldsymbol{v}\|^2$ 을 다음과 같아 정의한다.

$$\|\boldsymbol{v}\|^2 = \boldsymbol{v}^t \boldsymbol{v}$$

최대모형에 대한 계획행렬 $\boldsymbol{X}$ 에 대한 모자행렬을 이용하여 사영행렬 $\boldsymbol{P}$ 와 $\boldsymbol{Q}$를 다음과 같이 정의한다.

$$\boldsymbol{P} \equiv \boldsymbol{H}(\boldsymbol{X}) = \boldsymbol{X}(\boldsymbol{X}^t \boldsymbol{X})^{-1} \boldsymbol{X}^t, \quad \boldsymbol{Q} = \boldsymbol{I} - \boldsymbol{P}$$

또한 축소모형의 계획행렬 $\boldsymbol{X}_1$에 대한 모자행렬을 이용하여 사영행렬 $\boldsymbol{P}_1$ 와 $\boldsymbol{Q}_1$를 다음과 같이 정의한다.

$$\boldsymbol{P}_1 \equiv \boldsymbol{H}_1(\boldsymbol{X}_1) = \boldsymbol{X}_1(\boldsymbol{X}_1^t\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^t, \quad \boldsymbol{Q}_1 = \boldsymbol{I} - \boldsymbol{P}_1$$

분산에 대한 모수를 $\tau = \sigma^2$ 과 같이 쓰고 편의상 반응 벡터의 평균을 다음과 같이 표시하자

$$\boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\beta}, \quad \boldsymbol{\mu}_1 = \boldsymbol{X}_1\boldsymbol{\beta}_1$$

이제 최대모형 3.13 에 대한 최대 가능도 추정을 생각해 보자.

$$\begin{aligned}
\ell_n(\boldsymbol{\theta}_F; \boldsymbol{y}) &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\tau - \frac{1}{2\tau}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^t(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \\
&= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\tau - \frac{1}{2\tau}\left\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right\|^2 \\
&= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\tau - \frac{1}{2\tau}\left\|\boldsymbol{y} - \boldsymbol{\mu}\right\|^2 \\
&= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\tau - \frac{1}{2\tau}\left\|\boldsymbol{y} - \boldsymbol{P}\boldsymbol{y} + \boldsymbol{P}\boldsymbol{y} - \boldsymbol{\mu}\right\|^2 \\
&= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\tau - \frac{1}{2\tau}\left[\left\|\boldsymbol{y} - \boldsymbol{P}\boldsymbol{y}\right\|^2 + \left\|\boldsymbol{P}\boldsymbol{y} - \boldsymbol{\mu}\right\|^2\right] \\
&= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\tau - \frac{1}{2\tau}\left[\left\|\boldsymbol{Q}\boldsymbol{y}\right\|^2 + \left\|\boldsymbol{P}\boldsymbol{y} - \boldsymbol{\mu}\right\|^2\right]
\end{aligned}$$

위의 최대 모형에 대한 로그 가능도 함수 $\ell_n(\boldsymbol{\theta}; \boldsymbol{y})$ 에서 $\mu = \boldsymbol{X}\boldsymbol{\beta}$ 에 대한 최대가능도 추정량과 모분산 τ 에 대한 추정량은 다음과 같아 주어진다(Chapter 1 참조)

$$\hat{\mu} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{P}\boldsymbol{y}, \quad \hat{\tau}_F = \frac{1}{n}\left\|\boldsymbol{Q}\boldsymbol{y}\right\|^2 = \frac{1}{n}SSE(F)$$

따라서 최대 가능도 추정량 $\hat{\boldsymbol{\mu}}$ 와 $\hat{\tau}_F$를 로그 가능도 함수에 넣으면 다음과 같은 결과가 주어진다.

$$\max_{\boldsymbol{\mu},\boldsymbol{\tau}} \ell_n(\boldsymbol{\theta}_F; \boldsymbol{y}) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\hat{\tau}_F - \frac{n}{2}$$

위의 결과를 가능도 함수로 표시하면

$$\max_{\boldsymbol{\mu},\boldsymbol{\tau}} L_n(\boldsymbol{\theta}_F; \boldsymbol{y}) = (2\pi e)^{-n/2}[SSE(F)/n]^{-n/2} \tag{3.17}$$

축소모형 3.14 에 대해서도 같은 방법으로 최대 가능도 춘정량을 구하면 다음과 같이 주어지며

$$\hat{\mu}_1 = \boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1 = \boldsymbol{P}_1\boldsymbol{y}, \quad \hat{\tau}_R = \frac{1}{n}\left\|\boldsymbol{Q}_1\boldsymbol{y}\right\|^2 = \frac{1}{n}SSE(R)$$

로그 가능도 함수에 넣어면 다음과 같은 결과가 주어진다.

$$\max_{\boldsymbol{\mu}_1,\boldsymbol{\tau}} \ell_n(\boldsymbol{\theta}_R; \boldsymbol{y}) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\hat{\tau}_R - \frac{n}{2}$$

위의 결과를 가능도 함수로 표시하면

$$\max_{\boldsymbol{\mu},\boldsymbol{\tau}} L_n(\boldsymbol{\theta}_R;\boldsymbol{y}) = (2\pi e)^{-n/2}[SSE(R)/n]^{-n/2} \tag{3.18}$$

최대 가능도 추정에서 만약 최대 모형 F 가 축소모형 R 을 포함하면 가설 3.15 에 대한 검정이 가능하다.

위의 가설 검정에서 두 모형의 가능도 함수 3.17 와 3.18 의 비, 즉 가능도 비(Likelihood Ratio) Λ 는 다음과 같이 주어진다.

$$\Lambda = \frac{\max_{\boldsymbol{\mu}_1,\boldsymbol{\tau}} L_n(\boldsymbol{\theta}_R;\boldsymbol{y})}{\max_{\boldsymbol{\mu},\boldsymbol{\tau}} L_n(\boldsymbol{\theta}_F;\boldsymbol{y})} = \left[\frac{SSE(F)}{SSE(R)}\right]^{n/2} \tag{3.19}$$

위의 가능도 비 Λ가 작으면 귀무가설 H_0을 기각한다.

$$\text{if } \Lambda < c', \text{ then reject } H_0$$

식 3.19 에 나타난 가능도 비를 다시 표현해 보자

$$\Lambda = \left[\frac{SSE(F)}{SSE(R)}\right]^{n/2} = \left[1 + \frac{SSE(R) - SSE(F)}{SSE(F)}\right]^{-n/2} \equiv (1 + F^*)^{-n/2}$$

위의 식을 보면 Λ가 F^*에 반비례하므로 다음과 같이 F^*가 크면 H_0를 기각할 수 있다. 이제 회귀분석에서 주로 쓰이는 부분 F-검정 통계량을 위의 식에서 나타난 F^*로 표시해보면 다음과 같이 쓸수 있다.

$$\text{if } F = \frac{n-(p+q+1)}{q} F^* = \frac{[SSE(R) - SSE(F)]/q}{SSE(F)/[n-(p+q+1)]} > c, \text{ then reject } H_0 \tag{3.20}$$

위의 3.20 에 있는 가설검정의 절차는 추가제곱합을 이용한 부분 F-검정(교과서 p.158-161)과 동일한 검정이다. 따라서 부분 F-검정은 가능도비 검정이다.

3.3.1. 부분 F-검정 통계량의 분포

위 3.16 에서 나온 부분 F-검정 통계량 F_0는 $H_0 : \boldsymbol{\beta}_2 = \mathbf{0}$ 가 참일 때 자유도가 각각 q와 $n-(p+q+1)$인 F-분포를 따른다. 검정통계량의 분포를 부록에서 배운 이차형식의 분포 이론으로 유도해 보자.

일단 검정통계량 3.16 에서 분자 $SSE(R) - SSE(F)$의 분포를 귀무가설 3.15 하에서 구해 보자. 두 잔차제곱합의 차는 다음과 같이 주어진다.

$$\begin{aligned} SSE(R) - SSE(F) &= \boldsymbol{y}^t(\boldsymbol{I} - \boldsymbol{P}_1)\boldsymbol{y} - \boldsymbol{y}^t(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{y} \\ &= \boldsymbol{y}^t(\boldsymbol{P} - \boldsymbol{P}_1)\boldsymbol{y} \end{aligned}$$

이 때 $\boldsymbol{P} - \boldsymbol{P}_1$ 은 대칭이며 멱등행렬임을 보일 수 있다.

$$\begin{aligned}(\boldsymbol{P}-\boldsymbol{P}_1)^2 &= \boldsymbol{P}^2-\boldsymbol{P}\boldsymbol{P}_1-\boldsymbol{P}_1\boldsymbol{P}+\boldsymbol{P}_1^2\\ &= \boldsymbol{P}-\boldsymbol{P}_1-\boldsymbol{P}_1+\boldsymbol{P}_1\\ &= \boldsymbol{P}-\boldsymbol{P}_1\end{aligned}$$

위의 결과는 부록 Section B.9 에서 설명한 포함관계에 있는 두 열공간의 사영에 대한 사실을 이용하여 유도한 것이다.

따라서 $SSE(R)-SSE(F)$의 분포는 다음과 같은 비중심 카이제곱 분포로 주어진다.

$$\frac{SSE(R)-SSE(F)}{\sigma^2} \sim \chi^2(q,\lambda_1^2)$$

여기서 자유도 $tr(\boldsymbol{P}-\boldsymbol{P}_1)=q$ 이고 비중심모수 λ_1 는 다음과 같이 주어진다.

$$\lambda_1^2 = \tfrac{1}{\sigma^2}\boldsymbol{\beta}^t\boldsymbol{X}^t\left(\boldsymbol{P}-\boldsymbol{P}_1\right)\boldsymbol{X}\boldsymbol{\beta}$$

만약 H_0가 참이면 $\boldsymbol{\beta}_2=\boldsymbol{0}$ 이므로

$$\boldsymbol{X}\boldsymbol{\beta}=\boldsymbol{X}_1\boldsymbol{\beta}_1+\boldsymbol{X}_2\boldsymbol{\beta}_2=\boldsymbol{X}_1\boldsymbol{\beta}_1+\boldsymbol{0}=\boldsymbol{X}_1\boldsymbol{\beta}_1$$

따라서

$$\begin{aligned}\left(\boldsymbol{P}-\boldsymbol{P}_1\right)\boldsymbol{X}\boldsymbol{\beta} &= \left(\boldsymbol{P}-\boldsymbol{P}_1\right)\boldsymbol{X}_1\boldsymbol{\beta}_1\\ &= \boldsymbol{P}\boldsymbol{X}_1\boldsymbol{\beta}_1-\boldsymbol{P}_1\boldsymbol{X}_1\boldsymbol{\beta}_1\\ &= \boldsymbol{X}_1\boldsymbol{\beta}_1-\boldsymbol{X}_1\boldsymbol{\beta}_1\\ &= \boldsymbol{0}\end{aligned}$$

따라서 비중심모수 λ_1는 0이 된다. 즉 H_0가 참이면 $SSE(R)-SSE(F)$는 자유도가 q 인 중심 카이제곱 분포를 따른다.

또한 $SSE(F)$는 자유도가 $n-(p+q+1)$ 인 중심 카이제곱 분포를 따르며 $SSE(R)-SSE(F)$ 는 독립이다. 왜냐하면 Theorem G.2 에 의하여

$$\begin{aligned}(\boldsymbol{I}-\boldsymbol{P})(\boldsymbol{P}-\boldsymbol{P}_1) &= \boldsymbol{P}-\boldsymbol{P}_1-\boldsymbol{P}\boldsymbol{P}+\boldsymbol{P}\boldsymbol{P}_1\\ &= \boldsymbol{P}-\boldsymbol{P}_1-\boldsymbol{P}+\boldsymbol{P}_1\\ &= \boldsymbol{0}\end{aligned}$$

4. 모형의 진단

4.1. 등분산성 가정의 위반

일반적인 회귀분석모형에서

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$$

오차항이 다음과 같이 서로 독립이고 등분산성을 만족한다면

$$var(\boldsymbol{e}) = \sigma^2 \boldsymbol{I}_n$$

최소제곱법에 의한 회귀계수 추정량 $\hat{\boldsymbol{\beta}}$ 다음과 같고

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{y}$$

이는 최소분산선형추정량(BLUE)이다. 만약에 오차항에 대한 가정이 만족하지 않는다면 최소제곱법 추정량 $\hat{\boldsymbol{\beta}}$의 최적성이 유지되는가에 대한 질문이 생기게 된다.

여기서 오차항의 분산에 대하여 좀더 일반적인 모형을 생각해보자. 가장 일반적인 모형은 다음과 같은 임의의 양정치행렬(positive definite matrix) $\boldsymbol{V}$가 오차항의 공분산 행렬인 경우이다.

$$Var(\boldsymbol{e}) = \boldsymbol{V}$$

가장 일반적인 경우를 고려하기 전에 전형적인 가정을 약간 벗어나면서 실제 문제에서 흔히 접하는 경우를 생각해 보자.

일단 오차항이 서로 독립이지만 분산이 다른 경우이다.

$$Var(\boldsymbol{e}) = \text{diag}[\sigma_1^2, \sigma_2^2, \dots, \sigma_n^2]$$

4.1.1. 가중 최소제곱법

이러한 경우에 가중 최소제곱법(Weighted Leadt Square Estimator; WLSE)을 사용하면 최소제곱법 추정량의 최적성을 유지할 수 있다.

오차항의 분산 $\sigma_1^2, \sigma_2^2, \dots, \sigma_n^2$을 안다고 가정하면 각 관측값 y_i를 해당 오차의 표분편차 σ_i로 나누면 등분산성을 다시 얻을 수 있다.

$$Var(y_i) = Var(\boldsymbol{x}_i^t\boldsymbol{\beta} + e_i) = Var(e_i) = \sigma_i^2$$

$$\Rightarrow Var(y_i/\sigma_i) = Var(\boldsymbol{x}_i^t\boldsymbol{\beta}' + e_i/\sigma_i) = Var(e_i/\sigma_i) = 1$$

이 때 새로운 관측치 $y_i' = y_i/\sigma_i$를 사용하여 최소제곱법을 적용하면

$$\min_{\boldsymbol{\beta}'}\sum_{i=1}^{n}(y_i' - \boldsymbol{x}_i^t\boldsymbol{\beta}')^2 = \min_{\boldsymbol{\beta}}\sum_{i=1}^{n}\left[\frac{1}{\sigma_i^2}\right](y_i - \boldsymbol{x}_i^t\boldsymbol{\beta})^2 \equiv \min_{\boldsymbol{\beta}}\sum_{i=1}^{n} w_i(y_i - \boldsymbol{x}_i^t\boldsymbol{\beta})^2$$

여기서 $w_i = 1/\sigma_i^2$이다. 이러한 가중최소제곱법은 각 관측치에 대하여 서로 다른 가중치를 적용하여 최소제곱 추정량을 구하며 위의 경우에는 가중치가 반응값의 분산에 반비례한다. 따라서 분산이 큰 오차항을 가진 반응값의 가중치는 분산이 작은 반응값에 비해 상대적으로 작다. 이러한 가중치와 분산의 관계는 변이가 적은 반응값 근방에서 오차를 더욱 줄이려고 하는 직관적인 생각과 일치한다. 가중치를 적용한 최소제곱 추정량은 다음과 같이 나타낼 수 있다.

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^t\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{W}\boldsymbol{y}$$

여기서

$$\boldsymbol{W} = \text{diag}\left[\frac{1}{\sigma_1^2}, \frac{1}{\sigma_2^2}, \dots, \frac{1}{\sigma_n^2}\right] = [Var(\boldsymbol{e})]^{-1} = \boldsymbol{V}^{-1}$$

위에서 본 가중최소제곱법을 오차항이 일반적인 공분산 행렬 $\boldsymbol{V}$를 가질 때 적용하면 다음과 같이 목적 함수를 나타낼 수 있으며

$$\min_{\boldsymbol{\beta}}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^t\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$$

가중최소제곱추정량은 다음과 같이 나타낼 수 있다.

$$\hat{\boldsymbol{\beta}}_* = (\boldsymbol{X}^t\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{V}^{-1}\boldsymbol{y} \tag{4.1}$$

가중최소제곱추정량은 다음과 같은 성질은 만족한다.

- 가중최소제곱추정량은 불편추정량이다: $E(\hat{\boldsymbol{\beta}}_*) = \boldsymbol{\beta}$
- 가중최소제곱추정량 $\hat{\boldsymbol{\beta}}_*$는 분산이 가장 작은 불편선형 추정량이다 (Gauss-Markov Theorem)
- 가중최소제곱추정량 $\hat{\boldsymbol{\beta}}_*$의 분산:

$$Var(\hat{\boldsymbol{\beta}}_*) = (\boldsymbol{X}^t\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}$$

- 가중최소제곱추정량 $\hat{\boldsymbol{\beta}}_*$은 $\boldsymbol{y}$가 평균이 $\boldsymbol{X\beta}$이고 분산이 V인 정규분포에서 $\boldsymbol{\beta}$에 대한 최대우도추정량이다.
- 일반최소제곱추정량도 불편추정량이다: $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$

가중최소제곱법에서 유의할 점은 오차항의 공분산 행렬 $\boldsymbol{V}$에 대하여 모르는 경우 이를 추정해야 하며 가중최소제곱추정량에 공분산 행렬의 추정치
$\hat{\boldsymbol{V}}$을 사용할 경우 위에서 언급한 최적성은 더 이상 성립하지 않는다.

$$\hat{\boldsymbol{\beta}}_* = (\boldsymbol{X}^t\hat{\boldsymbol{V}}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^t\hat{\boldsymbol{V}}^{-1}\boldsymbol{y}$$

더 나아가 공분산행렬 $\boldsymbol{V}$에 대한 추정은 어려운 문제이므로 그 추정 방법과 통계적 성질을 잘 고려하여 사용해야 한다. 정확한 추정을 위해서 또는 자료의 특성을 이용하여 공분산 행렬 $\boldsymbol{V}$에 대한 모형을 어느 정ㅇ도 단순화하는 것이 바람직하다. 예를 들어 공분산 행렬 $\boldsymbol{V}$를 다음과 같이 나타낼 수 있다면 유용할 것이다.

$$\boldsymbol{V} = \sigma^2\text{diag}[v_1, v_2, \dots, v_n]$$

여기서 $(v_1, v_2, \dots, v_n)$은 알려진 값이고 σ^2는 추정해야하는 모수이다.

오차항의 등분산성에 대한 가정을 검토하기 위한 방법중 가장 유용하고 간단한 방법은 잔차그림을 이용하는 것이다. 잔차 r_i와 적합된 값 $\hat{y}_i$에 대한 잔차그림을 그려서 잔차의 퍼진 정도가 적합된 값 $\hat{y}_i$에 따라 변하면 등분산성에 대한 가정을 의심해봐야 한다. 실제 자료에서 반응값의 분산이 독립변수에 비례하여 나타나는 경우가 많다. 단순회귀모형을 고려하고

$$y_i = \beta_0 + \beta_1 x_i + e_i$$

오차항이 서로 독립이며 그 분산이 독립변수에 비례한다고 가정하자.

$$Var(e_i) = x_i\sigma^2$$

이러한 경우는 독립변수의 값이 양인 경우이며 독립변수의 값이 커지면 반응값의 분산도 커진다. 이러한 경우 독립변수와 종속변수의 관계, 회귀식, 잔차그림은 다음과 같이 나타난다.

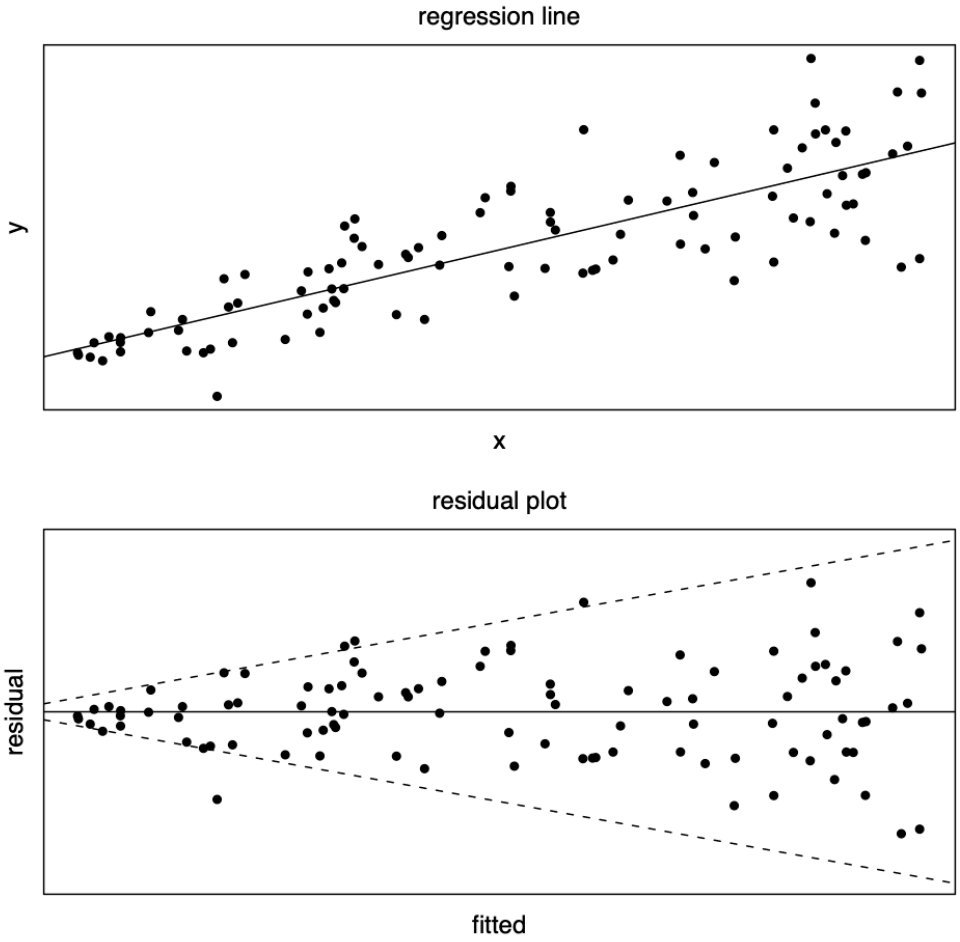


그림 4.1.: 오차항의 등분산성이 위반된 경우

4.2. 변수변환

변수변환(Variable transformation)은 독립변수와 종속변수를 변환함으로서 회귀식의 적합도를 향상시켜 예측력을 높일 수 있을뿐 아니라 최소제곱법에서의 등분산성 가정에 대한 만족도를 높일 수 있는 유용한 방법이다(variance stabilization). 이 절에서는 변수변환의 종류와 그 효과를 단순회귀식에서 살펴본다. 중회귀의 경우에는 변수변환의 적용을 복합적으로 고려해야 할 것이다.

4.2.1. 지수모형과 멱함수: 로그변환

회귀식에 대한 모형이 지수함수 모형인 경우, 즉 독립변수와 종속변수가 다음과 같은 경우

$$y = \beta_0 \exp(\beta_1 x)$$

종속변수에 대한 로그 변환(log transformation)을 하면 선형관계에 매우 가깝게 된다 (아래 그림 참조).

$$\log(y_i) = \beta_{,0} + \beta_1 x_i + e_i$$

여기서 주의할 점은 원래의 지수모형에 오차항의 지수함수가 곱해지는 형태가 되어야 로그 변환후에 등분산성의 가정을 만족하게 된다. 즉 오차항 e_i를 서로 독립이고 평균이 0, 분산이 σ^2이라고 하면 다음과 같은 관계가 성립된다.

$$y_i = \beta_0 \exp(\beta_1 x_i) \exp(e_i) \quad \Rightarrow \quad \log(y_i) = \beta_0' + \beta_1 x_i + e_i$$

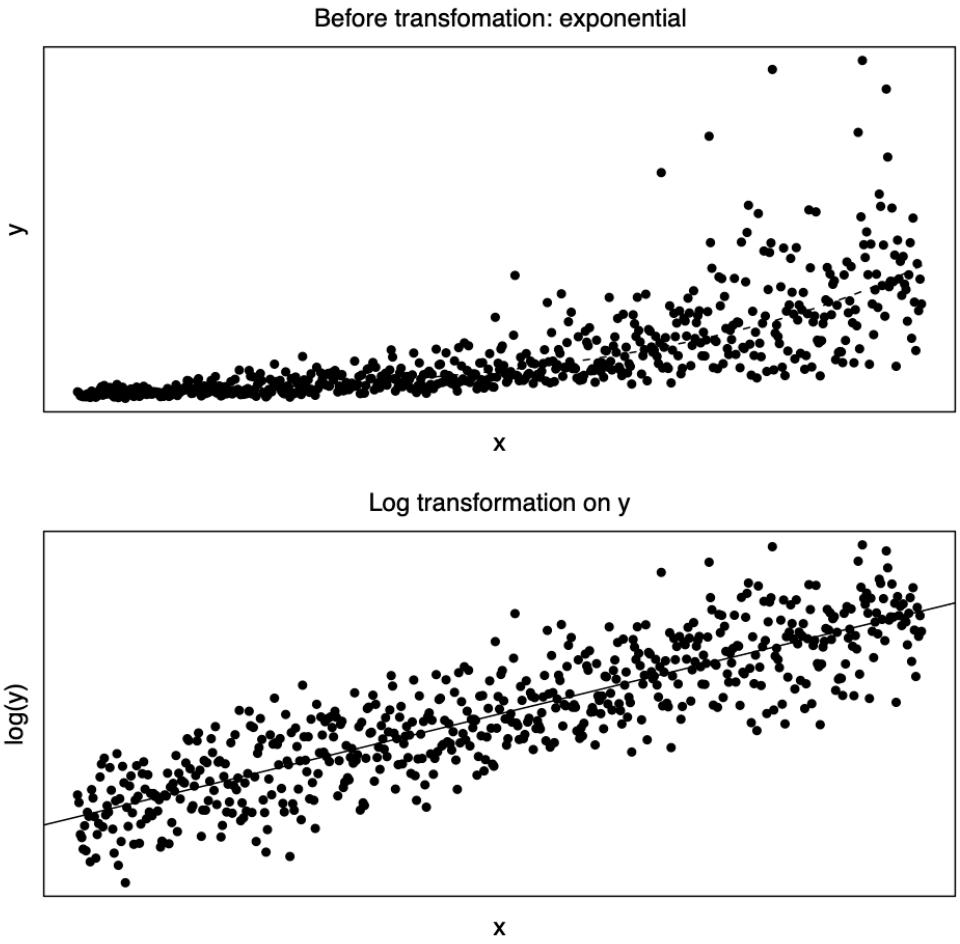


그림 4.2.: 지수모형과 로그변환

회귀식에 대한 모형이 멱함수모형인 경우, 즉 독립변수와 종속변수가 다음과 같은 경우

$$y = \beta_0 x_1^{\beta}$$

독립변수와 종속변수에 대한 로그 변환을 하면 선형관계에 매우 가깝게 된다.

$$\log(y_i) = \beta_{,0} + \beta_1 \log(x_i) + e_i$$

여기서 주의할 점도 원래의 멱함수모형에 오차항이 곱해지는 형태가 되어야 로그 변환후에 등분산성의 가정을 만족하게 된다. 즉 오차항 e_i를 서로 독립이고 평균이 0, 분산이 σ^2이라고 하면 다음과 같은 관계가 성립된다.

$$y_i = \beta_0 x_1^{\beta} \exp(e_i) \quad \Rightarrow \quad \log(y_i) = \beta_0' + \beta_1 \log(x_i) + e_i$$

회귀식에 대한 모형이 역지수함수 모형인 경우, 즉 독립변수와 종속변수가 다음과 같은 경우

$$y = \beta_0 \exp(\beta_1 / x)$$

종속변수에 대한 로그 변환과 독림변수에 대한 역변환을 하면 선형관계에 매우 가깝게 된다.

$$\log(y_i) = \beta_{,0} + \beta_1 \left(\frac{1}{x_i} \right) + e_i$$

여기서 주의할 점은 원래의 역지수모형에 오차항의 지수함수가 곱해지는 형태가 되어야 로그 변환후에 등분산성의 가정을 만족하게 된다.

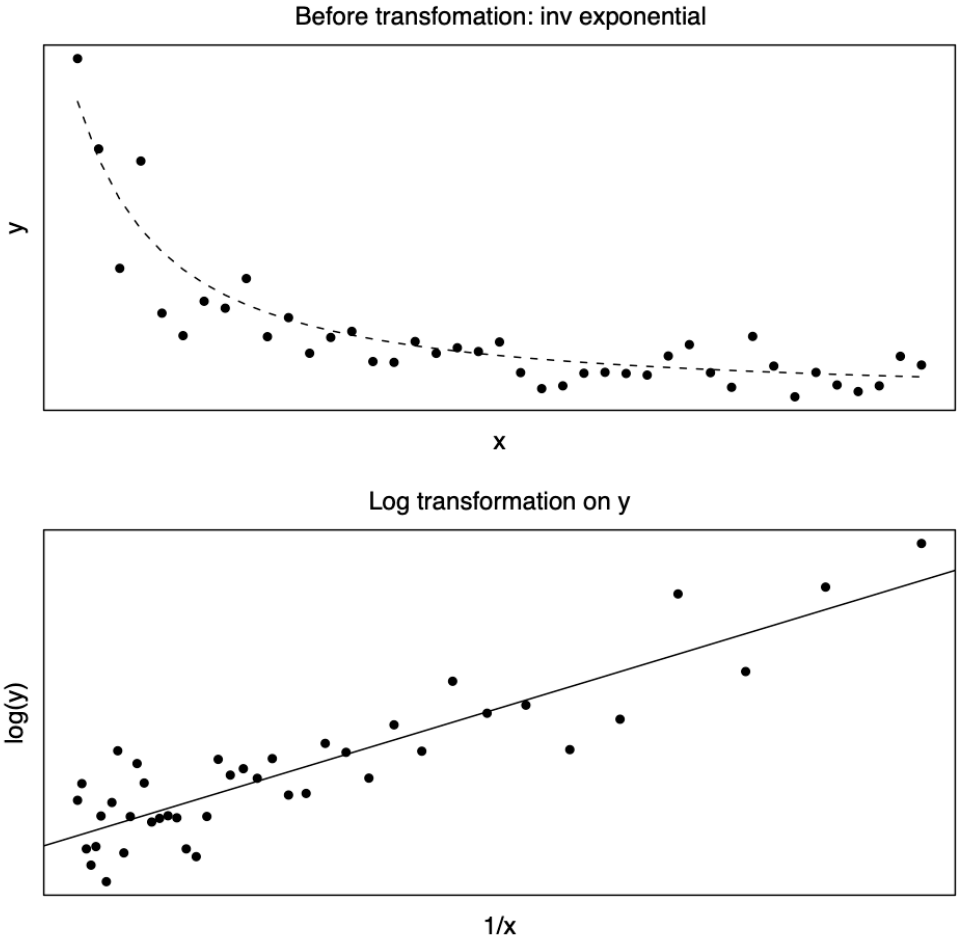


그림 4.3.: 역지수모형과 로그/역변환

4.2.2. 쌍곡선과 역변환

생물학, 경제학등의 문야에서 독립변수와 종속변수의 관계가 쌍곡선(Hyperbola) 형태인 경우가 많다. 독립변수가 증가함에 따라 종속변수의 값이 수렴하는 경우에 이러한 관계가 매우 유용하다.

$$y = \frac{x}{\beta_0 + \beta_1 x}$$

이러한 경우에 독립변수와 종속변수에 모두 역변환(Inverse transformation)을 취하면 선형관계에 매우 가깝게 된다 (아래 그림 참조).

$$\frac{1}{y_i} = \beta_0 + \beta_1 \left(\frac{1}{x_i}\right) + e_i$$

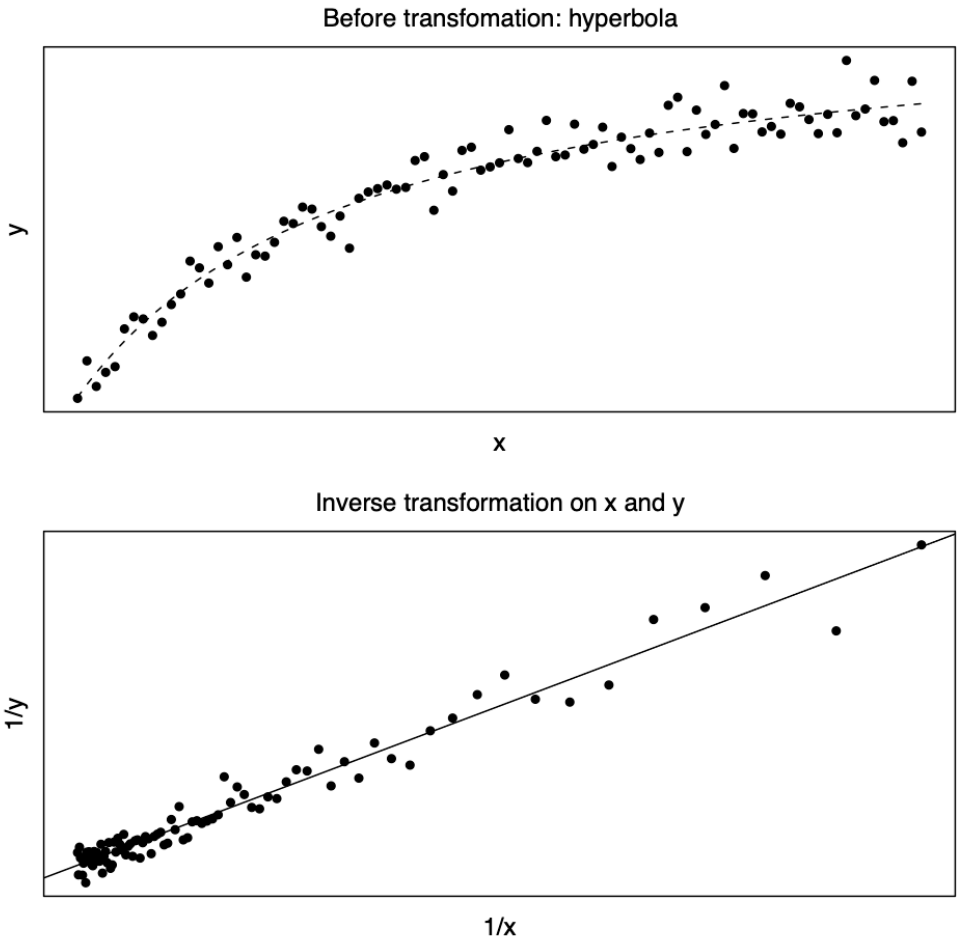


그림 4.4.: 쌍곡선과 역변환

위에서 살펴본 회귀모형들에서 변수변환후에 등분산성에 대한 가정을 만족하려면 대부분 변환전의 함수관계식에 오차항의 지수함수가 곱해지는 형태가 되어야 한다 (multiplicative error model). 이렇게 오차항이 함수관계식에 곱해지는 형태가 아니 다른 형태라면, 예를 들어 오차항이 원래의 함수관계식에 더해지는 형태 (additive error model), 등분산성의 가정이 상당히 위배될 수 있음을 주의해야 한다. 예를 들어 회귀식에 대한 모형이 지수함수 모형인 경우 서로 독립이고 평균이 0, 분산이 σ^2 인 오차항 e_i 를 함수 관계식에 더해졌다면 로그변환된 종속 변수와 독립변수는 선형관계를 보이지만 등분산성 가정은 만족하지 못하게된다.

$$y_i = \beta_0 \exp(\beta_1 x_i) + e_i \quad \Rightarrow \quad \log(y_i) \cong \beta_0' + \beta_1 x_i + e_i^*$$

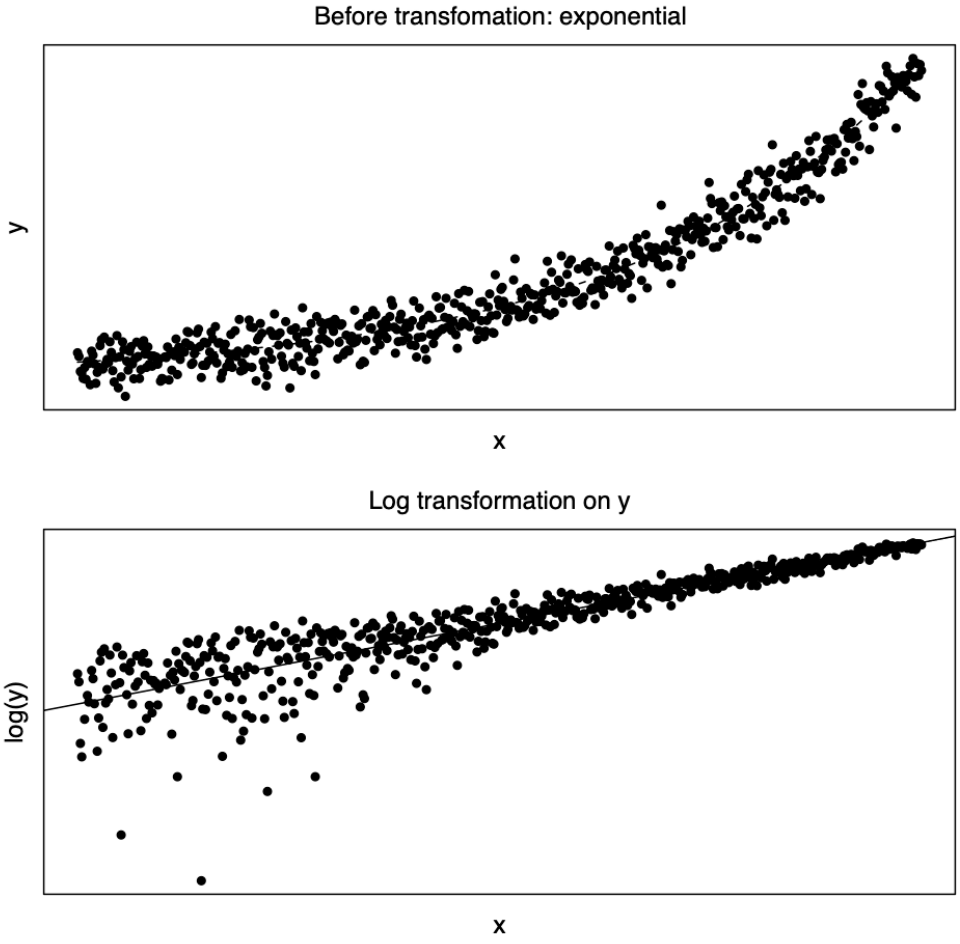


그림 4.5.: Additive error model 과 로그변환

4.2.3. Box-Cox 변환

앞 절에서 보았듯이 종속변수에 로그변환 등을 적용하면 여러 가지 유용한 점이 많다. 위에서 살펴본 종속변수에 변환은 여러 가지 비선형모형을 선형모형에 가깝게 만들어 주며 Multiplicative error model과 같이 반응값이 분산이 독립변수의 크기에 영향을 받는 모형을 등분산성을 가진 형태의 모형으로 버꾸어 준다 (Variance stabilization). 이렇게 종속변수에 대한 여러 가지 변환을 하나의 체계적인 형태로 결합한것을 Box-Cox 변환으로 부르며 다음과 같아 정의한다.

$$y^{(\lambda)} = \begin{cases} \log(y) & \text{if } \lambda = 0 \\ \frac{y^{\lambda}-1}{\lambda} & \text{if } \lambda \neq 0 \end{cases} \tag{4.2}$$

Box-Cox 변환은 로그변환과 멱변환을 모두 포함하고 있다. 또한 Box-Cox 변환된 $y^{(\lambda)}$가 정규분포를 따른다는 가정 하에 자료가 주어졌을때 λ에 대한 최대우도추정량을 구할 수 있다.

4.3. 다중공선성

회귀분석에서 독립변수들간의 강한 선형 관계의 경향이 있을 때 이를 다중공선성(multicollinearity)라고 한다. 즉, p개의 독립변수 $x_1, x_2, \dots, x_p$의 관계가 다음과 같은 선형관계에 가깝다면 다중공선성이 존재한다고 한다.

$$c_1 x_1 + c_2 x_2 + \cdots + c_p x_p \cong 0$$

다중공선성에 의해 발생하는 여러 가지 문제점들을 기술적으로 독립변수들의 강한 선형관계때문에 행렬 X^tX가 ill-conditioned 행렬이 되어 그 역행렬이 불안정하게 구해지는 결과 때문에 생기게 된다. 여기서 회귀계수의 공분산 행렬은 다음과 같이 주어짐에 유의하자.

$$Var(\boldsymbol{b}) = \sigma^2 (X^t X)^{-1}$$

예를 들어서 두 개의 독립변수가 있는 회귀 모형을 생각해 보자.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e$$

그림 4.6 에서 (a)의 경우는 두 개의 독립변수 x_1과 x_2의 관계가 독립적이어서 (linearly independent) 적합된 회귀식이(그림에서 2차원 평면) 안정적이다. 반면에 그림 4.6 에서 (b)의 경우는 두 개의 독립변수 x_1과 x_2가 완벽한 선형관계가 있기 때문에 (linearly dependent)

$$c_1 x_1 + c_2 x_2 = 0$$

적합된 회귀식이 여러가지 존재한다. 이러한 경우는 매우 드물지만 그림 4.6 에서 (c)의 경우는 두 개의 독립변수 x_1과 x_2가 선형관계에 매우 가깝기 때문에 적합된 회귀식이 불안정하다.

회귀분석에서 다중공선성의 정도를 측정할 수 있는 통계량은 다음과 같은 것들이 있다.

4.3.1. 독립변수간의 상관계수

독립변수간의 상관계수를 보아 강한 상관관계를 가지는 변수들이 있다면 다중공선성의 가능성이 크다.

4.3.2. X^tX 의 고유값(Eigenvalues)

행렬 $\boldsymbol{X^tX}$의 고유값를 구하여 큰 순서대로 나열했을 때 가장 작은 값이 0에 매우 가까우면 다중공선성의 가능성이 크다. 만약에 독립변수들간의 선형 관계가 있다면 행렬 $\boldsymbol{X^tX}$은 최대 계수(full rank) 행렬이 아니므로 하나 이상의 고유값이 0이 되게 된다.

이제 $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p$를 $\boldsymbol{X^tX}$의 고유값라고 하자. 이런 가정 하에서 $1/\lambda_i$는 $(\boldsymbol{X^tX})^{-1}$의 고유값이다. $\boldsymbol{X^tX}$의 고유값에 대한 고유벡터를 $\boldsymbol{p}_1, \boldsymbol{p}_2, \dots, \boldsymbol{p}_p$라고 하고 행렬 $\boldsymbol{P} = [\boldsymbol{p}_1, \boldsymbol{p}_2, \dots, \boldsymbol{p}_p]$로 정의하자. 이때 다음과 같이 스펙트럴 분해를 이용하여 $\boldsymbol{X^tX}$를 나타낼 수 있다.

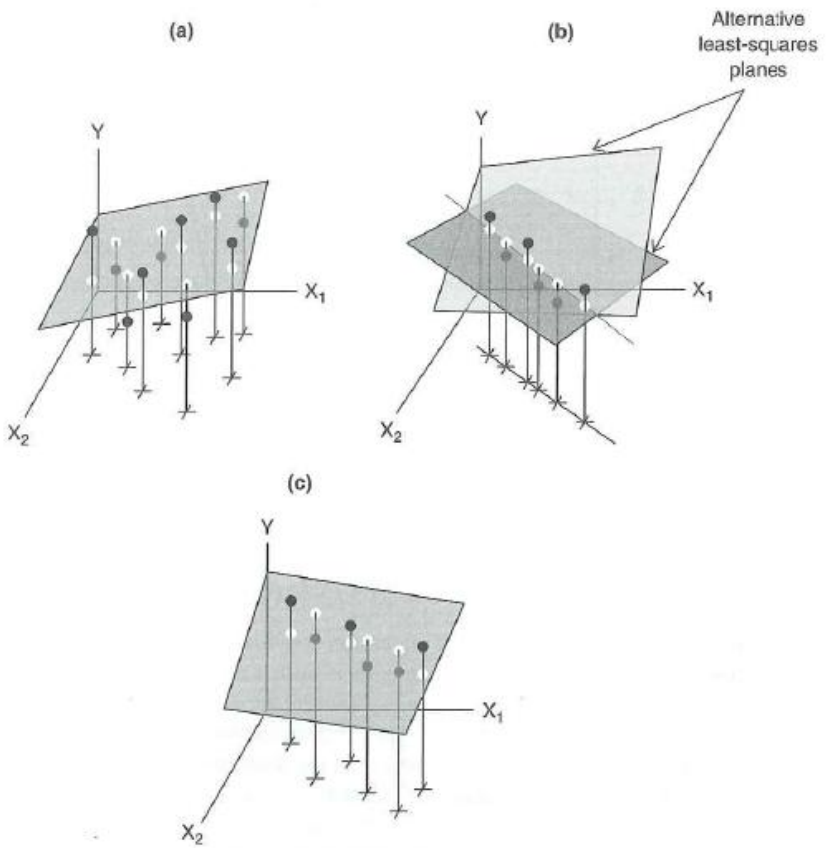


그림 4.6.: 회귀분석에서의 다중공선성의 정도(Fox 2008, p310)

$$\boldsymbol{P}^t(\boldsymbol{X}^t\boldsymbol{X})\boldsymbol{P} = \boldsymbol{\Lambda} = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_p)$$

또한

$$(\boldsymbol{X}^t\boldsymbol{X})^{-1} = \boldsymbol{P}\boldsymbol{\Lambda}^{-1}\boldsymbol{P}^t$$

따라서 가장 작은 고유값 λ_p가 매우 0에 가까우면 $(\boldsymbol{X}^t\boldsymbol{X})\boldsymbol{p}_p = \lambda_p\boldsymbol{p}_p \approx 0$이 성립하고 이는 $\boldsymbol{X}\boldsymbol{p}_p \approx 0$을 의미하여 독립변수간에 선형관계가 있다는 것을 암시한다.

일반적으로 최소제곱추정량의 공분산은 다음과 같아 나타내어지고

$$Var(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}^t\boldsymbol{X})^{-1} = \sigma^2\boldsymbol{P}\boldsymbol{\Lambda}^{-1}\boldsymbol{P}^t = \sigma^2\sum_{i=1}^{p}\frac{1}{\lambda_i}\boldsymbol{p}_i\boldsymbol{p}_i^t$$

최소제곱추정량의 분산의 합(total variance)은 다음과 같다.

$$\sum_{i=1}^{p}Var(\hat{\beta}_i) = tr[\sigma^2(\boldsymbol{X}^t\boldsymbol{X})^{-1}] = \sigma^2 tr[(\boldsymbol{X}^t\boldsymbol{X})^{-1}] = \sigma^2\sum_{i=1}^{p}\frac{1}{\lambda_i}$$

따라서 고유값의 값이 0에 가까와 지면 추정량의 분산의 합은 매우 크게 된다.

4.3.3. 조건지수 (condition number)

다중공선성의 판별을 위하여 행렬 $\boldsymbol{X}^t\boldsymbol{X}$의 고유값이 중요한 측도라고 했다. 고유값을 상대적으로 비교하면 다중공선성을 더 명확하게 알 수 있다. 조건지수는 가장 튼 고유값과 다른 고유값의 비율의 제곱근으로 나타내어진다.

$$\kappa_i = \sqrt{\frac{\lambda_1}{\lambda_i}}, \quad i = 2, 3, \cdots, p$$

중요한 역활을 하는 조건지수는 가장 작은 고유값에 의한 것이다.

$$\kappa_p = \kappa(\boldsymbol{X}) = \sqrt{\frac{\lambda_1}{\lambda_p}}$$

$\kappa(\boldsymbol{X})$의 값이 크면 클수록 다중공선성의 가능성이 높다. 일반적으로 $\kappa(\boldsymbol{X})$ 가 30 이상이면 다중공선성의 가능성이 크다고 본다.

4.3.4. 분산팽창계수 (Variance Inflation Factor ;VIF)

하나의 독립변수 x_i를 나머지 다른 $p-1$개의 독립변수 $x_1, \cdots, x_{i-1}, x_{i+1}, \cdots, x_p$를 이용하여 회귀식을 적합시킬 수 있다. 이때 다중공선성이 존재한다면 적합된 회귀식의 결정계수 R_i^2는 1에 매우 가까울 것이다

$$x_i = \beta_0 + \beta_1 x_1 + \cdots + \beta_{i-1} x_{i-1} + \beta_{i+1} x_{i+1} + \cdots + \beta_p x_p + e$$

이때 독립변수 x_i에 대한 VIF는 다음과 같이 정의되며 그 값이 5 또는 10보다 크면 다중공선성의 가능성이 크다.

$$VIF_i = \frac{1}{1 - R_i^2}$$

5. 관측값에 대한 진단

5.1. 서론

회귀분석을 포함한 통계적 자료분석에서 흔하게 접하는 문제는 자료 중의 일부가 통계적 모형에 의해 추정된 평균적인 경향에서 매우 벗어나 있는 점을 발견하게 되는 경우이다.

이러한 경우 평균적인 경향에서 매우 벗어난 자료를 분석에서 제외시킬 것인지에 대한 논의도 필요할 수 있으며 이러한 자료들이 모형의 모수에 대한 추정에 어떤 영향을 미칠 것인자에 대한 검토도 필요할 수 있다.

평균적인 경향에서 매우 벗어난 자료를 흔히 이상점(outlier)라고 부른다. 회귀분석에서 이러한 이상점은 회귀계수 추정과 그에 따른 여러 가지 통계적 추론에 많은 영향을 미친다. 따라서 이상점들이 회귀분석의 계수 추정에 어떤 영향을 얼만큼 끼치는가에 대한 검토는 매우 중요하다.

5.2. 이상점의 유형

일차원 자료에서는 평균적인 경향에서 매우 벗어난 자료의 식별이 단순하고 쉽다. 예를 들어 다음과 같이 일변량 자료 $\boldsymbol{x}$만 고려하면 이상점이 어떤 점인지는 쉽게 찾을 수 있다.

$$\boldsymbol{x}^t = (1, 2, 3, 4, 5, 10)$$

그러나 반응변수와 독립변수들을 고려해야 하는 회귀분석에서 이상점은 간단하게 파악하기 힘들고 상황에 따라 그 의미가 매우 다르다.

회귀분석에서 이상점의 다양한 종류와 그 영향을 알아보기 위하여 단순회귀분석을 고려하고 다음 그림을 보자.

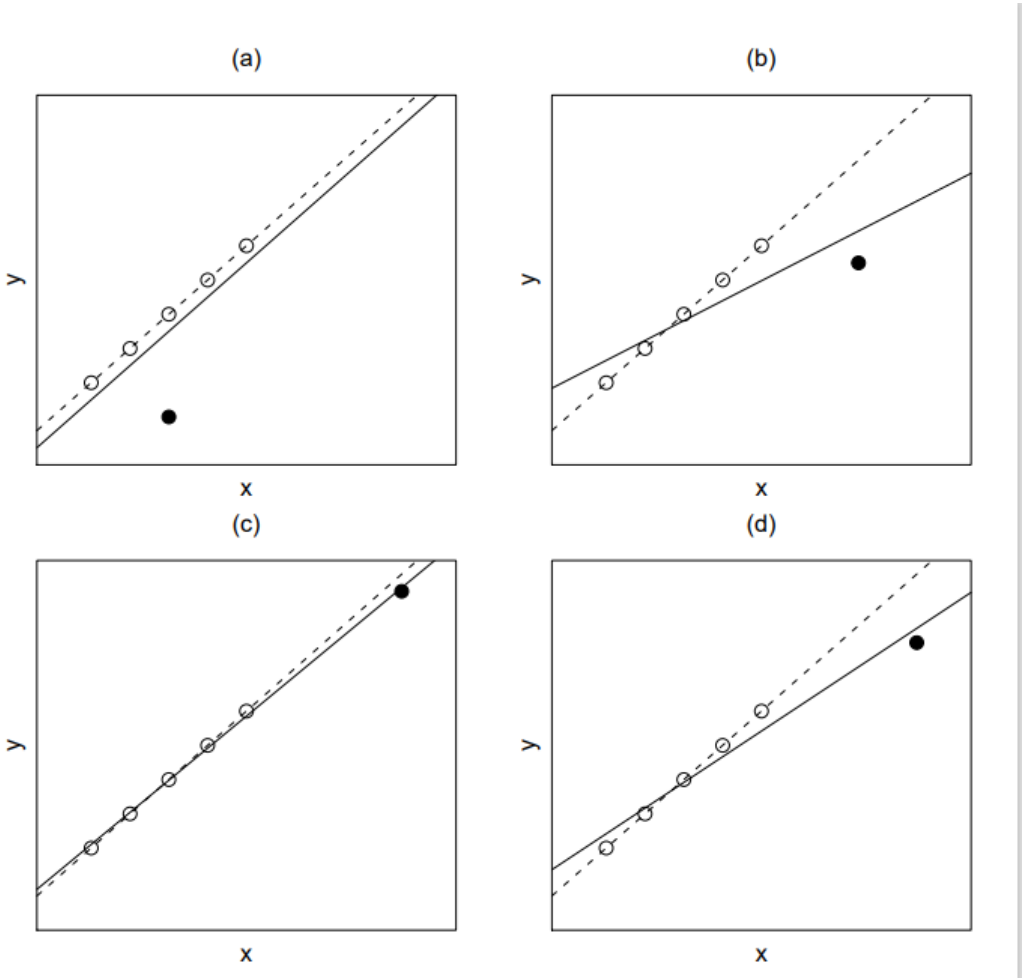


그림 5.1.: 여러가지 종류의 이상점과 그 영향

위의 그림에서 실선은 검정색 점을 포함해서 적합한 회귀직선이고 점선은 검정색 점을 제외했을 때의 회귀직선이다

그림의 (a)에서 검정색 점은 설명변수 x에 대해서는 이상점이 아니지만 x가 주어진 경우 반응변수 y에 대해서는 평균적인 경향에서 많이 벗어나 있기 때문에 이상점이다. 이러한 이상점을 회귀이상점(regression outlier)라고 부르기도 한다. (a)의 회귀이상점의 유무는 회귀계수의 추정에 크게 영향을 주지 않는다. 이렇게 어떤 관측점이 있고 없음에 따라 회귀계수의 값이 크게 변하지 않는다면 그 자료의 영향력(leverage)이 작다고 한다.

(b)에서의 검은 점은 설명변수 x에 대하여 이상점이며 또한 회귀이상점이다. 더 나아가 이 이상점을 제외하고 적합한 회귀계수는 이상점을 포함했을 때 적합한 회귀계수와 매우 다르다. 이 경우 이 이상점은 큰 영향력을 가졌다고 말한다.

(c)에서의 검은 점은 설명변수 x에 대해서 이상점이지만 회귀이상점은 아니다. 이러한 경우 이상점의 유무에 따라 회귀계수의 값이 크게 변하지 않으므로 이상점은 작은 영향력을 가졌다고 말한다. 하지만 (c)에서의 검은 점이 설명변수 x의 중심점으로부터 크게 멀어져있으므로 y의 값이 조그만 변해도 그림 (d)와 같이 큰 영향력을 가진다.

관측치 y_i 를 제외할 때와 포함할 때의 회귀계수 추정치가 매우 다르면 그 관측치를 영향점(influential point)라고 하며 그 영향의 크기는 그 점이 가진 영향력(leverage)의 크기와 평균에서 떨어진 정도에 비례한다.

자료가 선형모형의 계수 추정에 미치는 영향 $\propto$ 영향력의 크기 $\times$ 이상치의 특이한 정도

위에서 살펴보았듯이 회귀분석에서 이상점의 종류와 그 영향은 매우 다양하며 복잡하다. 이러한 이상점의 종류와 회귀계수의 영향에 대하여 분석할 때 유용하게 쓰이는 통계량이 잔차(residual)이다.

5.3. 지렛값

y가 반응변수이고 $p-1$개의 설명변수 $x_1, x_2, \dots, x_{p-1}$가 있을 때 회귀식은 다음과 같이 표현된다.

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$$

회귀계수 $\boldsymbol{\beta}$의 최소제곱 추정치 $\hat{\boldsymbol{\beta}}$는 다음과 같이 주어지며

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{y}$$

관측값 $\boldsymbol{y}$ 의 추정치 $\hat{\boldsymbol{y}}$는 다음과 같다.

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{y} \equiv \boldsymbol{H}\boldsymbol{y}$$

여기서 $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t$를 사영행렬(hat matrix 또는 projection matrix)라고 부르며 사영행렬 $\boldsymbol{H}$ 의 i 번째 대각원소를 h_{ii}라고 하며 이는 이상치 또는 영향치 분석에 중요한 역할을 한다.

i번째 관측치의 설명변수 벡터를 다음과 같이 표시하면

$$\boldsymbol{x}_i^t = (1, x_{i1}, x_{i2}, \dots, x_{i,p-1})$$

$\boldsymbol{H}$ 의 i 번째 대각원소를 h_{ii}는 다음과 같이 표현된다.

$$h_{ii} = \boldsymbol{x}_i^t(\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{x}_i \tag{5.1}$$

h_{ii}는 i 번째 관측치의 설명변수 $(x_{i1}, x_{i2}, \dots, x_{i,p-1})$ 들이 모든 관측치의 평균 $(\bar{x}_1, \bar{x}_2, \dots, \bar{x}_{p-1})$ 에서 얼마나 멀리 떨어져 있는가에 대한 상대적인 양을 나타낸다. 따라서 h_{ii} 를 지렛점(leverage point)라고 부른다.

지렛점 값이 클수로 영향점일 가능성이 크며 큰 값을 높은 지렛값(high leverage point)라고 부른다.

보통 h_{ii}값이 p/n보다 크면 영향력이 크다고 말한다. 참고로 단순 회귀식에서 h_{ii}는 다음과 같이 주어진다.

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum(x_i - \bar{x})^2}$$

또한 h_{ii} 값을 모두 더하면 설명변수의 개수와 같다.

$$tr(\boldsymbol{H}) = \sum_{i=1}^{k} h_{ii} = p$$

5.4. 내 표준화 잔차

잔차 r_i 는 y 의 실제 관측값과 그 추정치의 차이이며

$$r_i = y_i - \hat{y}_i = y_i - \boldsymbol{x}_i^t \hat{\boldsymbol{\beta}} \tag{5.2}$$

잔차벡터에 대한 식은 다음과 같다.

$$\boldsymbol{r} = \boldsymbol{y} - \hat{\boldsymbol{y}} = \boldsymbol{y} - \boldsymbol{H}\boldsymbol{y} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y} \tag{5.3}$$

잔차의 공분산 행렬을 살펴보면 다음과 같이 주어진다. 따라서 i 번째 잔차의 분산은 $Var(r_i) = (1 - h_{ii})\sigma^2$ 이다.

$$Var(\boldsymbol{r}) = \sigma^2(\boldsymbol{I} - \boldsymbol{H}) \tag{5.4}$$

위에서 언급한 잔차 r_i 를 보통 잔차(ordinary residual)이라고 하며 그 크기가 단위에 따라 바뀌므로 잔차분석에서는 표준화 잔차(standardized residual)을 더 많이 사용한다.

아래와 같이 잔차를 그 표준편차로 나눈 값을 **내 표준화 잔차(internally studentized residual)** 이라고 부른다.

$$r_i^s = \frac{r_i}{s\sqrt{1 - h_{ii}}} \tag{5.5}$$

위의 식에서 s는 오차항의 표준편차 σ의 추정량이며 h_{ii}는 사영행렬 $\boldsymbol{H}$ 의 i 번째 대각원소(즉 지렛값)이다.

잔차분석에서는 척도(scale)에 영향이 없는 표준화 잔차를 이용하는 것이 좋다. 그 값이 클수로 이상치일 가능성이 크다. 보통 내표준화 잔차의 절대값이 2보다 크면 이상치일 가능성이 크다.

5.5. 관측값의 영향: 계수 추정

회귀분석에서 하나의 관측치가 회귀계수의 추정에 영향을 미치는 정도를 알아볼 때 유용한 방법은 그 관측치를 제외했을 때의 최소제곱추정량과 포함했을 때의 추정량을 비교하는 것이다.

i번째 관측치에 대한 반응값과 설명변수들이 다음과 같은 때

$$y_i, \quad \boldsymbol{x}_i^t = (1, x_{i1}, x_{i2}, \dots, x_{i,p-1})$$

i번째 관측치를 제외한 자료에서 반응변수와 설명변수의 벡터식을 다음과 같이 표시한다.

$$\boldsymbol{y}_{-i}, \quad \boldsymbol{X}_{-i}$$

i번째 관측치를 제외했을 때 회귀계수의 최소제곱추정량을 $\hat{\boldsymbol{\beta}}_{-i}$라 하면 모든 관측치를 이용한 최소제곱추정량을 $\hat{\boldsymbol{\beta}}$와의 관계는 다음과 같이 나타낼 수 있다. 아래 식 세번째 중의 결과는 우드베리 공식 A.3 을 이용하였다.

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{-i} &= (\boldsymbol{X}_{-i}^t \boldsymbol{X}_{-i})^{-1} \boldsymbol{X}_{-i}^t \boldsymbol{y}_{-i} \\
&= (\boldsymbol{X}^t \boldsymbol{X} - \boldsymbol{x}_i \boldsymbol{x}_i^t)^{-1} (\boldsymbol{X}^t \boldsymbol{y} - \boldsymbol{x}_i y_i) \\
&= \left[(\boldsymbol{X}^t \boldsymbol{X})^{-1} + \frac{(\boldsymbol{X}^t \boldsymbol{X})^{-1} \boldsymbol{x}_i \boldsymbol{x}_i^t (\boldsymbol{X}^t \boldsymbol{X})^{-1}}{1 - \boldsymbol{x}_i^t (\boldsymbol{X}^t \boldsymbol{X})^{-1} \boldsymbol{x}_i} \right] (\boldsymbol{X}^t \boldsymbol{y} - \boldsymbol{x}_i y_i) \\
&= \hat{\boldsymbol{\beta}} + \frac{1}{1 - h_{ii}} (\boldsymbol{X}^t \boldsymbol{X})^{-1} \boldsymbol{x}_i \left[\boldsymbol{x}_i^t \hat{\boldsymbol{\beta}} - (1 - h_{ii}) y_i - h_{ii} y_i \right] \\
&= \hat{\boldsymbol{\beta}} - \frac{1}{1 - h_{ii}} (\boldsymbol{X}^t \boldsymbol{X})^{-1} \boldsymbol{x}_i (y_i - \boldsymbol{x}_i^t \hat{\boldsymbol{\beta}}) \\
&= \hat{\boldsymbol{\beta}} - \frac{r_i}{1 - h_{ii}} (\boldsymbol{X}^t \boldsymbol{X})^{-1} \boldsymbol{x}_i
\end{aligned} \tag{5.6}$$

또한 i번째 관측치를 제외했을 때 오차항 분산의 추정량을 s_{-i}^2로 나타낸다.

$$s_{-i}^2 = \frac{1}{n - p - 1} \sum_{j \neq i} (y_j - \boldsymbol{x}_j^t \hat{\boldsymbol{\beta}}_{-i})^2 \tag{5.7}$$

5.6. 외 표준화 잔차와 PRESS 잔차

잔차를 표준화 할 때 i번째 관측치를 제외했을 때 분산의 추정량을 s_{-i}^2을 이용하는 것이 합리적이다. 이는 반응값이 이상점인 경우 분산의 추정량이 커지게 된다. 식 5.5 에서 정의된 내 표준화 잔차에서는 이상점이 분산의 추정량에 영향을 주어 잔차의 크기가 작아지게 된다. 따라서 내 표본화 잔차는 이상점을 구별할 수 있는 능력이 떨어진다. 이러한 점을 보완하기 위하여 이상점의 영향을 약화시킬 수 있도록 s_{-i}^2를 이용하여 표준화 한 양이 아래와 같이 정의된 표준화 잔차이다.

$$r_i^* = \frac{r_i}{s_{-i} \sqrt{1 - h_{ii}}} \tag{5.8}$$

식 5.8 에서 정의된 차를 표준화 잔차(studentized residual) 또는 **외 표준화 잔차(externally studentized residual)**라고 부른다.

외 표준화 잔차는 i번째 관측치가 회귀식 적합에 미치는 영향을 내 표분화 잔차보다 더 민감하게 탐색할 수 있다. 보통 외 표준화 잔차의 절대값이 2보다 크면 이상치일 가능성이 크다.

PRESS 잔차 $r_{i,-i}$는 i 번째 관측값을 빼고 적합한 회귀식으로 부터 얻은 $E(y|\boldsymbol{x}_i)$의 추정치 $\hat{y}_{i,-i}$를 이용하여 만든 잔차이다. PRESS 잔차는 다음과 같이 정의된다.

$$r_{i,-i} = y_i - \hat{y}_{i,-i} = y_i - \boldsymbol{x}_i^t \hat{\boldsymbol{\beta}}_{-i} \tag{5.9}$$

실제 PRESS 잔차를 구할 경우 관측값을 제외하지 않고도 원래의 회귀식을 이용하여 아래와 같이 쉽게 구할 수 있다. 그 값이 클수로 이상치 또는 영향점일 가능성이 크다.

$$\begin{aligned} r_{i,-i} &= y_i - \hat{y}_{i,-i} \\ &= y_i - \boldsymbol{x}_i^t \hat{\boldsymbol{\beta}}_{-i} \\ &= y_i - \boldsymbol{x}_i^t \left[\hat{\boldsymbol{\beta}} - \frac{1}{1-h_{ii}} (\boldsymbol{X}^t \boldsymbol{X})^{-1} \boldsymbol{x}_i r_i \right] \\ &= (y_i - \boldsymbol{x}_i^t \hat{\boldsymbol{\beta}}) + r_i \frac{\boldsymbol{x}_i^t (\boldsymbol{X}^t \boldsymbol{X})^{-1} \boldsymbol{x}_i}{1-h_{ii}} \\ &= \frac{r_i}{1-h_{ii}} \end{aligned} \tag{5.10}$$

5.7. 관측값의 영향: 분산 추정

참고로 식 5.7 에서 정의된 s^2_{-i} 과 $s^2 = SSE/(n-p)$의 관계를 살펴보자. 먼저 SSE의 정의와 식 5.6 과 5.10 를 이용하여 다음과 같은 분해가 가능하다.

$$\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{-i} = (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) + \frac{r_i}{1-h_{ii}} \boldsymbol{X}(\boldsymbol{X}^t \boldsymbol{X})^{-1} \boldsymbol{x}_i$$

따라서

$$
\begin{aligned}
&\sum_{j\neq i}(y_j - \boldsymbol{x}_j^t\hat{\boldsymbol{\beta}}_{-i})^2 + (y_i - \hat{y}_{i,-i})^2 \\
&= \sum_{j\neq i}(y_j - \boldsymbol{x}_j^t\hat{\boldsymbol{\beta}}_{-i})^2 + (y_i - \boldsymbol{x}_i^t\hat{\boldsymbol{\beta}}_{-i})^2 \\
&= (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{-i})^t(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{-i}) \\
&= (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^t(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) - 2\frac{r_i}{1-h_{ii}}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^t\boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{x}_i \\
&\quad + \frac{r_i^2}{(1-h_{ii})^2}\boldsymbol{x}_i^t(\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{x}_i \\
&= (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^t(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) - 2\frac{r_i}{1-h_{ii}}\boldsymbol{y}^t(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{x}_i \\
&\quad + \frac{r_i^2}{(1-h_{ii})^2}\boldsymbol{x}_i^t(\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{x}_i \\
&= SSE + 0 + \frac{r_i^2}{(1-h_{ii})^2}h_{ii} \\
&= SSE + \frac{r_i^2 h_{ii}}{(1-h_{ii})^2}
\end{aligned}
\tag{5.11}
$$

이제 위의 식의 결과와 식 5.10 를 이용하면 다음과 같은 결과를 얻는다.

$$
\begin{aligned}
\sum_{j\neq i}(y_j - \boldsymbol{x}_j^t\hat{\boldsymbol{\beta}}_{-i})^2 &= SSE + \frac{r_i^2 h_{ii}}{(1-h_{ii})^2} - (y_i - \hat{y}_{i,-i})^2 \\
&= SSE + \frac{r_i^2 h_{ii}}{(1-h_{ii})^2} - (y_i - \hat{y}_{i,-i})^2 \\
&= SSE + \frac{r_i^2 h_{ii}}{(1-h_{ii})^2} - \frac{r_i^2}{(1-h_{ii})^2} \\
&= SSE - \frac{r_i^2}{1-h_{ii}}
\end{aligned}
\tag{5.12}
$$

따라서 다음 식을 이용하면 s_{-i}^2은 모든 관측값을 이용한 s^2으로부터 쉽게 유도할 수 있다.

$$
(n-p-1)s_{-i}^2 = (n-p)s^2 + -\frac{r_i^2}{1-h_{ii}} \tag{5.13}
$$

5.8. 영향력의 측도

하나의 관측값이 있는 경우 회귀계수 추정치와 없는 경우의 추정치의 차이가 크면 그 관측값이 큰 영향력을 가진다. 이러한 영향력을 측정할 수 있는 측조에 대하여 알아보자.

쿡의 거리(COOK's distance) C_i는 i번째 관측치가 회귀식 적합의 계수에 미치는 영향을 나타내는 양으로서 다음과 같이 정의된다.

$$
C_i = \frac{(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{-i})^t[\widehat{Cov}(\hat{\boldsymbol{\beta}}]^{-1}(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{-i})}{p} = \frac{(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{-i})^t(\boldsymbol{X}^t\boldsymbol{X})(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{-i})}{ps^2} \tag{5.14}
$$

여기서 $\hat{\boldsymbol{\beta}}_{-i}$는 i번째 관측치를 제외하고 적합한 회귀식에 의한 회귀계수이며 p는 설명변수의 개수이다. 그 값이 클수로 영향점일 가능성이 크다.

쿡의 거리 C_i과 내 표준화 잔차와의 관계는 다음과 같다.

$$C_i = \frac{(r_i^s)^2}{p}\left(\frac{h_{ii}}{1-h_{ii}}\right)$$

DFFITS는 n개의 모든 자료를 이용했을 때의 i 번째 관측값의 평균 $E(y|\boldsymbol{x}_i)$의 추정치 $\hat{y}_i$와 i 번째 관측값을 빼고 적합한 회귀식에 의한 추정치 $\hat{y}_{i,-i}$의 표준화된 차이을 말한다.

즉, $\hat{y}_{i,-i}$를 i번째 관측치를 제외하고 적합한 회귀식에 의한 예측치라고 한다면 두 예측치의 차이 $\hat{y}_i - \hat{y}_{i,-i}$를 표준화시키면 다음과 같다.

$$DFFITS_i = \frac{\hat{y}_i - \hat{y}_{i,-i}}{s_{-i}\sqrt{h_{ii}}} \tag{5.15}$$

DFFITS 는 그 값이 클수로 영향점일 가능성이 크다.

여기서 식 5.6 를 이용하면 다음 식를 얻고

$$\boldsymbol{x}_i^t\hat{\boldsymbol{\beta}}_{-i} = \boldsymbol{x}_i^t\hat{\boldsymbol{\beta}} - \frac{r_i h_{ii}}{1-h_{ii}}$$

DFFITS과 잔차와의 관계를 알 수 있다.

$$\begin{aligned} DFFITS_i &= \frac{\hat{y}_i - \hat{y}_{i,-i}}{s_{-i}\sqrt{h_{ii}}} \\ &= \frac{[h_{ii}/(1-h_{ii})]r_i}{s_{-i}\sqrt{h_{ii}}} \\ &= \frac{r_i}{s_{-i}\sqrt{1-h_{ii}}}[h_{ii}/(1-h_{ii})]^{1/2} \\ &= r_i^*\left[\frac{h_{ii}}{1-h_{ii}}\right]^{1/2} \end{aligned}$$

6. 모형의 선택

6.1. 서론

모형의 선택은 자료의 분석에서 고려하는 다수의 모형들(a family of models) 중에서 가장 적합한 모형(best model)을 찾는 것이다. 여기서 가장 적합하다는 의미는 다양한 기준이 있지만 일반적으로 선택된 모형의 예측력 또는 설명력이 다른 모든 모형들보다 더 좋다는 의미이다.

분석에서 고려하는 모형들의 집합을 모형 공간(model space)이라고 하며 이 공간에서 가장 적합한 모형을 찾는 것이 모형 선택(model selection)이다. 이 장에서 **모형의 예측력 또는 설명력** 을 정의하고 비교하는 방법에 대하여 배울 것이다.

주어진 모형 공간에서 가장 좋은 모형을 선택했을 때 다음과 같은 질문이 가능하다.

선택된 모형보다 더 좋은 모형이 있지 않을까? 더 좋은 모형이 주어진 모형공간에 포함되지 않을 수도 있다.

주어진 자료에서 반응변수와 설명변수의 관계를 더욱 잘 설명할 수 있는모형을 계속 찾는다면 결국에는 예측력을 높이기 위하여 더 많은 설명변수를 포함하는 모형을 찾게 될 것이다. 궁국적으로는 반응변수의 관측값 y 와 예측값 $\hat{y}$의 차이가 가장 작은 모형, 즉 설명력이 가장 좋은 모형을 선택하려는 노력을 계속한다면 **과적합(overfitting)** 이 발생할 수 있다.

과적합은 모형의 복잡도가 증가함에 따라 주어진 자료에 대한 모형의 예측력은 증가하지만 모형의 일반적인 예측의 효율은 오히려 감소하는 현상을 말한다.

이러한 과적합을 피하려면 모형의 복잡도와 예측력 사이의 적절한 균형을 찾아야 한다.

현실 세계의 상황에서는 진정한 모형이 알려지지 않거나 자료의 정확한 분포와 관계를 기술할 수 있는 모형을 파악하는 것은 매우 어렵다. 하지만 실제로 데이터를 생성하는 과정이나 현상을 정확하게 기술하는 가상의 모형이 존재한다고 가정할 수는 있다. 이렇게 자료의 분포와 관계를 정확하게 기술하는 가상의 모형을 참모형(true model)이라고 한다. 다시 강조하지만 가상의 모형이라고 말한 의미는 자료의 생성 과정을 정확하게 기술할 수 있는 모형을 구체화하여 표현하는 것이 매우 힘들기 때문이다.

모형의 선택하는 또 다른 기준은 가상의 **참모형에 제일 가까운 모형** 을 선택하는 것이다. 우리가 생각할 수 있는 대부분의 모형 공간은 참모형을 포함하지 않는 다고 가정할 수 있다. 이러한 경우 고려하는 다수의 모형들 중에서 참모형에 가장 가까운 모형을 최적의 모형이라고 할 수 있다.

6.2. 모형선택의 측도

회귀분석모형의 구축을 시작할 때는 될 수 있는 한 많은 독립변수들을 고려하고 그 중에 모형에 적합한 변수들과 그렇지 않은 변수들을 구별하여 최선의 모형을 찾으려고 많은 노력을 기울인다.

이 절에서는 설명변수의 조합으로 만들 수 있는 다양한 모형들을 비교할 수 있는 기준과 통계적 방법에 대하여 알아보고자 한다.

일반적인 회귀분석모형에서 다음과 같은 선형 회귀모형을 가정한다.

$$\boldsymbol{y} = \boldsymbol{X}_p\boldsymbol{\beta}_p + \boldsymbol{e}$$

오차항이 다음과 같이 서로 독립이고 등분산성을 만족한다면

$$V(\boldsymbol{e}) = \sigma^2\boldsymbol{I}_n$$

최소제곱법에 의한 회귀계수 추정량 $\hat{\boldsymbol{\beta}}_p$ 다음과 같고

$$\hat{\boldsymbol{\beta}}_p = (\boldsymbol{X}_p^t\boldsymbol{X}_p)^{-1}\boldsymbol{X}_p^t\boldsymbol{y}$$

> **!** Important
>
> 이 절의 선형 회귀모형에서는 독립변수의 개수가 p개인 것을 강조하기 위하여 $\boldsymbol{X}_p$ 와 $\boldsymbol{\beta}_p$ 를 사용하였다.

모든 가능한 회귀모형의 개수는 2^p-1개이므로 p가 크지 않다면 가능한 모든 회귀모형을 비교하여 하나의 모형을 선택하는 것이 좋을 것이다. 여러가지 모형들을 비교할 수 있는 모형 선택의 측도들을 알아보자.

6.2.1. 결정계수

총제곱합에서 회귀모형으로 설명할 수 있는 변동 모형 제곱합이 차지하는 부분의 비율, 즉 모형제곱합 SSR을 총제곱합 SST으로 나눈 비율을 결정계수(coefficient of determination)라 하며 R^2으로 표현한다.

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \tag{6.1}$$

결정계수 R^2는 언제나 0 이상 1 이하의 값을 갖는다. 회귀모형이 데이터에 아주 잘 적합되면 결정계수의 값은 1에 가깝게 된다.

주의할 점은 회귀식에 독립변수를 추가하면 결정계수는 언제나 증가한다. 즉 반응변수와 관련이 없는 변수도 회귀식에 추가하면 결정계수의 값이 증가하기 때문에 결정계수로 모형을 선택하면 언제나 모든 독립변수가 모형에 들어간 가장 큰 모델이 선택된다.

6.2.2. 결정계수의 수정

수정 결정계수 $\tilde{R}^2$는 독립변수의 개수가 증가함에 따라 증가하는 결정계수 R^2를 보정한 모형 선택의 척도이다.

$$\begin{aligned}\tilde{R}^2 &= 1 - \frac{SSE_p/(n-p)}{SST/(n-1)} \\ &= 1 - \frac{s_p^2}{SST/(n-1)}\end{aligned}$$

여기서 p는 회귀모형에 포함된 독립변수의 개수이다.

6.2.3. Mallow's C_p

모형의 적합도를 측정하기 위한 여러 가지 통계량중 가장 중요하고 자주 쓰이는 통계량이 평균제곱오차(mean squared error; MSE)이다 이 책에서는 평균제곱오차를 Δ_p^2 으로 표시할 것이다.

반응변수 y_i의 평균을 $\mu_i = E(y_i)$로 하고 독립 변수의 개수가 p개인 선형회귀 모형에서 최소제곱법에 의한 예측값을 $\hat{y}_{ip} = \boldsymbol{x}_{ip}^t \hat{\boldsymbol{\beta}}_p$라고 하면 MSE는 다음과 같이 주어진다.

$$\begin{aligned}E[(\hat{y}_{ip} - \mu_i)^2] &= E[(\boldsymbol{x}_{ip}^t \hat{\boldsymbol{\beta}}_p - \mu_i)^2] \\ &= E[(\boldsymbol{x}_{ip}^t \hat{\boldsymbol{\beta}}_p - E(\boldsymbol{x}_{ip}^t \hat{\boldsymbol{\beta}}_p) + E(\boldsymbol{x}_{ip}^t \hat{\boldsymbol{\beta}}_p) - \mu_i)^2] \\ &= Var(\boldsymbol{x}_{ip}^t \hat{\boldsymbol{\beta}}_p) + [E(\boldsymbol{x}_{ip}^t \hat{\boldsymbol{\beta}}_p) - \mu_i]^2 \\ &= \sigma^2 \boldsymbol{x}_{ip}^t (\boldsymbol{X}_p^t \boldsymbol{X}_p^t)^{-1} \boldsymbol{x}_{ip} + (\eta_{ip} - \mu_i)^2\end{aligned}$$

여기서 $\eta_{ip} = E(\boldsymbol{x}_{ip}^t \hat{\boldsymbol{\beta}}_p)$이다. 여기서 유의할 점은 반응변수 y_i의 평균을 μ_i와 η_{ip} 는 다를 수도 있으며 그 차이를 모형에 의한 편이(bias)라고 한다.

$$E(\boldsymbol{x}_{ip}^t \hat{\boldsymbol{\beta}}_p) - \mu_i = \eta_{ip} - \mu_i$$

이제 평균제곱오차를 구하기 위하여 각각의 관측값 $y_1, y_2, \ldots, y_n$에 대한 제곱합을 구해보자

$$\begin{aligned}\Delta_p^2 &= \sum_i E(\hat{y}_{ip} - \mu_i)^2 \\ &= \sigma^2 \sum_i \boldsymbol{x}_{ip}^t (\boldsymbol{X}_p^t \boldsymbol{X}_p^t)^{-1} \boldsymbol{x}_{ip} + \sum_i (\eta_{ip} - \mu_i)^2 \\ &= \sigma^2 tr(\boldsymbol{X}_p (\boldsymbol{X}_p^t \boldsymbol{X}_p^t)^{-1} \boldsymbol{X}_p^t) + \sum_i (\eta_{ip} - \mu_i)^2 \\ &= p\sigma^2 + SSB_p\end{aligned}$$

여기서 $SSB_p = \sum_i (\eta_{pi} - \mu_i)^2$이며 예측값의 편이들의 제곱합이다.평균제곱오차 Δ_p^2 은 모형에서 추정된 값이 실제 평균과 가까운 정도를 나타내는 측도이지만 실제로 자료를 이용하여 구할 수는 없는 양이다.

여기서 중요한 점은 평균제곱오차 Δ_p^2 는 분산과 편차 제곱들의 합이다.

실제 평균제곱오차 Δ_p^2는 계산할 수 있는 값이 아니므로 이를 적절히 추정할 수 있는 통계량으로 잔차제곱합 (SSE)를 생각해 보자. 독립 변수의 개수가 p개인 선형회귀 모형에 의한 잔차제곱합을 고려하고 그 기대값을 구해보면

$$\begin{aligned} E(SSE_p) &= E[\boldsymbol{y}^t(\boldsymbol{I} - \boldsymbol{X}_p(\boldsymbol{X}_p^t\boldsymbol{X}_p)^{-1}\boldsymbol{X}_p^t)\boldsymbol{y}] \\ &= tr(\sigma^2(\boldsymbol{I} - \boldsymbol{X}_p(\boldsymbol{X}_p^t\boldsymbol{X}_p)^{-1}\boldsymbol{X}_p^t)) + E(\boldsymbol{y})^t(\boldsymbol{I} - \boldsymbol{X}_p(\boldsymbol{X}_p^t\boldsymbol{X}_p)^{-1}\boldsymbol{X}_p^t)E(\boldsymbol{y}) \\ &= \sigma^2 tr((\boldsymbol{I} - \boldsymbol{X}_p(\boldsymbol{X}_p^t\boldsymbol{X}_p)^{-1}\boldsymbol{X}_p^t)) + E(\boldsymbol{y})^t(\boldsymbol{I} - \boldsymbol{H}_p)(\boldsymbol{I} - \boldsymbol{H}_p)E(\boldsymbol{y}) \\ &= \sigma^2(n-p) + [E(\boldsymbol{y}) - E(\hat{\boldsymbol{y}}_p)]^t[E(\boldsymbol{y}) - E(\hat{\boldsymbol{y}}_p)] \\ &= \sigma^2(n-p) + SSB_p \end{aligned}$$

위의 결과는 $\boldsymbol{I} - \boldsymbol{X}_p(\boldsymbol{X}_p^t\boldsymbol{X}_p)^{-1}\boldsymbol{X}_p^t = \boldsymbol{I} - \boldsymbol{H}_p$가 멱등행렬인 사실과 아래의 식을 이용하였다.

$$\eta_p = E(\hat{\boldsymbol{y}}_p) = E(\boldsymbol{X}_p^t\hat{\boldsymbol{\beta}}_p) = \boldsymbol{X}_p(\boldsymbol{X}_p^t\boldsymbol{X}_p)^{-1}\boldsymbol{X}_p^t E(\boldsymbol{y}) = \boldsymbol{H}_p E(\boldsymbol{y})$$

만약 σ^2의 불편추정량을 $\hat{\sigma}^2$라 하면 식 6.2 와 6.2 를 이용하여 다음과 같은 결과를 얻는다.

$$\begin{aligned} E[SSE_p - (n-2p)\hat{\sigma}^2] &= \sigma^2(n-p) + SSB_p - (n-2p)E(\hat{\sigma}^2) \\ &= p\sigma^2 + SSB_p \\ &= \Delta_p^2 \end{aligned}$$

따라서 평균제곱오차 Δ_p^2의 추정량으로 $SSE_p - (n-2p)\hat{\sigma}^2$ 을 사용할 수 있다. Mallow(1973)가 제안한 Mallow's C_p는 평균제곱오차를 분산의 추정량으로 나눈값 Δ_p^2/σ^2이며 이를 최소화는 모형을 선택할 것을 Mallow가 제안하였다.

$$C_p = \frac{SSE_p}{\hat{\sigma}^2} + (2p - n) \tag{6.2}$$

식 6.2 에서 주어진 C_p 에서 $\hat{\sigma}^2$은 고려하는 모든 변수를 포함하는 모형(full model)에서 구한 오차항 분산의 추정량이다. Mallow(1973)는 Δ_p^2이 SSB_p가 0일 때, 즉 $E(\hat{\boldsymbol{y}}_p) = E(\boldsymbol{y})$일 때 최소값 $p\sigma^2$를 같는다는 사실에 의거하여 C_p와 p에 대한 그림을 그리고 C_p의 값이 해당하는 p값에 가깝거나 작은 모형을 선택하는 탐색적 방법을 제안하였다.

여기서 주목할 점은 Mallow's C_p 에서 설명변수의 개수 p의 개수를 크게 하면 SSE_p 는 작아지지만 항 $2p-n$은 증가하게 된다. 따라서 SSE_p에 더해주는 항 $2p-n$ 은 설명변수의 증가에 따른 벌칙항(penalty term)으로 볼 수 있다.

6.2.4. PRESS

PRESS는 prediction error sum of square의 약자로 Cross-validation에 의거한 모형선택을 위한 척도이다. 전차분석에서 보았던 처럼 i번째 관측치 $(y_i, \boldsymbol{x}_i)$를 제외한 반응변수 벡터, 계획행렬, 회귀계수를 각각 $\boldsymbol{y}_{-i}$, $\boldsymbol{X}_{-i}$, $\hat{\boldsymbol{\beta}}_{-i}$와 같이 표시하고 그에 해당하는 예측값을 $\hat{y}_{ip,-i}$라 하면 RESS는 다음과 같이 정의된다.

$$PRESS_p = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_{ip,-i})^2 \tag{6.3}$$

여기서 잔차분석에서 유도한 것처럼

$$y_i - \hat{y}_{ip,-i} = \frac{r_i}{1-h_{ii}}$$

를 이용하면 PRESS를 다음과 같이 표현할 수 있다.

$$PRESS_p = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{y_i - \hat{y}_{ip}}{1-h_{ii}}\right]^2 \approx \frac{SSE_p}{n(1-p/n)^2}$$

위의 식 마지막 근사는 모든 h_{ii}가 그 평균값 p/n에 가깝다는 가정 하에 세워진 식이다.

6.3. AIC 와 BIC

통계모형을 선택하는 척도로서 가능도함수이론에 근거한 AIC(Akaike information criteria)와 베이지안 검정이론에 기초한 BIC(bayesian or schwartz information criteria)가 있다.

AIC와 BIC는 회귀분석뿐 아니라 일반적인 통계 모형에서 자주 사용하는 모형의 선택에 대한 척도이다. AIC와 BIC의 정의는 다음과 같다.

$$AIC = -2\log \ell(\hat{\boldsymbol{\theta}}) + 2k \tag{6.4}$$

$$BIC = -2\log \ell(\hat{\boldsymbol{\theta}}) + (\log n)k \tag{6.5}$$

여기서 k는 모형에 포함된 모수의 총 개수 이다. $\ell(\hat{\boldsymbol{\theta}})$은 최대가능도추정량 $\hat{\boldsymbol{\theta}}$에서 계산된 로그 가능도함수이다.

선형모형에 대한 가능도 추정에서 식 1.16 에서 보았듯이 정규분포 가정 하에서 회귀모형에 대한 로그 가능도함수는 다음과 같으므로

$$l_n(\hat{\boldsymbol{\theta}}) = l_n(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2} - \frac{n}{2}\log\frac{SSE_p}{n}$$

따라서 선형회귀 모형에서의 AIC와 BIC는 다음과 같이 주어진다.

$$\begin{aligned} AIC &= n\log(2\pi) + n + n\log\frac{SSE_p}{n} + 2(p+1) \\ BIC &= n\log(2\pi) + n + n\log\frac{SSE_p}{n} + (\log n)(p+1) \end{aligned} \tag{6.6}$$

여기서 p는 회귀모형에 포함된 독립변수의 개수이며 오차항의 분산까지 포함하여 모수의 총 개수는 $p+1$ 이다.

이제 잔차제곱합 SSE_p 가 작아지면 AIC 와 BIC의 SSE_p 부분이 작아지지만 각 측도의 벌칙항은 증가하게 된다. 여러 개의 모형을 비교하 때 AIC, BIC 의 값이 작은 모형이 좋은 모형이라고 할 수 있다. 또한 주목할 점은 AIC, BIC 의 벌칙항이 다르며 특히 BIC 의 벌칙항에 표본의 개수 n 이 로그스케일로 포함되어 있다.

AIC 와 BIC 에 대한 이론적인 설명은 Appendix I 에 제시되어 있으니 참고하자.

6.4. 변수 선택법

주어진 설명변수들 중에 반응변수에 유의한 영향을 미치는 변수들을 단계적으로 선택하는 방법(variable selection procedure)은 다음과 같이 세 종류의 방법이 있다.

- Forward selection: Forward selection 방법은 회귀모형에 독립변수를 하나 씩 추가하는 방법이다. 첫 번째 추가하는 변수는 설명변수가 한 개인 모형 중에 결정계수 R^2(또는 다른 측도)이 가장 큰 변수를 선택하며 두번째 부터는 추가되었을 때 R^2 의 증가가 가장 큰 값을 선택하게 된다. 변수의 추가가 멈추는 조건은 추가된 변수가 주어진 신뢰수준에서 유의하지 않을 때이다.
- Backward elimination: Backward elimination 방법은 모든 설명변수를 포함한 가장 큰 회귀모형(full model)에서 설명변수를 하나 씩 제거하는 방법이다. 제거하는 변수의 선택은 변수가 제거되었을 때 R^2 의 감소가 가장 작은 값을 선택하게 된다.
- Stepwise: Stepwise는 Forward selection과 Backward elimination을 조합하여 변수의 추가와 제거가 모두 가능한 방법이다.

변수선택법은 이 방법이 제안되었을 당시 매우 유용한 방법으로 여겨졌다. 그러나 변수선택법의 무리한 남용 등 여러 가지 단점들로 인하여 조심해서 사용해야 한다는 것이 현재의 공통된 의견이다. 변수선택법의 이용과 그 유의사항은 다음과 같이 요약할 수 있다.

- 미숙한 이용자에 의해 남용될 수 있다.
- 다중공선성이 존재할 때 불안정하다.
- Stepwise는 주어진 추가와 제거 시 사용되는 유의수준에 따라 최적의 모형이 다를 수 있다.
- 모든 가능한 회귀 모형(All possible regressions)을 사용하는 것이 대안이 될 수 있다.
- 과적합(overfitting)의 위험성이 크다.
- 변수의 추가나 제거에 통계적 검정법을 쓰는데 여러 가지 위험성이 존재한다 (예로 다중비교 문제)

7. 분산분석 모형

7.1. 서론

선형모형에서 설계행렬(design matrix) $\boldsymbol{X}$가 완전계수(full rank)행렬일 때 회귀계수의 추정치는 최소제곱법에서 구해진 정규방정식의 유일한 해로 구해진다.

$$(\boldsymbol{X}^t\boldsymbol{X})\boldsymbol{\beta} = \boldsymbol{X}^t\boldsymbol{y} \quad \Rightarrow \quad \hat{\boldsymbol{\beta}} = (\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{y}$$

그러나 여러 가지 실험이나 자료의 형태에서 설계행렬 $\boldsymbol{X}$의 계수가 완전하지 않을 때가 있으며(less than full rank)

$$rank(\boldsymbol{X}) = r < p = \text{ number of columns in } \boldsymbol{X}$$

이러한 경우에는 정규방정식에서 유일한 해가 존재하지 않는다. 이 장에서는 이러한 경우의 해결 방법을 알아 보고 일원 배치법에 어떻게 적용되는 자를 알아본다.

7.2. 일원배치법

이제 일원배치법에 대한 통계적 모형에서 모수에 대한 추정을 생각해 보자.

$$y_{ij} = \mu + \alpha_i + e_{ij}, \quad i = 1, 2, \dots, a, \;\; j = 1, 2, \dots, r \tag{7.1}$$

추정해야할 모수는 전체 평균 μ와 각 그룹의 처리 효과 $\alpha_1, \alpha_2, \dots, \alpha_a$ 그리고 분산 σ_E^2이다. 전체 평균과 그룹의 효과는 오차제곱합(Sum of Square Error; SSE)을 최소로 하는 모수를 추정하는 최소제곱법(Least Square method; LS)으로 구할 수 있다.

$$\min_{\mu,\alpha_1,\dots\alpha_a} \sum_{i=1}^{a}\sum_{j=1}^{r}(y_{ij} - \mu - \alpha_i)^2 = \min_{\mu,\alpha_1,\dots\alpha_a} SSE \tag{7.2}$$

위의 오차제곱합이 모든 모수에 대하여 미분가능한 이차식으므로 최소제곱 추정량은 제곱합을 모수에 대하여 미분하고 0 으로 놓아 방정식을 풀어서 얻을 수 있다.

오차제곱합을 모수 μ와 $\alpha_1, \alpha_2, \dots, \alpha_a$ 로 미분하여 0 으로 놓은 방정식은 다음과 같다.

$$\frac{\partial}{\partial \mu} SSE = -2 \sum_{i=1}^{a} \sum_{j=1}^{r} (y_{ij} - \mu - \alpha_i) = 0$$
$$\frac{\partial}{\partial \alpha_i} SSE = -2 \sum_{j=1}^{r} (y_{ij} - \mu - \alpha_i) = 0, \quad i = 1, 2, \dots, a$$

위의 방정식을 정리하면 다음과 같은 $a+1$개의 방정식을 얻는다.

$$\begin{aligned} \mu + \frac{\sum_{i=1}^{a} \alpha_i}{a} &= \bar{\bar{y}} \\ \mu + \alpha_1 &= \bar{y}_{1.} \\ \mu + \alpha_2 &= \bar{y}_{2.} \\ &\cdots\cdots \\ \mu + \alpha_a &= \bar{y}_{a.} \end{aligned} \tag{7.3}$$

위의 방정식에서 첫 번째 방정식은 다른 a개의 방정식을 모두 합한 방정식과 같다. 따라서 모수는 $a+1$개이지만 실제 방정식의 개수는 a개이므로 유일한 해가 얻어지지 않는다. 따라서 유일한 해를 구하려면 하나의 제약조건이 필요하며 일반적으로 다음과 같은 두 개의 조건 중 하나를 사용한다.

7.2.1. set-to-zero condition

첫 번째 효과 α_1를 0으로 놓는 조건을 주는 것이다 ($\alpha_1 = 0$). set-to-zero 조건 하에서는 다음과 같은 추정량이 얻어진다.

$$\hat{\mu} = \bar{y}_{1.}, \quad \hat{\alpha}_1 = 0, \;\; \hat{\alpha}_i = \bar{y}_{i.} - \bar{y}_{1.}, \;\; i = 2, \dots, a \tag{7.4}$$

7.2.2. sum-to-zero condition

처리들의 효과의 합은 0이라는 조건을 주는 것이다 ($\sum_{i=1}^{a} \alpha_i = 0$). sum-to-zero 조건에서는 계수의 추정치가 다음과 같이 주어진다.

$$\hat{\mu} = \bar{\bar{y}}, \quad \hat{\alpha}_i = \bar{y}_{i.} - \bar{\bar{y}}, \;\; i = 1, 2, \dots, a \tag{7.5}$$

여기서 유의할 점은 **개별 모수들의 추정량은 조건에 따라서 달라지지만 집단의 평균을 나타내는 모수 $\mu + \alpha_i$ 에 대한 추정량은 언제나 같다.**

$$\widehat{\mu + \alpha}_i = \hat{\mu} + \hat{\alpha}_i = \bar{y}_{i.}$$

만약에 자료를 아래와 같은 평균 모형으로 나타낼 경우에는 각 평균 μ_i 는 각 그룹의 표본 평균으로 추정된다.

$$y_{ij} = \mu_i + e_{ij}$$

평균 모형에서 각 그룹의 모평균에 대한 최소제곱 추정량은 $\hat{\mu}_i = \bar{y}_{i\cdot}$ 이며 이는 주효과 모형에서의 추정량과 동일하다.

또한 모형에 관계없이 오차항의 분산 σ_E^2 에 대한 추정량은 다음과 같이 주어진다.

$$\hat{\sigma}_E^2 = \frac{\sum_i \sum_j (y_{ij} - \hat{\mu} - \hat{\alpha}_i)^2}{a(r-1)}$$

7.3. 선형모형과 제약 조건

일원배치 모형 7.1 를 다음과 같은 벡터를 이용한 선형모형(linear model, regression model) 형태로 나타내고자 한다.

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e} \tag{7.6}$$

위의 선형모형식의 요소 $\boldsymbol{y}$, $\boldsymbol{X}$, $\boldsymbol{\beta}$, $\boldsymbol{e}$는 다음과 같은 벡터와 행렬로 표현된다.

$$\begin{bmatrix} y_{11} \\ y_{12} \\ \vdots \\ y_{1r} \\ y_{21} \\ y_{22} \\ \vdots \\ y_{2r} \\ \vdots \\ y_{a1} \\ y_{a2} \\ \vdots \\ y_{ar} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & . & . & 0 \\ 1 & 1 & 0 & . & . & 0 \\ 1 & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 0 & . & . & 0 \\ 1 & 0 & 1 & . & . & 0 \\ 1 & 0 & 1 & . & . & 0 \\ 1 & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 1 & . & . & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & . & . & 1 \\ 1 & 0 & 0 & . & . & 1 \\ 1 & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & . & . & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_a \end{bmatrix} + \begin{bmatrix} e_{11} \\ e_{12} \\ \vdots \\ e_{1r} \\ e_{21} \\ e_{22} \\ \vdots \\ e_{2r} \\ \vdots \\ e_{a1} \\ e_{a2} \\ \vdots \\ e_{ar} \end{bmatrix} \tag{7.7}$$

이제 위에서 논의한 최소제곱법을 선형 모형 7.6 에 적용하면 다음과 같이 표현할 수 있다.

$$\min_{\mu,\alpha_1,\ldots\alpha_a} \sum_{i=1}^{a} \sum_{j=1}^{r} (y_{ij} - \mu - \alpha_i)^2 = \min_{\boldsymbol{\beta}} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^t (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \tag{7.8}$$

최소제곱법의 기준을 만족하는 계수 $\boldsymbol{\beta}$는 다음과 같은 정규방정식(normal equation)의 해(solution)이다.

$$\boldsymbol{X}^t\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}^t\boldsymbol{y} \tag{7.9}$$

정규방정식 7.9 을 일원배치의 선형모형식 10.15 에 나타난 $\boldsymbol{y}$, $\boldsymbol{X}$로 이용하여 나타내면 다음과 같다.

$$\begin{bmatrix} ar & r & r & \cdot & \cdot & r \\ r & r & 0 & \cdot & \cdot & 0 \\ r & 0 & r & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ r & 0 & 0 & \cdot & \cdot & r \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \cdot \\ \cdot \\ \alpha_a \end{bmatrix} = \begin{bmatrix} ar\bar{\bar{y}} \\ r\bar{y}_{1.} \\ r\bar{y}_{2.} \\ \cdot \\ \cdot \\ r\bar{y}_{a.} \end{bmatrix} \tag{7.10}$$

정규방정식 7.10 는 위에서 구한 최소제곱법에서 유도된 방정식 7.3 과 같다.

여기서 유의할 점은 선형모형식 10.15 의 계획행렬 $\boldsymbol{X}$ 가 완전 계수(full rank) 행렬이 아니다. 계획행렬 $\boldsymbol{X}$의 첫 번째 열은 다른 열을 합한 것과 같다. 또한 정규 방정식 7.10 에서 $\boldsymbol{X}^t\boldsymbol{X}$ 행렬도 완전계수 행렬이 아니다. 따라서 $\boldsymbol{X}^t\boldsymbol{X}$ 행렬의 역행렬은 존재하지 않는다.

이러한 이유로 모수에 대한 유일한 추정량이 존재하지 않기 때문에 앞에서 언급한 제약 조건을 고려해야 정규방정식을 풀 수 있다.

7.3.1. Set-to-zero 조건에서의 모형과 최소제곱 추정량

만약 Set-to-zero 조건을 가정한다면 모수에서 α_1을 제외하고 선형모형식 10.15 를 다음과 같이 다시 표현할 수 있다.
효과 α_1을 0 으로 놓는다는 것은 α_1을 추정할 필요가 없으므로 모수벡터 $\boldsymbol{\beta}$ 에서 α_1를 빼고 계획행렬에서도 대응하는 열을 제거하는 것이다.

$$\begin{bmatrix} y_{11} \\ y_{12} \\ \vdots \\ y_{1r} \\ y_{21} \\ y_{22} \\ \vdots \\ y_{2r} \\ \vdots \\ y_{a1} \\ y_{a2} \\ \vdots \\ y_{ar} \end{bmatrix} = \begin{bmatrix} 1 & 0 & . & . & 0 \\ 1 & 0 & . & . & 0 \\ 1 & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & . & . & 0 \\ 1 & 1 & . & . & 0 \\ 1 & 1 & . & . & 0 \\ 1 & \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & . & . & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & . & . & 1 \\ 1 & 0 & . & . & 1 \\ 1 & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & . & . & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_2 \\ \alpha_3 \\ \vdots \\ \alpha_a \end{bmatrix} + \begin{bmatrix} e_{11} \\ e_{12} \\ \vdots \\ e_{1r} \\ e_{21} \\ e_{22} \\ \vdots \\ e_{2r} \\ \vdots \\ e_{a1} \\ e_{a2} \\ \vdots \\ e_{ar} \end{bmatrix} \tag{7.11}$$

이제 수정된 모형식 7.11 에 최소제곱법을 적용하여 정규방정식을 구하면 다음과 같은 방정식을 얻는다.

$$\begin{bmatrix} ar & r & r & \cdot & \cdot & r \\ r & r & 0 & \cdot & \cdot & 0 \\ r & 0 & r & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ r & 0 & 0 & \cdot & \cdot & r \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_2 \\ \alpha_3 \\ \cdot \\ \cdot \\ \alpha_a \end{bmatrix} = \begin{bmatrix} ar\bar{\bar{y}} \\ r\bar{y}_{2.} \\ r\bar{y}_{3.} \\ \cdot \\ \cdot \\ r\bar{y}_{a.} \end{bmatrix} \tag{7.12}$$

위의 정규방정 7.12 를 풀면 위에서 언급한 sum-to-zero 조건에서 구해지는 모수의 추정량 7.4 를 얻을 수 있다.

7.3.2. Sum-to-zero 조건에서의 모형과 최소제곱 추정량

이제 Sum-to-zero 조건에서 모수의 추정에 대해 알아보자. 조건 $\sum_{i=1}^{a} \alpha_i = 0$ 조건을 마지막 모수 α_a 에 대하여 표현하면 다음과 같다.

$$\alpha_a = -\alpha_1 - \alpha_2 - \cdots - \alpha_{a-1}$$

따라서 마지막 처리 α_a 에 대한 관측값에 대한 모형은 다음과 같아 쓸 수 있다.

$$y_{aj} = \mu + \alpha_a + e_{aj} = \mu + (-\alpha_1 - \alpha_2 - \cdots - \alpha_{a-1}) + e_{ij}$$

이러한 결과를 모형방정식에 반영한다. 즉, 모수벡터 $\boldsymbol{\beta}$ 에서 α_a 를 제거하고 계획행렬에 위의 마지막 처리에 대한 효과식을 반영하면 다음과 같은 선형모형식을 얻는다.

$$\begin{bmatrix} y_{11} \\ y_{12} \\ \vdots \\ y_{1r} \\ y_{21} \\ y_{22} \\ \vdots \\ y_{2r} \\ \vdots \\ y_{a1} \\ y_{a2} \\ \vdots \\ y_{ar} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & . & . & 0 \\ 1 & 1 & 0 & . & . & 0 \\ 1 & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 0 & . & . & 0 \\ 1 & 0 & 1 & . & . & 0 \\ 1 & 0 & 1 & . & . & 0 \\ 1 & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 1 & . & . & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & . & . & 1 \\ 1 & 0 & 0 & . & . & 1 \\ 1 & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & . & . & 1 \\ 1 & 0 & 0 & . & . & 1 \\ 1 & -1 & -1 & . & . & -1 \\ 1 & -1 & -1 & . & . & -1 \\ 1 & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & -1 & -1 & . & . & -1 \\ 1 & -1 & -1 & . & . & -1 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_{a-1} \end{bmatrix} + \begin{bmatrix} e_{11} \\ e_{12} \\ \vdots \\ e_{1r} \\ e_{21} \\ e_{22} \\ \vdots \\ e_{2r} \\ \vdots \\ e_{a1} \\ e_{a2} \\ \vdots \\ e_{ar} \end{bmatrix} \tag{7.13}$$

이제 수정된 모형식 7.13 에 최소제곱법을 적용하여 정규방정식을 구하면 다음과 같은 방정식을 얻는다.

$$\begin{bmatrix} ar & 0 & 0 & \cdot & \cdot & 0 \\ 0 & 2r & r & \cdot & \cdot & r \\ 0 & r & 2r & \cdot & \cdot & r \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & r & r & \cdot & \cdot & 2r \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \cdot \\ \cdot \\ \alpha_{a-1} \end{bmatrix} = \begin{bmatrix} ar\bar{\bar{y}} \\ r\bar{y}_{1.} - r\bar{y}_{a.} \\ r\bar{y}_{2.} - r\bar{y}_{a.} \\ \cdot \\ \cdot \\ r\bar{y}_{a-1,.} - r\bar{y}_{a.} \end{bmatrix} \tag{7.14}$$

위의 정규방정 7.14 를 풀면 위에서 언급한 sum-to-zero 조건에서 구해지는 모수의 추정량 7.5 를 얻을 수 있다.

7.3.3. 예제 7.3

교과서 예제 7.3 은 4년제 대학교의 학년별 영어시험 점수 자료이다. 이 자료는 4개의 학년에 대하여 각각 6명의 학생들에 대한 영어시험 점수이다. 이 자료를 이용하여 일원배치법을 적용하고 각 학년의 평균 점수에 대한 추정량을 구해보자.

```
head(english1)
```

```
  score grade
1    81     1
```

```
2    75     1
3    69     1
4    90     1
5    72     1
6    83     1
```

```
english1$grade <- factor(english1$grade)
summary(english1)
```

```
     score       grade
 Min.   :62.00   1:6
 1st Qu.:72.00   2:6
 Median :79.00   3:5
 Mean   :77.33   4:4
 3rd Qu.:81.00
 Max.   :94.00
```

이제 다음과 같이 일원배치 모형을 적합하고 결과를 보자.

```
fit_anova_set0 <- lm(score ~ grade, data=english1)
summary(fit_anova_set0)
```

```
Call:
lm(formula = score ~ grade, data = english1)

Residuals:
   Min     1Q Median     3Q    Max
-9.500 -5.500  0.600  4.667 11.667

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   78.333      2.868  27.312 1.75e-15 ***
grade2        -3.833      4.056  -0.945   0.3579
grade3        -6.933      4.254  -1.630   0.1215
grade4         9.167      4.535   2.021   0.0593 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.025 on 17 degrees of freedom
Multiple R-squared:  0.4341,    Adjusted R-squared:  0.3342
F-statistic: 4.347 on 3 and 17 DF,  p-value: 0.01905
```

함수 `lm()` 을 이용하여 일원배치 모형을 적합한 결과를 보면 default 로 `grade` 에 대하여 1학년 효과 `grade1` 는 0 으로 고정되고 `grade2`, `grade3`, `grade4` 에 대한 추정치는 각각 -3.83, - 6.933, 9.167 로 추정된다. 즉, 범주형 변수의 첫 번째 수준을 0으로 고정하고 이를 기준으로 다른 수준에 대한 효과를 추정한 것이다.

$$\hat{\mu} = 78.333, \quad \hat{\alpha}_1 = 0, \quad \hat{\alpha}_2 = -3.83, \quad \hat{\alpha}_3 = -6.933, \quad \hat{\alpha}_4 = 9.167$$

사용된 계획행렬을 보면 다음과 같다.

```
model.matrix(fit_anova_set0)
```

```
   (Intercept) grade2 grade3 grade4
1            1      0      0      0
2            1      0      0      0
3            1      0      0      0
4            1      0      0      0
5            1      0      0      0
6            1      0      0      0
7            1      1      0      0
8            1      1      0      0
9            1      1      0      0
10           1      1      0      0
11           1      1      0      0
12           1      1      0      0
13           1      0      1      0
14           1      0      1      0
15           1      0      1      0
16           1      0      1      0
17           1      0      1      0
18           1      0      0      1
19           1      0      0      1
20           1      0      0      1
21           1      0      0      1
attr(,"assign")
[1] 0 1 1 1
attr(,"contrasts")
attr(,"contrasts")$grade
[1] "contr.treatment"
```

이제 각 학년에 영어 평균 $\mu + \alpha_i$ 에 대한 추정값 $\widehat{\mu + \alpha_i}$ 를 구해보자.

$$\widehat{\mu + \alpha_i} = \hat{\mu} + \hat{\alpha}_i$$

```
emmeans::emmeans(fit_anova_set0, "grade")
```

```
 grade emmean   SE df lower.CL upper.CL
 1       78.3 2.87 17     72.3     84.4
 2       74.5 2.87 17     68.4     80.6
 3       71.4 3.14 17     64.8     78.0
 4       87.5 3.51 17     80.1     94.9

Confidence level used: 0.95
```

> **Note**
>
> 함수 lm 에서 default 로 설정되는 set-to-zero 조건은 다음과 명령어로 지정할 수 있다.
>
> options(contrasts=c("contr.treatment", "contr.poly"))

이제 일원배치 모형에서 sum-to-zero 조건을 적용하여 모수를 추정해 보자. sum-to-zero 조건을 적용하려면 다음과 같은 명령어를 실행해야 한다.

```
options(contrasts=c("contr.sum", "contr.poly"))
```

이제 sum-to-zero 조건을 적용하여 일원배치 모형을 적합하고 결과를 보자.

```
fit_anova_sum0 <- lm(score ~ grade, data=english1)
summary(fit_anova_sum0)
```

```
Call:
lm(formula = score ~ grade, data = english1)

Residuals:
   Min     1Q Median     3Q    Max
-9.500 -5.500  0.600  4.667 11.667

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   77.933      1.554  50.135   <2e-16 ***
grade1         0.400      2.555   0.157   0.8775
grade2        -3.433      2.555  -1.344   0.1967
grade3        -6.533      2.711  -2.410   0.0276 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 7.025 on 17 degrees of freedom
Multiple R-squared:  0.4341,    Adjusted R-squared:  0.3342
F-statistic: 4.347 on 3 and 17 DF,  p-value: 0.01905
```

sum-to-zero 조건하에서 계획행렬은 다음과 같다.

```
model.matrix(fit_anova_sum0)
```

```
   (Intercept) grade1 grade2 grade3
1            1      1      0      0
2            1      1      0      0
3            1      1      0      0
4            1      1      0      0
5            1      1      0      0
6            1      1      0      0
7            1      0      1      0
8            1      0      1      0
9            1      0      1      0
10           1      0      1      0
11           1      0      1      0
12           1      0      1      0
13           1      0      0      1
14           1      0      0      1
15           1      0      0      1
16           1      0      0      1
17           1      0      0      1
18           1     -1     -1     -1
19           1     -1     -1     -1
20           1     -1     -1     -1
21           1     -1     -1     -1
attr(,"assign")
[1] 0 1 1 1
attr(,"contrasts")
attr(,"contrasts")$grade
[1] "contr.sum"
```

sum-to-zero 조건하에서 일원배치 모형을 적합한 결과를 보면 다음과 같다.

$$\hat{\mu} = 77.933, \quad \hat{\alpha}_1 = 0.400, \quad \hat{\alpha}_2 = -3.433, \quad \hat{\alpha}_3 = -6.533, \quad \hat{\alpha}_4 = -(\hat{\alpha}_1 + \hat{\alpha}_2 + \hat{\alpha}_3) = 9.566$$

이렇게 set_to_zero 조건과 sum-to-zero 조건을 적용한 결과는 다르지만 각 학년의 평균에 대한 추정치는 동일하다.

```
emmeans::emmeans(fit_anova_sum0, "grade")
```

```
 grade emmean   SE df lower.CL upper.CL
 1       78.3 2.87 17     72.3     84.4
 2       74.5 2.87 17     68.4     80.6
 3       71.4 3.14 17     64.8     78.0
 4       87.5 3.51 17     80.1     94.9

Confidence level used: 0.95
```

> **Note**
>
> 참고로 교과서에서 4학년을 기준으로 다른 학년들과 비교하기 위하여 변수 grade 의 수준(level) 의 순서를 설정할 때 다음과 같이 4학년을 첫 번째 수준으로 설정하면 된다.

```
options(contrasts=c("contr.treatment", "contr.poly"))
english1$grade <- factor(english1$grade, levels = c(4,1,2,3))
str(english1)
```

```
'data.frame':   21 obs. of  2 variables:
 $ score: int  81 75 69 90 72 83 65 80 73 79 ...
 $ grade: Factor w/ 4 levels "4","1","2","3": 2 2 2 2 2 2 3 3 3 3 ...
```

```
fit_anova_set0_1 <- lm(score ~ grade, data=english1)
summary(fit_anova_set0_1)
```

```

Call:
lm(formula = score ~ grade, data = english1)

Residuals:
   Min     1Q Median     3Q    Max
-9.500 -5.500  0.600  4.667 11.667

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   87.500      3.513  24.910 8.06e-15 ***
grade1        -9.167      4.535  -2.021  0.05927 .
grade2       -13.000      4.535  -2.867  0.01069 *
grade3       -16.100      4.713  -3.416  0.00329 **
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.025 on 17 degrees of freedom
Multiple R-squared:  0.4341,    Adjusted R-squared:  0.3342
F-statistic: 4.347 on 3 and 17 DF,  p-value: 0.01905
```

7.4. 불완전 계수행렬에서의 추정

설계행렬 $\boldsymbol{X}$의 계수가 완전하지 않을 때 회귀 계수를 추정하기 위한 방법으로서 다음과 같은 세 가지 방법이 있다.

7.4.1. 모수의 재조정 (reparameterization)

$\boldsymbol{X}$의 계수가 완전하지 않을 때 설계행렬의 열을 다시 구성하여 계수를 완전하게 하는 방법이 있다. 즉 $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2)$으로 표시하고 $\boldsymbol{X}_1$을 $n \times r$ $(r < p)$라고 하며 어떤 행렬 $\boldsymbol{F}$가 존재하여 $\boldsymbol{X}_2 = \boldsymbol{X}_1 \boldsymbol{F}$의 관계를 가진다고 가정하자. 이러한 관계는 $\boldsymbol{X}_2$의 열들이 $\boldsymbol{X}_1$의 열들의 선형결합으로 표현될 수 있다는 것을 의미한다. 이러한 경우에 선형모형은 다음과 같이 표현될 수 있다.

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e} = \boldsymbol{X}_1(\boldsymbol{I}, \boldsymbol{F})\boldsymbol{\beta} + \boldsymbol{e} = \boldsymbol{X}_1\boldsymbol{\alpha} + \boldsymbol{e}$$

여기서 새롭게 조정된 계수 $\boldsymbol{\alpha}$와 처음의 계수 $\boldsymbol{\beta}$는 다음과 같은 관계가 있다.

$$\boldsymbol{\alpha} = (\boldsymbol{I}, \boldsymbol{F})\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$$

따라서 새롭게 구성된 선형모형 $\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\alpha} + \boldsymbol{e}$에서 새로운 계수의 추정치는 $\hat{\boldsymbol{\alpha}} = (\boldsymbol{X}_1^t \boldsymbol{X}_1)^{-1} \boldsymbol{X}_1^t \boldsymbol{y}$ 이다.

7.4.2. 부가 조건의 이용

회귀계수에 부가 조건(side condition)을 주면 유일한 계수의 추정치를 구할 수 있다. 즉 $(p-r) \times p$ 행렬 $\boldsymbol{H}$를 고려하고 $\boldsymbol{H}\boldsymbol{\beta} = 0$이라는 부가조건을 가정하자. 즉 모든 $\boldsymbol{\eta} = R(\boldsymbol{X})$에 대하여 $\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta}$와 $\boldsymbol{H}\boldsymbol{\beta} = 0$를 만족하는 $\boldsymbol{\beta}$는 유일하게 존재한다.

$$\begin{bmatrix} \boldsymbol{\eta} \\ \boldsymbol{0} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X} \\ \boldsymbol{H} \end{bmatrix} \boldsymbol{\beta} \equiv \boldsymbol{G}\boldsymbol{\beta}, \quad \text{with} \quad \boldsymbol{G} = \begin{bmatrix} \boldsymbol{X} \\ \boldsymbol{H} \end{bmatrix}$$

이러한 부가 조건 $\boldsymbol{H}\boldsymbol{\beta} = 0$과 정규방정식 $(\boldsymbol{X}^t\boldsymbol{X})\boldsymbol{\beta} = \boldsymbol{X}^t\boldsymbol{y}$를 동시에 만족하는 유일한 해를 구하고 이를 최소제곱 추정량으로 한다. 이러한 부가 조건을 주는 방법은 분산분석을 이용하는 여러 가지 선형 모형 (예: 일원 배치법)에 자주 사용된다.

7.4.3. 일반화 역행렬의 이용

$\boldsymbol{X}$의 계수가 완전하지 않을 때 일반화 역행렬(generalized inverse matrix)를 이용하면 회귀계수의 추정치를 구할 수 있다.

여기서 $m \times n$ 행렬 $\boldsymbol{A}$의 일반화 역행렬 $\boldsymbol{A}^-$는 다음을 만족하는 행렬이다.

$$\boldsymbol{A} = \boldsymbol{A}\boldsymbol{A}^-\boldsymbol{A}$$

일반화 역행렬은 일반적으로 유일하지 않다. $\boldsymbol{A}$가 정방행렬이고 정칙행렬일 때 유일하게 존재하며 $\boldsymbol{A}^- = \boldsymbol{A}^{-1}$이다. 정규방정식 $\boldsymbol{X}^t\boldsymbol{y} = \boldsymbol{X}^t\boldsymbol{X}\hat{\boldsymbol{\beta}}$의 양변에 $\boldsymbol{X}^t\boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{X})^-$를 곱하면

$$\boldsymbol{X}^t\boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{X})^-\boldsymbol{X}\boldsymbol{y} = \boldsymbol{X}^t\boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{X})^-\boldsymbol{X}^t\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}^t\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}^t\boldsymbol{y}$$

이므로 $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^t\boldsymbol{X})^-\boldsymbol{X}^t\boldsymbol{y}$는 정규방정식의 해가 된다. 앞에서 언급하였듯이 일반화 역함수를 이용한 계수의 추정량은 유일하지 않다. 그러나 반응변수의 추정량 $\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$는 추정된 계수에 관계없이 유일하다.

7.5. 추정 가능한 함수

7.5.1. 일원배치법에 추정가능한 모수

앞 절에서 보았듯이 일원배치법을 선형 모형식으로 표현하는 경우 평균에 대한 모수는 모두 $a+1$ 개가 있다.

$$\mu, \alpha_1, \alpha_2, \cdots, \alpha_a$$

하지만 모형식에서 계획행렬 $\boldsymbol{X}$가 완전 계수 행렬이 아니기 때문에 1개의 제약 조건을 가정하고 모수를 추정하였다. 하지만 제약 조건이 달라지면 각 모수의 추정량이 달라지기 때문에 각 모수는 유일한 값으로 추정이 불가능하다.

이렇게 각 모수들은 제약 조건에 따라서 유일하게 추정이 불가능하지만 앞 절에서 보았듯이 $\mu+\alpha_i$ 에 대한 추정량은 제약조건에 관계없이 표본 평균 $\bar{y}_{i\cdot}$으로 동일하게 추정되어 진다.

그러면 어떤 모수들은 유일하게 추정이 불가능하고 어떤 모수들이 유일하게 추정이 가능할까?

이제 제약조건이 달라도 유일하게 추정이 가능한 모수들의 형태를 살펴보자.

7.5.2. 추정가능한 모수의 함수

선형모형 $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{e}$ 에서 계획행렬 $\boldsymbol{X}$의 계수가 완전하지 않으면 모수 벡터 $\boldsymbol{\beta}$는 유일한 값으로 추정할 수 없다.

이제 임의의 벡터 $\boldsymbol{c}$가 있을 때 모수들의 선형결합 $\psi = \boldsymbol{c}^t\boldsymbol{\beta}$를 고려하자.

예를 들어 일원배치 모형에서는 다음과 같은 모수들의 선형결합을 고려하는 것이다.

$$\psi = \boldsymbol{c}^t\boldsymbol{\beta} = [c_0\ c_1\ c_2\ \cdots\ c_a] \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_a \end{bmatrix} = c_0\mu + c_1\alpha_1 + c_2\alpha_2 + \cdots + c_a\alpha_a$$

위에서 본 것처럼 하나의 모수 α_1에 대한 유일한 추정은 불가능하다.

$$\alpha_1 = (0)\mu + (1)\alpha_1 + (0)\alpha_2 + \cdots + (0)\alpha_a$$

하지만 모수의 조합 $\mu + \alpha_2$ 은 유일한 추정이 가능하다.

$$\mu + \alpha_1 = (1)\mu + (1)\alpha_1 + (0)\alpha_2 + \cdots + (0)\alpha_a$$

이제 문제는 선형조합 $\psi = \boldsymbol{c}^t\boldsymbol{\beta}$ 에서 계수들 $c_0, c_1, \ldots, c_a$가 어떤 값을 가지는 경우 유일한 추정이 가능한 지 알아내는 것이다.

이제 $\psi = \boldsymbol{c}^t\boldsymbol{\beta}$ 에 대한 유일한 추정량 $\hat{\psi}$ 이 있다고 가정하자. 선형 모형에서 추정량 $\hat{\psi}$의 형태는 관측값에 대한 선형함수가 되어야 한다. 따라서 추정량을 $\hat{\psi} = \boldsymbol{a}^t\boldsymbol{y}$ 로 나타낼 수 있다. 이제 추정량 $\hat{\psi}$의 기대값은 $\psi = \boldsymbol{c}^t\boldsymbol{\beta}$이어야 하므로 다음이 성립해야 한다.

$$E(\hat{\psi}|\boldsymbol{X}) = E(\boldsymbol{a}^t\boldsymbol{y}|\boldsymbol{X}) = \boldsymbol{a}^tE(\boldsymbol{y}|\boldsymbol{X}) = \boldsymbol{a}^t\boldsymbol{X\beta} = \boldsymbol{c}^t\boldsymbol{\beta}$$

위의 식에서 가장 마지막 두 항의 관계를 보면 다음이 성립해야 한다.

$$\boldsymbol{a}^t\boldsymbol{X} = \boldsymbol{c}^t \quad \text{equivalently} \quad \boldsymbol{c} = \boldsymbol{X}^t\boldsymbol{a} \tag{7.15}$$

즉 추정가능한 모수의 조합 $\psi = \boldsymbol{c}^t\boldsymbol{\beta}$에서 **계수 벡터 $\boldsymbol{c}$ 는 계획행렬에 있는 행들의 선형 조합**으로 표시되어야 한다는 것이다. 이렇게 유일하게 추정이 가능한 모수의 조합을 **추정가능한 함수(estimable function)**이라고 한다.

7.5.3. 예제

2개의 수준이 있고 반복이 2번 있는 일원배치 ($a = 2, r = 2$) 에 대한 선형모형 10.15 을 생각해보자. 이 경우 계획행렬 $\boldsymbol{X}$ 과 모수벡터 $\boldsymbol{\beta}$ 는 다음과 같다.

$$\boldsymbol{X} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix} \quad \boldsymbol{\beta} = \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \end{bmatrix}$$

이제 유일하게 추정 가능한 모수 조합 ψ 은 어떤 형태일까?

$$\psi = \boldsymbol{c}^t \boldsymbol{\beta} = c_0\mu + c_1\alpha_1 + c_2\alpha_2$$

위의 7.15 에서 추정가능한 모수의 조합에 대한 계수 벡터 $\boldsymbol{c}$ 는 다음과 같은 조건을 만족해야 한다.

$$\boldsymbol{c} = \boldsymbol{X}^t \boldsymbol{a}$$

이제 임의의 벡터 $\boldsymbol{a}$ 에 대하여 $\boldsymbol{c} = \boldsymbol{X}^t\boldsymbol{a}$의 형태를 보자.

$$\begin{aligned} \boldsymbol{c} &= \boldsymbol{X}^t\boldsymbol{a} \\ &= \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \end{bmatrix} \\ &= a_1 \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} + a_2 \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} + a_3 \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} + a_4 \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} \\ &= (a_1 + a_2) \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} + (a_3 + a_4) \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} \\ &= b_1 \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} + b_2 \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} \end{aligned} \tag{7.16}$$

이제 $\boldsymbol{X}^t\boldsymbol{a}$ **는 계획행렬** $\boldsymbol{X}$**에 있는 유일한 행들의 선형조합**임을 알 수 있다.

> **Note**
>
> 위의 7.16 에서 유의할 점은 벡터 $\boldsymbol{a} = [a_1\ a_2\ a_3\ a_4]^t$ 는 임의로 주어진 벡터이다.
> 7.16 에서 $a_1 = 1$, $a_2 = 1$ 인 경우는 $a_1 = 2$, $a_2 = 0$ 인 경우와 동일하다.

따라서 유일하게 추정 가능한 모수의 선형조합 $\psi = \boldsymbol{c}^t\boldsymbol{\beta}$ 에 대한 계수 벡터 $\boldsymbol{c} = [c_0\ c_1\ c_2]^t$ 는 계획행렬 $\boldsymbol{X}$ 의 유일한 행들의 선형 조합으로 구성되어야 한다.

$$\boldsymbol{c} = \begin{bmatrix} c_0 \\ c_1 \\ c_2 \end{bmatrix} = b_1 \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} + b_2 \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} \tag{7.17}$$

- 처리의 효과를 나타내는 모수 α_i 는 추정이 불가능하다.

첫 번째 처리에 대한 효과 모수 α_1 를 선형조합으로 나타내면

$$\alpha_1 = c_0\mu + c_1\alpha_1 + c_2\alpha_2 = (0)\mu + (1)\alpha_1 + (0)\alpha_2$$

따라서 조건 7.17 에서 $\boldsymbol{c}^t = [0\ 1\ 0]$ 을 만들수 있는 계수 b_1 과 b_2 를 찾아야 하는데 이는 불가능하다. 따라서 모수 α_1 은 추정 불가능하다.

$$\boldsymbol{c} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} = b_1 \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} + b_2 \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}$$

- 처리의 평균을 나타내는 모수의 조합 $\mu + \alpha_i$ 는 추정이 가능하다.

모수 조합 $\mu + \alpha_1$ 를 선형조합으로 나타내면

$$\mu + \alpha_1 = c_0\mu + c_1\alpha_1 + c_2\alpha_2 = (1)\mu + (1)\alpha_1 + (0)\alpha_2$$

따라서 조건 7.17 에서 $\boldsymbol{c}^t = [1\ 1\ 0]$ 을 만들수 있는 계수는 $b_1 = 1$과 $b_2 = 0$ 이므로 추정이 가능하다.

$$\boldsymbol{c} = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} = (1) \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} + (0) \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}$$

- 처리 효과의 차이를 나타내는 모수의 조합 $\alpha_1 - \alpha_2$ 는 추정이 가능하다.

$$\alpha_1 - \alpha_2 = c_0\mu + c_1\alpha_1 + c_2\alpha_2 = (0)\mu + (1)\alpha_1 + (-1)\alpha_2$$

따라서 조건 7.17 에서 $\boldsymbol{c}^t = [0\ 1\ -1]$ 을 만들수 있는 계수는 $b_1 = 1$과 $b_2 = -1$ 이므로 추정이 가능하다.

$$\boldsymbol{c} = \begin{bmatrix} 0 \\ 1 \\ -1 \end{bmatrix} = (1) \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} + (-1) \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}$$

7.6. 가설 검정

이제 효과모형 7.1 을 고려하면 집단들 사이에 차이가 있는지에 대한 가설을 다음과 같이 고려할수 있다. 집단에 대한 효과가 모두 0이 되면 집단 간의 평균에 대한 차이는 없다.

$$H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_a = 0 \quad \text{vs.} \quad H_1 : \text{ not } H_0 \tag{7.18}$$

7.6.1. 변동의 분해

이제 집단 간의 변동(각 집단의 평균의 차이가 얼마나 나는지에 대한 통계량)과 집단 내의 변동(각 집단내에서 관측값들의 퍼진 정도)를 측정하는 통계량을 찾아서 검정 통계량을 구성해 보자.

일단 각 집단의 반복 측정값의 횟수는 모두 같다고 가정하자($r_i = r$). 전체 평균과 집단의 평균을 정의하자.

$$\bar{y}_{..} = \frac{\sum_{i=1}^{a} \sum_{j=1}^{r} y_{ij}}{ar}, \quad \bar{y}_{i.} = \frac{\sum_{j=1}^{r} y_{ij}}{r}$$

이제 하나의 관측값 y_{ij}과 전체 평균 $\bar{y}_{..}$ 간의 편차(deviation)를 다음과 같이 분해해 보자.

$$\underbrace{y_{ij} - \bar{y}_{..}}_{\text{total deviation}} = \underbrace{(y_{ij} - \bar{y}_{i.})}_{\text{within-group deviation}} + \underbrace{(\bar{y}_{i.} - \bar{y}_{..})}_{\text{between-group deviation}} \tag{7.19}$$

7.19 에서 집단 평균과 총 평균의 편차 ($\bar{y}_{i.} - \bar{y}_{..}$)는 처리의 효과를 측정할 수 있는 통계량이다. 집단 간의 차이를 반영하는 양으로 처리 효과 α_i들에 의하여 발생한다.

집단 내의 관측값과 집단 평균의 차이 ($y_{ij} - \bar{y}_{i.}$)는 집단 내의 변동을 나타내는 통계량으로 측정 오차 e_{ij}에 의하여 발생한다.

7.19 의 각 편차들은 양수와 음수로서 부호를 가지기 때문에 이를 변동으로 표현하기 위하여 차이를 제곱하여 합친 제곱합(sum of squares)을 고려해 보자.

$$\begin{aligned}
\sum_{i=1}^{a} \sum_{j=1}^{r} (y_{ij} - \bar{y}_{..})^2 &= \sum_{i=1}^{a} \sum_{j=1}^{r} \left[(y_{ij} - \bar{y}_{i.}) + (\bar{y}_{i.} - \bar{y}_{..}) \right]^2 \\
&= \sum_{i=1}^{a} \sum_{j=1}^{r} (y_{ij} - \bar{y}_{i.})^2 + \sum_{i=1}^{a} \sum_{j=1}^{r} (\bar{y}_{i.} - \bar{y}_{..})^2 + 2 \sum_{i=1}^{a} \sum_{j=1}^{r} (y_{ij} - \bar{y}_{i.})(\bar{y}_{i.} - \bar{y}_{..}) \\
&= \sum_{i=1}^{a} \sum_{j=1}^{r} (y_{ij} - \bar{y}_{i.})^2 + \sum_{i=1}^{a} r(\bar{y}_{i.} - \bar{y}_{..})^2 + 0(why?)
\end{aligned}$$

결과적으로 다음과 같은 변동의 분해를 제곱합의 형식으로 얻을 수 있다.

$$\underbrace{\sum_{i=1}^{a}\sum_{j=1}^{r}(y_{ij}-\bar{y}_{..})^2}_{\text{total variation}} = \underbrace{\sum_{i=1}^{a}\sum_{j=1}^{r}(y_{ij}-\bar{y}_{i.})^2}_{\text{within-group variation}} + \underbrace{\sum_{i=1}^{a}\sum_{j=1}^{r}(\bar{y}_{i.}-\bar{y}_{..})^2}_{\text{between-group variation}} \tag{7.20}$$

분해식 7.20 에서 나타난 각 제곱합에 대한 이름과 의미를 살펴보자.

- SST를 총 제곱합(Total Sum of Squares)이라고 부르며 자료의 전체 변동을 의미한다.

$$SST = \sum_{i=1}^{a}\sum_{j=1}^{r}(y_{ij}-\bar{y}_{..})^2$$

- SSE를 잔차 제곱합(Residual Sum of Squares)이라고 부르며 관측 오차에 발생된 집단 내의 변동 또는 급내 변동(within-group variation)을 의미한다.

$$SSE = \sum_{i=1}^{a}\sum_{j=1}^{r}(y_{ij}-\bar{y}_{i.})^2$$

- SSA를 처리 제곱합(Treatment Sum of Squares)이라고 부르며 처리들의 차이로 발생하는 변동으로거 집단 간의 변동 또는 급간 변동(bwtween-group variation)을 의미한다.

$$SSA = \sum_{i=1}^{a}\sum_{j=1}^{r}(\bar{y}_{i.}-\bar{y}_{..})^2 = \sum_{i=1}^{a} r(\bar{y}_{i.}-\bar{y}_{..})^2$$

이제 분해식 7.20 을 다음과 같이 나타낼수 있다.

$$SST = SSA + SSE \tag{7.21}$$

위의 분해식에서 볼 수 있듯이 집단 간의 변동의 크기를 나타내는 처리제곱합이 커질수록, 또는 집단내의 변동의 크기를 나타내는 오차제곱합이 작아질수록 귀무가설에 반대되는(즉, 집단 간의 평균이 유의한 차이가 난다는) 증거가 강해진다.

7.6.2. 제곱합과 F-통계량

이제 가설 7.18 을 검정하기 위한 통계량을 구성해 보자. 먼저 다음과 같은 제곱합들을 각 자유도로 나눈 평균제곱합(Mean Sum of Squares)를 정의한다.

$$MS_A = \frac{SSA}{\phi_A}, \quad MSE = \frac{SSE}{\phi_E} \tag{7.22}$$

집단 간의 변동과 집단 내의 변동의 상대적 비율로 그룹 간의 차이를 검정할 수 있다는 개념을 확장하여 다음과 같은 F-통계량 F_0 를 만들어 보자.

$$F_0 = \frac{MSA}{MSE} = \frac{\text{between-group variation}}{\text{within-group variation}} \tag{7.23}$$

위 7.23 에서 정의된 F-통계량은 그룹 간에 평균의 차이가 클수록, 그룹 내의 차이가 작을 수록 그 값이 커진다. 따라서 F-통계량의 값이 크면 클수록 귀무가설에 반대되는 증거가 강해진다.

이렇게 전체의 변동을 집단 간의 변동과 집단 내의 변동으로 나누어 집단 간의 평균의 차이를 추론하는 방법을 **분산분석**(Analysis of Variance, **ANOVA**)이라고 한다.

이제 7.23 에서 정의된 F-통계량을 이용하여 가설 @ref(eq:hypo1)를 검정하는 통계적 방법을 만들어 보자. 일단 두 제곱합의 통계적 성질은 다음과 같다.

- 잔차 제곱합을 오차항의 분산으로 나눈 통계량은 자유도가 ϕ_E 를 가지는 카이제곱 분포를 따른다.

$$\frac{SSE}{\sigma_E^2} \sim \chi^2(\phi_E)$$

- **귀무가설이 참인 경우** 처리 제곱합을 오차항의 분산으로 나눈 통계량은 자유도가 ϕ_A 를 가지는 카이제곱 분포를 따른다.

$$\frac{SSA}{\sigma_E^2} \sim \chi^2(\phi_A) \quad \text{under } H_0$$

- 잔차 제곱합과 처리 제곱합은 서로 독립이다.

따라서 **귀무가설이 참인 경우** F-통계량은 자유도가 ϕ_A, ϕ_E를 가지는 F-분포를 따른다.

$$F_0 = \frac{MSA}{MSE} = \frac{\frac{SSA/\sigma_E^2}{\phi_A}}{\frac{SSE/\sigma_E^2}{\phi_E}} \sim F(\phi_A, \phi_E) \quad \text{under } H_0$$

{#eq-anova-ftest)

유의수준 α에서 F-통계량이 기각역을 벗어나면 귀무가설을 기각한다.

$$\text{Reject } H_0 \text{ if } F_0 > F(1-\alpha, \phi_A, \phi_E)$$

또는 다음과 같이 계산된 p-값이 유의수준 α 보다 작으면 귀무가설을 기각한다.

$$p - value = P[F(\phi_A, \phi_E) > F_0]$$

F-통계량을 정의할 때 편리하고 유용하게 사용되는 것이 다음과 같은 분산분석표(ANOVA table)이다.

요인	제곱합	자유도	평균제곱합	F_0	p-값
처리	SSA	$\phi_A = a-1$	$MSA = SSA/\phi_A$	$F_0 = MSA/MSE$	$P[F(\phi_A, \phi_E) > F_0]$
잔차	SSE	$\phi_E = a(r-1)$	$MSE = SSE/\phi_E$		
총합	SST	$\phi_T = ar-1$			

7.7. 다중비교

7.7.1. 일원배치에서 평균의 비교

분산분석표를 이용한 F-검정으로 귀무가설을 기각하면 모든 처리 수준의 평균이 같지 않다는 결론을 내리고 어떤 집단 간에 평균의 차이가 유의한지 더 분석해야 한다. 평균 차이에 대한 신뢰구간과 가설 검정은 아래와 같이 주어진다.

두 수준 평균의 차이 $\delta_{ij} = \mu_i - \mu_j$ 에 대한 $100(1-\alpha)$ % 신뢰구간은 다음과 같이 주어진다.

$$(\bar{x}_{i.} - \bar{x}_{j.}) \pm t(1-\alpha/2, \phi_E)\sqrt{\frac{2MSE}{r}} \tag{7.24}$$

두 평균의 차이 δ_{ij} 에 대한 가설을 검정은 유의 수준 α에서 다음과 같은 조건을 만족하면 위의 귀무가설을 기각한다.

$$\left|\bar{x}_{i.} - \bar{x}_{j.}\right| > t(1-\alpha/2, \phi_E)\sqrt{\frac{2MSE}{r}} \tag{7.25}$$

7.25 에서 검정을 위한 조건의 우변을 최소유의차(least significant difference; LSD) 라고 부른다. 두 수준의 차이가 유의하려면 두 평균 차이의 절대값이 최소한 최소유의차의 값보다 커야한다.

$$\text{LSD} = t(1-\alpha/2, \phi_E)\sqrt{\frac{2MSE}{r}}$$

7.7.2. 두 개 이상의 가설

일원배치 계획에서 수준의 개수가 a 개 인 경우 처리 수준들의 차이에 대하여 비교를 한다면 $\binom{a}{2}$ 개의 가설검정을 수행해야 한다. 예를 들어 처리 수준이 3개 있는 경우 다음과 같이 3개의 조합에 대하여 가설 검정을 수행할 수 있다.

$$H_{01} : \mu_1 = \mu_2, \quad H_{02} : \mu_2 = \mu_3, \quad H_{03} : \mu_3 = \mu_1 \tag{7.26}$$

가설검정에서 사용되는 유의수준(significance level, α)에 대하여 생각해 보자. 지금까지 가설검정을 수행할 때 **유의수준 5%** 라는 말을 사용해 왔는데 이것이 무슨 의미를 가지는지 알아보자.

유의수준 5%라는 것은 수행하는 가설검정에서 귀무가설이 옳은 경우에 기각하는 확률을 말한다. 예를 들어 7.26의 3개의 검정에 대하여 각각 t-검정을 수행하는 경우 귀무가설이 옳은데 우연하게 자료가 극단적으로 나와서 귀무가설을 기각하고 대립가설을 채택하는 확률이 유의수준이며 보통 5%를 사용한다. 이러한 오류를 제 1종의 오류(Type I error; false discovery error;false positve error)라고 한다.

위 7.26 에서 처럼 3개의 가설 검정을 동시에 실시한다면 각각의 가설검정에서 제 1 종의 오류를 범할 확률은 5% 이다. 그런데 3개의 가설 검정을 동시에 실행하므로 다음과 같이 3개의 검정을 합쳐서 다음과 같은 확률에 관심이 있을 수 있다.

3개의 가설검정을 동시에 수행할 때 제 1종의 오류가 최소한 1번 발생할 확률은 얼마인가?

세 개의 가설검정을 동시에 수행하는 경우 세 검정 모두 제 1 종의 오류를 범하거나 두 개 또는 하나의 검정에서 제 1 종의 오류를 범할 사건의 확률은 얼마나 될까? 5%보다 작을까 아니면 클까? 또는 5%인가? 간단한 확률 공식을 이용하여 알아보자.

7.7.3. 실험단위 오류

일단 두 개의 검정 H_{01} 과 H_{02}을 각각 유의수준 $\alpha = 0.05$로서 동시에 수행 한다고 가정하고 다음과 같은 사건을 정의한다.

- A_1: H_{01} 검정에서 제 1 종의 오류를 범하는 사건
- A_2: H_{02} 검정에서 제 1 종의 오류를 범하는 사건

각 검정에서 제 1 종의 오류를 범할 확률을 α라고 가정하자.

$$P(A_1) = P(A_2) = \alpha = 0.05$$

이제 두개의 가설검정을 동시에 수행하는 경우 **제 1 종의 오류를 최소한 1번 범하는 사건**은 $P(A_1 \cup A_2)$ 이며 여사건의 확률공식을 이용하면 다음과 같이 나타낼 수 있다.

$$P(A_1 \cup A_2) = 1 - P(A_1^c \cap A_2^c)$$

여기서 우리는 $P(A_1^c) = P(A_2^c) = 1 - 0.05 = 0.95$를 알 수 있지만 두 사건의 교집합에 대한 확률은 계산하기 쉽지 않다. 왜냐하면 두 사건 A_1과 A_2가 일반적으로 독립이 아니어서 두 확률의 곱으로 쉽게 나타낼 수 없다.

만약에 두 사건이 독립이라면 다음과 같은 결과가 나온다. 즉 두 개의 독립인 가설검정을 동시에 수행하는 경우 최소한 1번의 제 1 종의 오류를 범하는 사건의 학률은 0.0975로 5%의 두 배 정도가 된다.

$$P(A_1 \cup A_2) = 1 - P(A_1^c \cap A_2^c) = 1 - P(A_1^c)P(A_2^c) = 1 - (1 - 0.05)^2 = 0.0975 > 0.05$$

만약 k 개의 독립인 가설검정을 동시에 수행하는 경우 제 1 종의 오류를 최소한 1번 이라도 범하는 사건의 학률은 $1-(1-0.05)^k$ 으로 급격하게 증가한다. 예를 들어 $k=6$인 경우 26.5% 로 5%의 5 배가 된다. 여기서 유의할 점은 이러한 결과는 모든 가설검정이 독립이고 여러 개의 가설검정들을 동시에 고려하는 경우이다.

즉, 두 개 이상의 가설검정을 동시에 고려해서 **제 1 종의 오류를 최소한 1번 범할 경우**를 오류라고 한다면 그 확률은 고려하는 검정의 개수가 증가함에 따라 빠르게 커진다.

이렇게 두 개 이상의 가설검정을 동시에 고려해서 계산하는 오류의 확률을 **실험단위 오류(Experiment-wise error 또는 Family-wise error)**라고 하며 반대로 가설검정을 동시에 고려하지 않고 개별적로 생각하는 오류를 **개별단위 오류(Individual-wise error)**라고 한다.

7.7.4. 예제: 2개의 가설을 가진 임상실험

임상실험에서 신약(처리 1)의 효과가 위약(처리 2)보다는 우월하다는 사실을 입증하는 것이 일반적이다. 그런데 기존의 약(처리 3)보다 우월하다는 사실을 동시에 입증하려고 하는 경우도 있다. 이러한 경우 다음과 같은 두 개의 가설을 동시에 수행해야 한다.

$$H_{01}: \mu_1 = \mu_2, \quad H_{02}: \mu_1 = \mu_3$$

3개의 수준(신약, 위약, 기본의 약)을 가진 일원배치법으로 실험을 수행한 경우 첫 번째 가설 H_{01}은 $\bar{x}_{1.} - \bar{x}_{2.}$를 이용하고 두 번째 가설 H_{02}은 $\bar{x}_{1.} - \bar{x}_{3.}$을 이용하여 가설검정을 한다.

이러한 경우 각 검정에 대하여 유의 수준을 5% (개별단위 오류를 범할 확률이 5%) 라고 해도 실험단위 오류를 범할 확률은 5% 보다 크다.

여기서 유의할 점은 두 개의 가설에 대한 검정 통계량 $\bar{x}_{1.} - \bar{x}_{2.}$과 $\bar{x}_{1.} - \bar{x}_{3.}$는 독립이 아니므로(why?) 실험단위 오류를 범할 확률은 5% 보다 크고 9.75% 보다는 작다.

7.7.5. 다중비교

다시 실험 단위 오류의 계산으로 돌아가서 만약에 두 사건이 독립이 아닌 경우에 실험적 오류를 통제할 수 있는, 즉 5%보다 작거나 같게 하는 방법에 대해서 알아보자 두 사건이 독립이 아닌 일반적인 경우에 확률 공식을 이용하여 다음과 같은 부등식을 얻을 수 있다.

$$P(A_1 \cup A_2) \leq P(A_1) + P(A_2) = (2)(0.05) = 0.1$$

위의 결과를 보면 만약에 두 개의 가설검정을 동시에 수행하는 경우 각 가설검정에 대한 개별단위의 제 1 종 오류에 대한 확률을 반으로 줄이면(0.05/2=0.025) 실험적 오류가 5%보다 작거나 같게 된다.

$$P(A_1 \cup A_2) \leq P(A_1) + P(A_2) = (2)(0.05/2) = 0.05$$

위에서 보인 같은 논리로서 k 개의 가설검정을 동시에 수행하는 경우 각 가설검정에 대한 개별적 1종 오류의 확률을 k배 줄이면$(0.05/k)$ 실험단위 오류가 5%보다 작거나 같게 된다.

$$P(A_1 \cup A_2 \cup ... \cup A_k) \leq (k)(0.05/k) = 0.05$$

여기서 한 가지 유의할 점은 만약 두 개의 가설이 완전히 종속이거나$(A_1 = A_2)$ 거의 종속이면 실험적 오류는 거의 변하지 않는다. 따라서 개별단위 1종 오류에 대한 수정은 거의 필요하지 않다.

$$P(A_1 \cup A_2) = 1 - P(A_1^c \cap A_2^c) \approx 1 - P(A_1^c) = 0.05$$

이렇게 실험단위 오류를 통제하기 위하여(5%보다 작거나 같게) 각 가설에 대한 개별단위 1 종 오류의 확률(유의수준)를 보정하는 방법을 **다중비교(mutiple comparison)** 라고 한다.

위에서 제시한 개별단위 1종 오류를 k배로 줄이는(0.05/k) 방법을 특별하게 본페로니 수정(Bonferroni correction)이라고 부른다. 본페로니 수정은 가장 보수적인 수정(most conservative correction)이라고 불리는데 그 이유는 실험적 오류가 가질 수 있는 가장 큰 값을 가정하고 보정하기 때문에 각각 수정한 개별단위 오류에 대한 유의수준이 너무 작게 되어$(0.05/k)$ 귀무가설의 기각이 매우 힘들기 때문이다.

만약 k개의 가설 검정에 본페로니 수정을 적용한다면 신뢰구간과 가설검정은 다음과 같이 수정된다.

두 수준 평균의 차이 $\delta_{ij} = \mu_i - \mu_j$ 에 대한 본페로니 수정 신뢰구간은 다음과 같이 주어진다.

$$(\bar{x}_{i.} - \bar{x}_{j.}) \pm t(1 - \alpha/(2k), \phi_E)\sqrt{\frac{2MSE}{r}} \tag{7.27}$$

두 평균의 차이 δ_{ij} 에 대한 가설을 본페로니 수정 검정은 다음과 같은 조건을 만족하면 귀무가설을 기각한다.

$$\left|\bar{x}_{i.} - \bar{x}_{j.}\right| > t(1 - \alpha/(2k), \phi_E)\sqrt{\frac{2MSE}{r}} \tag{7.28}$$

기각역에 본페로니 수정을 하는 것은 원래의 p-값에 가설의 개수 k 를 곱하여 수정 p-값을 사용하는 것과 같다.

$$\text{Bonferoni adjusted p-value } = k \times \text{ unadjusted p-value} \tag{7.29}$$

일반적으로 각 가설검정들은 완전히 독립도 아니고 또한 완전한 종속도 아니다. 따라서 실험단위 오류는 각 가설검정들이 어떻게 확률적으로 관련되어 있느냐에 따라 매우 달라진다. 이러한 이유로 인하여 다중비교의 방법은 매우 다양하며, 선택한 방법에 따라서 검정의 결과도 매우 달라질 수있는 사실에 유의해야 한다. 다중비교의 방법을 선택하는 것은 매우 어려운 일이다.

> **Note**
>
> 가설이 2개 이상 있는 경우 실험단위의 오류의 확률을 제어해야 하는지에 대한 판단은 상황에 따라서 달라진다. 앞에서 살펴본 임상실험의 예와 같이 **중요한 의사 결정**을 동시에 수행하는 2개 이상의 검정 결과에 따라서

해야할 경우 주로 다중 비교를 적용한다.
또한 다중 비교 방법은 실험의 설계와 목적에 따라서 많은 방법들이 존재한다. 주어진 실험 계획과 목적에 부합하는 다중 비교법을 선택해야 한다.
반면 **탐색적인 목적**으로 여러 개의 가설 검정을 동시에 수행하는 경우에는 다중비교를 적용하지 않거나 다중 비교보다 더 유연한 False Discovery Rate 방법(참조) 을 사용한다.

8. 공분산분석

8.1. 공분산분석 개요

서로 다른 집단을 비교하는 실험이나 관측연구에서 관심이 있는 처리(treatment)나 요인(factor)뿐만 아니라 다른 예측변수들도 반응변수에 영향을 미친다. 이러한 예측변수들의 영향을 제거하기 위한 방법은 여러가지가 있지만 실험인 경우 임의화 방법(randomization)으로 그 영향을 상쇄시킬 수 도 있고 관측연구인 경우에는 사례-대조연구 방법을 이용하여 그 영향을 최소화하려고 노력을 한다.

하지만 일반적으로 관측연구(observational study)는 피실험자가 임의로 처리에 배정되지 않고, 다양한 요인을 제어하기 어렵기 때문에 전체변동에서 관심있는 요인(처리)의 효과를 제외한 오차에 의한 변동이 크다.

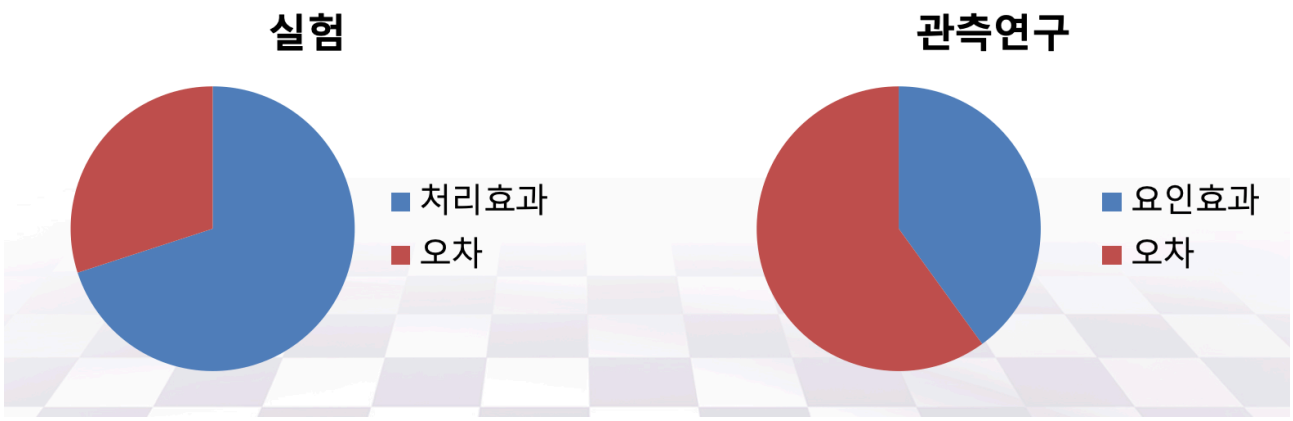


그림 8.1.: 실험과 관측연구에서의 전체 변동의 설명

많은 경우에 처리 이외의 여러 가지 변수들이 반응변수에 영향을 미친다. 이러한 경우에 이러한 독립변수(또는 예측변수)를 모형에 포함시켜서 그 영향을 반영하고 동시에 자료의 변동을 부가적으로 설명해주는 방법이 공분산 분석(analysis of covariance; ANCOVA)이다.

공분산 분석에서 고려되는 예측변수를 공변량(covarite)라고 부른다. 대부분의 실험연구에는 실험 전에 여러 가지 점수를 측정하는데 이 경우 이러한 점수를 공변량으로 모형에 포함시켜주는 것이 좋다 (예: 실험 전 상태에 대한 점수, 시험점수, IQ 점수). 또한 임상실험을 여러 개의 병원에서 진행하는 경우 병원 효과를 공변량으로 자주 사용한다.
공분산 모형의 주요한 장점은 **반응변수에 대해 설명력이 높은 공변량을 사용하게 되면 잔차제곱합이 감소하여 처리의 효과에 대한 검정력을 높일 수 있다.** 또한 공변량에 의한 편이(bias)도 줄이는 효과도 있다

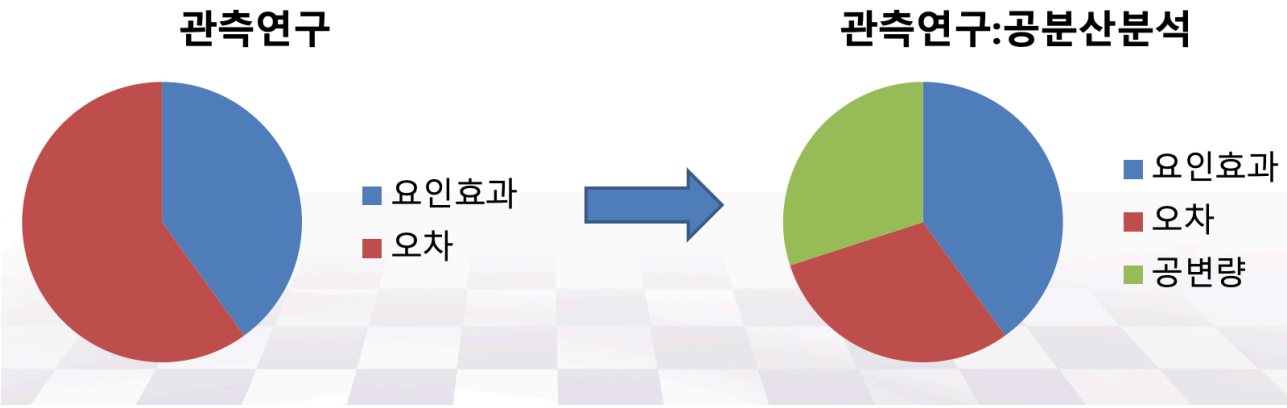


그림 8.2.: 공분산분석에 의한 전체 변동의 설명

간단한 예제로 분산분석 모형과 공분산분석 모형의 효율을 비교해 보자. 아래 왼쪽 그림은 3개의 집단을 비교하는 경우 반응변수의 분포를 상자그림으로 나타낸 것이다. 만약 반응변수가 나이와 상관관계가 높을 때 나이를 공변량으로 포함한 공분산분석 모형을 사용하면 오차에 의한 변동, 즉 잔차제곱합의 크기가 상재적으로 작아져서 처리의 효과에 대한 검정력이 높아진다.(오른쪽 그림)

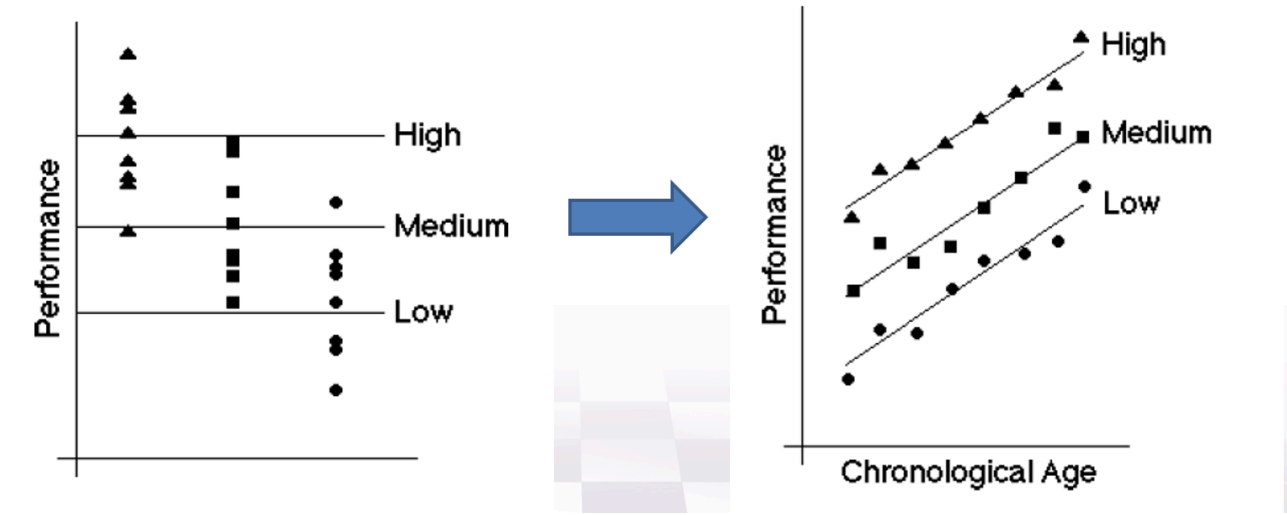


그림 8.3.: 공분산분석의 효과

공변량으로 사용할 수 있는 대표적이; 변수는 다음과 같다.

- 기저 상태(baseline)
  - 임상실험에서 환자가 실험에 참여하기 전에 환자의 상태(baseline condition, severity)가 처리의 효과 외에 반응값에 영향을 줄 수 있으므로 공변량으로 모형에 고려한다.
  - 교육 방법의 비교 실험에서 처리 전 학생의 점수는 중요한 공변량이 될 수 있다.
- 그룹 정보
  - 다기관(multicenter) 임상실험에서 각 의료기관에 대한 효과가 처리 효과 외에 반응값에 영향을 줄 수 있으므로 공변량으로 모형에 고려 (일반적으로 대부분의 분석에서 기관을 공변량으로 고려)

Caution

여기서 주의해야 할 점은 공변량과 처리는 독립이 되야한다는 점이다.
만약 처리의 결과가 공변량에 영향을 미치게 되면 이러한 공변량은 모형에 포함시키는 것이 부적절하다.
예를 들어 자동차정비 교육을 위한 두 가지 학습법을 비교하는 실험을 생각해 보자. 학생들을 임의로 두 가지 학습법중 하나를 선택하여 3개월 동안 교육을 받게 하고 시험을 보아 평균 점수(반응변수)의 차이를 알아보았다. 이 때 공변량으로 총 학습시간을 고려하였는데 학습법의 선택(처리)이 총 학습시간(공변량)에 영향을 줄 수 있다. 즉 고려된 학습법 중 하나는 컴퓨터를 이용한 학습법이며 이 학습법에 배정된 학생들은 컴퓨터 사용를 익히는 시간까지 학습시간에 포함되는 것이 나타났다. 이렇게 공변량이 처리에 의해 영향을 받는 경우(교호작용이 있는 경우)는 이를 모형에 포함시키는 것은 위험하다.
이러한 실험에서는 실험 전에 측정한 학생의 정비 능력은 공변량으로 사용하는 것이 적절할 수 있다.

8.2. 공분산분석의 모형

이제 일원배치에서 하나의 공변량이 있는 공분산분석의 모형은 일원배치 모형에 공변량 x의 효과를 다음과 같이 더해주는 것이다.

$$y_{ij} = \mu + \alpha_i + \beta(x_{ij} - \bar{x}_{..}) + e_{ij}, \quad i = 1, 2, \dots, a, \;\; j = 1, 2, \dots, r \tag{8.1}$$

모형 8.1 에서 x_{ij}는 관측값 y_{ij}의 공변량이며 $\bar{x}_{..} = \sum_{i=1}^{a} \sum_{j=1}^{r} x_{ij}$로 공변량의 전체 평균이다. 위의 효과모형은 다음과 같이 평균모형으로 나타낼 수 있다. 어떤 모형이든 모수에 대한 가설 검정의 결과는 동일하다.

$$\begin{aligned} y_{ij} &= \mu + \alpha_i + \beta(x_{ij} - \bar{x}_{..}) + e_{ij} \\ &= \beta_0 + \alpha_i + \beta x_{ij} + e_{ij} \\ &= \beta_{0i} + \beta x_{ij} + e_{ij} \end{aligned}$$

일반적으로 공분산분석 모형을 다루는 경우 8.1 모형과 같이 평균을 빼고 중심화된 공변량을 모형에서 고려한다.

8.3. 모수의 추정

모형 8.1 에서 각 모수의 추정은 ANOVA 모형에서와 같이 최소제곱법을 이용하여 추정하며 부가조건 $\sum_i \alpha_i = 0$을 이용하면 다음과 같은 추정량을 얻을 수 있다

$$\begin{aligned} \hat{\mu} &= \bar{y}_{..} \\ \hat{\alpha}_i &= \bar{y}_{i.} - \bar{y}_{..} - \hat{\beta}(\bar{x}_{i.} - \bar{x}_{..}) \\ \hat{\beta} &= \frac{\sum_i \sum_j (x_{ij} - \bar{x}_{i.})(y_{ij} - \bar{y}_{i.})}{\sum_i \sum_j (x_{ij} - \bar{x}_{i.})^2} \end{aligned}$$

8.4. 가설검정

공분산 분석 모형에서는 다음과 같은 두 가지 가설을 검정할 수 있다. 분산 분석 모형에서와 같이 각 그룹의 평균에 대한 검정을 할 수 있고

$$H_0 : \alpha_1 = \alpha_2 = ... = \alpha_a = 0 \quad \text{vesus} \quad H_1 : \text{ not } H_0$$

또한 공변량의 효과에 대한 검정도 할 수 있다.

$$H_0 : \beta = 0 \quad \text{vesus} \quad H_1 : \beta \neq 0 \tag{8.2}$$

가설검정을 위한 제곱합들을 다음과 같이 정의하자.

$$\begin{aligned}
S_{xx(i)} &= \sum_{i=1}^{a} (x_{ij} - \bar{x}_{i.})^2 \\
S_{yy(i)} &= \sum_{i=1}^{a} (y_{ij} - \bar{y}_{i.})^2 \\
S_{xy(i)} &= \sum_{i=1}^{a} (x_{ij} - \bar{x}_{i.})(y_{ij} - \bar{y}_{i.}) \\
S_{xx} &= \sum_{i=1}^{a} \sum_{j=1}^{r} (x_{ij} - \bar{x})^2 \\
S_{yy} &= \sum_{i=1}^{a} \sum_{j=1}^{r} (y_{ij} - \bar{y})^2 \\
S_{xy} &= \sum_{i=1}^{a} \sum_{j=1}^{r} (x_{ij} - \bar{x})(y_{ij} - \bar{y}) \\
SST &= \sum_{i=1}^{a} \sum_{j=1}^{r} (y_{ij} - \bar{y})^2 \\
SSA &= \frac{(S_{xx})(S_{yy}) - (S_{xy})^2}{S_{xx}} - SSE \\
SSX &= \sum_{i=1}^{a} S_{yy(i)} - SSE \\
SSE &= \frac{(\sum_i S_{xx(i)})(\sum_i S_{yy(i)}) - (\sum_i S_{xy(i)})^2}{\sum_i S_{xx(i)}}
\end{aligned}$$

이제 위의 두 검정은 다음과 같은 분산분석표를 이용한 F-검정법으로 수행할 수 있다. $N = ar$으로 총 관측값의 개수이다.

표 8.1.: 공분산분석 모형의 분산분석표

요인	제곱합	자유도	평균제곱합	F
공변량	SSX	1	$MSX = SSX/1$	$F_1 = MSX/MSE$
처리	SSA	a-1	$MSA = SSA/(a-1)$	$F_2 = MSA/MSE$
오차	SSE	N-a-1	$MSE = SSE/(N-a-1)$	
총합	SST	N-1		

위의 분산분석표에서 공변량 효과에 대한 가설 8.2 은 다음과 같이 p-값을 계산하여 검정할 수 있다.

$$p - value = P[F(1, N-a-1) > F_1]$$

또한 그룹의 평균에 대한 검정은 ANOVA 검정과 유사하게 p-값을 계산하여 검정할 수 있다.

$$p - value = P[F(a-1, N-a-1) > F_2]$$

위의 두 F-검정에 쓰이는 F-분포의 두 번째 자유도가 ANOVA 검정에서 사용되는 자유도$(N-a)$보다 하나가 작음$(N-a-1)$을 유의하자.

8.5. 예제 1: 영어교육 방법

교과서 예제 7.4에 나타난 영어교육 방법을 비교하는 실험을 공분산분석을 이용하여 분석해 보자. 실험은 3개의 교육방법(`method`)을 비교하는 실험이며 반응변수는 교육 후 시험 점수(`postscore`)이다. 실험 전에 학생들의 영어능력을 측정한 실험 전 점수(`prescore`)를 공변량으로 사용하고자 한다.

먼저 함수 `relevel` 을 이용하여 `method` 변수의 수준에서 `C` 를 기준 수준으로 지정하자.

```
english2$method <- relevel(english2$method, ref="C")
```

이제 공변량을 사용하지 않는 분산분석 모형을 적합해 보자.

```
fit_anova <- lm(postscore~method, data=english2 )
summary(fit_anova)
```

```
Call:
lm(formula = postscore ~ method, data = english2)

Residuals:
```

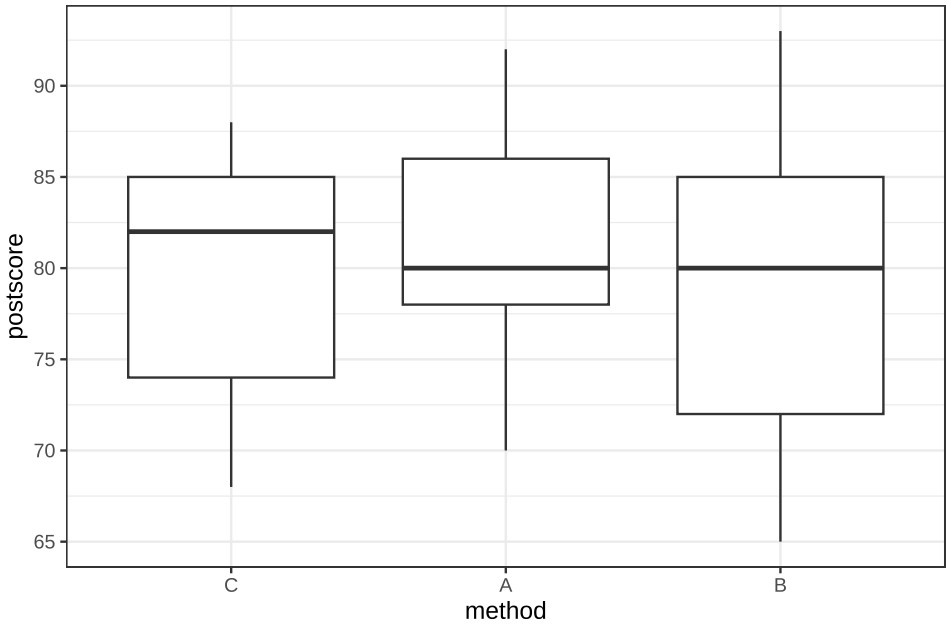
```
    Min      1Q  Median      3Q     Max
-14.000  -6.444   1.111   5.556  14.000

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  79.8889     2.7181  29.392   <2e-16 ***
methodA       1.0000     3.8439   0.260    0.797
methodB      -0.8889     3.8439  -0.231    0.819
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.154 on 24 degrees of freedom
Multiple R-squared:  0.009972,  Adjusted R-squared:  -0.07253
F-statistic: 0.1209 on 2 and 24 DF,  p-value: 0.8867
```

```
anova(fit_anova)
```

```
Analysis of Variance Table

Response: postscore
          Df  Sum Sq Mean Sq F value Pr(>F)
method     2   16.07   8.037  0.1209 0.8867
Residuals 24 1595.78  66.491
```

위에서 일원배치 분산분석 결과에서 method의 효과가 유의하지 않다고 나타났다.

```
ggplot(english2, aes(method, postscore))+geom_boxplot() + theme_bw()
```



이제 공변량을 포함한 공분산분석 모형을 적합해 보자.

```
fit_ancova <- lm(postscore~prescore + method, data=english2 )
summary(fit_ancova)
```

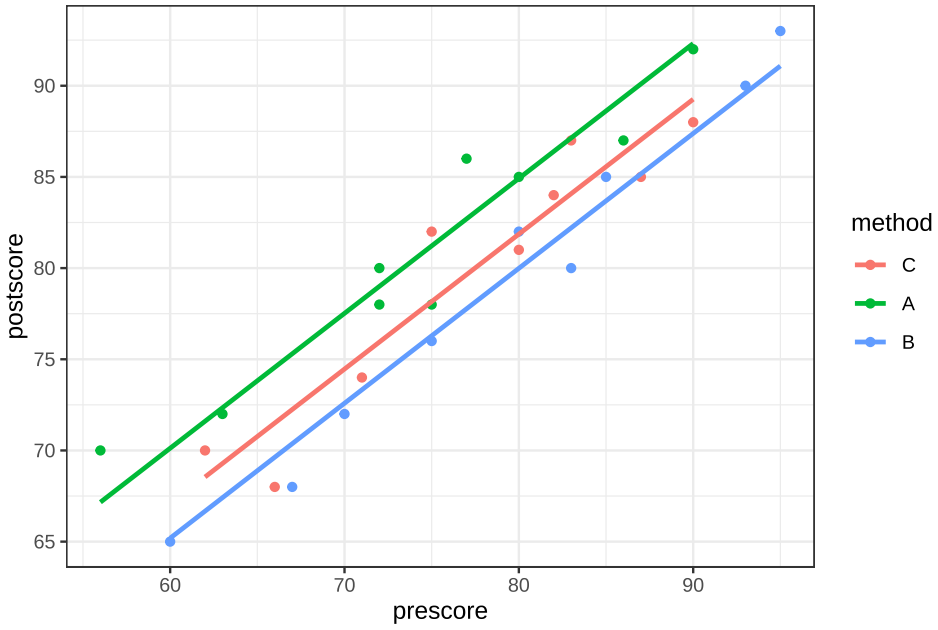
```
Call:
lm(formula = postscore ~ prescore + method, data = english2)

Residuals:
    Min      1Q  Median      3Q     Max
-3.5044 -1.2316 -0.2874  1.3847  3.8373

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 22.67685    3.22300   7.036 3.61e-07 ***
prescore     0.73981    0.04066  18.195 3.76e-15 ***
methodA      3.05503    1.00713   3.033  0.00591 **
methodB     -1.87530    1.00225  -1.871  0.07411 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.123 on 23 degrees of freedom
Multiple R-squared:  0.9357,    Adjusted R-squared:  0.9273
F-statistic: 111.5 on 3 and 23 DF,  p-value: 7.563e-14
```

```
anova(fit_ancova)
```

```
Analysis of Variance Table

Response: postscore
          Df  Sum Sq Mean Sq F value    Pr(>F)
prescore   1 1399.85 1399.85 310.593 7.472e-15 ***
method     2  108.34   54.17  12.019 0.0002671 ***
Residuals 23  103.66    4.51
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

실험 전 영어 점수를 공변량으로 포함하면 교육 후 점수에 유의한 영향을 미치는 것으로 나타났다. 또한 교육 방법의 효과도 유의하게 나타났다.

```
english2$pred = predict(fit_ancova)

ggplot(english2, aes(x=prescore, y=postscore, colour = method)) +
  geom_point() +
  geom_line(aes(y = pred), size = 1) +
  theme_bw()
```



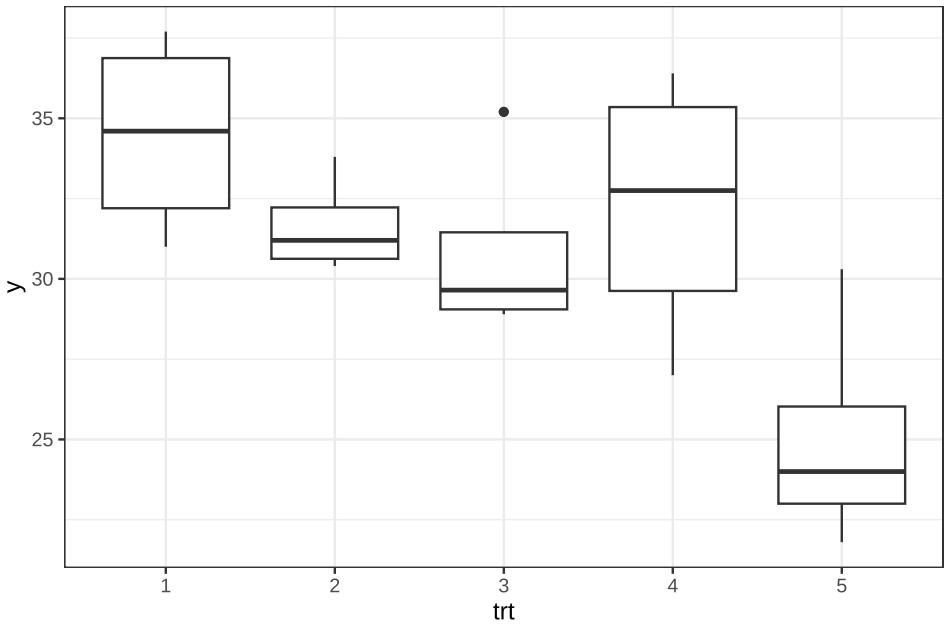
8.6. 예제 2: 저혈당 실험

```
dd <- read.table ("../data/chapter-5-data.txt",sep="",header = FALSE)
colnames(dd) <- c("trt","x","y")
dd$trt <- factor(dd$trt)
head(dd)
```

```
  trt    x    y
1   1 27.2 32.6
2   1 33.0 37.7
3   2 28.6 33.8
4   2 26.5 30.7
5   3 28.6 35.2
6   3 23.2 28.9
```

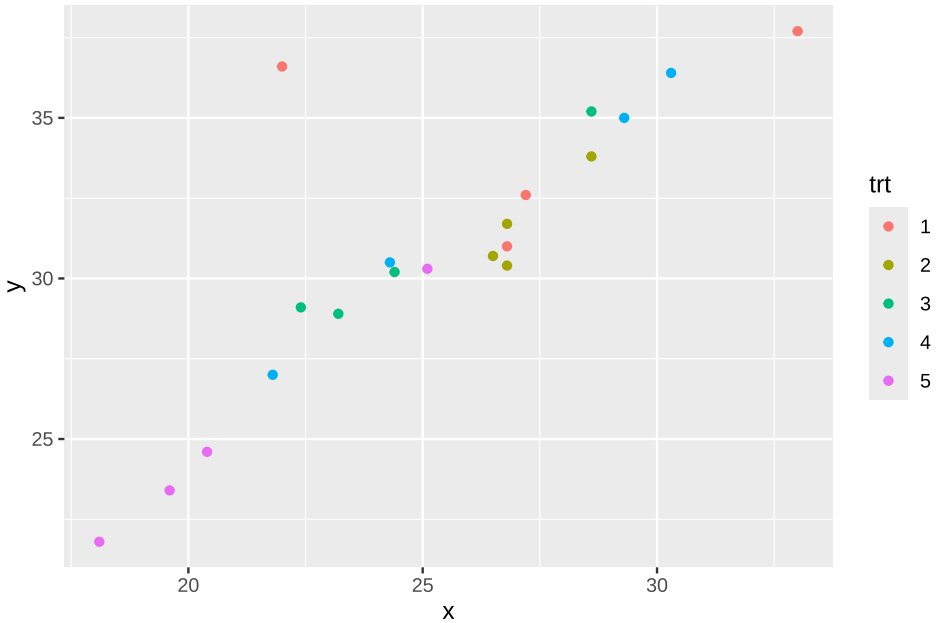
혈당을 감소시키기 위한 다섯 개의 처리(`trt`)를 비교하려고 한다. 반응 변수 y는 치료 적용 1달 후 혈당량 수치이며 각 처리그룹에 대한 자료와 상자그림은 다음과 같다.

```
ggplot(dd, aes(trt, y))+geom_boxplot() + theme_bw()
```



공변량 x 는 치료 전 측정한 혈당량 수치이며 치료 후의 수치 y의 관계는 다음과 같은 산점도로 나타낼 수 있으며 강한 양의 선형관계가 있다는 것을 알 수 있다.즉 초기 형당량 수치가 크면 1달 후 혈당량 수치도 평균적으로 크다.

```
ggplot(dd, aes(x, y))+geom_point(aes(colour = trt))
```



이제 공변량을 사용하지 않는 분산분석 모형을 적합해 보자.

```
diab1 <- lm(y~trt, data=dd )
summary(diab1)
```

```
Call:
lm(formula = y ~ trt, data = dd)

Residuals:
   Min     1Q Median     3Q    Max
-5.225 -1.781 -0.800  2.306  5.275

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   34.475      1.634  21.094 1.46e-12 ***
trt2          -2.825      2.311  -1.222 0.240469
trt3          -3.625      2.311  -1.568 0.137642
trt4          -2.250      2.311  -0.973 0.345753
trt5          -9.450      2.311  -4.089 0.000968 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.269 on 15 degrees of freedom
Multiple R-squared:  0.5532,    Adjusted R-squared:  0.434
F-statistic: 4.643 on 4 and 15 DF,  p-value: 0.01224
```

```r
anova(diab1)
```

```
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value  Pr(>F)
trt        4 198.41  49.602  4.6425 0.01224 *
Residuals 15 160.26  10.684
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

위의 결과를 보면 5개의 처리가 유의하게 다르다고 나타났다.

다음으로 치료 전의 혈당 x 를 공변량으로 포함한 공분석 결과를 적합해 보자.

```r
diab2 <- lm(y~x + trt, data=dd )
summary(diab2)
```

```
Call:
```

```
lm(formula = y ~ x + trt, data = dd)

Residuals:
    Min      1Q  Median      3Q     Max
-3.1360 -1.0024 -0.2827  0.7257  6.0806

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  13.9437     4.8219   2.892 0.011834 *
x             0.7534     0.1723   4.373 0.000637 ***
trt2         -2.7685     1.5554  -1.780 0.096793 .
trt3         -1.6660     1.6186  -1.029 0.320776
trt4         -1.6284     1.5618  -1.043 0.314787
trt5         -4.5903     1.9115  -2.401 0.030788 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.2 on 14 degrees of freedom
Multiple R-squared:  0.8112,    Adjusted R-squared:  0.7437
F-statistic: 12.03 on 5 and 14 DF,  p-value: 0.0001164
```

```
anova(diab2)
```

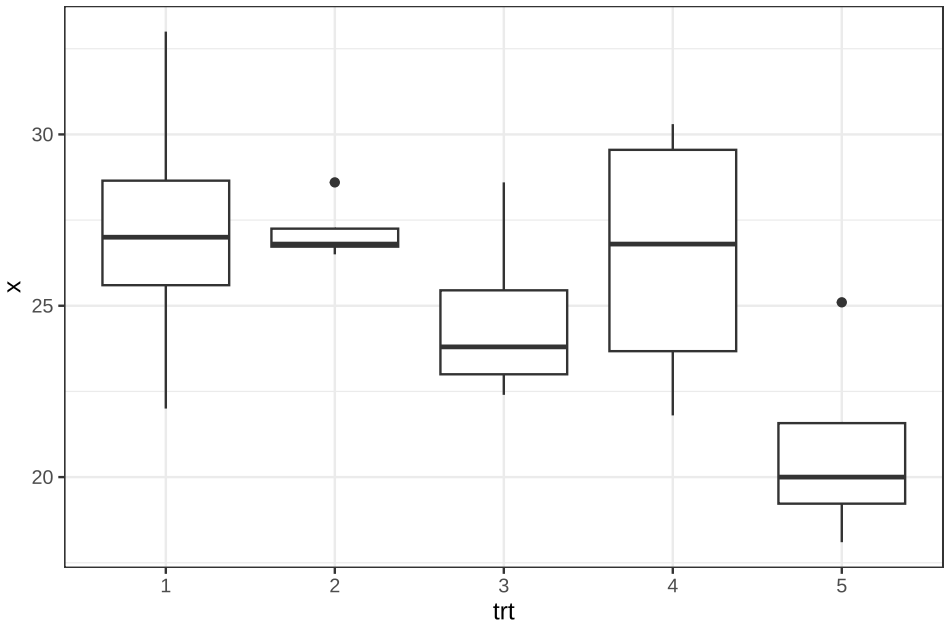
```
Analysis of Variance Table

Response: y
          Df  Sum Sq Mean Sq F value    Pr(>F)
x          1 256.747 256.747 53.0673 4.003e-06 ***
trt        4  34.188   8.547  1.7666    0.1917
Residuals 14  67.734   4.838
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

분산분석 모형에서는 평균 잔차제곱합이 $MSE = 10.684$ 이지만 공분산분석에서는 $MSE = 4.838$ 로 감소하였다.

이는 공변량인 치료 전의 혈당이 치료 후의 혈당을 설명하는데 중요한 요인이므로 모형의 설명력이 높아진 결과이다.

하지만 처리의 효과를 보면 공분산분석에서는 유의한 차이가 나타나지 않았다. 이는 공변량인 치료 전의 혈당이 치료 후의 혈당을 설명하는데 매우 유의한 요인이기 때문이다. 즉 치료 후의 혈당의 집단별 변동이 큰 이유는 치료 효과때문이 아니라 치료 전의 혈당과 처리의 관계 때문으로 보여진다 (아래 그림).

```
ggplot(dd, aes(trt, x))+geom_boxplot() + theme_bw()
```



이러한 결과의 교훈은 여러 개의 집단을 비교하는 경우, 처리를 배정할 때 임의화 방법(randomization)을 사용하여 실험 전 환자들의 인구적 특성과 증상의 정도가 처리그룹간에 큰 차이가 없이 만들어야 한다. 우리가 살펴본 저혈당 임상실험은 처리를 비교하기 위한 공정한 실험이라고 보기 힘들다.

9. 최소제곱 추정량의 계산

이제 선형모형에서 회귀계수를 구하는 계산 방법에 대하여 알아보자. 최소제곱법(동시에 최대 가능도 추정법)에 의한 회귀계수의 추정치를 구하려면 다음과 같은 최적화 문제를 풀어야한다.

$$\min_{\boldsymbol{\beta}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 \tag{9.1}$$

위의 최적화 문제로 부터 유도된 회귀계수에 대한 정규 방정식(normal equation)은 다음과 같다.

$$\boldsymbol{X}^t\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}^t\boldsymbol{y} \tag{9.2}$$

9.2 에 대한 해를 구하는 경우 대수적으로는 $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{y}$로 표시하지만 실제 계산에서 역행렬 $(\boldsymbol{X}^t\boldsymbol{X})^{-1}$을 실제로 구하지는 않는다. 9.2 에서 나타난 것과 같은 선형방정식을 푸는 계산적 방법은 가우스 소거법(Gauss elimination), 교환 연산(sweep operator) 등에 의한 전통적인 방법들이 있다. 교환 연산(sweep operator) 방법은 SAS 프로그램에서 최소제곱법을 푸는 방법으로 사용된다. 이러한 전통적인 방법들은 가우스 소거법에 근거하여 핼과 열 연산에 의한 소거법을 기반으로 한다. 하지만 최근에 나온 통계 계산 패키지에서는 소거법을 사용하지 않는다.

이 장에서는 최근 통계 계산 프로그램이 사용하는 행렬 분해에 의한 방법을 살펴보고자 한다. 다음에 나오는 방법들은 계획행렬 $\boldsymbol{X}$가 최대 계수(full rank)가 아닌 경우에도 적용할 수 있지만 여기서는 최대 계수인 경우만 고려하겠다.

아래 방법들을 공부하기 전에 Appendix D 에 제시된 행렬의 분해 방법을 먼저 공부하자.

9.1. 촐레스키 분해의 이용

방정식 9.2 에서 완쪽 항에 있는 $\boldsymbol{A} = \boldsymbol{X}^t\boldsymbol{X}$ 가 양정치 행렬이라고 하자. 양정치 행렬 A은 다음과 같이 유일하게 촐레스키 분해(Cholesky decomposition)가 가능하다.

$$\boldsymbol{A} = \boldsymbol{X}^t\boldsymbol{X} = \boldsymbol{L}\boldsymbol{L}^t \tag{9.3}$$

위의 분해에서 행렬 $\boldsymbol{L}$은 양수의 대각원소를 가지는 하삼각 행렬(lower triangular matrix)이다.

이제 촐레스키 분해를 정규 방정식에 적용해보자.

$$\boldsymbol{L}\boldsymbol{L}^t\boldsymbol{\beta} = \boldsymbol{X}^t\boldsymbol{y}$$

위의 식에서 $\boldsymbol{L}^t\boldsymbol{\beta} = \boldsymbol{\beta}_*$라고 하면 다음과 같은 방정식을 얻게 되고 이 방정식은 행렬 $\boldsymbol{L}$이 하삼각 행렬이기 때문에 대수적으로 축차식을 이용하여 쉽게 해 $\hat{\boldsymbol{\beta}}_*$를 얻을 수 있다.

$$\boldsymbol{L}\boldsymbol{\beta}_* = \boldsymbol{X}^t\boldsymbol{y}$$

다음으로 관계식 $\boldsymbol{L}^t\boldsymbol{\beta} = \boldsymbol{\beta}_*$ 이용하여 다음의 방정식을 풀면 회귀계수의 추정치 $\hat{\boldsymbol{\beta}}$를 얻게 된다. 아래의 방정식 또한 $\boldsymbol{L}^t$가 상삼각 행렬이기 때문에 축차적인 계산이 가능하다.

$$\boldsymbol{L}^t\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}_*$$

9.2. QR 분해의 이용

차원이 $n \times p$인 계획 행렬 $\boldsymbol{X}$의 QR 분해가 다음과 같이 주어졌다고 하자. $rank(\boldsymbol{X}) = p < n$ 이라고 가정하자.

$$\boldsymbol{X} = \boldsymbol{Q}_1\boldsymbol{R}_1$$

위의 QR 분해에서 $\boldsymbol{Q}_1$는 p 개의 직교하는 열을 가진 $n \times p$ 행렬이다 ($\boldsymbol{Q}_1^t\boldsymbol{Q}_1 = \boldsymbol{I}$). 또한 행렬 $\boldsymbol{R}_1$은 차원이 $p \times p$인 상삼각 행렬(upper diagonal matrix)이다.

그러면 행렬 $\boldsymbol{X}$는 다음과 같은 확장된 분해를 가진다.

$$\boldsymbol{X} = \boldsymbol{Q}\boldsymbol{R} = \begin{bmatrix}\boldsymbol{Q}_1 & \boldsymbol{Q}_2\end{bmatrix}\begin{bmatrix}\boldsymbol{R}_1 \\ \boldsymbol{0}\end{bmatrix}$$

위에서 $\boldsymbol{Q}_2$ 는 행렬 $\boldsymbol{Q}_1$의 열들과 직교하는 $n-p$ 개의 추가 정규직교 벡터들로 이루어진 행렬이다. $\boldsymbol{Q}_1$과 $\boldsymbol{Q}_2$는 각각 $n \times p$, $n \times (n-p)$의 행렬이다. 따라서 행렬 $\boldsymbol{Q} = [\boldsymbol{Q}_1\ \boldsymbol{Q}_2]$는 $n \times n$ 직교행렬이다 ($\boldsymbol{Q}^t\boldsymbol{Q} = \boldsymbol{Q}\boldsymbol{Q}^t = \boldsymbol{I}$).

또한 $\boldsymbol{R}$는 $n \times p$ 이며 $\boldsymbol{0}$은 차원이 $(n-p) \times p$인 영행렬이다.

$$\boldsymbol{R} = \begin{bmatrix}\boldsymbol{R}_1 \\ \boldsymbol{0}\end{bmatrix}$$

이제 $\boldsymbol{Q}^t\boldsymbol{Q} = \boldsymbol{I}$를 이용하여 잔차제곱합 $\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2$를 다음과 같이 분해할 수 있다.

$$\begin{aligned}
\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|^2 &= (\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^t(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}) \\
&= (\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^t\boldsymbol{Q}\boldsymbol{Q}^t(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}) \\
&= (\boldsymbol{Q}^t\boldsymbol{y}-\boldsymbol{Q}^t\boldsymbol{X}\boldsymbol{\beta})^t(\boldsymbol{Q}^t\boldsymbol{y}-\boldsymbol{Q}^t\boldsymbol{X}\boldsymbol{\beta}) \\
&= (\boldsymbol{c}-\boldsymbol{R}\boldsymbol{\beta})^t(\boldsymbol{c}-\boldsymbol{R}\boldsymbol{\beta}) \\
&= \left(\begin{bmatrix}\boldsymbol{c}_1\\ \boldsymbol{c}_2\end{bmatrix}-\begin{bmatrix}\boldsymbol{R}_1\\ \boldsymbol{0}\end{bmatrix}\boldsymbol{\beta}\right)^t\left(\begin{bmatrix}\boldsymbol{c}_1\\ \boldsymbol{c}_2\end{bmatrix}-\begin{bmatrix}\boldsymbol{R}_1\\ \boldsymbol{0}\end{bmatrix}\boldsymbol{\beta}\right) \\
&= (\boldsymbol{c}_1-\boldsymbol{R}_1\boldsymbol{\beta})^t(\boldsymbol{c}_1-\boldsymbol{R}_1\boldsymbol{\beta})+\boldsymbol{c}_2^t\boldsymbol{c}_2 \\
&= \|\boldsymbol{c}_1-\boldsymbol{R}_1\boldsymbol{\beta}\|^2+\|\boldsymbol{c}_2\|^2
\end{aligned}$$

결과를 요약하면 다음과 같은 분해를 얻는다.

$$\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|^2 = \|\boldsymbol{c}_1-\boldsymbol{R}_1\boldsymbol{\beta}\|^2+\|\boldsymbol{c}_2\|^2 \tag{9.4}$$

여기서

$$\boldsymbol{c}=\boldsymbol{Q}^t\boldsymbol{y}=\begin{bmatrix}\boldsymbol{Q}_1^t\\ \boldsymbol{Q}_2^t\end{bmatrix}\boldsymbol{y}=\begin{bmatrix}\boldsymbol{Q}_1^t\boldsymbol{y}\\ \boldsymbol{Q}_2^t\boldsymbol{y}\end{bmatrix}=\begin{bmatrix}\boldsymbol{c}_1\\ \boldsymbol{c}_2\end{bmatrix}$$

위의 9.4 를 보면 벡터 $\boldsymbol{c}_2=\boldsymbol{Q}_2^t\boldsymbol{y}$는 $\boldsymbol{\beta}$와 관계가 없으므로 잔차제곱합 $\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|^2$ 을 최소화하는 $\boldsymbol{\beta}$는 $\|\boldsymbol{c}_1-\boldsymbol{R}_1\boldsymbol{\beta}\|^2$ 을 0으로 만드는 것이다. 즉 $\boldsymbol{R}_1\boldsymbol{\beta}=\boldsymbol{c}_1$를 만족하는 $\boldsymbol{\beta}$가 최소제곱 추정량이다.

$$pmbin_{\boldsymbol{\beta}}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|^2 = pmbin_{\boldsymbol{\beta}}\|\boldsymbol{c}_1-\boldsymbol{R}_1\boldsymbol{\beta}\|^2+\|\boldsymbol{c}_2\|^2$$

$\boldsymbol{X}$가 완전계수 행렬이므로 상삼각행렬인 $\boldsymbol{R}_1$도 완전계수 행렬이다. 따라서 방정식 $\boldsymbol{R}_1\boldsymbol{\beta}=\boldsymbol{c}_1$는 유일한 해는 상삼각행렬의 성질을 이용하여 축차식으로 쉽게 구할 수 있다.

$$\hat{\boldsymbol{\beta}}=\boldsymbol{R}_1^{-1}\boldsymbol{c}_1$$

추정량은 $\hat{\boldsymbol{\beta}}$은 실제 역행렬을 구하지 않고 구할 수 있다.

참고로 잔차제곱합 SSE는 다음과 같이 계산된다.

$$SSE=\left\|\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\right\|^2=\|\boldsymbol{c}_2\|^2=\boldsymbol{y}^t\boldsymbol{Q}_2\boldsymbol{Q}_2^t\boldsymbol{y}$$

9.3. SVD 을 이용

이제 최소제곱법을 SVD (Singular Value Decomposition) 으로 푸는 방법을 살펴보자. 차원이 $n \times p$인 계획행렬 $\boldsymbol{X}$ 이 주어지고 $rank(\boldsymbol{X}) = p < n$ 이라고 가정하자.

이제 계획행렬 $\boldsymbol{X}$ 은 SVD 를 이용하여 다음과 같이 분해할 수 있다.

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{R}\boldsymbol{V}^t$$

위의 분해에서 각 행렬의 특성은 다음과 같다.

- $\boldsymbol{U}$: $n \times n$ 직교행렬
- $\boldsymbol{V}$: $p \times p$ 직교행렬

$\boldsymbol{R}$은 $n \times p$ 행렬이며 윗 부분 $p \times p$ 행렬 R_1은 대각 행렬이다. $\boldsymbol{R}$의 아래 부분 $(n-p) \times p$는 영행렬이다.

$$\boldsymbol{R} = \begin{bmatrix} \boldsymbol{R}_1 \\ \boldsymbol{0} \end{bmatrix} \quad \boldsymbol{R}_1 = diag(r_1, r_2, \dots, r_p)$$

따라서 계획행렬의 분해는 다음과 같이 축소할 수 있다.

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{R}\boldsymbol{V}^t = \begin{bmatrix} U_1 & U_2 \end{bmatrix} \begin{bmatrix} \boldsymbol{R}_1 \\ \boldsymbol{0} \end{bmatrix} \boldsymbol{V}^t = \boldsymbol{U}_1 \boldsymbol{R}_1 \boldsymbol{V}^t \tag{9.5}$$

위의 9.5 에서 행렬 $\boldsymbol{U}_1$은 $n \times p$, $\boldsymbol{U}_2$은 $n \times (n-p)$ 행렬이며 $\boldsymbol{U}_1^t \boldsymbol{U}_1 = \boldsymbol{I}, \boldsymbol{U}_2^t \boldsymbol{U}_2 = \boldsymbol{I}$ 이다.

이제 최소제곱법의 해를 SVD 로 구해보자.

$$\begin{aligned}
\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 &= (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^t (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \\
&= (\boldsymbol{y} - \boldsymbol{U}\boldsymbol{R}\boldsymbol{V}^t\boldsymbol{\beta})^t (\boldsymbol{y} - \boldsymbol{U}\boldsymbol{R}\boldsymbol{V}^t\boldsymbol{\beta}) \\
&= (\boldsymbol{y} - \boldsymbol{U}\boldsymbol{R}\boldsymbol{V}^t\boldsymbol{\beta})^t \boldsymbol{U}\boldsymbol{U}^t (\boldsymbol{y} - \boldsymbol{U}\boldsymbol{R}\boldsymbol{V}^t\boldsymbol{\beta}) \\
&= (\boldsymbol{U}^t\boldsymbol{y} - \boldsymbol{U}^t\boldsymbol{U}\boldsymbol{R}\boldsymbol{V}^t\boldsymbol{\beta})^t (\boldsymbol{U}^t\boldsymbol{y} - \boldsymbol{U}^t\boldsymbol{U}\boldsymbol{R}\boldsymbol{V}^t\boldsymbol{\beta}) \\
&= (\boldsymbol{U}^t\boldsymbol{y} - \boldsymbol{R}\boldsymbol{V}^t\boldsymbol{\beta})^t (\boldsymbol{U}^t\boldsymbol{y} - \boldsymbol{R}\boldsymbol{V}^t\boldsymbol{\beta}) \\
&= \|\boldsymbol{U}^t\boldsymbol{y} - \boldsymbol{R}\boldsymbol{V}^t\boldsymbol{\beta}\|^2 \\
&= \|\boldsymbol{c} - \boldsymbol{R}\boldsymbol{\beta}_*\|^2 \\
&= \left(\begin{bmatrix} \boldsymbol{c}_1 \\ \boldsymbol{c}_2 \end{bmatrix} - \begin{bmatrix} \boldsymbol{R}_1 \\ \boldsymbol{0} \end{bmatrix} \boldsymbol{\beta}_* \right)^t \left(\begin{bmatrix} \boldsymbol{c}_1 \\ \boldsymbol{c}_2 \end{bmatrix} - \begin{bmatrix} \boldsymbol{R}_1 \\ \boldsymbol{0} \end{bmatrix} \boldsymbol{\beta}_* \right) \\
&= (\boldsymbol{c}_1 - \boldsymbol{R}_1\boldsymbol{\beta}_*)^t (\boldsymbol{c}_1 - \boldsymbol{R}_1\boldsymbol{\beta}_*) + \boldsymbol{c}_2^t \boldsymbol{c}_2 \\
&= \|\boldsymbol{c}_1 - \boldsymbol{R}_1\boldsymbol{\beta}_*\|^2 + \|\boldsymbol{c}_2\|^2
\end{aligned}$$

이제 위의 분해에서 다음과 같이 새로운 벡터를 정의한다.

$$\boldsymbol{c} = \boldsymbol{U}^t\boldsymbol{y}, \quad \boldsymbol{\beta}_* = \boldsymbol{V}^t\boldsymbol{\beta} \tag{9.6}$$

그리고

$$\boldsymbol{c} = \boldsymbol{U}^t\boldsymbol{y} = \begin{bmatrix} \boldsymbol{U}_1^t \\ \boldsymbol{U}_2^t \end{bmatrix} \boldsymbol{y} = \begin{bmatrix} \boldsymbol{U}_1^t\boldsymbol{y} \\ \boldsymbol{U}_2^t\boldsymbol{y} \end{bmatrix} = \begin{bmatrix} \boldsymbol{c}_1 \\ \boldsymbol{c}_2 \end{bmatrix}$$

이제 QR 분해와 유사하게 최소제곱의 오차제곱합은 다음과 같이 분해된다.

$$\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 = \|\boldsymbol{c}_1 - \boldsymbol{R}_1\boldsymbol{\beta}_*\|^2 + \|\boldsymbol{c}_2\|^2 = \sum_{i=1}^{p}(c_{1i} - r_i\beta_{*i})^2 + \|\boldsymbol{c}_2\|^2 \tag{9.7}$$

위의 분해 9.7 는 행렬 $\boldsymbol{R}_1$ 이 대각행렬임을 이용한 것이다. 이제 9.7 를 최소화하는 벡터 $\hat{\boldsymbol{\beta}}_*$ 의 원소들은 다음과 같이 구할 수 있고

$$\hat{\beta}_{*i} = \frac{c_{1i}}{r_i}, \quad i = 1, 2, \dots, p$$

최종적으로 9.6 의 관계를 이용하면 최소제곱 추정량은 다음과 같이 주어진다.

$$\hat{\boldsymbol{\beta}} = \boldsymbol{V}\hat{\boldsymbol{\beta}}_*$$

참고로 QR 분해 방법과 유사하게 잔차제곱합 SSE는 다음과 같이 계산된다.

$$SSE = \left\|\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right\|^2 = \|\boldsymbol{c}_2\|^2 = \boldsymbol{y}^t\boldsymbol{U}_2\boldsymbol{U}_2^t\boldsymbol{y}$$

> **Note**
>
> 위에서 논의한 촐레스키, QR, SVD 를 이용한 최소제곱 추정량 $\hat{\beta}$를 구하는 방법은 계획행렬이 완전 계수가 아닌 경우에도 $(rank(X) < p)$ 쉽게 적용할 수 있다.

10. 회귀모형의 확장

10.1. 개요

이번 장에는 특별한 목적을 가지는 다양한 확장 회귀모형에 대하여 논의한다. 다음과 같은 모형들을 논의할 것이다.

- 로버스트 회귀분석 (교과서 8.3절)
- 비선형 회귀 (교과서 9.1 절)

10.2. 자료 만들기

개인정보 보호 문제로 실제 데이터를 사용하는데 어려움이 있어서 다음과 같이 유사한 데이터를 인공작으로 만들어서 사용하고자 한다.

```
set.seed(4531)
# 비선형 회귀분석을 위한 데이터
logifun <- function(t,beta, sigma) {

  n <- length(t)
  y <- beta[1]/(1 + exp(-10 -t)/beta[2]) + rnorm(n,0,sigma)
  y[y>beta[1]] <- beta[1]
  y[y<0] <- 0
  y <- round(y,0)
  data.frame(days=t, reserved = y)
}

true_beta <- c(50,2)
true_sigma <- 6
t <- seq(-20,0)
df <- logifun(t, true_beta, true_sigma)

# 로버스트 회귀분석을 위한 데이터
mean1 <- c(20,45)
sigma1 <- matrix(c(10,7,7,10), 2,2)
df20 <- round(rmvnorm(20, mean1,sigma1),0)
```

```
df20[df20[,2] > 50, 2] <- 50
df2 <- as.data.frame(df20)
colnames(df2) <- c("reserved","boarding")
df3 <- rbind(df2, c(49,44))
```

10.3. 로버스트 회귀

회귀분석을 수행하는 경우 가장 어려운 상황은 이상점 또는 영향력이 큰 관측값이 자료에 포함되어 설명변수와 반응변수의 일반적인 관계를 왜곡시키는 경우이다.

물론 앞 장에서 논의한 다양한 통계적 측도들(쿡의 거리 등)을 이용하여 이상점과 영향점(지렛점)을 참색한 후에 적저한 방법으로 자료를 정리한 후에 회귀식을 적합할 수 있다. 이렇게 회귀모형을 구축하는 경우 이상점을 탐색하고 제거하는 절차를 거쳐야 하지만 자료의 수가 매우 많거나 많은 수의 회귀모형을 동시에 고려해야 하는 경우 번거롭고 어려운 작업을 거쳐야 한다.

이러한 이상점을 판별하여 제거하는 것보다 회귀식을 적합할 때 이상점의 영향을 덜 받는 추정 방법을 적용하면 번거로운 탐색과 제거 작업을 하지 않아도 된다.

일반적으로 추정량 $\hat{\theta}$ 가 모형의 가정에 민감하지 않거나 이상점의 영향을 덜 받는 경우 로버스트(robust) 하다고 한다. 예를 들어 분포의 중심을 추정할 때 일반적으로 중앙값(median) 이 평균(average) 보다 로버스트 하다고 말할 수 있다.

회귀분석는 최소제곱 추정량(또는 정규분포 가정하의 최대가능도 추정량)은 중앙값보다 평균에 가까운 추정량으로 이상점에 로버스트한 추정량이 아니다. 로버스트 회귀(robust regression)은 이상점의 영향을 덜 받는 방법으로 회귀계수를 추정하는 회귀분석을 말한다.

통계학에서 나타나는 대부분의 추정 방법은 모수 θ 와 자료 $\boldsymbol{y} = (y_1, y_2, \dots, y_n$ 으로 구성된 목적함수(objective function 또는 loss function) $L(\theta, \boldsymbol{y})$을 최대화하거나 또는 최소화하는 추정량을 구하는 방법이다.

$$\hat{\theta} = \arg\min_{\theta} L(\theta, \boldsymbol{y}) \tag{10.1}$$

예를 들어 평균은 다음과 같은 2차식의 목적함수를 이용하여 구하는 추정량이다.

$$\bar{y} = \arg\min_{\mu} L(\mu, \boldsymbol{y}) \quad \text{where} \quad L(\mu, \boldsymbol{y}) = \sum_i (y_i - \mu)^2$$

중앙값은 목적함수를 절대값을 이용한 추정량이다.

$$med(y_i) = \arg\min_{\theta} L(\theta, \boldsymbol{y}) \quad \text{where} \quad L(\theta, \boldsymbol{y}) = \sum_i |y_i - \theta|$$

회귀분석에서 최소제곱 추정량(또는 정규분포 가정하의 최대가능도 추정량)은 목적함수로 제곱함수를 이용한다.

$$\hat{\beta}_{LS} = \arg\min_{\beta} L(\beta, \boldsymbol{y}; \boldsymbol{X}) \quad \text{where} \quad L(\beta, \boldsymbol{y}; \boldsymbol{X}) = \sum_i (y_i - \boldsymbol{x}_i^t \boldsymbol{\beta})^2$$

10.3.1. M-추정량

M-추정량은 Huber(1973) 가 제안한 추정량으로 이상점의 영향에 덜 민감한 목적함수를 사용하여 추정하는 방법이다.

$$\hat{\beta}_{M} = \arg\min_{\beta} L(\beta, \boldsymbol{y}, \boldsymbol{X}) \quad \text{where} \quad L(\beta, \boldsymbol{y}, \boldsymbol{X}) = \sum_i \rho\left(\frac{y_i - \boldsymbol{x}_i^t \boldsymbol{\beta}}{\sigma}\right) \tag{10.2}$$

위의 10.2 에서 목적함수 L 를 구성하는 함수 $\rho(u)$ 를 다음과 성질은 만족해야 한다.

- $\rho(u) \geq 0$ (언제나 0 또는 양수)
- $\rho(0) = 0$
- $\rho(u) = \rho(-u)$ (대칭)
- $\rho(u_1) \geq \rho(u_2)$, $|u_1| > |u_2|$ (단조성)

회귀분석에서 M-추정량을 이용한 추정은 $\rho(u)$가 제곱함수가 아닌 다른 함수를 이용하여 이상점을 영향을 작게한다. 이상점의 영향을 줄이는 방법은 $\rho(u)$ 가 원점으로 부터 멀어지는 경우 증가속도를 제곱함수보다 완만하게 해주는 것이다. M-추정에 사용될 수 있는 여러 가지 함수는 강근석 and 유형조 (2016) 의 표 8.1 에 있다. 예를 들어 가장 대표적인 huber 함수는 다음과 같다.

$$\rho(u) = \begin{cases} \frac{1}{2}u^2 & |u| \leq c \\ c(|u| - c/2) & |u| > c \end{cases}$$

M-추정량을 구하기 위해서는 10.2 의 목적함수 L 을 회귀계수 β_j 에 대하여 미분한 값을 0으로 놓는다. 이렇게 p 개의 회귀계수로 미분한 p 개의 방정식을 풀어서 M-추정량을 구한다. 함수 $\rho(u)$ 의 미분한 함수를 $\rho'(u) = \psi(u)$ 라고 하면 다음과 같은 방정식을 푸는 것이다.

$$\sum_{i=1}^{n} x_{ij} \psi\left(\frac{y_i - \boldsymbol{x}_i^t \boldsymbol{\beta}}{\sigma}\right) = 0, \quad j = 0, 1, 2, \ldots p-1 \tag{10.3}$$

위에서 구한 M-추정량의 방정식은 다음과 같이 가중치를 가진 최소제곱법의 방법식과 같이 변형할 수 있다.

$$\sum_{i=1}^{n} w_i x_{ij} (y_i - \boldsymbol{x}_i^t \boldsymbol{\beta}) = 0, \quad j = 0, 1, 2, \ldots p-1 \tag{10.4}$$

여기서 가중치 w_i 는 다음과 같이 정의된다. 주의할 점은 가중치 w_i 는 회귀 계수 $\boldsymbol{\beta}$ 의 함수 $w_i = w_i(\boldsymbol{\beta})$ 라는 것이다. 또한 잔차 $e_i = y_i - \boldsymbol{x}_i^t \boldsymbol{\beta}$ 이다.

$$w_i = \frac{\psi[(y_i - \boldsymbol{x}_i^t\boldsymbol{\beta})/\sigma]}{(y_i - \boldsymbol{x}_i^t\boldsymbol{\beta})/\sigma} = \frac{\psi(e_i/\sigma)}{e_i/\sigma} \tag{10.5}$$

위의 방정식 10.4 을 다음과 같이 행렬식으로 표기할 수 있으며 가중치 행렬 $\boldsymbol{W}$ 는 대각행렬로 대각원소는 10.5 와 같이 주어진다.

$$\boldsymbol{X}^t\boldsymbol{W}\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}^t\boldsymbol{W}\boldsymbol{y} \tag{10.6}$$

10.6 에 주어진 방정식은 먼저 회귀계수의 초기값 $\boldsymbol{\beta}_0$ 를 이용하여 가중치 행렬을 계산하고 회귀계수의 추정치 $\hat{\beta}_1$ 을 구한다. 다시 추정치 $\hat{\beta}_1$ 을 이용하여 가중치 행렬을 계산하고 회귀계수의 추정치 $\hat{\beta}_2$ 를 구한다. 이렇게 축차적으로 방정식을 풀면 궁극적으로 최종 주정량 $\hat{\beta}_M$ 에 수렴하게 된다. 이러한 축차적인 추정법을 반복 가중최소제곱법 (iteratively reweighted least square; IRLS, IWLS) 이라고 부른다.

또한 방정식의 가중치 10.5 를 계산하기 위해서는 오차항의 표준편차 σ 의 추정이 필요하다. M-추정량에서는 σ 의 로버스트 추정량인 중위절대편차(median absolute deviation; MAD)의 표준화 값을 사용한다.

$$\hat{\sigma} = \frac{med|e_i - med(e_i)|}{0.6745}$$

10.3.2. 가중치 함수

로버스트 회귀에서는 이상점에 크기에 반비례하는 가중치를 주어 그 영향을 축소한다. 이러한 이유에서 10.5 에서 정의된 가중치 함수 $w(u)$ 의 선택이 중요하다.

$$w(u) = \frac{\psi(u)}{u}$$

최소제곱법에서는 가중치 함수가 $w(u) = 1$ 이며 huber 함수를 이용한 M-추정량에서는 가중치 함수가 0 근처에서 1 이며 0으로부터 **특정한 값**만큼 멀어지면 가중치가 감소한다. 이렇게 가중치에 대한 영향을 조절하는 기준이 되는 특정한 값을 조율상수(tuning constant) 라고 한다.

Tukey 의 이중제곱(bi-square 또는 biweight) 함수는 가중치 함수가 0 으로 부터 멀어지면서 감소하기 시작하고 조율상수보다 멀어지먄 가중치가 0이 된다.

일반적으로 로버스트 회귀는 이러한 조율상수의 값에 따라서 추정량의 값이 달라진다. 조율상수의 값이 너무 작으면 로버스트 성질이 강해지지만 정보의 손실이 높으므로 상황에 맞게 적절하게 선택해야 한다.

아래 그림은 최소 제곱법, huber 함수를 이용한 M-추정량, Bi-setion 함수를 이용한 M-추정량의 $\rho(u)$, $\psi(u)$, $w(u)$ 함수를 그림으로 나타낸 것이다.

```
knitr::include_graphics(here::here("myimages","robust.png"))
```

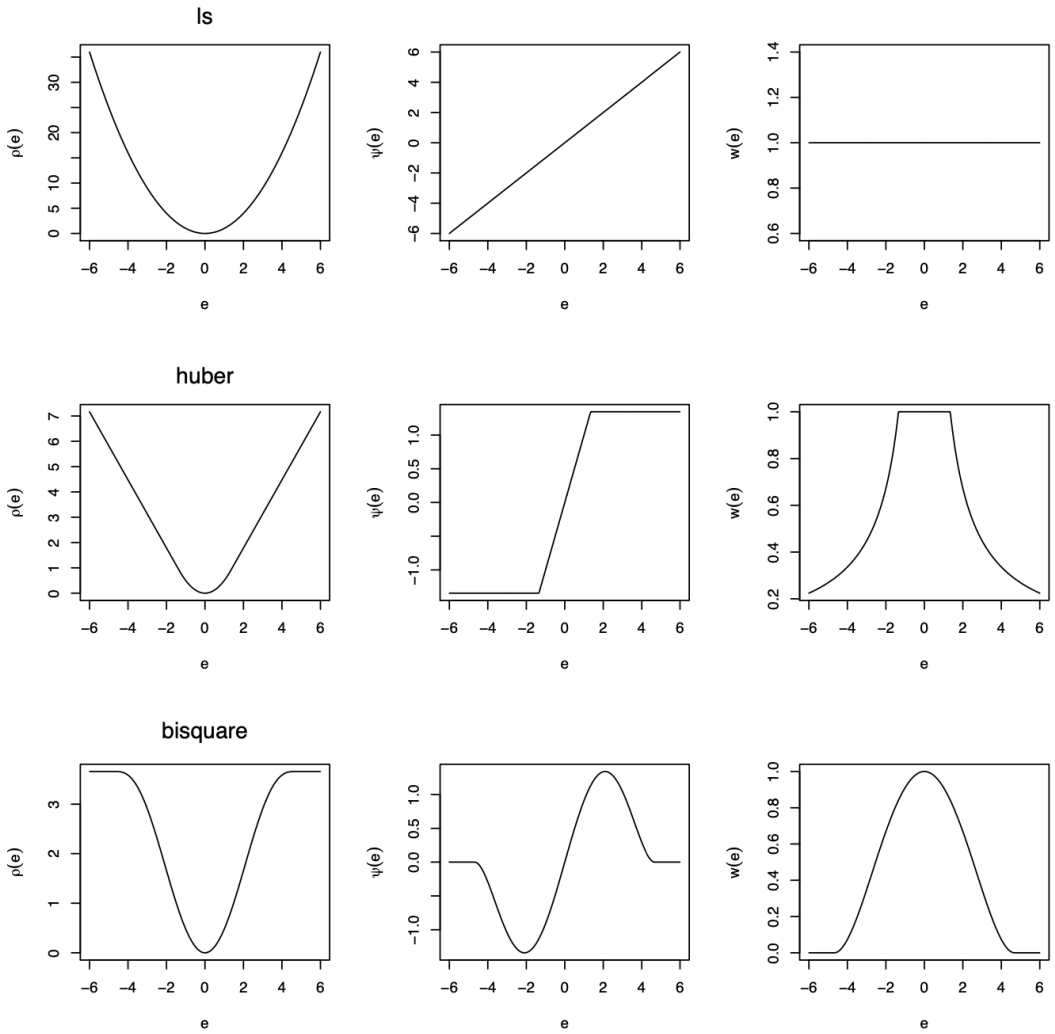


Figure 1: Objective, ψ, and weight functions for the least-squares (top), Huber (middle), and bisquare (bottom) estimators. The tuning constants for these graphs are $k = 1.345$ for the Huber estimator and $k = 4.685$ for the bisquare. (One way to think about this scaling is that the standard deviation of the errors, σ, is taken as 1.)

그림 10.1.: M-추정량의 이용되는 함수

그림의 출처: Fox and Weisberg (2018)

10.3.3. 기타 방법

최소절사제곱 추정량(least trimmed square; LTS)은 평균을 추정할 경우 절삭평균(trimmed mean)의 개념을 적용한 추정량이다. 잔차의 크기를 절대값 순으로 정렬하고 일정 비율의 큰 잔차를 가지는 관측치를 제외한 후에 추정량을 계산한다. 절삭하는 비율은 일반적으로 다음과 같이 설정한다.

$$\frac{3n+p+1}{4}$$

MM-추정량은 M-추정법과 LTS-추정법을 결합한 추정법이다.

10.3.4. 예제

열차 여객 운송에서 최종 탑승객의 수를 예측하는 경우를 고려하자. 출발 10일 전의 예약한 사람의 수(reserved)와 당일 실제 탑승객의 수(boarding) 의 관계를 이용하여 탑승객의 수를 10일 전에 예측하려고 한다.

예약한 사람의 수(reserved) 와 당일 실제 탑승객의 수(boarding) 에 대한 과거 자료(20개의 자요)를 얻어서 다음과 같이 데이터프레임으로 만들고 관계를 산포도로 그려 보았다.

```
df2
```

```
   reserved boarding
1        17       42
2        19       43
3        27       50
4        25       45
5        18       45
6        23       45
7        20       43
8        21       46
9        17       44
10       13       42
11       26       48
12       24       44
13       21       44
14       23       47
15       21       43
16       18       43
17       21       40
18       17       43
19       14       40
20       24       48
```

```
df2 %>% ggplot(aes(x=reserved, y=boarding)) + geom_point() + theme_bw()
```

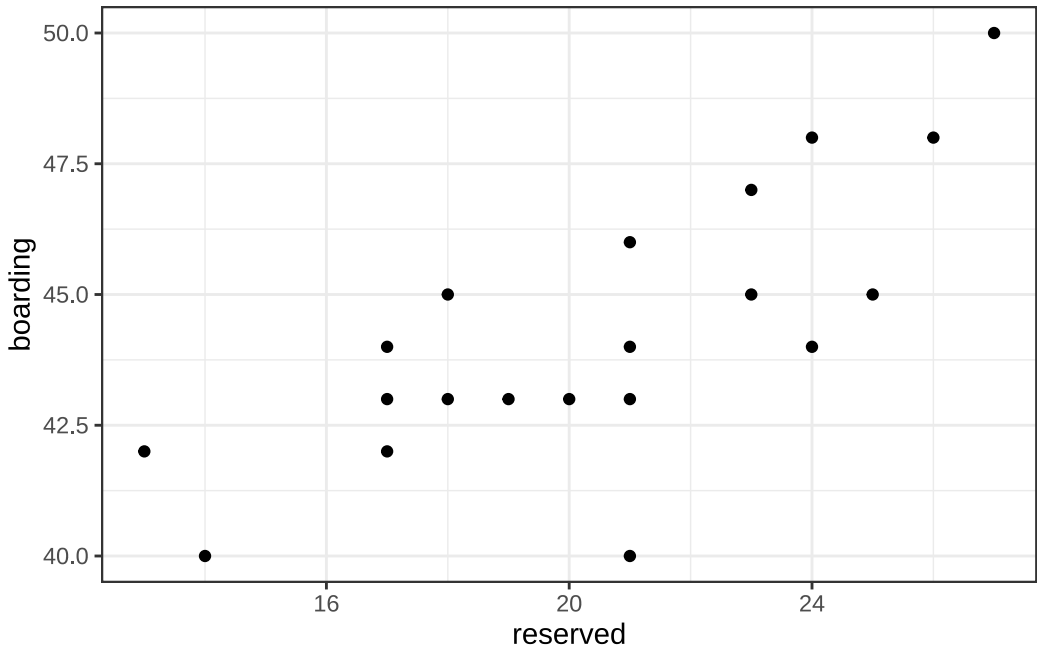


그림 10.2.: 10일전 예약 인원수와 실제 탑승객수의 관계

위의 그림에서 실제 탑승객의 수(boarding) 를 반응변수로 하는 단순회귀모형을 고려하였다. 추정된 회귀식을 이용한 예측식은 아래 그림에서 파란 선으로 나타난다.

```
roblm1 <- lm(boarding ~ reserved , data = df2)
df2$pred <- predict(roblm1)
summary(roblm1)
```

```
Call:
lm(formula = boarding ~ reserved, data = df2)

Residuals:
    Min      1Q  Median      3Q     Max
-4.5275 -1.0025 -0.2616  1.4770  2.4453

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  33.9322     2.1923  15.478 7.62e-12 ***
reserved      0.5045     0.1054   4.785 0.000148 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.767 on 18 degrees of freedom
Multiple R-squared:  0.5598,    Adjusted R-squared:  0.5354
```

```
F-statistic: 22.89 on 1 and 18 DF,  p-value: 0.0001483
```

```
df2 %>% ggplot(aes(x=reserved, y=boarding)) +
  geom_point() +
  stat_smooth(method = "lm", se = FALSE, data = df2,
              formula = y ~x ) +
  theme_bw()
```

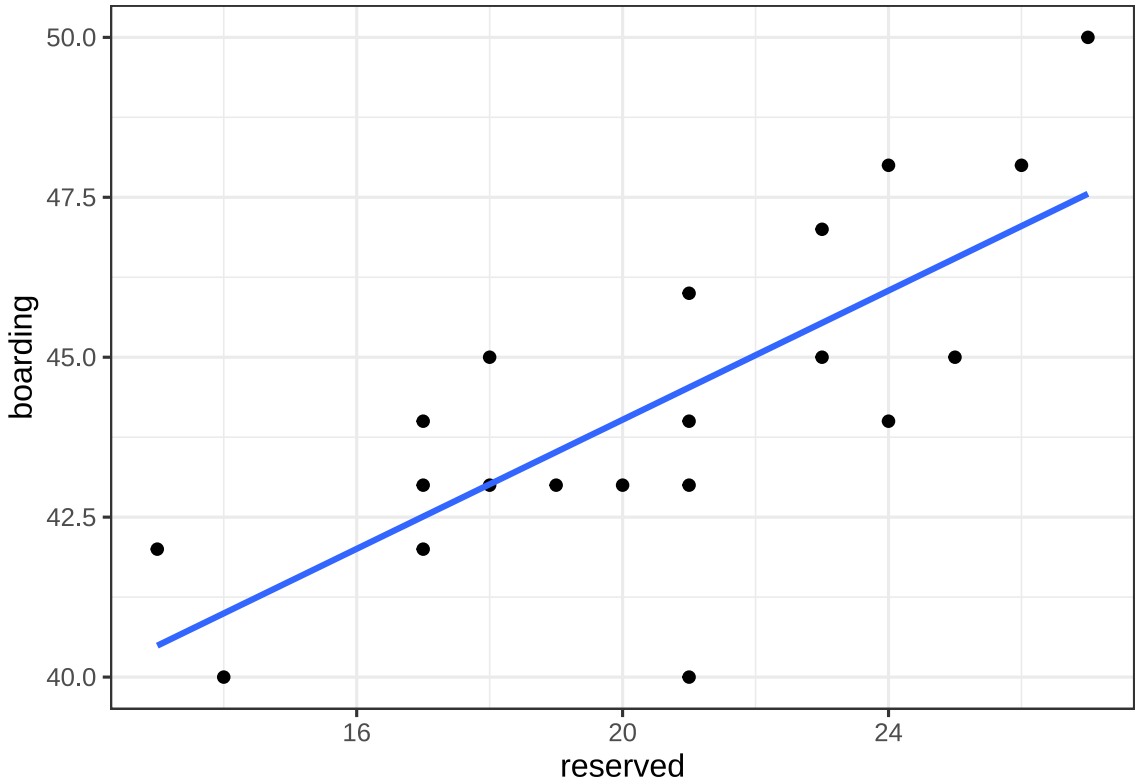


그림 10.3.: 탑승객수 추정 모형

이제 예약-탑승객 자료에 새로운 자료가 추가되었다고 가정하자. 그런데 추가된 새로운 자료는 10일 전의 예약한 사람의 수가 다른 자료보다 월등하게 많다. 이러한 이상점은 도착역에 근처에서 큰 행사(예를 들면 지방축제, 공무원시험 등)가 있는 날에 흔히 나타난다.

```
df3
```

```
   reserved boarding
1        17       42
2        19       43
3        27       50
4        25       45
5        18       45
6        23       45
```

7	20	43
8	21	46
9	17	44
10	13	42
11	26	48
12	24	44
13	21	44
14	23	47
15	21	43
16	18	43
17	21	40
18	17	43
19	14	40
20	24	48
21	49	44

```
df3 %>% ggplot(aes(x=reserved, y=boarding)) + geom_point() + annotate("text", x = 49, y = 43,
```

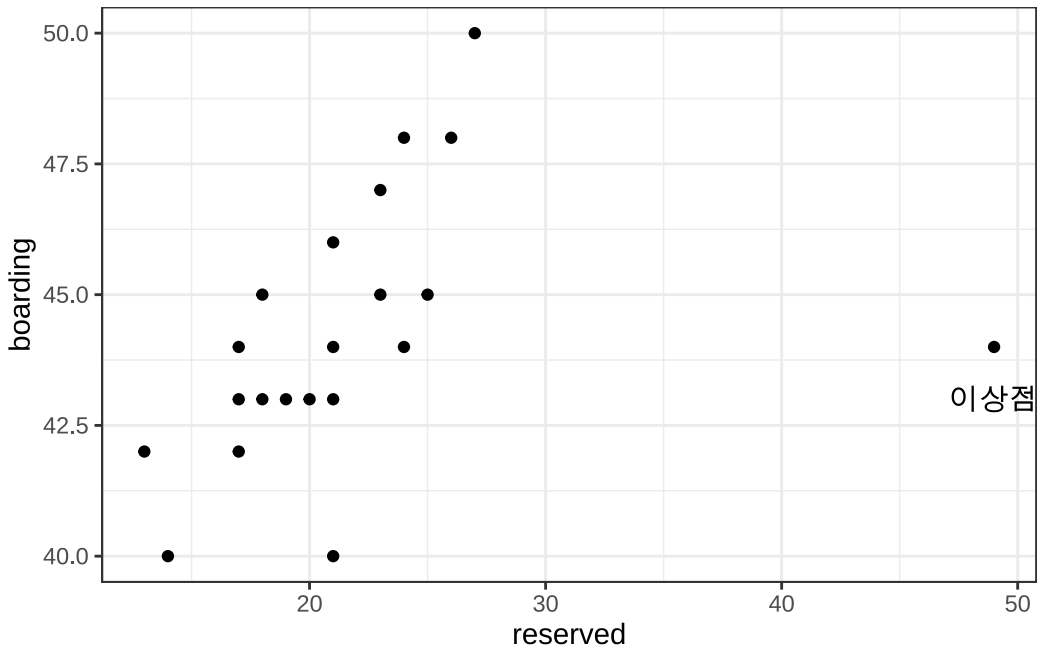


그림 10.4.: 10일전 예약 인원수와 실제 탑승객수의 관계 - 이상점이 포함된 경우

이렇게 이삼점이 자료에 포함되면 최종 탑승객을 추정하는 예측식에 큰 변화가 생겨서 예측의 정확성에 문제가 발생하게 된다.

```
roblm2 <- lm(boarding ~ reserved , data = df3)
df3$predLM <- predict(roblm2)
summary(roblm2)
```

```
Call:
lm(formula = boarding ~ reserved, data = df3)

Residuals:
    Min      1Q  Median      3Q     Max
-4.1348 -1.1136 -0.5177  1.2482  5.0994

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  41.4542     1.7017   24.36 8.59e-16 ***
reserved      0.1277     0.0742    1.72    0.102
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.413 on 19 degrees of freedom
Multiple R-squared:  0.1348,    Adjusted R-squared:  0.08924
F-statistic:  2.96 on 1 and 19 DF,  p-value: 0.1016
```

아래 그림의 파란 선은 이상점이 없을 때 예측식이고 빨간 선은 이상점이 포함된 경우의 예측식이다.

```
df3 %>% ggplot(aes(x=reserved, y=boarding)) +
  geom_point() +
  stat_smooth(method = "lm", se = FALSE, data = df3,
              formula = y ~x , colour="red" ) +
  geom_line(data=df2, aes(reserved, pred ), colour="blue"   ) +
  theme_bw()
```

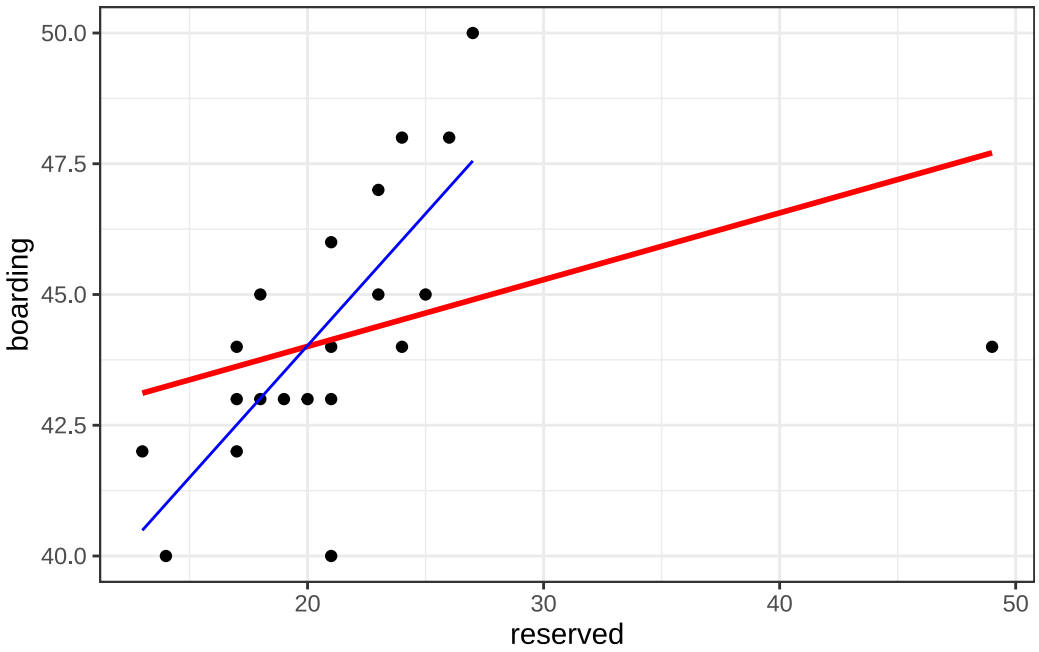


그림 10.5.: 탑승객수 추정 모형 - 이상점이 포함된 경우 -단순회귀

이렇게 많은 수의 회귀분석을 계속 수행하는 경우 이상점이 포함된 경우에는 회귀식을 적합할 때마다 잔차분석을 수행하기 어렵다. 따라서 이러한 경우는 로버스트 회귀식을 적용하면 잔차분석을 일일이 수행하지 않아도 이상점의 영향을 자동적으로 축소할 수 있다.

이제 위에서 언급한 MM-추정법을 이용하여 예측식을 적합해 보자.

```
roblm3 <- rlm(boarding ~ reserved , method= "MM", data = df3)
df3$predMM <- predict(roblm3)
summary(roblm3)
```

```
Call: rlm(formula = boarding ~ reserved, data = df3, method = "MM")
Residuals:
     Min       1Q   Median       3Q      Max
-14.7871  -1.0982  -0.5787   1.3953   2.3562

Coefficients:
            Value   Std. Error t value
(Intercept) 33.9679  1.2762    26.6165
reserved     0.5065  0.0556     9.1026

Residual standard error: 2.038 on 19 degrees of freedom
```

아래 그림는 MM-추정량을 이용하여 최종 탑승객수에 대한 추정모형을 적합한 결과이다. 그림의 파란 선은 이

상점이 없을 때 예측식이고 빨간 선은 이상점이 포함된 경우의 MM-추정에 의한 예측식이다. 예측 결과를 보면 MM-추정량은 추기된 이상치의 영향을 받지 않는 것으로 나타난다.

```
df3 %>% ggplot(aes(x=reserved, y=boarding)) +
  geom_point() +
  geom_line(data=df3, aes(reserved, predMM ), colour="red"  ) +
  geom_line(data=df2, aes(reserved, pred ), colour="blue"  ) +
  theme_bw()
```

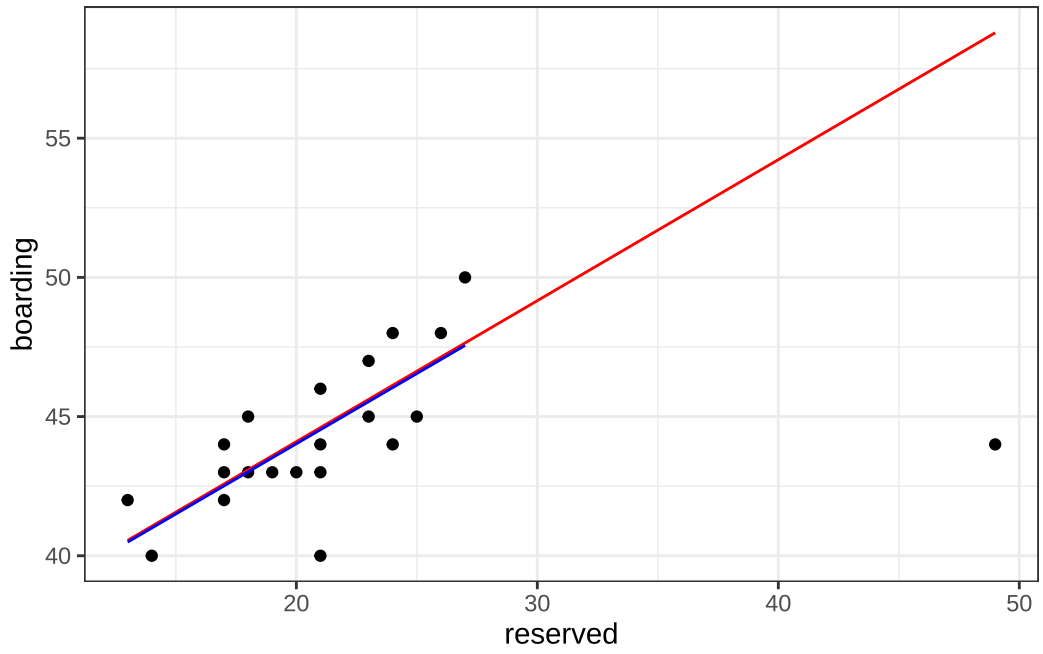


그림 10.6.: 탑승객수 추정 모형 - 이상점이 포함된 경우 - MM 추정

10.4. 비선형 회귀분석

10.4.1. 예제: 철도 여객 운송

철도 여객 운송은 사람들이 장거리 이동을 하는 경우 많이 이용한다. 하루에고 백 대가 넘는 열차가 운행되며 하나의 열차는 출발역에서 시작하여 중간에 여러 역에 정차함으로서 여러 지역 간의 여행을 가능하게 한다. 이렇게 하루에도 수 천개의 출발지와 목적지를 가지는 기차 노선이 운영되고 있다. 따라서 여객 운송을 관리하는 주체는 여객 노선의 수요를 예측하고 좌석을 합리적이고 효율적으로 할당하는 작업이 필요하다.

다음은 A역에서 B역으로 승객을 수송하는 고속열차의 예약 현황을 일별로 나타낸 자료이다. 출발일 20일 전부터 예약을 받으며 출발일에 수송할 수 있는 승객의 최대 수는 50명이다.

예약 자료와 그에 대한 그림을 그려보면 다음과 같다.

```
df %>% head()
```

```
  days reserved
1  -20        1
2  -19        0
3  -18        0
4  -17        2
5  -16        0
6  -15        0
```

```
df %>% ggplot(aes(x=days, y=reserved)) +
       geom_point()  +
       labs(x="출발전 일수", y="예약 인원수") +
       theme_bw()
```

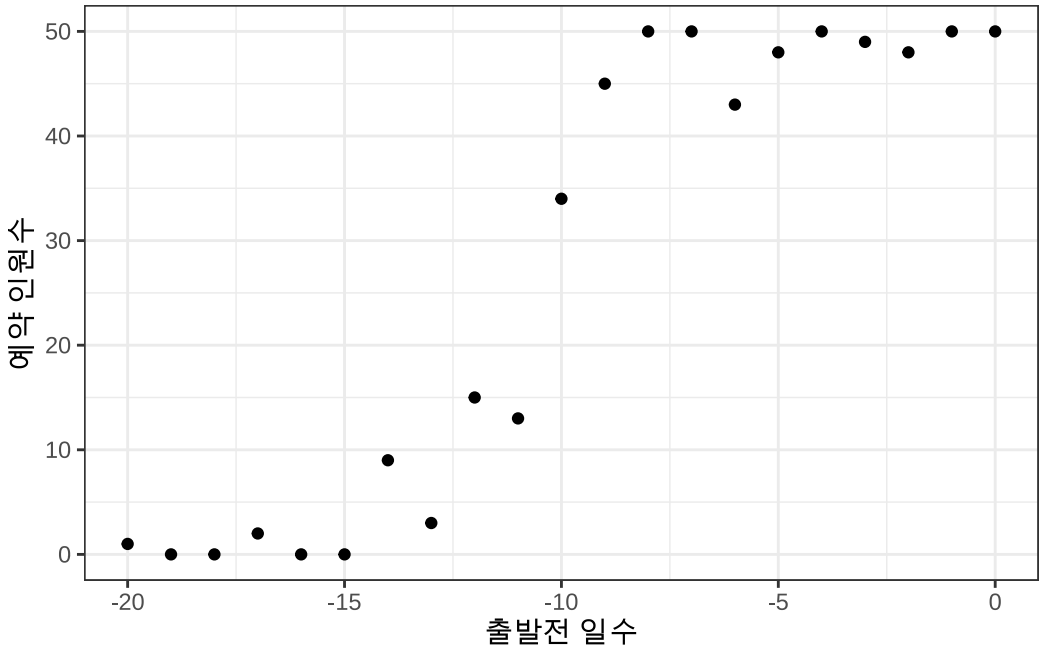


그림 10.7.: 예약 20일 전에서 출발일까지의 열차 예약 현황

예약 현황을 일별로 나타낸 그림 그림 10.7 의 특징을 보면 다음과 같다.

- 예약 인원수는 가장 작은 값이 0명, 가장 큰 값이 50명 이라는 제약이 있다.
- 예약 인원수는 일반적으로 출발일에 가까와 지면서 증가를 하는 경향이 있다.
- 예약 인원수는 특정 시점에서 급격하게 증가한다.
- 증가를 하지만 할당된 50 좌석을 넘을 수 는 없다 (수렴성).

10.4.2. 선형모형의 한계

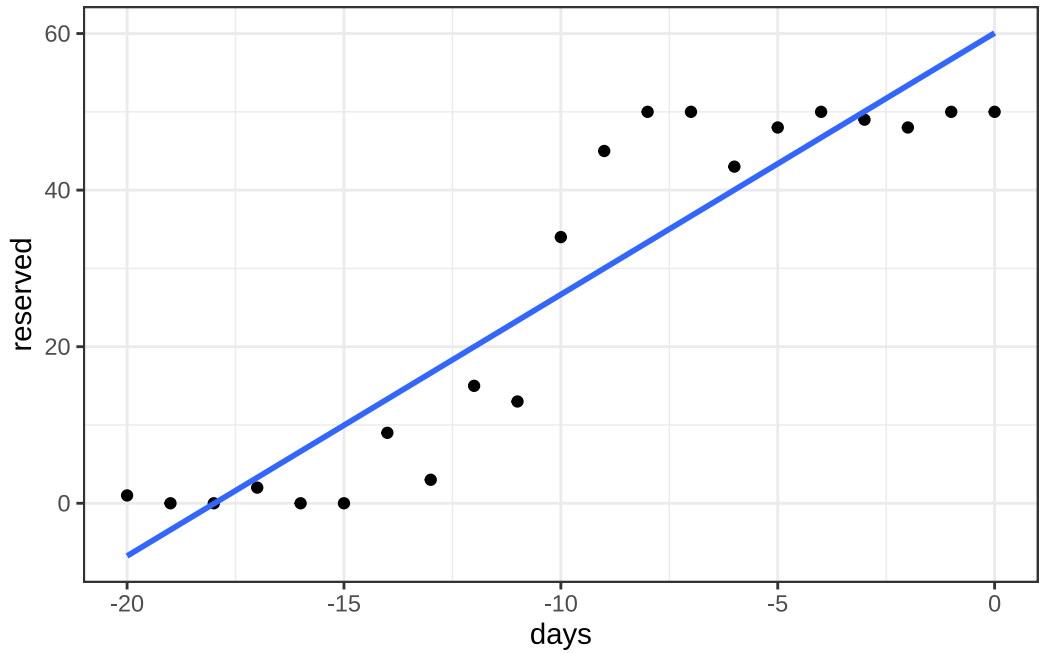
이제 열차 예약 자료를 예약전 일수(days)을 설명변수로 하고 예약 인원 수(reserved)를 반응변수로 하는 단순 회귀모형을 적합해보자.

아래 그림에서 볼 수 있듯이 단순 회귀모형은 예약 자료의 특성(최소/최대 예약 인원, 수렴성 등)을 전혀 반영하지 못한다.

```
lm1 <- lm(reserved ~ days, data = df)
summary(lm1)
```

```
Call:
lm(formula = reserved ~ days, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-13.646  -6.625  -1.048   4.632  16.653

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  60.0693     3.7477   16.03 1.71e-12 ***
days          3.3403     0.3206   10.42 2.70e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.896 on 19 degrees of freedom
Multiple R-squared:  0.8511,    Adjusted R-squared:  0.8432
F-statistic: 108.6 on 1 and 19 DF,  p-value: 2.704e-09
```

```
df %>% ggplot(aes(x=days, y=reserved)) +
  geom_point() +
  stat_smooth(method = "lm", se = FALSE, data = df,
              formula = y ~x ) +
  theme_bw()
```



이제 더 복잡한 모양인 다항식 모형(polynomial regression)을 생각해보자. 다항식 모형은 반응변수가 가진 비선형적인 특성을 다소 고려할 수 있다. 이제 열차 예약 자료에 대하여 다음과 같은 3차 다항식 모형을 고려해 보자.

$$y = \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3 + e$$

3차 다항식 모형을 적합한 결과와 예측식의 그림은 다음과 같다.

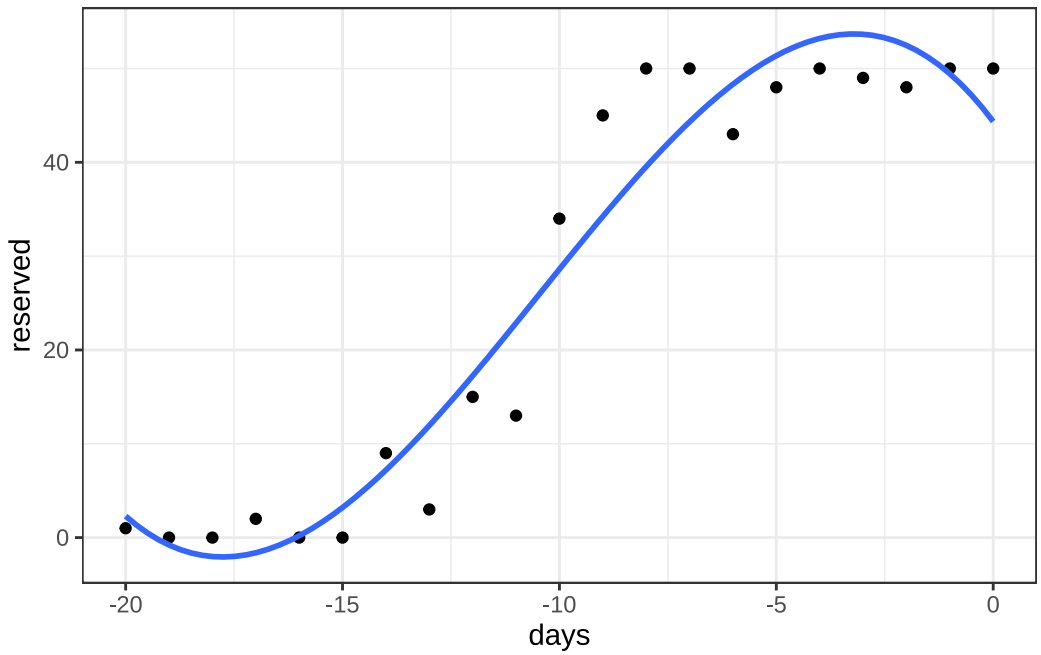
3차 다항식을 고려한 모형은 일차식에 비하여 자료의 특성을 어느 정도 반영하였으나 예약 인원의 제약과 수렴성은 반영할 수 없다. 주목할 점은 3차 다항식 또는 고차 다항식도 기본적으로 회귀계수에 대한 선형모형이다.

```
lm2 <- lm(reserved ~ days+I(days^2) + I(days^3), data = df)
summary(lm2)
```

```
Call:
lm(formula = reserved ~ days + I(days^2) + I(days^3), data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-9.8672 -3.3632 -0.1931  3.6205 10.7646

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 44.352532   4.475318   9.910 1.77e-08 ***
days        -6.187604   1.986322  -3.115 0.006297 **
I(days^2)   -1.137947   0.234147  -4.860 0.000147 ***
```

```
I(days^3)   -0.036170   0.007687  -4.705 0.000204 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.067 on 17 degrees of freedom
Multiple R-squared:  0.938, Adjusted R-squared:  0.9271
F-statistic: 85.74 on 3 and 17 DF,  p-value: 1.813e-10
```

```
df %>% ggplot(aes(x=days, y=reserved)) +
  geom_point() +
  stat_smooth(method = "lm", se = FALSE, data = df,
              formula = y ~x+I(x^2) + I(x^3) ) +
  theme_bw()
```



10.4.3. 비선형 모형

비선형 모형은 반응변수의 변화를 설명변수에 대한 선형모형으로 표현할 수 없는 경우 사용되는 모형이다.

반응변수의 변화가 설명변수에 대한 단순한 선형모형으로 표현이 안되는 경우 변수의 변환(예를 들어 로그 변환)이나 고차원 항을 고려해서 변환된 모형으로 적합할 수 있다. 하지만 이러한 변수 변환이나 고차식의 포함 등 으로 반응변수의 변화를 설명할 수 없는 경우도 있다.

이렇게 반응변수의 변화가 가지는 특성을 반영할 수 있는 비선형 함수를 사용한 모형을 비선형 회귀모형이라고 부른다. 비선형 회귀모형은 반응변수의 중요한 특성을 미리 파악할 수 있는 경우 주로 사용된다. 반응값의 중요한 특성을 반영할 수 있는 비선형 함수를 모형으로 사용하는 것이다. 따라서 비선형 회귀분석은 ㅈ자료가 가지는 중요한 특성과 자료가 생성되는 자세한 과정에 대한 과학적인 모형을 알 고 있는 경우에 주로 사용된다.

$$y_i = f(x_i, \boldsymbol{\beta}) + e_i \tag{10.7}$$

예를 들어 강근석 and 유형조 (2016) 예제 9.1 은 식물의 성장 속도에 대하여 다음과 같은 비선형 모형인 미캘리스-멘텐(Michaelis-Menten) 모형을 고려한다.

$$f(x; \theta_1, \theta_2) = \frac{\theta_1 x}{\theta_2 + x}$$

위의 미캘리스-멘텐 모형에서 모수 θ_1 과 θ_2 는 다음과 같은 특별한 의미를 가지고 있다.

표 10.1.: 미캘리스-멘텐(Michaelis-Menten) 비선형 함수의 모수와 의미

모수	설명	R `SSmicmen()` 함수의 인자
θ_1	수평 수렴 한계(horizontal asymptote)	`Vm`
θ_2	$y = \theta_1/2$ 가 되는 x 값	`k`

위의 열차 예약에 대한 예제에서 나타나는 특성은 반응 변수의 최소값과 최대값이 존재한다는 것이며 독립변수가 증가하면 반응값이 수평적으로 수렴상태(horizontal asymtote)에 이른다는 것이다. 또한 어느 시점이 되면 반응변수가 매우 빠르게 증가한다. 이러한 미리 파악된 반응값의 특성을 반영할 수 있는 비선형 함수를 모형으로 선택해야 한다.

열차 예약 자료와 같은 반응값의 한계와 수렴성이 있는 모형을 설명할 때 자주 사용되는 모형이 아래와 같은 로지스틱(logistic) 함수이다.

$$y = f(x; \phi_1, \phi_2, \phi_3) = \frac{\phi_1}{1 + \exp[(\phi_2 - x)/\phi_3]} \tag{10.8}$$

위의 로지스틱 함수 10.8 에서 나타난 3개의 모수 ϕ_1, ϕ_2, ϕ_3 는 반응변수의 변화에 대한 특별한 의미를 지니고 있으며 그 설명은 다음 표와 그림과 같다.

표 10.2.: 3개의 모수로 이루어진 로지스틱 비선형 함수의 모수와 의미

모수	설명	R `SSlogis()` 함수의 인자
ϕ_1	수평 수렴 한계(horizontal asymptote)	`Asym`
ϕ_2	$y = \phi_1/2$ 가 되는 x 값	`xmid`
ϕ_3	크기 모수 (scale parameter)	`scal`

```
knitr::include_graphics(here::here("myimages","logistic.png"))
```

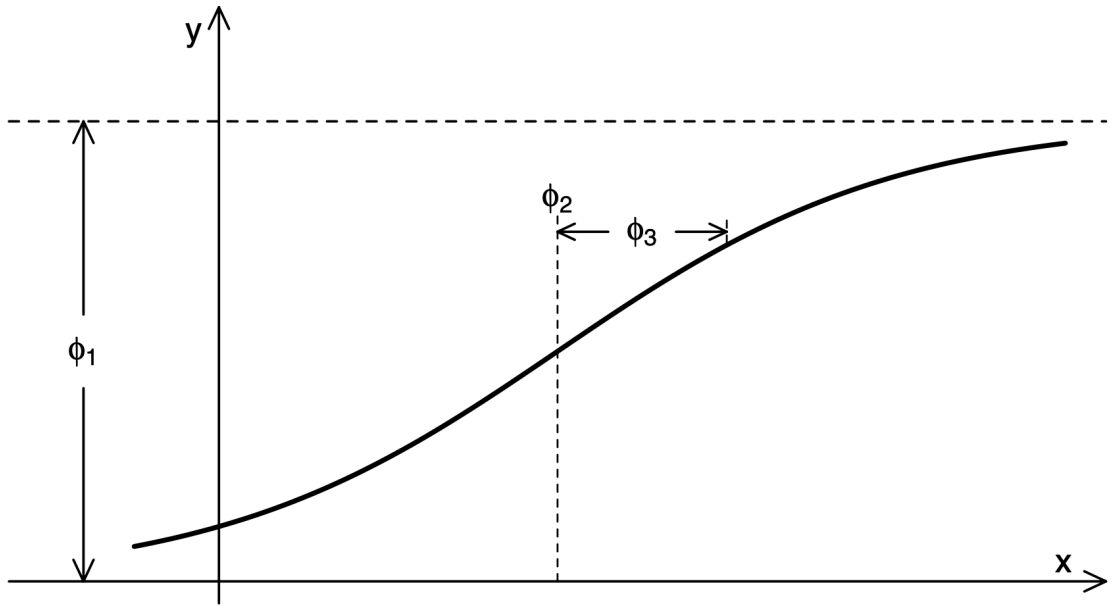


FIGURE C.7. The simple logistic model showing the parameters ϕ_1, the horizontal asymptote as $x \to \infty$, ϕ_2, the value of x for which $y = \phi_1/2$, and ϕ_3, a scale parameter on the x-axis. If $\phi_3 < 0$ the curve will be monotone decreasing instead of monotone increasing and ϕ_1 will be the horizontal asymptote as $x \to -\infty$.

그림 10.8.: 로지스틱 함수와 모수

위의 그림의 출처는 Pinheiro and Bates (2006) 이며 로지스틱 함수 외의 다양한 비선형 모형의 정의와 그에 대한 설명도 Pinheiro and Bates (2006) 또는 강근석 and 유형조 (2016) 의 표 9.2 에서 찾아볼 수 있다.

10.4.4. 비선형모형의 적합

비선형 모형의 적합은 함수 `nls()` 를 이용한다. 10.8 에서 정의된 로지스틱 함수는 R 함수 `SSlogis()` 함수로 미리 정의되어 있다. 함수 `SSlogis()` 의 인자는 다음과 같이 4개가 필요하며 인자 `input` 은 독립변수이고 나머지 3개의 인자 `Asym`, `xmid`, `scal` 은 표 @ref(tab:logistic)) 에 모수와 관계가 설명되어 있다.

```
SSlogis(input, Asym, xmid, scal)
```

함수 `SSlogis()` 는 Self-Starting 함수라고 부르며 주어진 자료에 대하여 모수의 초기값을 지정해주지 않아도 내부에서 자동적으로 계산해 주는 기능이 있다. 예를 들어서 열차 예약 자료에 대하여 로지스틱 함수를 함수 `SSlogis()` 로 적합하는 경우, 사용되는 초기값은 함수 `getInitial()` 를 사용하여 다음과 같이 계산해준다.

```
getInitial(reserved ~ SSlogis(days, Asym, xmid, scal), data = df)
```

```
      Asym       xmid       scal 
49.2546737 -10.6613959   0.9315841 
```

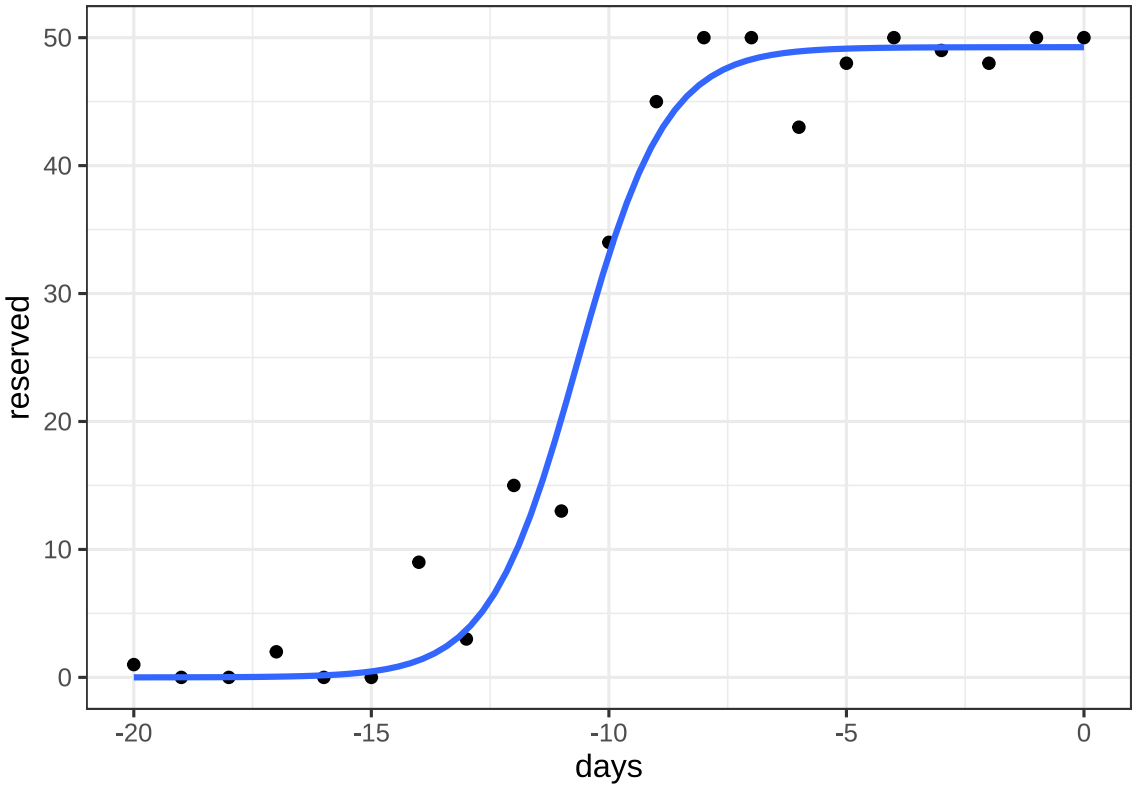
이제 함수 nls() 와 SSlogis() 를 이용하여 로지스틱 함수 10.8 로 열차 예약 자료를 적합해 보자.

```
fm1 <- nls(reserved ~ SSlogis(days, Asym, xmid, scal), data = df)
summary(fm1)
```

```
Formula: reserved ~ SSlogis(days, Asym, xmid, scal)

Parameters:
     Estimate Std. Error t value Pr(>|t|)
Asym  49.2547     1.2276  40.122  < 2e-16 ***
xmid -10.6614     0.1782 -59.824  < 2e-16 ***
scal   0.9316     0.1542   6.042 1.03e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.417 on 18 degrees of freedom

Number of iterations to convergence: 0
Achieved convergence tolerance: 2.56e-06
```

로지스틱 함수 10.8 로 적합한 모형의 예측값을 그림으로 다음과 같이 나타낼 수 있다. 로지스틱 함수가 가지고 잇는 특성으로 인하여 예약 자료에 대한 적합이 적절한 것을 알 수 있다.

```
df %>% ggplot(aes(x=days, y=reserved)) +
  geom_point() +
  stat_smooth(method = "nls", se = FALSE, data = df,
              formula = y ~ SSlogis(x, Asym, xmid, scal),
              method.args  = list( start=list ( Asym=coef(fm1)[1],xmid=coef(fm1)[2], scal=coef
  theme_bw()
```



10.4.5. 비선형 회귀의 추론

- 비선형 회귀분석에서도 모수에 대한 추론, 즉 가설검정과 신뢰구간 등을 구할 수 있다. 단 모든 추론은 비선형 함수의 선형 근사(Gauss-Newton method)와 점근적 방법(asymptotic methods)에 기반한다. 선형화와 점근적 방법에 기반한 추론은 그에 대한 가정이 어느 정도 만족한 경우 유효하므로 분석 시 이러한 가정이 적합한지 판단해서 추론 결과를 사용해야 한다. 비선형 회귀분석에 대한 추론의 기초 이론과 주의할 점은 강근석 and 유형조 (2016) 의 363–367 페이지 에 설명되어 있다.

열차 예약 자료를 적합한 로지스틱 모형에서 각 모수에 대한 점근적 95% 신뢰구간은 다음과 같다.

```
confint(fm1)
```

```
             2.5%       97.5%
Asym   46.7686165   51.829073
xmid -11.0484643 -10.299248
scal   0.5971395   1.329766
```

10.4.6. 유의할 사항

비선형 회귀 모형을 고려한 분석을 수행하는 경우 다음과 같은 사항들에 대하여 유의해야 한다.

- 비선형 회귀모형을 적합하는 경우 모수의 초기값을 설정해 주어야 한다. 사용하는 함수에 따라서 초기값이 자동적으로 계산되는 경우도 있다. 일반적으로 모수의 초기값은 경험과 지식을 바탕으로 자료에서 적절하게 정해져야 한다. 잘못 선정된 초기값은 가끔 모수의 부정확한 추정을 일으킬 수 있다. 모수의 초기값 설정에 대한 설명은 강근석 and 유형조 (2016) 의 9.2.3 절에 있다.
- 비선형 회귀 모형은 주로 반응값 변화에 대한 과학적인 모형이 존재하는 경우 주로 사용된다. 예를 들어 약동력학(pharmacokinetics) 에서 약의 성분이 인체에 퍼지는 속도 등을 미분 방정식으로 유도한 칸막이 모형(compartment model)이 대표적인 예이다. 따라서 자료의 특성을 반영하는 적절한 비선형 모형을 선택해야 한다.
- 비선형 회귀모형에서는 모수를 표현하는 형식에 따라서 모형의 적합이 영향을 받을 수도 있다. 강근석 and 유형조 (2016) 의 369 페이지에 설명된 것처럼 미켈리스-멘텐(Michaelis-Menten) 비선형 모형식은 다음과 같이 두 개의 서로 다른 형태의 모수로서 표현될 수 있다.

$$f(x;\theta_1,\theta_2) = \frac{\theta_1 x}{\theta_2 + x} \quad \text{or} \quad f(x;\beta_1,\beta_2) = \frac{x}{\beta_1 + \beta_2 x}$$

10.5. 비모수 회귀모형

반응변수 Y 와 설명변수 X 가 다음과 같은 관계를 가진다고 하자.

$$E(Y|X = x) = m(x)$$

이러한 관계를 회귀모형(regression model)이라고 하며 Y 의 평균이 X 에 따라서 변하는 괸계를 설정하는 모형이다. 만약 $m(x)$ 의 형태를 회귀계수의 선형식으로 나타낼 수 있다면 우리는 이를 선형회귀모형(linear regression model)이라고 한다.

$$m(x) = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

설명변수는 고정된 값이거나 또는 확률변수일 수도 있다.

비모수 회귀모형(nonparametric regression model)에서는 $m(x)$ 의 형태에 특별한 제한을 두지 않는다. 따라서 함수 $m(x)$ 는 무수히 많고 다양한 형태를 가질 수 있다. n 개의 독립표본 $(Y_1, X_1), (Y_2, X_2), \ldots, (Y_n, X_n)$ 이 주어지면 비모수 회귀모형을 다음과 같이 기술할 수 있다.

$$Y_i = m(X_i) + e_i$$

여기서 X_i 와 독립인 오차항 e_i 는 서로 독립이고 평균이 0이며 분산이 τ^2 인 확률변수로 가정한다. 따라서 설명변수의 값이 x 라면 아래와 같이 나타낼 수 있다.

$$E(Y_i|X_i = x) = E[m(X_i) + e_i|X_i = x] = m(x)$$

10.5.1. 편이-분산의 관계

만약 $\hat{m}(x)$가 $m(x)$의 추정량이라면 예측위험함수(prediction risk, prediction error) $R(m, \hat{m})$는 다음과 같이 정의되고분해할 수 있다. 아래의 식을 유도할 때 편의상 X는 확률변수가 아닌 고정된 값이라고 할 것이며 확률변수인 경우에도 유사한 결과를 얻는다.

$$\begin{aligned} R(m, \hat{m}) &= E[Y - \hat{m}(X)]^2 \\ &= E[(Y - m(X) + m(X) - E(\hat{m}(X)) + E(\hat{m}(X)) - \hat{m}(X)]^2 \\ &= E[(Y - m(X)]^2 + E[m(X) - E(\hat{m}(X))]^2 + E[\hat{m}(X) - E(\hat{m}(X))]^2 \\ &\quad + cross - product - terms \end{aligned}$$

위의 식에서 다음과 같은 결과를 이용하면 교차항들(cross-product terms)은 0이 됨을 보일 수 있으며

$$E[Y - m(X)] = E(e) = 0, \quad E[\hat{m}(X) - E(\hat{m}(X))] = 0$$

또한 다음과 같은 관계를 이용하면

$$\begin{aligned} E[Y - m(X)]^2 &= E(e^2) = \tau^2 \\ bias(\hat{m}(X)) &= m(X) - E(\hat{m}(X)) \\ E[\hat{m}(X) - E(\hat{m}(X))]^2 &= Var(\hat{m}(X)) \end{aligned}$$

다음과 같이 예측위험함수의 분해가 가능하다.

$$R(m, \hat{m}) = E[Y - \hat{m}(X)]^2 = \tau^2 + [bias(\hat{m}(X))]^2 + Var(\hat{m}(X))$$

위의 식에서 X가 확률변수이면 먼저 X가 주어진 조건부 기대값을 생각하고 위와 같이 유도하면 다음과 같이 나타낼 수 있다.

$$\begin{aligned} R(m, \hat{m}) &= E[Y - \hat{m}(X)]^2 \\ &= E\left\{E[(Y - \hat{m}(X))^2 | X = x]\right\} \\ &= \tau^2 + \int [bias(\hat{m}(x))]^2 dP(x) + \int Var(\hat{m}(x)) dP(x) \end{aligned}$$

10.5.2. Regressogram

Regressogram은 histogrm을 회귀모형에 적용한 방법이다. 아래의 함수 `regressogram`은 설명변수 벡터 `x`와 반응변수 벡터 `y` 를 인자로 받으며 구간 `(left, right)` 를 `k`개의 구간으로 나누어 각 구간마다 반응값들의 평균을 구해주는 함수이다. 이렇게 주어진 구간에서 평균값을 이용하여 회귀식을 추정하는 모형을 국소회귀(local regression)이라고 한다.

```
regressogram = function(x,y,left,right,k,plotit,xlab="",ylab="",sub=""){
  ### assumes the data are on the interval [left,right]
  n = length(x)
  B = seq(left,right,length=k+1)
  WhichBin = findInterval(x,B)
  N = tabulate(WhichBin)
  m.hat = rep(0,k)
  for(j in 1:k){
    if(N[j]>0)m.hat[j] = mean(y[WhichBin == j])
  }
  if(plotit==TRUE){
    a = min(c(y,m.hat))
    b = max(c(y,m.hat))
    plot(B,c(m.hat,m.hat[k]),lwd=3,type="s",
         xlab=xlab,ylab=ylab,ylim=c(a,b),col="blue",sub=sub)
    points(x,y)
  }
  return(list(bins=B,m.hat=m.hat))
}
```

위의 프로그램에서 `findInterval(x,B)`은 전체 구간 `B`의 각 구간에 벡터 `x` 의 값들이 속해있는 정보를 계산해 준다. 예를 들어서 $(0, 10)$ 구간을 4개로 나누고 $(1, 2, 2, 4, 6, 7, 8)$의 값이 어떤 구간에 속해있는지 다음과 같이 알 수 있다.

```
B = seq(0,10,length=5)
B
```

```
[1]  0.0  2.5  5.0  7.5 10.0
```

```
x = c(1,2,2,4,6,7,8)
findInterval(x,B)
```

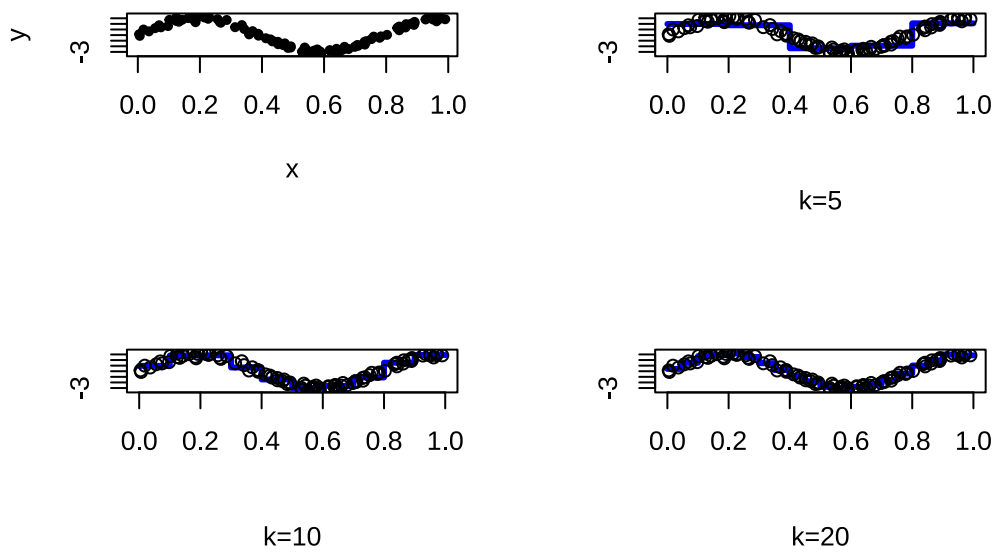
```
[1] 1 1 1 2 3 3 4
```

이제 다음과 같은 $m(x)$를 고려하고 오차항 e가 정규분포 $N(0, 3^2)$을 따른다고 가정하고 임의로 100개의 독립표본을 만든다. x의 값들은 구간 $(0, 1)$에서 균등분포를 따른다.

$$m(x) = 3\sin(8x)$$

이제 구간의 수를 $k = 5, 10, 20$ 으로 바꾸면서 regressogram이 어떤 형태로 $m(x)$를 추정하는지 알아보자.

```
par(mfrow=c(2,2))
n = 100
x = runif(n)
y = 3*sin(8*x) + rnorm(n,0,.3)
plot(x,y,pch=20)
out = regressogram(x,y,left=0,right=1,k=5,plotit=TRUE, sub="k=5")
out = regressogram(x,y,left=0,right=1,k=10,plotit=TRUE, sub="k=10")
out = regressogram(x,y,left=0,right=1,k=20,plotit=TRUE, sub="k=20")
```



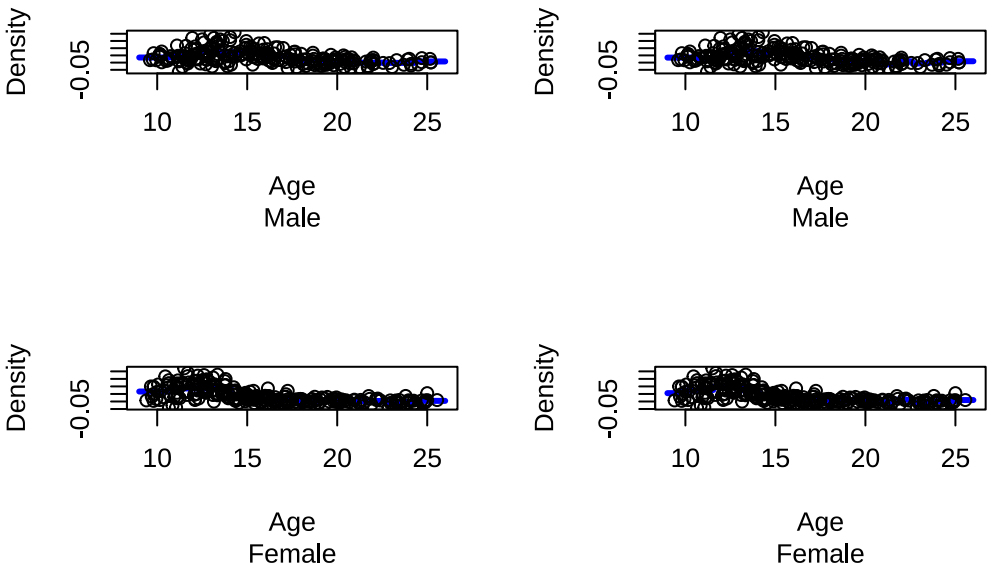
이제 라이브러리 ElemStatLearn에 있는 실제 자료 bone를 이용하여 regressogram 을 그려보자. 자료 bone의 반응변수는 density (Measurements in the bone mineral density) 이며 설명변수는 나이(age)이다. 남자와 여자를 따로 그렸으며 구간의 크기는 각각 $k = 10$과 $k = 20$을 사용해 보았다.

라이브러리 ElemStatLearn는 더 이상 R 라이브러리에 없으므로 다음 사이트에서 다운로드 받아서 직접 화일로 설치해야 한다

```
# 다운로드 사이트: https://cran.r-project.org/src/contrib/Archive/ElemStatLearn/
install.packages("~/Downloads/ElemStatLearn_2015.6.26.2.tar.gz", repos = NULL, type = "source"
```

```
par(mfrow=c(2,2))
#install.packages("ElemStatLearn")
library(ElemStatLearn)
attach(bone)
age.male = age[gender == "male"]
density.male = spnbmd[gender == "male"]
out = regressogram(age.male,density.male,left=9,right=26,k=10,plotit=TRUE,
```

```
                    xlab="Age",ylab="Density",sub="Male")
out = regressogram(age.male,density.male,left=9,right=26,k=20,plotit=TRUE,
                    xlab="Age",ylab="Density",sub="Male")
age.female = age[gender == "female"]
density.female = spnbmd[gender == "female"]
out = regressogram(age.female,density.female,left=9,right=26,k=10,plotit=TRUE,xlab="Age",ylab=
out = regressogram(age.female,density.female,left=9,right=26,k=20,plotit=TRUE,xlab="Age",ylab=
```



10.5.3. 커널추정

10.5.3.1. 커널추정법

국소회귀에서 가장 많이 사용되는 방법이 커널추정(kernel estimation) 방법이다. 다음의 함수 `kernel(x,y,grid,h)`은 표준편차의 크기가 h인 정규분포 확률밀도함수를 커널로 이용하여 주어진 값 x에서 반응변수 y들의 가중평균을 구해주는 함수이다. 변수 `grid`는 `x`축에서 커널을 계산할 때 사용되는 점들 모아놓은 벡터이다.

$$\hat{m}(x) = \frac{\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right) Y_i}{\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right)}, \quad K\left(\frac{x-y}{h}\right) = (2\pi h^2)^{-1/2} \exp\left(-\frac{(x-y)^2}{2h^2}\right)$$

```
kernel = function(x,y,grid,h){
  ### kernel regression estimator at a grid of values
  n = length(x)
  k = length(grid)
  m.hat = rep(0,k)
  for(i in 1:k){
```

```
    w = dnorm(grid[i],x,h)
    m.hat[i] = sum(y*w)/sum(w)
  }
  return(m.hat)
}
```

다음의 함수 `kernel.fitted(x,y,h)`는 주어진 자료벡터 `x,y`와 구간의 크기 `h`에 대하여 반응변수의 예측값 $\hat{Y}_i = \hat{m}_h(X_i)$ 을 커널추정으로 구해준다.

$$\hat{Y}_i = \hat{m}(X_i) = \frac{\sum_{j=1}^n K\left(\frac{X_i - X_j}{h}\right) Y_j}{\sum_{j=1}^n K\left(\frac{X_i - X_j}{h}\right)} = \sum_{j=1}^n l_j(X_i) Y_j$$

또한 위의 식에서 $L_{ij} = l_j(X_i)$라고 하면 다음과 같은 선형예측식에서 $n \times n$ 행렬 $L = \{L_{ij}\}$의 대각원소 L_{ii}도 계산해준다.

$$\hat{Y} = LY$$

```
kernel.fitted = function(x,y,h){
  ### fitted values and diaginal of smoothing matrix
  n = length(x)
  m.hat = rep(0,n)
  S = rep(0,n)
  for(i in 1:n){
    w = dnorm(x[i],x,h)
    w = w/sum(w)
    m.hat[i] = sum(y*w)
    S[i] = w[i]
  }
  return(list(fitted=m.hat,S=S))
}
```

10.5.3.2. 최적 구간의 길이 선택

다음의 함수 `CV(x,y,H)`는 자료벡터 `x,y`와 서로 다른 h의 값들을 모아 놓은 벡터 `H`를 받아서 다음과 같은 cross-validation의 값 CV, GCV, ν를 계산해 주는 함수이다.

$$CV = \frac{1}{n}\sum_i^n [Y_i - \hat{m}_h^{(-i)}(X_i)]^2 = \frac{1}{n}\sum_i^n \left(\frac{Y_i - \hat{m}_h(X_i)}{1 - L_{ii}}\right)^2$$

$$GCV = \frac{1}{(1-\frac{\nu}{n})^2}\frac{1}{n}\sum_i^n (Y_i - \hat{m}_h(X_i))^2$$

$$\nu = trace(L)$$

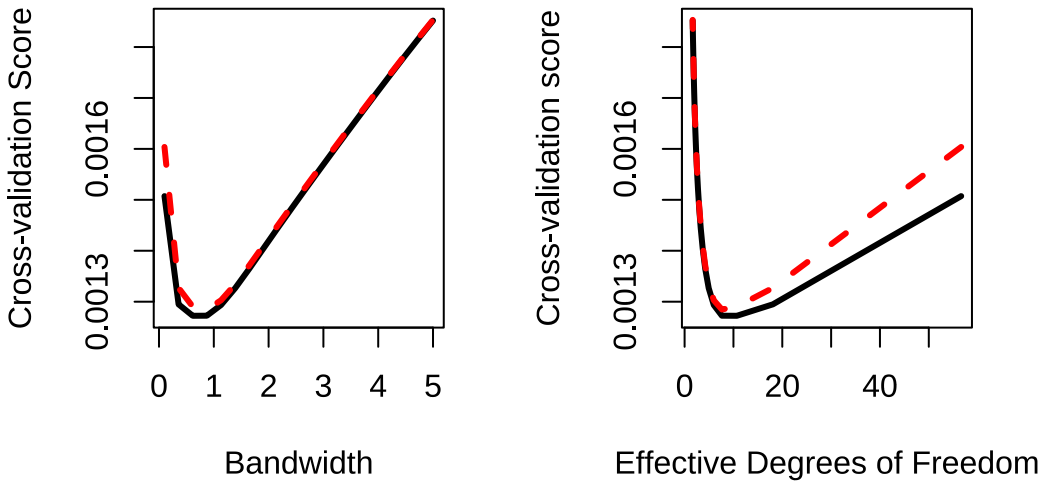
```
CV = function(x,y,H){
  ### H is a vector of bandwidths
  n = length(x)
  k = length(H)
  cv = rep(0,k)
  nu = rep(0,k)
  gcv = rep(0,k)
  for(i in 1:k){
    tmp = kernel.fitted(x,y,H[i])
    cv[i] = mean(((y - tmp$fitted)/(1-tmp$S))^2)
    nu[i] = sum(tmp$S)
    gcv[i] = mean((y - tmp$fitted)^2)/(1-nu[i]/n)^2
  }
  return(list(cv=cv,gcv=gcv,nu=nu))
}
```

이제 위에서 보았던 bone 자료에 대한 회귀식을 커널로 추정하려고 한다. 이때 CV 또는 GCV를 최소화하는 구간의 크기(bandwidth) h를 찾기위하여 구간 $(0.1, 5)$안에서 20개의 h값에 대한 CV 또는 GCV 값을 계산한다. 아래는 CV 또는 GCV 값을 최소로 하는 h값을 찾는 프로그램이다.

```
par(mfrow=c(1,2))
H = seq(.1,5,length=20)
H
```

```
 [1] 0.1000000 0.3578947 0.6157895 0.8736842 1.1315789 1.3894737 1.6473684
 [8] 1.9052632 2.1631579 2.4210526 2.6789474 2.9368421 3.1947368 3.4526316
[15] 3.7105263 3.9684211 4.2263158 4.4842105 4.7421053 5.0000000
```

```
out = CV(age.female,density.female,H)
plot(H,out$cv,type="l",lwd=3,xlab="Bandwidth",ylab="Cross-validation Score")
lines(H,out$gcv,lty=2,col="red",lwd=3)
plot(out$nu,out$cv,type="l",lwd=3,xlab="Effective Degrees of Freedom",ylab="Cross-validation s
lines(out$nu,out$gcv,lty=2,col="red",lwd=3)
```

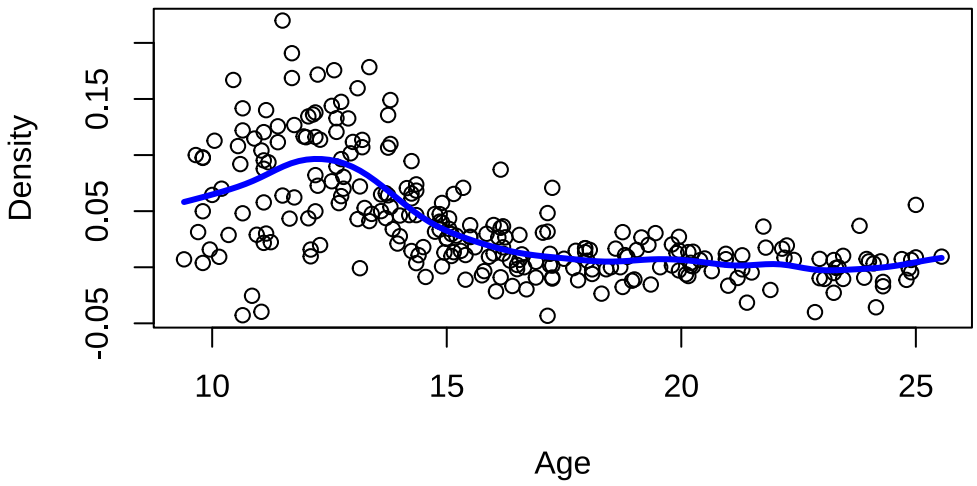



```
j = which.min(out$cv)
h = H[j]
h
```

```
[1] 0.6157895
```

위의 그래프에서 왼쪽은 h값에 대한 CV(검정 점선)과 GCV(빨간 점선)의 값에 대한 그래프이며 CV를 최소로 하는 h의 값은 0.6157895이다. 오른쪽 그래프는 유효 자유도 (Effective Degrees of Freedom) ν에 대한 CV(검정 점선)과 GCV(빨간 점선)의 그래프이다.

```
par(mfrow=c(1,1))
grid = seq(min(age.female),max(age.female),length=100)
m.hat = kernel(age.female,density.female,grid,h)
plot(age.female,density.female,xlab="Age",ylab="Density")
lines(grid,m.hat,lwd=3,col="blue")
```



위의 그래프는 선택된 $h = 0.6157895$을 이용하여 추정한 반응변수 `density` 의 커널회귀 함수 $\hat{m}(x)$ 이다.

10.5.3.3. 붓스트랩 신뢰 구간

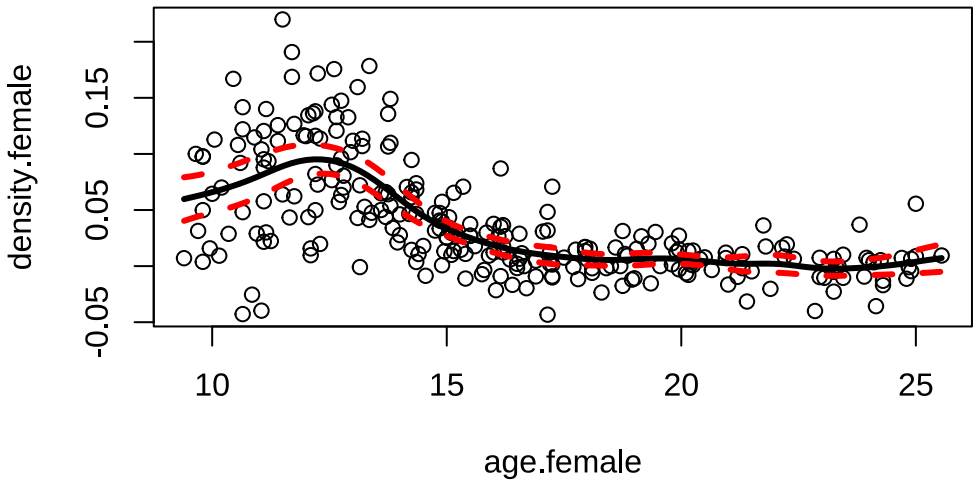
다음 프로그램은 붓스트랩(Bootstrap)을 이용하여 주어진 x값에서 추정량 $\hat{m}(x)$의 표준오차 $SE(x)$를 구하는 프로그램이다.

```
boot = function(x,y,grid,h,B){
  ### pointwise standard error for kernel regression using the bootstrap
  k = length(grid)
  n = length(x)
  M = matrix(0,k,B)
  for(j in 1:B){
    I = sample(1:n,size=n,replace=TRUE)
    xx = x[I]
    yy = y[I]
    M[,j] = kernel(xx,yy,grid,h)
  }
  s = sqrt(apply(M,1,var))
  return(s)
}
```

아래는 추정량 $\hat{m}(x)$의 95% 신뢰구간을 구하는 프로그램이다. 붓스트랩의 반복수는 $B = 1000$이다.

$$CI(x) = [\hat{m}(x) - 2SE(x), \hat{m}(x) + 2SE(x)]$$

```
par(mfrow=c(1,1))
h = .7
grid = seq(min(age.female),max(age.female),length=100)
plot(age.female,density.female)
mhat = kernel(age.female,density.female,grid,h)
lines(grid,mhat,lwd=3)
B = 1000
se = boot(age.female,density.female,grid,h,B)
lines(grid,mhat+2*se,lwd=3,lty=2,col="red")
lines(grid,mhat-2*se,lwd=3,lty=2,col="red")
```



10.6. 스플라인 회귀모형

10.6.1. 함수의 기저

반응변수 Y와 설명변수 X가 다음과 같은 관계를 가진다고 하자.

$$E(Y|X=x)=m(x)$$

이러한 관계를 회귀모형(regression model)이라고 하며 Y의 평균이 X에 따라서 변하는 관계를 설정하는 모형이다. 만약 $m(x)$의 형태를 회귀계수의 선형식으로 나타낼 수 있다면 우리는 이를 선형 회귀모형(linear regression model)이라고 한다.

$$m(x)=a+b_1x_1+b_2x_2+\cdots+b_px_p$$

설명변수는 고정된 값이거나 또는 확률변수일 수도 있다.

비모수 회귀모형(nonparametric regression model)에서는 $m(x)$의 형태에 특별한 제한을 두지 않는다. 따라서 함수 $m(x)$는 무수히 많고 다양한 형태를 가질 수 있다.

이제 설명변수의 수가 1개라고 가정하면 함수 $m(x)$를 가장 단순하게 표현할 수 있는 모형이 1차 회귀모형이다.

$$y = a + bx + e$$

만약 반응변수와 설명변수의 관계가 선형이 아니라 비선형이라면 $m(x)$를 p-차 다항식으로 사용할 수 있다.

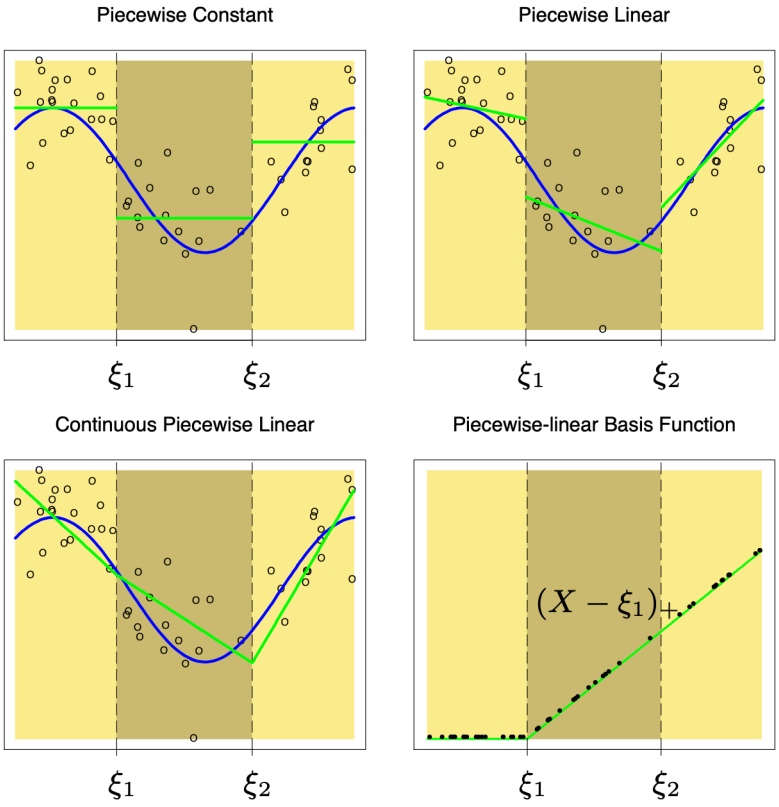
$$m(x) = a + b_1 x + b_2 x^2 + \cdots + b_p x^p \tag{10.9}$$

이렇게 $m(x)$를 다항식으로 표현하는 것은 우리가 알 수 없는 함수를 $p+1$개의 기저들(basis), 즉 1, x, x^2, ..,x^p의 선형조합으로 근사하는 것이며 다항식의 경우는 각 기저 $N_k(x)$ 는 설명변수의 k-차 항 x^k이다.

$$\beta_0 N_0(x) + \beta_1 N_1(x) + \beta_2 N_2(x) + \cdots + \beta_p N_p(x) \tag{10.10}$$

일반적으로 임의의 함수는 다양한 형태의 기저들로 표현할 수 있으며 기저들의 형태에 따라 다음과 같은 것들이 있다.

- 다항식(polynimials)
- 퓨리에 함수(Fourier series)
- 웨이블릿 함수(Wavelet series)



10.6.2. 직선 스플라인

이제 n개의 자료 $(y_1, x_1), (y_2, x_2), \ldots, (y_n, x_n)$이 주어졌다고 하자. 편의상 설명변수 x가 구간 $(0, 1)$의 값이라고 가정한다. 설명변수들이 포함되는 구간 $(0, 1)$를 $K+1$개의 구간 $\{B_j = (\xi_{k-1}, \xi_k) | k = 1, 2, .., K+1\}$ 으로 나누어 보자. 즉

$$\xi_0 = 0 < \xi_1 < \xi_2 < \cdots < \xi_{K-1} < \xi_K < 1 = \xi_{K+1}$$

위의 구간에서 K개의 내부점들 $\xi_1 < \xi_2 < \cdots < \xi_{K-1} < \xi_K$은 일반적으로 knots(연결점)이라고 부른다,

이제 반응변수들의 평균 $E(Y|x)$ 을 단순하게 각 구간에서 상수라고 가정하면

$$E(Y|x) = a_k \quad \text{if } x \in B_k$$

각 구간에 속하는 반응변수들의 평균으로 추정할 수 있다. 이러한 추정법을 우리는 Regressogram이라고 한다.

$$\hat{E}(Y|x) = \frac{1}{n_k} \sum_{i=1}^{n} y_i I_{B_k}(x_i) \quad \text{if } x \in B_k \tag{10.11}$$

이제 Regeressogram의 개념을 확장시켜서 각 구간의 회귀모형을 직선식으로 확장하여 보자. 즉

$$E(Y|x) = a_k + b_k x \quad \text{if } x \in B_k \tag{10.12}$$

모형 10.12 는 각 구간마다 회귀직선을 적합하는 것과 같은 모형이다. 이러한 모형에서 추정된 각 회귀식들은 각 구간의 연결점 $\xi_1, \xi_2, \cdots, \xi_K$ 에서 불연속이다. 각 연결점 ξ_k 에서 연속인 회귀식들을 구할 수 있을까?

이제 구간 $(0,1)$을 다음과 같이 $K+1=3$개의 구간으로 나누어 보자. 이 경우 연결점은 ξ_1과 ξ_2, 두 개가 있다.

$$B_1 = (0, \xi_1) \quad B_2 = (\xi_1, \xi_2) \quad B_3 = (\xi_2, 1)$$

모형 10.12 에서 주어진 회귀계수는 모두 $2(K+1)=6$ 개이며 추정값을 구하는 방법은 다음과 같은 오차제곱합을 구하는 최소제곱법을 사용할 수 있다.

$$SSE = \sum_{x_i \in B_1} (y_i - a_1 - b_1 x_i)^2 + \sum_{x_i \in B_2} (y_i - a_2 - b_2 x_i)^2 + \sum_{x_i \in B_3} (y_i - a_3 - b_3 x_i)^2 \tag{10.13}$$

이제 모형 10.12 에서 구한 직선회귀식이 연결점 ξ_1과 ξ_2에서 연속이 되려면 다음과 같은 두 조건을 만족해야 한다.

$$a_1 + b_1 \xi_1 = a_2 + b_2 \xi_1, \quad a_2 + b_2 \xi_2 = a_3 + b_3 \xi_2 \tag{10.14}$$

각 연결점의 연속을 만족하려면 위의 두 식을 만족해야 하므로 이제 추정해야 하는 모수의 개수는 4개이다. 왜냐하면 원래의 회귀계수의 개수 6 에서 제약식 2개의 개수를 제외해야 하기 때문이다.

제약식 10.14 을 다시 쓰면 다음과 같이 쓸수 있으며

$$\begin{aligned} a_2 &= a_1 + (b_1 - b_2)\xi_1 \\ a_3 &= a_2 + (b_2 - b_3)\xi_2 \\ &= a_1 + (b_1 - b_2)\xi_1 + (b_2 - b_3)\xi_2 \end{aligned}$$

이 제약식을 이용하여 오차제곱합 10.13 에서 두 번째 항과 세 번째 항을 다음과 같이 전개할 수 있다

$$
\begin{aligned}
\sum_{x_i \in B_2} (y_i - a_2 - b_2 x_i)^2 &= \sum_{x_i \in B_2} (y_i - a_1 - (b_1 - b_2)\xi_1 - b_2 x_i)^2 \\
&= \sum_{x_i \in B_2} [y_i - a_1 - b_1 \xi_1 - b_2 (x_i - \xi_1)]^2 \\
&= \sum_{x_i \in B_2} [y_i - a_1 - b_1 x_i - (b_2 - b_1)(x_i - \xi_1)]^2 \\
\sum_{x_i \in B_3} (y_i - a_3 - b_3 x_i)^2 &= \sum_{x_i \in B_3} (y_i - a_1 - (b_1 - b_2)\xi_1 - (b_2 - b_3)\xi_2 - b_3 x_i)^2 \\
&= \sum_{x_i \in B_3} [y_i - a_1 - b_1 \xi_1 + b_2 \xi_1 - b_2 \xi_2 - b_3 (x_i - \xi_2)]^2 \\
&= \sum_{x_i \in B_3} [y_i - a_1 - b_1 x_i + b_1 x_i - b_1 \xi_1 + b_2 \xi_1 - b_2 \xi_2 - b_3 (x_i - \xi_2)]^2 \\
&= \sum_{x_i \in B_3} [y_i - a_1 - b_1 x_i + b_1 (x_i - \xi_1) + b_2 (\xi_1 - x_i + x_i) - b_2 \xi_2 - b_3 (x_i - \xi_2)]^2 \\
&= \sum_{x_i \in B_3} [y_i - a_1 - b_1 x_i - (b_2 - b_1)(x_i - \xi_1) - (b_3 - b_2)(x_i - \xi_2)]^2
\end{aligned}
$$

위의 전개식에서 모수를 다음과 같이 다시 정의하면

$$
\beta_0 = a_1, \quad \beta_1 = b_1, \quad \beta_2 = b_2 - b_2, \quad \beta_3 = b_3 - b_2
$$

오차제곱합 SSE는 다음과 같이 전개할 수 있다.

$$
\begin{aligned}
SSE = &\sum_{x_i \in B_1} (y_i - \beta_0 - \beta_1 x_i)^2 + \sum_{x_i \in B_2} [y_i - \beta_0 - \beta_1 x_i - \beta_3 (x_i - \xi_1)]^2 \\
&+ \sum_{x_i \in B_3} [y_i - \beta_0 - \beta_1 x_i - \beta_2 (x_i - \xi_1) - \beta_3 (x_i - \xi_2)]^2
\end{aligned}
$$

10.15 에서 얻은 오차제곱합은 다음과 같은 회귀모형을 적합한 경우에 얻을 수 있는 오차제곱합이다.

$$
E(y|x) = \beta_0 + \beta_1 x + \beta_2 (x - \xi_1)_+ + \beta_3 (x - \xi_2)_+ \tag{10.15}
$$

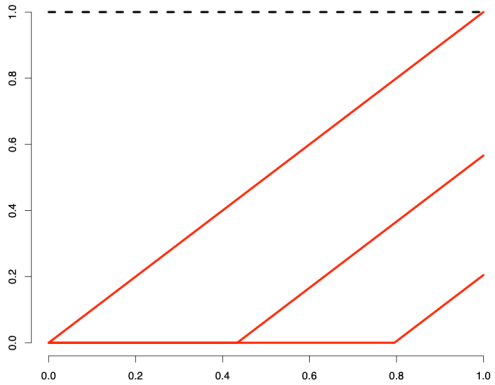
위의 식에서 함수 $(x)_+$ 는 다음과 같이 정의된 함수이다.

$$
(x)_+ = \begin{cases} x & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases}
$$

이제 연결점(knots)에서 연속인 조건을 만족하는 직선들을 추정하는 문제는 10.15 에 주어진 회귀식을 추정하는 문제와 같음을 보였다. 즉, 주어진 구간에서 서로 연결되는 최적의 직선식을 구하는 문제는 함수를 다음과 같은 기저로 조합된 것으로 보고 최적의 계수를 구하는 것과 동일한 문제이다.

$$N_0(x) = 1, \quad N_1(x) = x, \quad N_2(x) = (x - \xi_1)_+, \quad N_3(x) = (x - \xi_2)_+$$

위의 기저의 특징은 일부 기저함수의 값이 주어진 구간의 전 ξ_1 과 ξ_2 에 의존한다는 것이다. 아래 그림은 연결점이 2개인 경우 4개의 기저를 나타내는 그림이다.

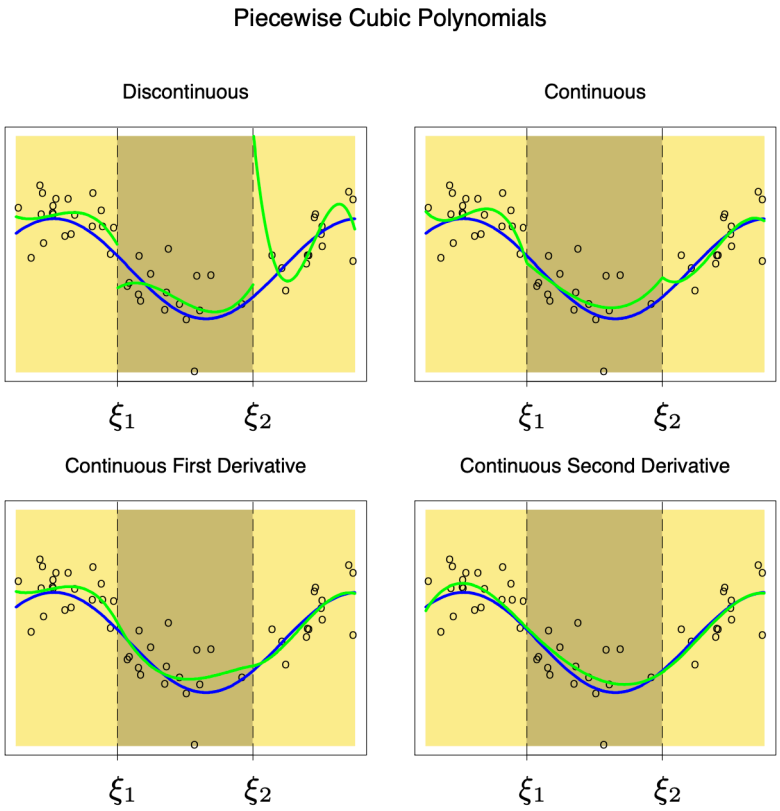


이제 위의 문제에서 만약 $K+1$ 개의 구간이 있다면 원래 회귀계수의 개수 $2(K+1)$ 에서 제약식의 수 K 를 제외한 총 $K+2$ 개의 기저가 필요하며 다음과 같다.

$$N_0(x) = 1, \quad N_1(x) = x, \quad N_k(x) = (x - \xi_k)_+, k = 1, 2, \dots, K$$

10.6.3. 스플라인 회귀

이제 주어진 구간에서 직선식이 아닌 p-차 다항식 10.9 을 고려하자. $K+1$ 개의 구간에서 p-차 다항식이 매우 부드럽게 연결되기 위한 조건은 연결점들에서 연속이며 더 나아가 $1, 2, .., p-1$ 차의 미분값이 동일한 것이다.



이러한 조건을 만족하는 최적의 p-차 다항식을 구하는 문제는 다음과 같은 $K+p+1$개의 기저로 표현된 함수식에서 최적의 함수를 구하는 문제와 같다. 아래의 기저들을 절단된 다항함수(truncated power) 기저라고 부른다.

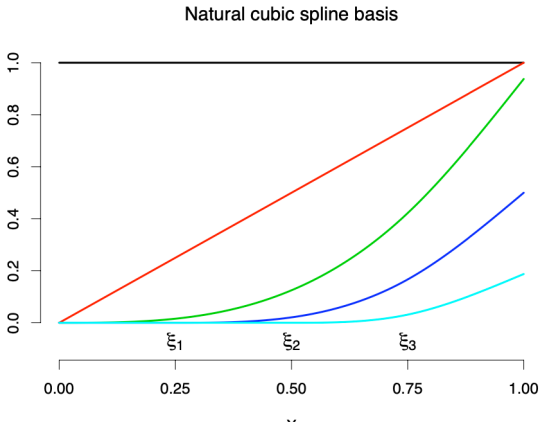
$$N_j(x) = x^j, \quad j = 0, 1, 2, \dots, p$$
$$N_{p+k}(x) = (x - \xi_k)_+^p, \quad k = 1, 2, \dots, K$$

위와 같은 스플라인 회귀에서 가장 자주 사용되는 것은 3차 스플라인 회귀(Cubic spline, $p = 3$)이다.

각 연결점에서 연속이고 1차와 2차 미분값이 같은 조건을 주고 각 구간에서 다항식을 구하는 것이다. 이때 구간의 하한과 상한, 즉 경계점(boundary)에서 직선의 성질을 가지는 조건, 즉

$$m''(0) = m'''(0) = 0, \quad m''(1) = m'''(1) = 0 \tag{10.16}$$

을 만족하는 스플라인을 자연 스플라인(natural spline)이라고 부르며 가장 자주 사용된다.



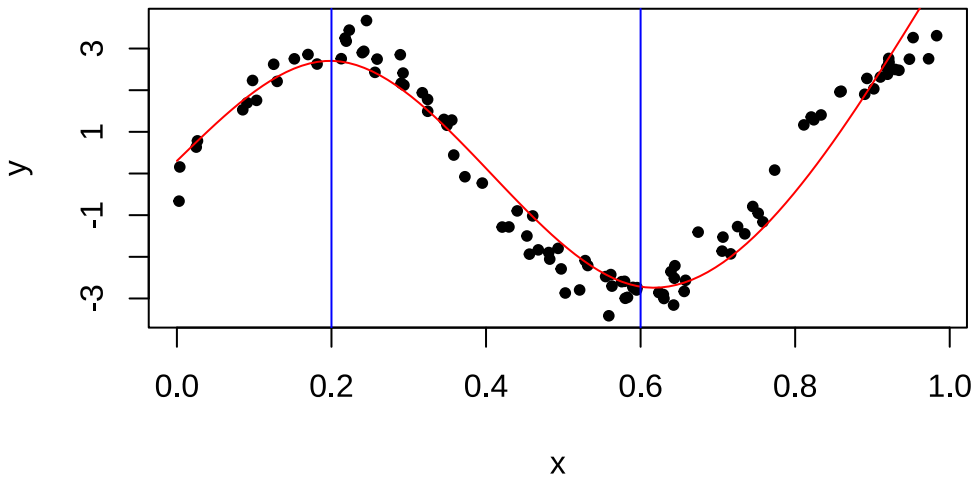
예제로서 다음과 같은 모형에서 생성된 자료를 가지고 자연 스플라인을 적합해 보자.

$$y_i = 3\sin(8x_i) + e_i, \quad e_i \sim N(0, (0.3)^2)$$

다음은 100개의 자료를 가지고 2개의 연결점 $\xi_1 = 0.2$와 $\xi_2 = 0.6$을 이용한 3차 스플라인 함수를 적합하고 그리는 프로그램이다.

스플라인 함수를 사용하려면 `splines` 패키지가 필요하다.

```
library(splines)
n = 100
x = runif(n)
x.grid = seq(0,1,length.out = 100)
y = 3*sin(8*x) + rnorm(n,0,.3)
fit = lm(y ~ ns(x,knots = c(0.2,0.6)) )
plot(x,y,pch=20)
pred  = predict(fit,newdata=list(x=x.grid))
lines(x.grid, pred,col="red")
abline(v=c(0.2,0.6), col = 'blue')
```



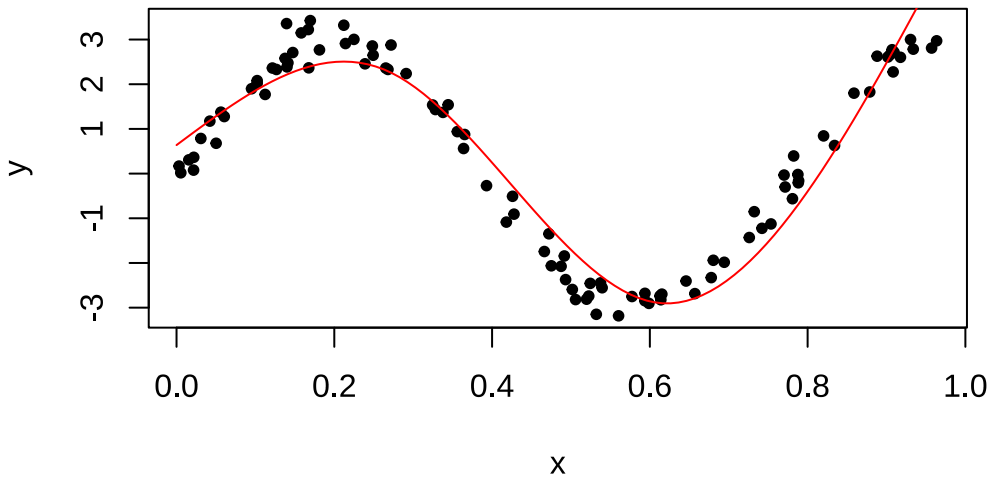
다음 프로그램에서는 연결점을 자동으로 선택하고 차수를 3차(자유도=3)로 사용하는 방법이다.

```
x = runif(n)
y = 3*sin(8*x) + rnorm(n,0,.3)
fit = lm(y ~ ns(x,df=3) )
summary(fit)
```

```
Call:
lm(formula = y ~ ns(x, df = 3))

Residuals:
     Min       1Q   Median       3Q      Max
-1.55542 -0.39760  0.03705  0.38510  1.22418

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)      0.6797     0.1824   3.726 0.000328 ***
ns(x, df = 3)1  -8.5991     0.2437 -35.279  < 2e-16 ***
ns(x, df = 3)2   3.2399     0.4682   6.919 5.08e-10 ***
ns(x, df = 3)3   1.7166     0.1908   8.995 2.15e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5706 on 96 degrees of freedom
Multiple R-squared:  0.9306,    Adjusted R-squared:  0.9284
```

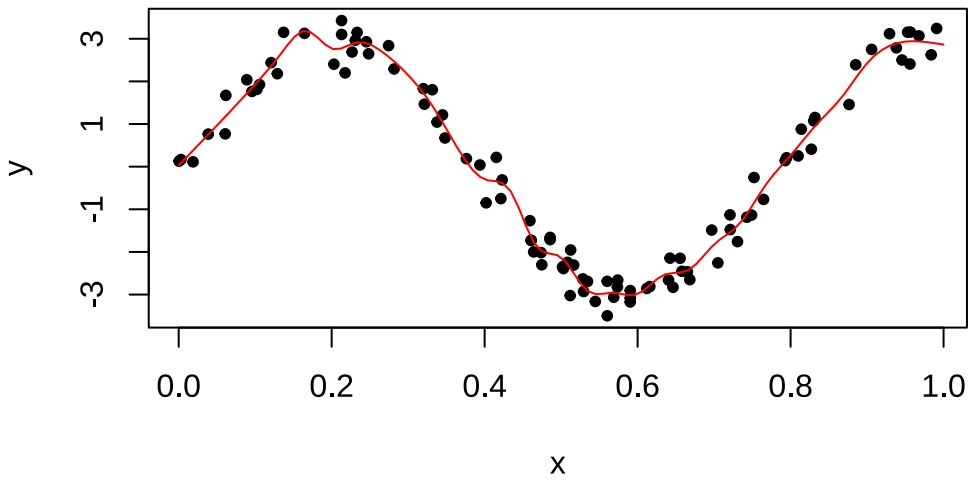
```
F-statistic: 429.1 on 3 and 96 DF,  p-value: < 2.2e-16
```

```
plot(x,y,pch=20)
pred  = predict(fit,newdata=list(x=x.grid))
lines(x.grid, pred,col="red")
```



다음 프로그램에서는 25개의 연결점을 선택하고 차수를 3차(자유도=3)로 사용하는 방법이다. 과적합(overfitting)이 발생하였다.

```
kknots=seq(0+0.1,1-0.1,len=25)
x = runif(n)
y = 3*sin(8*x) + rnorm(n,0,.3)
fit = lm(y ~ ns(x,df=3,knots = kknots ))
plot(x,y,pch=20)
pred  = predict(fit,newdata=list(x=x.grid))
lines(x.grid, pred,col="red")
```



10.6.4. 평활 스플라인

스플라인 회귀에서 추정된 평균 함수 $\hat{m}(x)$의 부드러움(smoothness)는 연결점의 수 K와 기저함수의 차수 p로 결정된다. 일반적으로 기저함수의 차수는 3차를 사용하는 자연 3차 스플라인 회귀식을 사용하며 더 높은 고차식을 사용해도 큰 차이가 나지 않는다. 따라서 추정된 평균함수의 부드러움을 결정하는 주요한 요인는 연결점의 수이다.

연결점의 수가 너무 적으면 너무 부드러워서 전체경향을 파악하기 힘들고(large bias, small variance) 반면에 너무 많으면 과적합(over fitting, small bias, large variance)이 된다.

평활 스플라인(smoothing sapline)은 자료의 모든 점을 연결점으로 하면서 동시에 부드러움을 조절하는 방법이다. 이제 $E(y|x)$를 추정하는 함수 $f(x)$에 대한 벌칙항을 포함한 오차제곱합(reisgual sum of squares with penalty term)을 고려하자.

$$SSE(f,\lambda) = \sum_{i=1}^{n}(y_i - f(x_i))^2 + \lambda \int [f''(x)]^2 dx \tag{10.17}$$

위의 식에서 $f''(x)$는 함수 $f(x)$의 이차 미분이다. 따라서 고려하는 함수 $f(x)$는 두 번 미분이 가능한 함수이다. 벌칙항을 포함한 오차제곱합에서 첫번째 항은 자료에 얼마나 가까운지를 측정하는 양이며 두 번째 항은 함수의 부드러운 정도를 측정하는 양이다.

10.17 에서 $\lambda \in (0,\infty)$는 평활 모수(smoothing parameter)라고 부르며 스플라인 추정 함수의 부드러움을 조절해 주는 모수이다.

- $\lambda = 0$: 이 경우는 10.17 가 일반적인 오차제곱함과 같다. 따라서 함수 $f(x)$는 제약이 없다면 각 자료의 점들을 이어주는 아주 매우 거칠은 모양의 함수를 추정치로 얻게 된다.
- $\lambda = \infty$: 이 경우는 $f(x)$가 가장 단순한(부드러운) 직선식으로 주어진다. 왜냐하면 직선식의 이차미분은 0 이기 때문이다.

벌칙항을 포함한 오차제곱합 10.17 를 최소로 하는 함수는 자료의 점 (x_i, y_i)를 지나는, 즉 n개의 연결점을 자기는 자연 3차 스플라인이다. 따라서 모수는 모두 n개이며 평활 모수의 값에 따라 계수들이 변한다.

이제 10.17 를 최소로 하는 함수를 다음과 같이 자연 3차 스플라인의 기저 $N_i(x), i = 0, 1, 2, \dots, n-1$와 대응하는 계수 β_i로 다음과 같이 나타내어 보자.

$$y_i = \sum_{i=0}^{n-1} \beta_i N_i(x)$$

그러면 벌칙항을 포함한 오차제곱합 10.17 를 다음과 같이 벡터식으로 나타낼 수 있다.

$$SSE(f, \lambda) = (\boldsymbol{y} - \boldsymbol{N\beta})^t(\boldsymbol{y} - \boldsymbol{N\beta})) + \lambda \boldsymbol{\beta}^t \boldsymbol{\Omega}_n \boldsymbol{\beta} \tag{10.18}$$

위의 식에서 $\boldsymbol{y}$, $\boldsymbol{N}$, $\boldsymbol{\beta}$, $\boldsymbol{\Omega}$는 다음과 같이 정의된다.

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad \boldsymbol{N} = \begin{bmatrix} N_0(x_1) & N_1(x_1) & \cdots & N_{n-1}(x_1) \\ N_0(x_2) & N_1(x_2) & \cdots & N_{n-1}(x_2) \\ \vdots & \vdots & \cdots & \vdots \\ N_0(x_n) & N_1(x_n) & \cdots & N_{n-1}(x_n) \end{bmatrix} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{n-1} \end{bmatrix}$$

$$\boldsymbol{\Omega} = \{\omega_{ij}\} \text{ where } \omega_{ij} = \int N_i^{''}(t) N_j^{''}(t) dt$$

벌칙항이 있는 오차제곱합 10.18 를 최소화하는 추정량 $\hat{\boldsymbol{\beta}}$는 다음과 같이 주어진다.

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{N}^t \boldsymbol{N} + \lambda \boldsymbol{\Omega})^{-1} \boldsymbol{N}^t \boldsymbol{y}$$

그리고 반응변수들에 대한 예측치 $\hat{\boldsymbol{y}}$는 다음과 같이 반응변수 벡터 $\boldsymbol{y}$의 선형식으로 나타난다.

$$\hat{\boldsymbol{y}} = \boldsymbol{N}(\boldsymbol{N}^t \boldsymbol{N} + \lambda \boldsymbol{\Omega})^{-1} \boldsymbol{N}^t \boldsymbol{y} = \boldsymbol{S}_\lambda \boldsymbol{y}$$

위의 식에서 $\boldsymbol{S}_\lambda$는 $n \times n$ 의 행렬이며 평활행렬(smoothing matrix)라고 부른다.

또한 평활행렬 $\boldsymbol{S}_\lambda$의 대각원소의 합 df_λ 을 유효 자유도(effective degrees of freedom)이라고 부른다.

$$df_\lambda = trace(\boldsymbol{S}_\lambda)$$

유효 자유도는 일반 선형모형에서 나타나는 독립변수의 개수를 확장한 개념이다. 다음과 같은 일반선형모형에서

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{e}$$

계획행렬 $\boldsymbol{X}$의 열의 개수는 상수항을 포함한 독립변수의 개수 $p+1$이다. 이 때 반응변수에 대한 예측은 반응변수 벡터 $\boldsymbol{y}$의 선형변환으로 나타나며

$$\hat{\boldsymbol{y}} = \boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{y} = \boldsymbol{H}\boldsymbol{y}$$

이때 사영행렬(projection matrix) $\boldsymbol{H}$의 대각합은 $p+1$이고 이는 반응변수의 평균을 추정할 때 필요한 자료의 크기이며 이를 자유도라고 한다.

$$trace(\boldsymbol{H}) = trace[\boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t] = trace[(\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{X}] = trace(\boldsymbol{I}_{p+1}) = p+1$$

참고로 일반 선형모형에서 오차의 분산 σ^2를 추정할 때 필요한 자유도는 전체 자료의 개수 n에서 반응변수의 평균을 추정할 때 필요한 자료의 개수 $p+1$를 뺀 $n-p-1$이다.

스플라인 회귀에서 정의된 유효 자유도는 일반 선형모형에서 사용된 사영행렬 $\boldsymbol{H}$ 의 대각합을 일반화한 개념이다. 일반 선형모형에서는 자유도, 즉 사영행렬의 대각합은 상수항을 포함한 독립변수의 개수 $p+1$이 된다. 스플라인 회귀와 같은 비모수회귀에서는 평활행렬 $\boldsymbol{S}_\lambda$ 의 대각합을 유효 자유도라고 일반화한 것이다. 따라서 유효 자유도는 정확하게 독립변수의 개수를 나타내는 것이 아니지만 반응변수의 평균을 추정할 때 사용되는 모수(계수)의 개수로 해석할 수 있다.

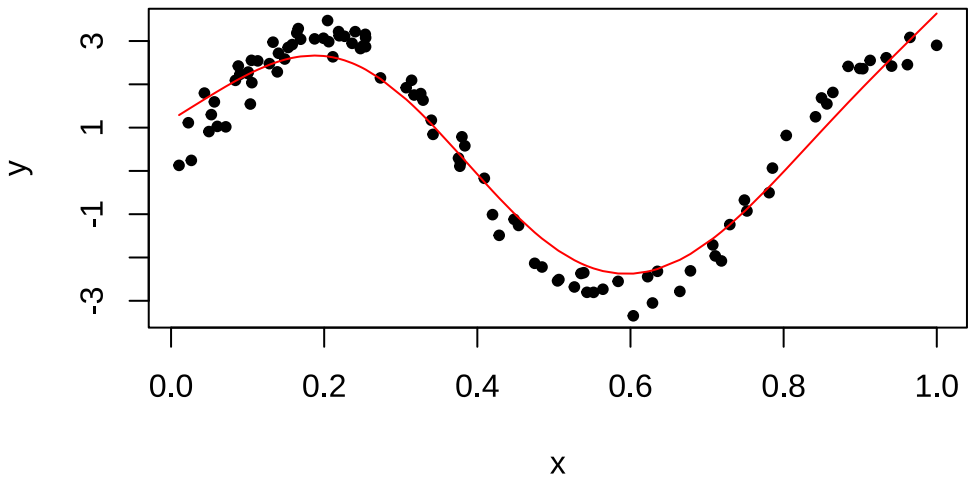
예를 들어 3차 스플라인(cubic spline)을 고려하고 연결점을 K개를 사용하면 총 $K+3+1 = K+4$개의 기저가 필요하다. 더 나아가 자연 3차 스플라인은 10.16 에서 주어진 4개의 제약조건때문에 전체적으로 K개의 모르는 모수가 필요하다고 할 수 있다. 따라서 자연 3차 스플라인의 유효 자유도는 K개이다.

다음 프로그램은 R 에서 평활 스플라인을 적합하는 방법이다. 유효 자유도를 5로 사용한 예이다.

```
n = 100
x = runif(n)
y = 3*sin(8*x) + rnorm(n,0,.3)
fitsm5 = smooth.spline(y ~ x, all.knots=T, cv=T, df=5)
fitsm5
```

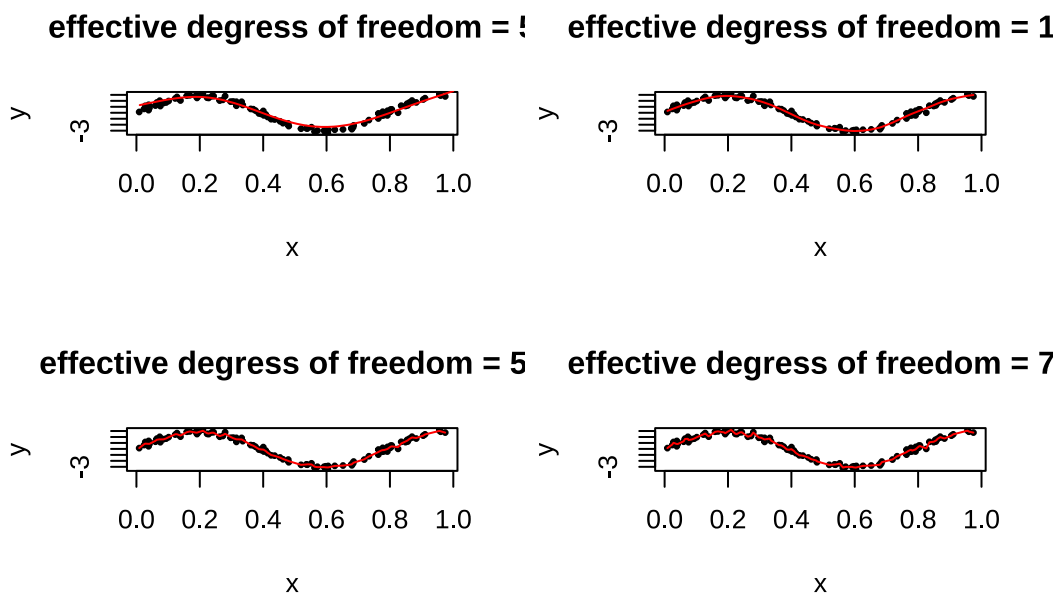
```
Call:
smooth.spline(x = y ~ x, df = 5, cv = T, all.knots = T)

Smoothing Parameter  spar= 1.157577  lambda= 0.005691548 (16 iterations)
Equivalent Degrees of Freedom (Df): 4.999385
Penalized Criterion (RSS): 25.27925
PRESS(l.o.o. CV): 0.2858942
```

```
plot(x,y,pch=20)
lines(fitsm5,col="red")
```



유효 자유도를 증가시키면 추정된 평활곡선은 다음과 같이 나타난다.

```
par(mfrow = c(2,2))
n = 100
x = runif(n)
y = 3*sin(8*x) + rnorm(n,0,.3)
fitsm15= smooth.spline(y ~ x, all.knots=T, cv=T, df=15)
fitsm50= smooth.spline(y ~ x, all.knots=T, cv=T, df=50)
fitsm70= smooth.spline(y ~ x, all.knots=T, cv=T, df=70)
plot(x,y,pch=20, cex=0.6)
title("effective degress of freedom = 5")
lines(fitsm5,col="red")
plot(x,y,pch=20, cex=0.6)
title("effective degress of freedom = 15")
lines(fitsm15,col="red")
plot(x,y,pch=20, cex=0.6)
title("effective degress of freedom = 50")
lines(fitsm50,col="red")
plot(x,y,pch=20, cex=0.6)
title("effective degress of freedom = 70")
lines(fitsm70,col="red")
```

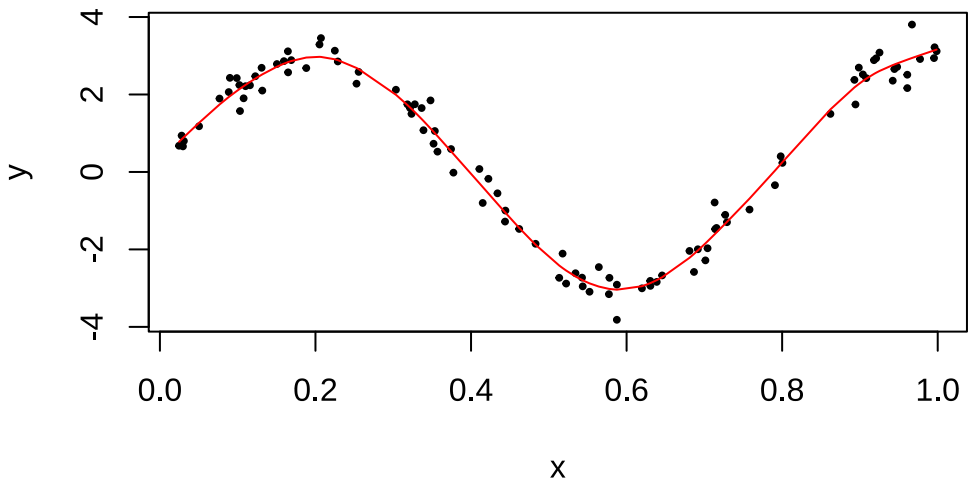


만약 유효 자유도를 지정해주지 않으며 CV를 최소화해주는 자유도를 구하여 자동으로 스플라인함수를 추정해준다. 아래에서 구해진 평활모수는 $\lambda = 0.000122$ 이고 자유도는 $df = 11.34031$ 이다.

```
par(mfrow = c(1,1))
n = 100
x = runif(n)
y = 3*sin(8*x) + rnorm(n,0,.3)
fitsmauto= smooth.spline(y ~ x, all.knots=T, cv=T)
fitsmauto
```

```
Call:
smooth.spline(x = y ~ x, cv = T, all.knots = T)

Smoothing Parameter  spar= 0.8712038  lambda= 0.0001931802 (15 iterations)
Equivalent Degrees of Freedom (Df): 10.32359
Penalized Criterion (RSS): 9.280936
PRESS(l.o.o. CV): 0.1129098
```

```
plot(x,y,pch=20, cex=0.6)
lines(fitsmauto,col="red")
```


y
4
2
0
-2
-4
0.0
0.2
0.4
0.6
0.8
1.0
x

11. 일반화 선형모형

> **Caution**
>
> 이 장에서는 벡터와 스칼라의 표기를 구분없이 모두 보통체로 사용한다.

11.1. 일반화 선형모형

선형모형의 의미는 모형에서 고려하는 설명변수가 변할 때 반응값의 평균이 변하는 관계가 선형이라는 것이다. 즉, 반응값의 평균을 설명하는 회귀식이 회귀계수에 대하여 선형이라는 의미이다.

$$E(y|x_1, x_2, \dots, x_p) = \beta_1 x_1 + \cdots + \beta_p x_p \equiv \eta \tag{11.1}$$

참고로 11.1 의 오른쪽에 나타나는 식을 선형예측식(linear predictor, η)라고 부른다.

이러한 평균의 선형성의 가정이 적절하지 않은 경우가 있다. 예를 들어 공학이나 생물학에서 사용되는 비선형 회귀모형(nonlinear regression model)처럼 반응변수 평균의 변화가 설명변수들의 복잡한 비선형 관계(예를 들어 미분방정식의 관계)로 나타나는 경우로 흔히 나타난다. 이러한 비선형 회귀모형은 Section 10.4 장에서 다루었다.

반응변수가 가질 수 있는 평균값의 범위에 제한이 있을 수 있다. 예를 들어 베르누이 분포의 경우 평균이 성공확률이기 때문에 0과 1사이에 있으며 포아송 분포의 경우 반응값은 음의 값을 가질 수 없다. 따라서 11.1 의 선형예측식 η와 반응값의 평균 $E(y|x)$의 관계를 선형모형 11.1 처럼 정의할 수 없다.

이렇게 반응변수의 평균과 선형에측식의 범위가 일치하지 않는 경우 임의의 단조증가 함수 g를 사용하여 그 범위를 일치하게 만들어 줄 수 있다. 예를 들어 베르누이 분포의 경우 표준 정규분포의 누적분포함수 Φ를 사용하여 확률의 범위와 선형에측식의 범위를 맞추어 줄 수 있다. 이러한 회귀모형을 프로빗(probit)모형이라고 부른다.

$$\Phi^{-1}[p(y|x)] = \beta_1 x_1 + \cdots + \beta_p x_p = \eta \tag{11.2}$$

이제부터 정규분포하에서 정의되는 선형모형을 다른 분포들로 확장한 모형인 일반화 선형모형(Generalized Linear Model; GLM)을 살펴보기로 하자.

11.1.1. 지수군 분포와 일반화 선형모형

지수군 분포에 대한 자세한 내용은 Appendix H 에 있으니 참고하자.

일변량 확률변수 지수군(exponential family) 분포를 따른다고 가정하자.

$$f(y|\theta,\phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\}$$

충분통계량이 관측값 y 이고 1차원의 기본형 모수 θ 로 정의된 분포임을 유의하자.

확률변수 y의 평균을 $\mu = E(y)$ 이라고 하고 독립변수 벡터 x와 회귀계수 벡터 β로 구성된 선형예측식을 η라고 하자.

$$\eta = x^t\beta \tag{11.3}$$

일반화 선형모형은 분포의 특성에 따라 주어진 단조증가함수 g 를 이용하여 y 의 평균과 선형예측식의 관계를 설정하는 모형이다. 이러한 함수 g 를 연결함수(link function)라고 부른다.

$$g(\mu) = g(E[y|x]) = x^t\beta \tag{11.4}$$

반응값의 분포가 주어진 경우 연결함수는 평균의 범위와 선형예측식의 범위를 연속적으로 1-1 대응하게 해주는 함수이며 사용할 수 있는 가능한 함수는 무한히 많다. 예를 들어 베르누이 분포의 경우 위에서 정의된 프로빗 모형 11.2 에서 Φ^{-1} 같이 (0,1) 에서 실수전체 집합으로 단조증가하는 함수는 모두 연결함수로 고려할 수 있다.

지수군 분포에서 모수 θ를 기본형모수(canonical parameter)라고 부르며 일반적으로 θ 는 평균 μ의 비선형 함수로 나타난다. 만약 다음과 같은 관계를 나타내는 연결함수 g 가 있다면 그 함수를 기본형 연결함수(canonical link function) 또는 자연결합함수(natural link function)라고 부른다.

$$\theta = \eta \tag{11.5}$$

예를 들어 이항분포의 확률밀도함수를 지수군 분포의 형태로 표현했을 때 그 형태를 보면 다음과 같은 관계를 알 수 있다.

$$\theta = \log\frac{\mu}{1-\mu} = \log\frac{p}{1-p}$$

따라서 11.5 을 만족하는 연결함수는 다음과 같고

$$\log\frac{p}{1-p} = \eta = x^t\beta \tag{11.6}$$

이를 로짓 연결함수(logit link function)이라고 부르며 이는 이항분포의 기본형 연결함수이다.

> **Note**
>
> 만약 y의 분포가 정규분포이며 연결함수 g가 $g(\mu)=\mu$이면 선형회귀모형이 된다.
>
> $$E[y|x]=x^t\beta$$

11.2. 일반화 선형모형의 가능도함수

하나의 표본 y에 대하여 기본형 모수 θ 하나인 로그가능도함수(log likelihood function) ℓ은 다음과 같이 정의된다.

$$\ell=\log f(y)=\frac{y\theta-b(\theta)}{a(\phi)}+\log c(y,\phi)$$

표본 $y_1,y_2,\dots,y_n$ 가 각각 설명변수 벡터$x_1,x_2,\dots,x_n$에서 독립적으로 얻어졌다면 로그가능도함수 ℓ_n 은 다음과 같다.

$$\ell_n=\log\prod_{i=1}^n f(y_i)=\sum_{i=1}^n\frac{y_i\theta_i-b(\theta_i)}{a(\phi_i)}+\sum_{i=1}^n\log c(y_i,\phi) \tag{11.7}$$

여기서 θ_i는 i 번째 관측값에 대한 모수로서 첨자 i 를 붙이는 이유는 관측치의 기대값 $\mu_i=E(y_i|x_i)$가 독립변수의 값 x_i에 따라 다를 수 있고 θ_i는 평균 μ_i의 함수이기 때문이다.

이제 11.4 과 같이 설명변수와 반응변수 평균과의 관계가 연결함수 g로 정의되었다고 하자.

$$g(\mu_i)=g(E[y_i|x_i])=x_i^t\beta\equiv\eta_i,\quad i=1,2,\dots,n \tag{11.8}$$

Example 11.1 (이항분포)**.** 주어진 예측변수 $\boldsymbol{x}_i=(x_{1i},\dots,x_{pi})^t$에서 실행횟수가 m_i인 이항분포$B(m_i,p(x_i))$를 생각하자. m_i의 시행 중에 성공의 횟수가 y_i라고 하면 y_i의 평균과 분산은 다음과 같다.

$$E(y_i|x_i)=m_ip(x_i),\qquad Var(y_i|x_i)=m_ip(x_i)(1-p(x_i)),\quad i=1,2,\dots,n$$

이항분포를 위한 로지스틱 회귀방정식은 선형예측식과 성공의 확률의 관계를 다음과 같이 정한다.

$$\log\left[\frac{p(x_i)}{1-p(x_i)}\right]=\boldsymbol{x}_i^t\boldsymbol{\beta}=\beta_0+\beta_1x_{1i}+\dots\beta_px_{pi} \tag{11.9}$$

서로 독립인 관측값 $(y_1,y_2,\dots,y_n)$의 가능도함수(likelihood function) L은 이항분포들의 결합확률밀도함수와 같고 아래와 같이 주어지며

$$L=\prod_{i=1}^n f(y_i|p(x_i))=\prod_{i=1}^n\left[\binom{m_i}{y_i}\right]p(x_i)_i^y(1-p(x_i))^{m-y_i}$$

로그가능도함수(log likelihood function) l 은 다음과 같이 나타낼 수 있다.

$$\begin{aligned} l = \log L &= \sum_i \log \binom{m_i}{y_i} + \sum_i y_i \log p(x_i) + \sum_i (m_i - y_i) \log(1 - p(x_i)) \\ &= c(\boldsymbol{y}, \boldsymbol{m}) + \sum_i y_i \log \left[\frac{p(x_i)}{1 - p(x_i)} \right] + \sum_i m_i \log(1 - p(x_i)) \end{aligned}$$

위의 로그가능도함수에서 볼 수 있듯이 충분통계량인 성공의 횟수 y_i와 곱으로 나타내어진 함수가 로짓함수이며 이렇게 가능도함수에서 얻어진 결합함수를 기본형 연결함수라고 한다.

> **Note**
>
> Nelder and Wedderburn (1972) 에서 연결 함수(link function)의 개념이 제시할 때 다음과 같은 작업 변량(working variate) z_i를 이용하여 선형모형을 일반화하고자 하였다. 작업 변량 z_i 은 다음과 같이 정의된다.
>
> $$\begin{aligned} z_i &= g(y_i) \\ &\simeq g(\mu_i) + g_\mu(\mu_i)(y_i - \mu_i) \\ &= x_i^t \beta + g_\mu(\mu_i)(y_i - \mu_i) \\ &\simeq x_i^t \beta + r_i \end{aligned}$$
>
> 위의 식에서 g_μ 은 연결함수 g 를 미분한 함수이다. 위의 식에서 오른쪽 식의 두번째 항의 기대값이 0 이므로
>
> $$E(y_i - \mu_i) = 0$$
>
> $y_i - \mu_i$ 를 오차항과 같이 생각하면 위의 모형을 오차항의 분산이 다른 선형모형으로 생각할 수 있다. 지수군 분포의 성질을 이용하면 작업 변량 z_i의 분산은 다음과 같다.
>
> $$Var(z_i) = Var(r_i) = [g_\mu(\mu_i)]^2 Var(y_i) = [g_\mu(\mu_i)]^2 [a(\phi_i) v(\mu_i)] \tag{11.10}$$
>
> 이러한 가정하에서 작업 변량 z_i를 반응변수로 놓고 분산의 역수를 가중치로하는 가중 선형모형(wighted linear regression) 을 최소제곱법으로 적합하고 계수의 값이 수렴할 때까지 반복적으로 수행하는하는 계산법을 제공하였다. 이러한 방법을 반복가중최소최곱법(iterative weighted least square method; IWLS) 라고 부른다.
> (Searle and McCulloch 2001 의 136 쪽 참조)

11.3. 최대가능도추정

이제 회귀계수 β를 최대가능도추정법(Maximum Likelihood Estimation)으로 구하기 위하여 로그가능도함수 11.7 를 회귀계수 벡터 β로 미분한 가능도함수 방정식을 고려하자.

$$\frac{\partial \ell_n}{\partial \beta} = 0 \tag{11.11}$$

! Important

여기서 일반화 선형모형에서 나타나는 모수들 β, μ_i, θ_i의 관계를 살펴보자.

1. 회귀계수 벡터 β는 설명변수 벡터 x_i와 내적 형태로 연결되어 있으며 이를 선형 예측식이라고 한다.

$$\eta_i = x_i^t \beta$$

2. 선형 예측식 η_i는 관측값의 평균 μ_i와 연결함수 g로 연결되어 있다.

$$g(\mu_i) = \eta_i$$

3. 관측값의 평균 μ_i는 기본형 모수 θ_i와 함수 b로 연결되어 있다.

$$b'(\theta_i) = \mu_i$$

일반화 선형모형에서 최종적으로 추정해야 하는 모수는 회귀계수 벡터 β 이며 모수 β, μ_i, θ_i 다음과 연결되어 있음을 알 수 있다.

$$\beta \underset{g}{\longrightarrow} \mu_i \underset{b}{\longrightarrow} \theta_i \tag{11.12}$$

최대가능도 추정량을 구하는 방정식을 유도할 때 다음과 같은 일반화 선형모형의 지수군 분포에서 나타나는 미분 공식이 적용된다.

1. **선형 예측식의 미분**:

$$\frac{\partial \eta}{\partial \beta} = \frac{\partial x^t \beta}{\partial \beta} = x$$

2. **연결함수의 역함수에 대한 미분**: $g(\mu) = \eta$ 의 관계를 이용하면

$$\frac{\partial \mu}{\partial \eta} = \frac{\partial \mu}{\partial g(\mu)} = \left[\frac{\partial g(\mu)}{\partial \mu}\right]^{-1} = [g'(\mu)]^{-1} = g_\mu^{-1}(\mu)$$

여기서 $g_\mu(\mu) = g'(\mu)$로서 연결함수의 미분을 나타내는 기호이다.

3. **평균과 기본형모수의 미분, 분산함수**: 평균과 분산의 관계식을 이용하면

$$\frac{\partial \theta}{\partial \mu} = \left[\frac{\partial \mu}{\partial \theta}\right]^{-1} = [b''(\theta)]^{-1} = \frac{1}{v(\mu)}$$

11.3.1. 가능도 방정식의 유도: 첫 번째 방법

이제 가능도 함수 11.7 의 형태를 이용하여 방정식 11.11 를 유도해 보자.

$$
\begin{aligned}
0 &= \frac{\partial \ell_n}{\partial \beta} \\
&= \sum_{i=1}^{n} \left[\frac{1}{a(\psi_i)} \right] \left[y_i \frac{\partial \theta_i}{\partial \beta} - \frac{\partial b(\theta_i)}{\partial \beta} \right] \\
&= \sum_{i=1}^{n} \left[\frac{1}{a(\psi_i)} \right] \left[y_i \frac{\partial \theta_i}{\partial \beta} - \frac{\partial \theta_i}{\partial \beta} \frac{\partial b(\theta_i)}{\partial \theta_i} \right] \\
&= \sum_{i=1}^{n} \left[\frac{1}{a(\psi_i)} \right] \left[\frac{\partial \theta_i}{\partial \beta} (y_i - \mu_i) \right] \\
&= \sum_{i=1}^{n} \left[\frac{1}{a(\psi_i)} \right] \left[\frac{\partial \eta_i}{\partial \beta} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \theta_i}{\partial \mu_i} (y_i - \mu_i) \right] \\
&= \sum_{i=1}^{n} \left[\frac{1}{a(\psi_i)} \right] \left[x_i \frac{(y_i - \mu_i)}{v(\mu_i) g_\mu(\mu_i)} \right] \\
&= \sum_{i=1}^{n} \left[x_i w_i g_\mu(\mu_i)(y_i - \mu_i) \right] \\
&= \begin{bmatrix} \sum_{i=1}^{n} x_{i1} w_i g_\mu(\mu_i)(y_i - \mu_i) \\ \sum_{i=1}^{n} x_{i2} w_i g_\mu(\mu_i)(y_i - \mu_i) \\ \vdots \\ \sum_{i=1}^{n} x_{ip} w_i g_\mu(\mu_i)(y_i - \mu_i) \end{bmatrix}
\end{aligned}
\tag{11.13}
$$

여기서 가중치 w_i는 다음과 정의한다. 가중치 w_i는 앞에서 설명한 작업 변량의 분산의 역수와 동일하다. 11.10 을 참조하자.

$$
w_i \equiv \frac{1}{g_\mu^2(\mu_i) a(\phi_i) v(\mu_i)}
$$

11.3.2. 가능도 방정식의 유도: 두 번째 방법

위의 방정식 11.11 에 미분의 연쇄법칙(chain rule)을 적용하면 다음과 같은 방정식을 얻는다.

$$
\frac{\partial \ell_n}{\partial \beta} = \frac{\partial \eta}{\partial \beta} \frac{\partial \mu}{\partial \eta} \frac{\partial \theta}{\partial \mu} \frac{\partial \ell_n}{\partial \theta} = 0 \tag{11.14}
$$

위의 식에서 η, μ, θ는 다음과 같이 n개의 대응되는 원소로 이루어진 벡터이다.

$$
\eta = \begin{bmatrix} \eta_1 \\ \eta_2 \\ \vdots \\ \eta_n \end{bmatrix}, \quad \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{bmatrix}, \quad \theta = \begin{bmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_n \end{bmatrix}
$$

이제 11.14 에서 나타난 도함수들을 각각 구해보자. 먼저

$$\begin{aligned}\frac{\partial \eta}{\partial \beta} &= \begin{bmatrix} \frac{\partial \eta_1}{\partial \beta} & \frac{\partial \eta_2}{\partial \beta} & \cdots & \frac{\partial \eta_n}{\partial \beta} \end{bmatrix} \\ &= \begin{bmatrix} \frac{\partial x_1^t \beta}{\partial \beta} & \frac{\partial x_2^t \beta}{\partial \beta} & \cdots & \frac{\partial x_n^t \beta}{\partial \beta} \end{bmatrix} \\ &= [x_1 \;\; x_2 \;\; \ldots \;\; x_n] \\ &= X^t \end{aligned}$$

위의 식에서 행렬 X는 $n \times p$ 계획행렬이다. 또한

$$\begin{aligned}\frac{\partial \mu}{\partial \eta} &= \begin{bmatrix} \frac{\partial \mu_1}{\partial \eta} & \frac{\partial \mu_2}{\partial \eta} & \cdots & \frac{\partial \mu_n}{\partial \eta} \end{bmatrix} \\ &= \begin{bmatrix} \frac{\partial g^{-1}(\eta_1)}{\partial \eta} & \frac{\partial g^{-1}(\eta_2)}{\partial \eta} & \cdots & \frac{\partial g^{-1}(\eta_2)}{\partial \eta} \end{bmatrix} \\ &= \begin{bmatrix} \frac{1}{g'(\mu_1)} & 0 & \ldots & 0 \\ 0 & \frac{1}{g'(\mu_2)} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \frac{1}{g'(\mu_n)} \end{bmatrix} \\ &= \begin{bmatrix} g_\mu^{-1}(\mu_1) & & & \\ & g_\mu^{-1}(\mu_2) & & \\ & & \ddots & \\ & & & g_\mu^{-1}(\mu_n) \end{bmatrix}\end{aligned}$$

위에서 $\frac{\partial \mu}{\partial \eta}$는 n-차원 대각행렬이며 $g_\mu(\mu) = g'(\mu)$로서 연결함수 g를 1차 미분한 함수이다.

또한 다음과 같은 결과를 얻는다.

$$\begin{aligned}\frac{\partial \theta}{\partial \mu} &= \begin{bmatrix} \frac{\partial \theta_1}{\partial \mu} & \frac{\partial \theta_2}{\partial \mu} & \cdots & \frac{\partial \theta_n}{\partial \mu} \end{bmatrix} \\ &= \begin{bmatrix} \frac{1}{b''(\theta_1)} & 0 & \ldots & 0 \\ 0 & \frac{1}{b''(\theta_2)} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \frac{1}{b''(\theta_n)} \end{bmatrix} \\ &= \begin{bmatrix} v^{-1}(\mu_1) & & & \\ & v^{-1}(\mu_2) & & \\ & & \ddots & \\ & & & v^{-1}(\mu_n) \end{bmatrix}\end{aligned}$$

마지막으로 11.7 를 이용하면 로그가능도함수 ℓ_n 을 θ로 미분한 n-차원 벡터는 다음과 같이 얻어진다.

$$\frac{\partial \ell_n}{\partial \theta} = \begin{bmatrix} \frac{y_1 - b'(\theta_1)}{a(\phi_1)} \\ \frac{y_2 - b'(\theta_2)}{a(\phi_2)} \\ \vdots \\ \frac{y_n - b'(\theta_n)}{a(\phi_n)} \end{bmatrix} = \begin{bmatrix} \frac{y_1 - \mu_1}{a(\phi_1)} \\ \frac{y_2 - \mu_2}{a(\phi_2)} \\ \vdots \\ \frac{y_n - \mu_n}{a(\phi_n)} \end{bmatrix}$$

이제 가능도추정을 위한 방정식 11.14 을 위에서 유도한 도함수 벡터와 행렬을 이용하여 다시 쓰면 다음과 같다.

$$\begin{aligned} 0 = \frac{\partial \ell_n}{\partial \beta} &= \frac{\partial \eta}{\partial \beta} \frac{\partial \mu}{\partial \eta} \frac{\partial \theta}{\partial \mu} \frac{\partial \ell}{\partial \theta} \\ &= X^t \begin{bmatrix} \frac{1}{g_\mu(\mu_1)} & & & \\ & \frac{1}{g_\mu(\mu_1)} & & \\ & & \ddots & \\ & & & \frac{1}{g_\mu(\mu_n)} \end{bmatrix} \begin{bmatrix} \frac{1}{v(\mu_1)} & & & \\ & \frac{1}{v(\mu_2)} & & \\ & & \ddots & \\ & & & \frac{1}{v(\mu_n)} \end{bmatrix} \begin{bmatrix} \frac{y_1 - \mu_1}{a(\phi_1)} \\ \frac{y_2 - \mu_2}{a(\phi_2)} \\ \vdots \\ \frac{y_n - \mu_n}{a(\phi_n)} \end{bmatrix} \\ &= X^t \begin{bmatrix} \frac{1}{g_\mu^2(\mu_1) a(\phi_1) v(\mu_1)} & & & \\ & \frac{1}{g_\mu^2(\mu_1) a(\phi_2) v(\mu_2)} & & \\ & & \ddots & \\ & & & \frac{1}{g_\mu^2 a(\phi_n)(\mu_n) v(\mu_n)} \end{bmatrix} \\ &\quad \times \begin{bmatrix} g_\mu(\mu_1) & & & \\ & g_\mu(\mu_2) & & \\ & & \ddots & \\ & & & g_\mu(\mu_n) \end{bmatrix} \begin{bmatrix} y_1 - \mu_1 \\ y_2 - \mu_2 \\ \vdots \\ y_n - \mu_n \end{bmatrix} \\ &= X^t W \Delta (y - \mu) \end{aligned} \tag{11.15}$$

위의 식에서 가중값 대각행렬 W, 연결함수 미분값 대각행렬 Δ, 관측값 벡터 y, 평균 벡터 μ는 다음과 같이 정의된다.

$$W = \begin{bmatrix} w_1 & & & \\ & w_2 & & \\ & & \ddots & \\ & & & w_n \end{bmatrix} = \begin{bmatrix} \frac{1}{g_\mu^2(\mu_1) a(\phi_1) v(\mu_1)} & & & \\ & \frac{1}{g_\mu^2(\mu_1) a(\phi_2) v(\mu_2)} & & \\ & & \ddots & \\ & & & \frac{1}{g_\mu^2 a(\phi_n)(\mu_n) v(\mu_n)} \end{bmatrix}$$

$$\Delta = \begin{bmatrix} \delta_1 & & & \\ & \delta_2 & & \\ & & \ddots & \\ & & & \delta_n \end{bmatrix} = \begin{bmatrix} g_\mu(\mu_1) & & & \\ & g_\mu(\mu_2) & & \\ & & \ddots & \\ & & & g_\mu(\mu_n) \end{bmatrix}$$

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{bmatrix}$$

11.4. 최대가능도추정량의 계산

이제 회귀계수 β를 최대가능도추정법(Maximum Likelihood Estimation)으로 가능도 방정식을 11.15 를 이용하면 다음과 같은 행렬 방정식으로 표시된다.

$$X^t W \Delta y = X^t W \Delta \mu \tag{11.16}$$

위의 방정식은 일반적으로 회귀계수 벡터 β에 대하여 선형방정식이 아니므로 최소제곱법과 같이 최대가능도 추정량을 직접 구할 수 없다.

> **Note**
>
> 1. 정규분포 가정 하에서 선형회귀 모형에서는 11.16 이 최소제곱법의 방정식 $X^t y = X^t X \beta$로 유도되고 직접적으로 구할 수 있다.
> 2. 많은 경우 스케일 모수 $a(\phi_i)$는 관측값 y_i에 따라 변하지 않고 상수인 경우가 흔하다. 즉 $a(\phi_i) \equiv a(\phi)$. 이러한 경우 가능도 방정식 11.13 또는 11.15 에서 스케일 모수 $a(\phi_i)$를 1로 놓고 방정식을 푼다.

최대가능도추정량을 실제 계산하기 위하여 로그 가능도 함수의 2차 도함수(헤시안) 행렬을 구해보자. 11.13 에서 얻은 1차 도함수를 한번 더 미분하면 다음과 같은 결과를 얻는다.

$$\begin{aligned}
\frac{\partial^2 \ell_n}{\partial \beta \partial \beta^t} &= \sum \frac{\partial}{\partial \beta} \left[x_i w_i g_\mu(\mu_i)(y_i - \mu_i) \right] \\
&= \sum \frac{\partial}{\partial \beta} \left[x_i c_i (y_i - \mu_i) \right] \quad [c_i \equiv w_i g_\mu(\mu_i)] \\
&= \sum \left[\frac{\partial c_i (y_i - \mu_i)}{\partial \beta} x_i^t \right] \\
&= \sum \left[\frac{\partial c_i}{\partial \beta}(y_i - \mu_i) + \frac{\partial (y_i - \mu_i)}{\partial \beta} c_i \right] x_i^t \\
&= \sum \left[\frac{\partial c_i}{\partial \beta}(y_i - \mu_i) - \frac{\partial \eta_i}{\partial \beta} \frac{\partial \mu_i}{\partial \eta_i} c_i \right] x_i^t \\
&= \sum \left[\frac{\partial c_i}{\partial \beta}(y_i - \mu_i) - x_i [g_\mu(\mu)]^{-1} c_i \right] x_i^t \quad [c_i [g_\mu(\mu)]^{-1} = w_i] \\
&= \sum \left[\frac{\partial c_i}{\partial \beta}(y_i - \mu_i) \right] x_i^t - \sum x_i w_i x_i^t \\
&= \sum \left[\frac{\partial c_i}{\partial \beta}(y_i - \mu_i) \right] x_i^t - X^t W X
\end{aligned}$$

그러므로 피셔정보 $I(\beta)$는 다음과 같이 얻어진다.

$$\begin{aligned}
I(\beta) &= -E \left[\frac{\partial^2 \ell_n}{\partial \beta \partial \beta^t} \right] \\
&= E \left[- \sum \left[\frac{\partial c_i}{\partial \beta}(y_i - \mu_i) \right] x_i^t + X^t W X \right] \\
&= 0 + X^t W X
\end{aligned} \tag{11.17}$$

또는 11.15 에서 얻은 1차 도함수 방정식을 를 한번 더 미분하여 기대값을 취하면 11.17 와 동일한 결과를 얻는다.

$$\begin{aligned}
\frac{\partial^2 \ell_n}{\partial\beta\partial\beta^t} &= \frac{\partial}{\partial\beta}\left[X^t W\Delta(y-\mu)\right] \\
&= \left\{\frac{\partial}{\partial\beta}\left[X^t W\Delta\right]\right\}(y-\mu) + \left\{\frac{\partial}{\partial\beta}\left[(y-\mu)^t\right]\right\}\Delta WX \\
&= \left\{\frac{\partial}{\partial\beta}\left[X^t W\Delta\right]\right\}(y-\mu) - \left[\frac{\partial\mu}{\partial\beta}\right]\Delta WX \\
&= \left\{\frac{\partial}{\partial\beta}\left[X^t W\Delta\right]\right\}(y-\mu) - \left[\frac{\partial\eta}{\partial\beta}\frac{\partial\mu}{\partial\eta}\right]\Delta WX \\
&= \left\{\frac{\partial}{\partial\beta}\left[X^t W\Delta\right]\right\}(y-\mu) - \left[X^t\Delta^{-1}\right]\Delta WX \\
&= \left\{\frac{\partial}{\partial\beta}\left[X^t W\Delta\right]\right\}(y-\mu) - X^t WX
\end{aligned}$$

최대가능도추정량 $\hat{\beta}$는 가능도 방정식 11.16 을 직접 풀어서 계산할 수 있지만 대부분의 경우 직접해(explicit solution)를 구하는 것이 불가능 하다. 따라서 보통의 경우 선형화된 작업변량에 반복가중최소제곱법(iterative weighted least square; IWLS)을 적용하여 최대가능도 추정량을 구하며 IWLS로 구하는 해는 가능도 방정식 11.16 의 해와 동일하다.

주어진 분포에서 기본 연결함수를 g 라고 하고 관측값 y를 변환한 작업변량 $z = g(y)$의 테일러 전개를 다음과 같이 고려해 보자.

$$z \equiv g(y) \cong g(\mu) + g_\mu(\mu)(y-\mu)$$

작업 변량 z 의 분산은 11.10 와 같이 다음으로 주어진다.

$$var(z) = v(\mu)g_\mu^2(\mu) \equiv w^{-1}$$

회귀계수 벡터의 초기값을 β_0라고 하자. 그러면 작업 변량의 초기값 z_0는 β_0 로 계산된 μ_0를 이용하여 다음과 같이 구할 수 있다.

$$z_0 = g(\mu_0) + g_\mu(\mu_0)(y-\mu_0) = \eta_0 + g_\mu(\mu_0)(y-\mu_0)$$

IWLS 추정량 $\hat{\beta}$는 z_0 를 설명변수 벡터 x로 선형회귀분석을 적합할 때 가중치를 w_0 로 이용하는 가중최소제곱법으로 반복적으로 적용하여 개선할 수 있다.

실제로 IWLS 추정량은 피셔정보를 이용한 스코어 방법(Fisher scoring method)로 구한 최대가능도 추정량과 동일함을 보일 수 있다. 일단 회귀계수 벡터의 초기값 $\hat{\beta}^0$ 으로 계산된 피셔정보 행렬을 A로 아래와 같이 정의하자.

$$A = -E\left[\frac{\partial^2 \ell_n}{\partial\beta\partial\beta^t}\right]_{\beta=\hat{\beta}^0}$$

$$A = -E\left[\frac{\partial^2 \ell_n}{\partial\beta\partial\beta^t}\right]_{\beta=\hat{\beta}^0}$$

새로운 추정량 $\hat{\beta}^1$ 가 이전의 추정량 $\hat{\beta}^0$에서 다음과 같은 축차식으로 계산되는 방법이 피셔 스코링 방법이다.

$$0 = \frac{\partial \ell_n}{\partial\beta}|_{\beta=\hat{\beta}^0} - A(\hat{\beta}^1 - \hat{\beta}^0) \quad \Leftrightarrow \quad A(\hat{\beta}^1 - \hat{\beta}^0) = \frac{\partial \ell_n}{\partial\beta}|_{\beta=\hat{\beta}^0}$$

위의 피셔의 스코링 방법을 더 정리하면 다음과 같이 유도할 수 있다.

$$\begin{aligned} A(\hat{\beta}^1 - \hat{\beta}^0) &= \frac{\partial \ell_n}{\partial\beta}|_{\beta=\hat{\beta}^0} \\ \Leftrightarrow X^t W_0 X(\hat{\beta}^1 - \hat{\beta}^0) &= X^t W_0 \Delta_0 (y - \mu_0) \\ \Leftrightarrow X^t W_0 X\hat{\beta}^1 &= X^t W_0 [X\hat{\beta}^0 + \Delta_0(y - \mu_0)] \\ \Leftrightarrow X^t W_0 X\hat{\beta}^1 &= X^t W_0 [\eta_0 + \Delta_0(y - \mu_0)] \\ \Leftrightarrow X^t W_0 X\hat{\beta}^1 &= X^t W_0 z_0 \end{aligned}$$

위의 방정식은 $z_0 = \eta_0 + g_\mu(\mu_0)(y-\mu_0)$를 가중치 w_0를 사용하여 얻은 가중최소제곱법에서 나온 방정식임을 알 수 있다. 따라서 최대가능도 추정량을 구하는 피셔의 스코링 방법은 앞에서 알아본 반복가중최소제곱법과 동일하다.

11.5. 편차

선형모형에서 잔차제곱합(residual sum of square; SSE)에 대한 의미를 살펴보고 이를 일반화 선형모형에 확장하는 개념인 편차(deviance)의 정의를 알아보자.

먼저 다음과 같은 선형회귀식을 고려한다.

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots \beta_p x_{pi} + e_i, \quad i = 1, 2, \dots, n$$

여기서 오차항 e_i를 서로 독립이며 평균이 0이고 분산이 σ^2 인 정규분포를 따른다고 가정하고 (σ^2는 알고있다고 가정하자) 각 관측변수의 평균을 다음과 같이 μ_i로 하자.

$$\mu_i = E(y_i|x_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi}$$

서로 독립인 관측변수 y_i의 분포는 정규분포를 따르므로

$$y_i \sim N(\mu_i, \sigma^2)$$

관측치 $\boldsymbol{y} = (y_1, y_2, \dots, y_n)^t$의 로그가능도함수는 다음과 같이 나타낼 수 있다.

$$l(\boldsymbol{\mu}|\boldsymbol{y}) = C - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu_i)^2$$

예측변수 $x_1, x_2, \dots, x_p$를 고려한 선형회귀모형에서 각 반응변수 평균의 예측에 선형예측식을 사용하며 모형에서 모수의 개수는 $p+1$로 보통 모수의 수가 관측값의 수 n보다 작다.

$$\hat{\mu}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \cdots + \hat{\beta}_p x_{pi} \equiv \hat{y}_i$$

이 때 선형회귀모형의 로그가능도함수의 최대값는 다음과 같다.

$$\begin{aligned} l(\hat{\boldsymbol{\mu}}|\boldsymbol{y}) &= l_{regession}(\hat{\boldsymbol{\beta}}|y) \\ &= C - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2 \\ &= C - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \\ &= C - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}SSE \end{aligned}$$

이제 위의 선형회귀모형에서 예측변수 $x_1, x_2, \dots, x_p$를 고려하지 않는 포화 모형을 생각해보자.

$$y_i = \beta_{0i} + e_i \quad \text{or} \quad E(y_i) = \beta_{0i}$$

이러한 포화 모형은 n개의 반응변수의 평균을 n개의 모수를 가진 모형으로 추정하는 것으로 위와 같은 모형을 포화 모형(saturated model)이라고 한다. 포화 모형에서 모수 β_{0i}의 최소제곱 추정량(또는 최대가능도 추정량)는 관측값 y_i임을 쉽게 알수 있다.

$$\min_{\beta_{0i}} \sum_{i=1}^{n}(y_i - \beta_{0i})^2 \quad \Rightarrow \hat{\beta}_{0i} = y_i, \quad i = 1, 2, \dots, n$$

포화 모형의 의미는 우리가 생각할 수 있는 모형 중에 가장 큰 모형으로 포화모형보다 큰 모형을 생각할 수 없다. 위에서 언급한 바와 같이 n개의 관측값에 대하여 모수의 수가 n개보다 큰 모형을 생각하면 유일한 모수의 추정이 불가능하다.

선형회귀모형에서 포화모형은 $\hat{\beta}_{0i} = y_i$이며 로그가능도함수의 최대값은 $l(\boldsymbol{y}|\boldsymbol{y})$로 표시하며 다음과 같다.

$$\begin{aligned} l(\boldsymbol{y}|\boldsymbol{y}) &= l_{saturated}(\hat{\boldsymbol{\beta}}_0|y) \\ &= C - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \hat{\beta}_{0i})^2 \\ &= C - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - y_i)^2 \\ &= C - \frac{n}{2}\log\sigma^2 + 0 \end{aligned}$$

포화모형은 설정할 수 있는 최대의 모수를 가진 가장 큰 모형이므로 우리가 생각할 수 있는 모형 중에서 관측값을 예측하는 예측력은 가장 좋다는 것을 알 수 있다(하지만 과적합모형이다). 따라서 예측변수 $\boldsymbol{x}$들을 사용하는 선형회귀모형의 예측력이 포화모형이 가지는 예측력에 가까우면 좋은 모형이라고 생각할 수 있다. 반응변수의 평균을 예측하는 예측력은 로그가능도함수의 크기로서 나타낼 수 있다. 포화모형과 선형회귀모형의 로그가능도함수를 비교하면 포화모형의 로그가능도함수가 크다는 것을 알수 있고 (why?) 두 로그가능도함수의 의 차이를 비교하면 다음과 같다.

$$l(\boldsymbol{y}|\boldsymbol{y}) - l(\hat{\boldsymbol{\mu}}|\boldsymbol{y}) = l_{saturated}(\hat{\boldsymbol{\beta}}_0|y) - l_{regession}(\hat{\boldsymbol{\beta}}|y) = \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{1}{2\sigma^2}SSE$$

따라서 포화모형과 로그가능도함수의의 차이가 작다는 것은 선형회귀모형의 잔차제곱합(SSE)이 작다는 것을 의미한다. 보통 잔차제곱합이 작으면 선형회귀모형의 예측력이 좋은 모형이며 이는 선형회귀모형의 가능도함수가 포화모형의 가능도함수에 가깝다는 의미이다.

이렇게 모형의 예측능력을 평가하는 측도로서 편차(deviance)를 포화모형과 고려한 회귀모형의 로그가능도함수의 차이에 2를 곱한 양으로 정의한다. 따라서 편차는 작을 수록 좋다.

$$deviance \equiv D(\boldsymbol{y};\hat{\boldsymbol{\mu}}) = 2\left[l(\boldsymbol{y}|\boldsymbol{y}) - l(\hat{\boldsymbol{\mu}}|\boldsymbol{y})\right] \tag{11.18}$$

정규분포인 경우 편차는 다음과 같이 주어진다.

$$D(\boldsymbol{y};\hat{\boldsymbol{\mu}}) = 2\left[l(\boldsymbol{y}|\boldsymbol{y}) - l(\hat{\boldsymbol{\mu}}|\boldsymbol{y})\right] = \frac{1}{\sigma^2}SSE$$

이제 이항분포들에서 나온 관측값에 대한 포화모형을 생각해 보자.

$$y_i \sim B(m_i, p_i(x_i)), \quad i = 1, 2, \ldots, n$$

위의 모형에서 포화모형은 어떤 모형일까? 포화모형은 n개의 관측변수의 평균, 여기서 $E(y_i/m_i) = p(\boldsymbol{x}_i)$를 n개의 관측치 y_i를 이용하여 추정한 모형으로서 각 성공확률은 해당하는 관측된 성공의 비율에 의해 추정된다. 즉,

$$\hat{p}(x_i) = \frac{y_i}{m_i}$$

이러한 경우의 로그가능도함수의 값은 다음과 같이 주어진다.

$$\begin{aligned} l(\boldsymbol{y}|\boldsymbol{y}) &= l_{saturated} \\ &= \sum_i \log\binom{m_i}{y_i} + \sum_i y_i \log\hat{p}(x_i) + \sum_i (m_i - y_i)\log(1 - \hat{p}(x_i)) \\ &= \sum_i \log\binom{m_i}{y_i} + \sum_i y_i \log\frac{y_i}{m_i} + \sum_i (m_i - y_i)\log(1 - \frac{y_i}{m_i}) \end{aligned}$$

따라서 위에서 주어진 포화함수의 로그가능도함수에서 로지스틱회귀식의 로그가능도함수 11.10 를 빼고 2를 곱해서 편차를 정의할 수 있다.

$$
\begin{aligned}
D(\boldsymbol{y};\hat{\boldsymbol{\mu}}) &= 2\left[l(\boldsymbol{y}|\boldsymbol{y}) - l(\hat{\boldsymbol{\mu}}|\boldsymbol{y})\right] \\
&= 2(l_{saturated} - l_{regession}) \\
&= 2\left[\sum_i y_i \log\frac{y_i}{m_i} + \sum_i (m_i - y_i)\log(1-\frac{y_i}{m_i}) - \sum_i y_i \log\hat{p}(x_i) - \sum_i (m_i - y_i)\log(1-\hat{p}(x_i))\right] \\
&= 2\left[\sum_i y_i \log\frac{y_i}{m_i\hat{p}(x_i)} + \sum_i (m_i - y_i)\log\frac{1-y_i/m_i}{1-\hat{p}(x_i)}\right] \\
&= 2\left[\sum_i y_i \log\frac{y_i}{m_i\hat{p}(x_i)} + \sum_i (m_i - y_i)\log\frac{m_i-y_i}{m_i-m_i\hat{p}(x_i)}\right] \\
&= 2\left[\sum_i y_i \log\frac{y_i}{\hat{y}_i} + \sum_i (m_i - y_i)\log\frac{m_i-y_i}{m_i-\hat{y}_i}\right]
\end{aligned}
$$

위에서 $\hat{y}_i = m_i\hat{p}(x_i)$으로 로지스틱 회귀에서 성공의 횟수의 평균에 대한 예측값이다.

위의 논의에서 알 수 있듯이 로지스틱 회귀에서의 편차는 선형회귀 분석에서 잔차 제곱합 SSE의 의미로 해석할 수 있으며 작을 수록 모형의 예측력이 좋다는 것을 알 수 있다.

또한 편차는 표본의 개수가 충분히 크면 자유도가 $n-p$ 인 카이제곱분포를 따른다. 여기서 p는 회귀계수 벡터 $\boldsymbol{\beta}$의 크기이다.

정규분포와 이항분포의 편차를 비교하면 정규분포의 편차에는 산포를 나타내는 모수 σ^2 이 포함되어 있지만 이항분포의 편차에는 다른 모수가 나타나지 않는다. 11.18 에서 주어진 편차를 척도 모수(scaled parameter) 또는 산포 모수(dispersion parameter) ϕ 를 곱해준 값을 척도화 편차(scaled deviance) $D^*(\boldsymbol{y};\hat{\boldsymbol{\mu}})$ 라고 부른다.

$$
D^*(\boldsymbol{y};\hat{\boldsymbol{\mu}}) = \phi D(\boldsymbol{y};\hat{\boldsymbol{\mu}}) \tag{11.19}
$$

정규분포에서 산포 모수가 분산 $\phi = \sigma^2$ 이므로 척도화 편차는 잔차제곱합 $D^*(\boldsymbol{y};\hat{\boldsymbol{\mu}}) = SSE$ 가 되며 이항분포에서는 편차와 척도화 편차가 같다.

이제 선형모형에서 고려하는 다음과 같은 가설 검정을 고려해 보자.

$$
H_0: \text{ reduced model} \quad vs. \quad H_1: \text{full model} \tag{11.20}
$$

예를 들어 다음과 같은 가설검정을 자주 고려하게 된다.

$$
H_0: \beta_1 = \beta_2 = \cdots = \beta_q = 0 \quad vs. H_1: \text{ not } H_0
$$

만약 축소모형(reduced moldel) 에 대한 편차를 D_0 라고 하고 큰 모형(full model)에 대한 척도화 편차를 D_1 라고 하면 귀무가설이 참인 경우 두 편차의 차이 $D_0 - D_1$ 은 근사적으로 자유도가 d 인 카이제곱분포를 따른다.

여기서 자유도 d 는 두 모형의 회귀계수의 갯수 차이이다. 이러한 두 편차의 차이의 점근적 분포 가능도비검정 이론에 의하여 유도할 수 있다.

11.6. 이항변수에 대한 회귀모형

11.6.1. 단순 로지스틱 회귀모형

이제 반응변수의 값이 연속형 변수가 아니라 두 개의 가능한 결과만을 가지는 이항변수라면 선형 회귀식은 적절하지 못하다. 왜냐하면 반응변수의 기대값이 0과 1사이의 확률로 나타나기 때문이다.

$$E(y|x) = 1 \cdot P(y=1|x) + 0 \cdot P(y=0|x) = P(y=1|x)$$

따라서 반응변수의 기대값의 범위와 예측변수가 있는 선형예측식(linear predictor) $\beta_0 + \beta_1 x$의 범위가 일치하지 않아서 선형회귀식을 그대로 사용할 수 없다.

위의 문제를 해결하기 위한 방법중의 하나는 다음과 같은 함수 m를 생각하여 변환된 선형예측식의 범위를 $[0,1]$로 만드는 것이다.

$$m : \Re \rightarrow [0,1] \quad \text{and } g(x) \text{ is monotone function.}$$

따라서 다음과 같은 이항변수를 반응변수로 하는 새로운 회귀식을 만들 수있다.

$$E(y|x) = m(\beta_0 + \beta_1 x) \tag{11.21}$$

주로 쓰이는 변환함수로 다음과 같은 로지스틱 함수(logistic function)가 있다.

$$m(x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)} \tag{11.22}$$

반응변수가 베르누이분포를 따를 때 위의 로지스틱함수를 사용하는 회귀식을 로지스틱 회귀식이라고 한다.

$$P(y=1|x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)} = \{1 + \exp[-(\beta_0 + \beta_1 x)]\}^{-1} \tag{11.23}$$

위의 로지스틱 회귀식을 다시 역으로 정리하면 다음과 같은 식을 얻을 수 있다.

$$\log\left[\frac{P(y=1|x)}{1 - P(y=1|x)}\right] = \log\frac{p(x)}{1-p(x)} = \beta_0 + \beta_1 x \tag{11.24}$$

11.24 에서 나타난 함수 $g(p) = \log[p/(1-p)]$를 로짓함수(logit function)이라고 부르며 이는 로지스틱 함수의 역함수로서 0과 1 사이의 값을 가지는 확률을 실수 전체로 변환하는 함수로서 선형예측식의 범위와 일치하게 한다.

이렇게 관측값의 평균 (베르누이분포에서는 성공확률)과 선형예측식의 관계를 설정하는 함수는 11.4 에서 정의한 결합함수(link function)라고 한다. 아래 로짓함수는 특별히 식 11.5 에 정의된 것과 같이 기본형 연결함수라고 부른다.

$$g[E(y|x)] = g(p(x)) = \log \frac{p(x)}{1-p(x)} = \beta_0 + \beta_1 x$$

다른 종류의 결합함수도 생각할 수 있다. 11.2 에 나타난 프로빗 함수 $\Phi(x) = P(Z \leq x)$ 또한 결합함수로 생각할 수 있다.

11.6.2. 오즈비

일반적인 선형 회귀분석의 모형에서 기울기 계수 β_1은 기울기로서 예측변수 x의 단위가 1 증가할 때 반응변수의 평균이 β_1만큼 증가하는 것으로 해석할 수 있다. 하지만 로지스틱 회귀모형 11.23 에서는 이러한 해석을 할 수 없다.

로지스틱 회귀모형에서 기울기 β_1의 의미를 알아보기 위하여 다음과 같은 주어진 확률 p에 대한 몇 가지 함수를 알아야 한다.

먼저 하나의 확률 p 에 대한 **오드(odd)**는 다음과 같이 정의된다.

$$\text{odd} = \frac{p}{1-p}$$

예로 성공의 확률이 1/3일 때 오드는 1/2가 되며 이는 평균적으로 세 번의 시행할 때 한 번 성공하고 두 번 실패한다는 의미이다. 반대로 성공의 확률이 2/3일 때 오드는 $2 = 2/1$가 되며 이는 평균적으로 세 번의 시행할 때 두 번 성공하고 한 번 실패한다는 의미이다. 성공의 확률이 1/2일 때 오드는 1이 된다.

두 개의 확률 p_1과 p_2 에 대한 **오즈비(odds ratio)** 는 다음과 같이 정의된다.

$$\text{odds ratio} = \frac{p_1/(1-p_1)}{p_2/(1-p_2)}$$

오즈비는 두 개의 성공 확률 p_1과 p_2를 비교할 때 쓰는 양이다. 두 개의 오드를 비율로서 비교하는 양이며 오즈비가 1일 경우에 두 확률은 같다.

이제 단순 로지스틱 회귀식 11.24 을 생각하고 예측변수 x를 0과 1의 값을 가지는 이항변수로 가정한다. $x = 1$인 경우는

$$\frac{P(y=1|x=1)}{1-P(y=1|x=1)} = \exp(\beta_0 + \beta_1)$$

이며 $x = 0$인 경우는

$$\frac{P(y=1|x=0)}{1-P(y=1|x=0)} = \exp(\beta_0)$$

표 11.1.: 나이와 만성심장질환의 관계

	나이 ≥ 55 $(x=1)$	나이 ≥ 55 $(x=1)$	합
CHD 있음 $(y=1)$	21	22	43
CHD 있음 $(y=1)$	6	51	57
합	27	73	100

위에서 주어진 두 개의 오드, 즉 $x=1$인 경우와 $x=0$인 경우의 두 오드의 비를 구하면 다음과 같다.

$$\frac{\frac{P(y=1|x=1)}{1-P(y=1|x=1)}}{\frac{P(y=1|x=0)}{1-P(y=1|x=0)}} = \exp(\beta_1)$$

이는 다시 쓰면

$$\frac{P(y=1|x=1)}{1-P(y=1|x=1)} = \exp(\beta_1)\frac{P(y=1|x=0)}{1-P(y=1|x=0)}$$

위의 식에서 볼 때 예측변수 x가 1 의 값을 가질 때 반응 변수의 오드가 예측변수 x가 0일 경우의 오드의 $\exp(\beta_1)$ 배로 변하는 것을 알 수 있다. 따라서 $\exp(\beta_1)$는 반응변수의 오드의 증가량으로 볼 수 있다. 이는 두 성공확률의 오즈 비가 $\exp(\beta_1)$을 말한다. 위의 식에 로그를 취하면 다음과 같은 관계를 얻는다.

$$\log\left[\frac{P(y=1|x=1)}{1-P(y=1|x=1)} \Big/ \frac{P(y=1|x=0)}{1-P(y=1|x=0)}\right] = \beta_1$$

즉 오즈 비의 로그값이 단순 로지스틱 회귀식에서 기울기 β_1으로 나타난다.

간단한 예제를 통하여 오즈비와 로지스틱 회귀의 기울기의 관계를 명확히 해보자. 100명의 사람들을 55세 이상의 사람$(x=1)$과 55세 미만의 사람$(x=0)$의 그룹으로 나누었을 때 각 그룹에서 만성심장질환(CHD)가 있는 사람$(y=1)$과 없는 사람$(y=0)$의 수가 다음 표에 주어져있다.

표 11.2.: 나이와 만성심장질환의 관계

	나이 ≥ 55 $(x=1)$	나이 < 55 $(x=0)$	합
CHD 있음 $(y=1)$	21	22	43
CHD 없음 $(y=0)$	6	51	57
합	27	73	100

여기서 나이에 대한 CHD유무의 오즈비는 다음과 같이 계산된다.

$$\text{Odds Ratio } = \frac{\frac{21/27}{6/27}}{\frac{22/73}{51/73}} = \frac{\frac{21}{6}}{\frac{22}{51}} = 8.11$$

위의 표 **??**의 자료를 이용하여 로지스틱회귀를 적합시키면 결과가 아래와 같고 회귀계수 β_1의 추정값은 오즈비의 로그값임을 알 수 있다.

$$\hat{\beta}_1 = \log(8.11) = 2.094$$

```
yes <- c(21,22)
no <- c(6,51)
x <- c (1,0)
m1 <- glm(cbind(yes,no) ~ x,family=binomial())
summary(m1)
```

```
Call:
glm(formula = cbind(yes, no) ~ x, family = binomial())

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.8408     0.2551  -3.296  0.00098 ***
x             2.0935     0.5285   3.961 7.46e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1.8704e+01  on 1  degrees of freedom
Residual deviance: 1.4211e-14  on 0  degrees of freedom
AIC: 11.987

Number of Fisher Scoring iterations: 3
```

11.6.3. 로지스틱 회귀식 추정 및 예측

주어진 예측변수 $\boldsymbol{x}_i = (x_{1i}, \dots, x_{pi})^t$에서 실행횟수가 m_i인 이항분포$B(m_i, p(x_i))$를 생각하자. m_i의 시행 중에 성공의 횟수가 y_i라고 하면 y_i의 평균과 분산은 다음과 같다.

$$E(y_i|x_i) = m_i p(x_i), \qquad Var(y_i|x_i) = m_i p(x_i)(1-p(x_i)), \quad i = 1, 2, \dots, n$$

이항분포를 위한 로지스틱 회귀방정식은 선형예측식과 성공의 확률의 관계를 다음과 같이 정한다.

$$\log\left[\frac{p(x_i)}{1-p(x_i)}\right] = \boldsymbol{x}_i^t\boldsymbol{\beta} = \beta_0 + \beta_1 x_{1i} + \dots \beta_p x_{pi} \tag{11.25}$$

서로 독립인 관측값 $(y_1, y_2, \dots, y_n)$의 가능도함수 L은 이항분포들의 결합확률밀도함수와 같고 아래와 같이 주어지며

$$L = \prod_{i=1}^{n} f(y_i|p(x_i)) = \prod_{i=1}^{n} \left[\binom{m_i}{y_i} \right] p(x_i)_i^y (1 - p(x_i))^{m-y_i}$$

로그가능도함수 l 은 다음과 같이 나타낼 수 있다.

$$\begin{aligned} l = \log L &= \sum_i \log \binom{m_i}{y_i} + \sum_i y_i \log p(x_i) + \sum_i (m_i - y_i) \log(1 - p(x_i)) \\ &= c(\boldsymbol{y}, \boldsymbol{m}) + \sum_i y_i \log \left[\frac{p(x_i)}{1 - p(x_i)} \right] + \sum_i m_i \log(1 - p(x_i)) \end{aligned}$$

로지스틱 회귀식이 추정된 후에 새로운 예측변수 $x = x^*$에 대하여 성공의 확률을 예측하고 싶다면 다음과 같은 식을 써서 예측할 수 있다.

$$P(y = 1|x = x^*) = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 x_1^* + \cdots + \hat{\beta}_p x_p^*)}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 x_1^* + \cdots + \hat{\beta}_p x_p^*)} = \{1 + \exp[-(\hat{\beta}_0 + \hat{\beta}_1 x_1^* + \cdots + \hat{\beta}_p x_p^*)]\}^{-1}$$

11.7. 발생 횟수에 대한 모형

11.7.1. 포아송 회귀모형

반응변수 y가 어떤 사건이 일어난 횟수(count)라면 주로 포아송분포를 확률 모형으로 사용한다.

$$P(Y = y) = f(y|\mu) = \frac{e^{-\mu} \mu^y}{y!}, \quad y = 0, 1, 2, \dots \tag{11.26}$$

이러한 포아송 분포에서 나온 반응변수에 대하여 예측변수 x의 영향에 대한 회귀분석을 포아송 회귀식이라고 한다. 포아송 분포의 평균 μ는 양의 실수이고 선형예측식 $\eta = \boldsymbol{x}^t \boldsymbol{b}eta$의 범위는 실수이기 때문에 로그함수를 결합함수(link function)으로 이용하여 회귀식을 세운다.

$$\log E(y|\boldsymbol{x}_i) = \log \mu(\boldsymbol{x}_i) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p \tag{11.27}$$

사실 포아송분포의 로그 가능도함수에서 로그함수가 기본형 결합함수임을 쉽게 알 수 있다. 즉, $\boldsymbol{y} = (y_1, y_2, \dots, y_n)^t$를 서로 독립이고 평균이 $\mu_i = \mu(\boldsymbol{x}_i)$인 포아송 확률변수라고 한다면 로그가능도함수는 다음과 같다.

$$\begin{aligned} l &= \log \prod_{i=1}^{n} f(y_i|\mu_i) \\ &= \sum_{i=1}^{n} [y_i \log \mu_i - \mu_i - \log y_i!] \end{aligned}$$

위에서 볼 수 있듯이 충분통계량 y_i에 대응하는 모수에 대한 항은 $\log \mu_i$로서 이는 로그함수가 기본형 결합함수임을 나타낸다.

포아송 회귀분석는 다음과 같은 몇 가지 특징을 가지고 있다.

- 만약에 어떤 사건이 일어난 횟수가 몇 가지 가능한 수중 하나라면 (예: $y \leq M$) 포아송분포를 이항분포의 근사(approximation)로 생각할 수 있다. 즉 n이 크고 성공확률 p가 작으면 이항분포는 평균이 $\mu = np$인 포아송 분포와 매우 가깝기 때문에 가능한 횟수가 제한되었다 하더라도 포아송 회귀식을 이용할 수 있다.
- 사건의 일어난 횟수가 주어진 시간의 길이에 비례하고 다른 사건과 독립이면 포아송 분포를 따른다. 또한 포아송 분포는 두 개의 사건이 일어날 때 시간 간격이 지수분포(exponential distribution)을 따른다면 주어진 시간 간격동안 일어난 사건의 횟수는 포아송 분포를 따른다.
- y_i가 서로 독립이고 평균이 μ_i인 포아송 분포를 따른다면 합 $\sum_i y_i$는 평균이 $\sum_i \mu_i$인 포아송분포를 따른다

포아송 회귀분석에서 11.18 에서 정의된 편차 $D(\boldsymbol{y}; \hat{\boldsymbol{\mu}})$를 구해보기 위하여 포화모형을 생각해보자. 각 관측값의 평균 μ_i를 자신의 관측값 y_i로 추정하는 것이 포화모형이다. 따라서 포화모형의 로그가능도함수는 다음과 같이 주어지고

$$l(\boldsymbol{y}|\boldsymbol{y}) = \sum_{i=1}^{n} [y_i \log y_i - y_i - \log y_i!]$$

포아송 회귀분석의 로그가능도함수에 적용하면 $D(\boldsymbol{y}; \hat{\boldsymbol{\mu}})$를 얻을 수 있다.

$$\begin{aligned} D(\boldsymbol{y}; \hat{\boldsymbol{\mu}}) &= 2\left[l(\boldsymbol{y}|\boldsymbol{y}) - l(\hat{\boldsymbol{\mu}}|\boldsymbol{y})\right] \\ &= 2\left\{\sum_{i=1}^{n}[y_i \log y_i - y_i - \log y_i!] - \sum_{i=1}^{n}[y_i \log \hat{\mu}_i - \hat{m}u_i - \log y_i!]\right\} \\ &= 2\sum_{i=1}^{n}[y_i \log(y_i/\hat{\mu}_i) - (y_i - \hat{\mu}_i)] \end{aligned}$$

Example 11.2 (포아송 회귀분석)**.** Galapagos 군도에 있는 30개의 섬에서 사는 거북이의 개체 수를 반응변수 y로하고 5개의 지리적 변수를 예측변수로 하는 포아송 회귀식을 적합하려고 한다 (자료 출처는 Faraway (2016)).

```
library(faraway)
data(gala)
gala <- gala[,-2]
modp <- glm(Species ~ .,family=poisson, gala)
summary(modp)
```

```
Call:
glm(formula = Species ~ ., family = poisson, data = gala)

Coefficients:
```

```
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.155e+00  5.175e-02  60.963  < 2e-16 ***
Area        -5.799e-04  2.627e-05 -22.074  < 2e-16 ***
Elevation    3.541e-03  8.741e-05  40.507  < 2e-16 ***
Nearest      8.826e-03  1.821e-03   4.846 1.26e-06 ***
Scruz       -5.709e-03  6.256e-04  -9.126  < 2e-16 ***
Adjacent    -6.630e-04  2.933e-05 -22.608  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 3510.73  on 29  degrees of freedom
Residual deviance:  716.85  on 24  degrees of freedom
AIC: 889.68

Number of Fisher Scoring iterations: 5
```

11.7.2. 비율 모형

어떤 사건이 일어날 횟수(count)는 집단이나 시간의 크기(group size)에 의존할 수 있다. 예를 들어 각 도시의 1년 범죄 발생 횟수는 그 도시의 인구수나 크기에 비례하게 된다. 이러한 모형은 이항분포를 이용하여 분석할 수 도 있지만 사건의 발생 확률이 매우 작고 집단의 크기가 크면 포아송 근사를 통한 분석도 가능하다. 또한 어떤 경우에는 집단의 크기에 대한 정보가 부족할 수 있다.

이러한 비율에 대한 모형(Rate Models)을 나타내면 아래와 같고

$$\log \frac{\text{count}}{\text{group size}} = \boldsymbol{x}^t\boldsymbol{\beta}$$

이는 다시 발생횟수에 대한 포아송 회귀모형의 형태로 나타내면 다음과 같이 쓸 수 있다.

$$\log\ \text{count}\ = (1)(\log\ \text{group size}\ ) + \boldsymbol{x}^t\boldsymbol{\beta}$$

따라서 발생횟수에 대한 포아송 회귀분석을 적합할 때 집단의 크기를 안다면 그 로그 변환값을 회귀식에 포함하여 적합할 수 있다. 위의 식에서 알 수 있듯이 크기의 로그 변환변수는 회귀계수를 강제로 1 로 놓는 제약을 둘 수 있다. 이러한 변수를 오프셋(offset)이라고 한다.

Example 11.3 (비율 모형)**.** 세포(`cells`)에 감마 방사능을 쏘였을 때 비정상성(`ca`)를 나타내는 횟수에 대하여 비율 모형을 적합시켰다. 독립변수는 방사능의 양(`doseamt`)와 비율(`doserate`)이다. 여기서 세포의 수(`cells`)를 오프셋(offset) 변수로 사용한다 (자료 출처는 Faraway (2016)).

```
data(dicentric)
round(xtabs(ca/cells ~ doseamt+doserate, dicentric),2)
```

```
       doserate
doseamt  0.1 0.25  0.5    1  1.5    2  2.5    3    4
    1   0.05 0.05 0.07 0.07 0.06 0.07 0.07 0.07 0.07
    2.5 0.16 0.28 0.29 0.32 0.38 0.41 0.41 0.37 0.44
    5   0.48 0.82 0.90 0.88 1.23 1.32 1.34 1.24 1.43
```

```
dicentric$dosef <- factor(dicentric$doseamt)
rmod <- glm(ca ~ offset(log(cells))+log(doserate)*dosef, family=poisson,dicentric)
summary(rmod)
```

```
Call:
glm(formula = ca ~ offset(log(cells)) + log(doserate) * dosef,
    family = poisson, data = dicentric)

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)           -2.74671    0.03426 -80.165  < 2e-16 ***
log(doserate)          0.07178    0.03518   2.041 0.041299 *
dosef2.5               1.62542    0.04946  32.863  < 2e-16 ***
dosef5                 2.76109    0.04349  63.491  < 2e-16 ***
log(doserate):dosef2.5 0.16122    0.04830   3.338 0.000844 ***
log(doserate):dosef5   0.19350    0.04243   4.561  5.1e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 4753.00  on 26  degrees of freedom
Residual deviance:   21.75  on 21  degrees of freedom
AIC: 209.16

Number of Fisher Scoring iterations: 4
```

11.7.3. 음이항 분포

베르누이 독립시행에서 k번째의 성공까지의 시행회수 z는 음이항 분포(negative bionomial)을 따른다. 음이항분포는 포아송 분포에서 모수가 감마를 따를 때 근사분포로 사용될 수 있다.

$$P(z) = \binom{z-1}{k-1} p^k (1-p)^{z-k}, \quad z = k, k+1, ... \tag{11.28}$$

위의 분포에서 확률 변수와 모수를 다시 아래와 같이 정의하면

$$y = z - k, \quad p = \frac{1}{1+\alpha}$$

y의 확률분포는 다음과 같고

$$P(y) = \binom{y+k-1}{k-1} \frac{\alpha^y}{(1+\alpha)^{y+k}}, \quad y = 0, 1, 2, ...$$

따라서 y의 평균과 분산은 다음과 같이 주어진다.

$$E(y) = \mu = k\alpha, \quad Var(y) = k\alpha + k\alpha^2 = \mu + \mu^2/k$$

또한 로그가능도함수는 다음과 같이 주어지고

$$l = \sum_{i=1}^{n} \left(y_i \log \frac{\alpha}{1+\alpha} - k \log(1+\alpha) + \sum_{j=0}^{y_i - 1} \log(j+k) - \log y_i! \right)$$

결합함수는 다음과 같다.

$$\log \frac{\alpha}{1+\alpha} = \log \frac{\mu}{\mu+k} = \eta = \boldsymbol{x}^t \boldsymbol{\beta}$$

보통의 경우 k는 고정된 상수로 생각할 수도 있고 또는 모수로 보고 추정할 수 도 있다.

11.7.4. 영과잉모형

어떤 사건의 발생 횟수에 대한 자료를 수집할 때 영(0, zero)이 비정상적으로 많이 나타나는 경우가 있다. 만약 발생횟수의 분포를 포아송분포 11.26 으로 가정하면 0이 관측될 확률은 크지 않다.

$$P(y = 0) = e^{-\mu}$$

자료에서 0의 발생 빈도가 비정상적으로 많은 자료를 영과잉 자료(zero inflated data)라고 하며 이러한 자료에 포아송 분포를 그대로 적용하면 회귀 계수의 추정량에 편이(bias)가 발생할 수 있는 여러 가지 문제가 생긴다.

발생 횟수에 0이 많은 이유는 매우 다양하다. 0이 많이 발생하는 대표적인 이유를 살펴보자.

- 외부 요인에 의하여 사건의 발생이 제약을 받는 경우

- 발생은 했는데 관측이 안된 경우
- 원래 0이 많은 경우

이렇게 영과잉 자료를 분석할 수 있는 대표적인 모형은 영과잉 포아송 모형(zero inflated poission model; ZIP)이다. 확률변수 y_i를 사건의 발생 회수라고 하면 ZIP 모형에서 0이 관측될 확률을 다음과 같이 나타낼 수 있다.

$$P(y=0) = P(\text{ Extra zeros }) + [1 - P(\text{ Extra zeros })]P(\text{ count process gives a zero })$$

즉 0이 관측될 확률은 별도로 나타난 0이 관찰 될 확률과 원래 확률 과정에서 0이 관찰 될 확률의 조합(mixture)으로 나타난다. 이제 i번째 관측에서 별도로 나타난 0이 관찰될 확률을 π_i라 하면

$$P(y_i=0) = \pi_i + (1-\pi_i)P(\text{ count process gives a zero })$$

더 나아가 확률 과정이 평균이 μ_i인 포아송 분포를 따른다고 가정하면

$$\begin{aligned} P(y_i|y_i=0) &= \pi_i + (1-\pi_i)e^{-\mu_i} \\ P(y_i|y_i>0) &= (1-\pi_i)\frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!} \end{aligned}$$

위의 분포에서 y의 평균과 분산을 구해보면 다음과 같이 주어진다.

$$\begin{aligned} E(y_i) &= (1-\pi_i)\mu_i \\ Var(y_i) &= (1-\pi_i)\mu_i + (1-\pi_i)\pi_i\mu_i^2 \end{aligned}$$

위의 식에서 볼 수 있듯이 영과잉 포아송 모형은 과포화(overdispersion)을 보인다 ($Var(y_i) > E(y_i)$). 영과잉 포아송 모형에 대한 회귀분석은 별도로 나타난 0이 관측될 확률 π_i에 대한 로지스틱회귀와 발생회수에 대한 포아송회귀의 결합으로 분석할 수 있다.

$$\begin{aligned} \log\frac{\pi_i}{1-\pi_i} &= \boldsymbol{x}_b^t\boldsymbol{\beta}_b \\ \log\mu_i &= \boldsymbol{x}_p^t\boldsymbol{\beta}_p \end{aligned}$$

References

Faraway, Julian J. 2016. *Extending the Linear Model with r: Generalized Linear, Mixed Effects and Nonparametric Regression Models*. Chapman; Hall/CRC.

Fox, John, and Sanford Weisberg. 2018. *An r Companion to Applied Regression*. Sage publications.

Nelder, John Ashworth, and Robert WM Wedderburn. 1972. "Generalized Linear Models." *Journal of the Royal Statistical Society: Series A (General)* 135 (3): 370–84.

Pinheiro, José, and Douglas Bates. 2006. *Mixed-Effects Models in s and s-PLUS*. Springer Science & Business Media.

Searle, Shayle Robert, and Charles E McCulloch. 2001. *Generalized, Linear and Mixed Models*. Wiley.

강근석, and 유형조. 2016. *R*을 활용한 선형회귀분석. 1st ed. 교우사. https://github.com/regbook/regbook.

A. 행렬의 기초

이 장에서는 회귀분석의 이론 전개에 필요한 행렬 이론과 선형 대수의 기초에 대하여 알아볼 것이다.

A.1. 벡터와 행렬

다음 p-차원 벡터(vector) 또는 열벡터(column vector) $\boldsymbol{a}$ 는 p개의 원소 $a_1, a_2, \dots, a_p$ 를 하나의 열(column)에 배치한 형태를 가진 개체이다.

$$\boldsymbol{a} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix} \tag{A.1}$$

차원이 $n \times p$ 인 행렬 $\boldsymbol{A}$ 는 다음과 같이 n개의 행과 p 개의 열에 원소 a_{ij}를 다음과 같이 배치한 형태를 가진다.

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1p} \\ a_{21} & a_{22} & \dots & a_{2p} \\ \vdots & \vdots & & \dots \\ a_{n1} & a_{n2} & \dots & a_{np} \end{bmatrix}$$

A.2. 두 행렬의 덧셈

두 행렬 $\boldsymbol{A}$ 와 $\boldsymbol{B}$ 를 더하는 규칙은 다음과 같다.

- 두 행렬 $\boldsymbol{A}$ 와 $\boldsymbol{B}$ 는 행과 열의 갯수가 같아야 한다.
- $\boldsymbol{A} + \boldsymbol{B} = \boldsymbol{C}$ 라고 하면, 덧셈의 결과로 만들어진 행렬 $\boldsymbol{C}$는 두 행렬과 같은 수의 행과 열을 가지면 각 원소는 다음과 같다.

$$\boldsymbol{A} + \boldsymbol{B} = \boldsymbol{C} \quad \rightarrow \quad c_{ij} = a_{ij} + b_{ij}$$

A.3. 스칼라곱

임의의 실수 λ (스칼라)가 주어졌을 때, λ 와 행렬 $\boldsymbol{A}$의 스칼라곱(scalar product) 는 행렬의 모든 원소에 λ 를 곱해준 행렬로 정의된다.

예를 들어 $\lambda = 2$, $\boldsymbol{A} \in \mathbb{R}^{2\times 3}$ 인 경우

$$\lambda \boldsymbol{A} = 2 \begin{bmatrix} 1 & 2 & 3 \\ -1 & 0 & 2 \end{bmatrix} = \begin{bmatrix} 2 & 4 & 6 \\ -2 & 0 & 4 \end{bmatrix}$$

A.4. 벡터와 행렬의 곱셈

$n \times p$ 인 행렬 $\boldsymbol{A}$ 와 p-차원 벡터(vector) $\boldsymbol{b}$는 다음과 같이 두 개의 서로 다른 형태로 나타낼 수 있다.

A.4.1. 행과 열의 내적

먼저 행렬과 벡터의 곱셈은 행렬 $\boldsymbol{A}$ 의 행벡터와 벡터 $\boldsymbol{b}$ 의 내적(inner product)로 나타낼 수 있다.

$$\begin{aligned}
\boldsymbol{Ab} &= \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1p} \\ a_{21} & a_{22} & \dots & a_{2p} \\ \vdots & \vdots & & \dots \\ a_{n1} & a_{n2} & \dots & a_{np} \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_p \end{bmatrix} \\
&= \begin{bmatrix} \boldsymbol{r}_1^t \\ \boldsymbol{r}_2^t \\ \vdots \\ \boldsymbol{r}_n^t \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_p \end{bmatrix} \qquad \text{where } \boldsymbol{r}_i^t = \begin{bmatrix} a_{i1} & a_{i2} & \dots & a_{ip} \end{bmatrix} \\
&= \begin{bmatrix} \boldsymbol{r}_1^t \boldsymbol{b} \\ \boldsymbol{r}_2^t \boldsymbol{b} \\ \vdots \\ \boldsymbol{r}_n^t \boldsymbol{b} \end{bmatrix} = \begin{bmatrix} \sum_{j=1}^p a_{1j} b_j \\ \sum_{j=1}^p a_{2j} b_j \\ \vdots \\ \sum_{j=1}^p a_{nj} b_j \end{bmatrix} \\
&= \begin{bmatrix} < \boldsymbol{r}_1, \boldsymbol{b} > \\ < \boldsymbol{r}_2, \boldsymbol{b} > \\ \vdots \\ < \boldsymbol{r}_n, \boldsymbol{b} > \end{bmatrix}
\end{aligned}$$

위에서 $< \boldsymbol{a}, \boldsymbol{b} >$ 는 다음과 같은 두 벡터의 내적(inner product)을 의미한다.

$$< \boldsymbol{a}, \boldsymbol{b} > = \boldsymbol{a}^t \boldsymbol{b} = \sum_{i=1}^{p} a_i b_i$$

A.4.2. 열벡터의 선형조합

이제 행렬과 벡터의 곱셈을 행렬을 구성하는 열벡터들의 선형조합(linear combination)으로 나타낼 수 있다.

$$
\begin{aligned}
\boldsymbol{Ab} &= \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1p} \\ a_{21} & a_{22} & \dots & a_{2p} \\ \vdots & \vdots & & \dots \\ a_{n1} & a_{n2} & \dots & a_{np} \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_p \end{bmatrix} \\
&= \begin{bmatrix} \boldsymbol{c}_1 & \boldsymbol{c}_2 & \dots & \boldsymbol{c}_p \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_p \end{bmatrix} \quad \text{where } \boldsymbol{c}_j = \begin{bmatrix} a_{1j} \\ a_{2j} \\ \vdots \\ a_{nj} \end{bmatrix} \\
&= b_1 \begin{bmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{n1} \end{bmatrix} + b_2 \begin{bmatrix} a_{12} \\ a_{22} \\ \vdots \\ a_{n2} \end{bmatrix} + \cdots + b_p \begin{bmatrix} a_{1p} \\ a_{2p} \\ \vdots \\ a_{np} \end{bmatrix} \\
&= b_1 \boldsymbol{c}_1 + b_2 \boldsymbol{c}_2 + \cdots + b_p \boldsymbol{c}_p
\end{aligned}
$$

A.5. 행렬의 전치

$\boldsymbol{A}^t$는 행렬의 전치(transpose)를 나타낸다. 행렬의 전치는 원소의 행과 열을 바꾸어 만든 행렬이다.

$$
\boldsymbol{A} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1p} \\ a_{21} & a_{22} & \dots & a_{2p} \\ \vdots & \vdots & & \dots \\ a_{n1} & a_{n2} & \dots & a_{np} \end{bmatrix} = \{a_{ij}\}_{n \times p} \quad \rightarrow \quad \boldsymbol{A}^t = \begin{bmatrix} a_{11} & a_{21} & \dots & a_{n1} \\ a_{12} & a_{22} & \dots & a_{n2} \\ \vdots & \vdots & & \dots \\ a_{1p} & a_{2p} & \dots & a_{np} \end{bmatrix} = \{a_{ji}\}_{p \times n}
$$

A.6. 행렬의 곱셈

먼저 두 행렬 $\boldsymbol{A}$ 와 $\boldsymbol{B}$ 의 곱셈

$$
\boldsymbol{A} \times \boldsymbol{B} \equiv \boldsymbol{AB}
$$

을 정의하려면 다음과 같은 조건이 만족되어야 한다.

- 행렬 $\boldsymbol{A}$ 의 열의 갯수와 행렬 $\boldsymbol{B}$ 의 행의 갯수가 같아야 한다

따라서 두 행렬의 곱셈은 순서를 바꾸면 정의 자체가 안될 수 있다.

이제 두 행렬 $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ 와 $\boldsymbol{B} \in \mathbb{R}^{n \times k}$ 의 곱셈은 다음과 같이 정의된다.

$$\boldsymbol{AB} = \boldsymbol{C}$$

행렬 $\boldsymbol{C}$ 는 m 개의 행과 k개의 열로 구성된 행렬이며($\boldsymbol{C} \in \mathbb{R}^{m \times k}$) 각 원소 c_{ij}는 다음과 같이 정의된다.

$$c_{ij} = \sum_{l=1}^{n} a_{il} b_{lk}, \quad i = 1, 2, \dots, m; j = 1, 2, \dots, k$$

먼저 간단한 예제로 다음과 같은 두 개의 행렬의 곱을 생각해 보자.

$$\boldsymbol{AB} = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ -1 & 2 \end{bmatrix} = \begin{bmatrix} (1)(0) + (2)(-1) & (1)(1) + (2)(2) \\ (3)(0) + (4)(-1) & (3)(1) + (4)(2) \end{bmatrix} = \begin{bmatrix} -2 & 5 \\ -4 & 11 \end{bmatrix}$$

곱하는 순서를 바꾸어 계산해 보자.

$$\boldsymbol{BA} = \begin{bmatrix} 0 & 1 \\ -1 & 2 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} = \begin{bmatrix} (0)(1) + (1)(3) & (0)(2) + (1)(4) \\ (-1)(1) + (2)(3) & (-1)(2) + (2)(4) \end{bmatrix} = \begin{bmatrix} 3 & 4 \\ 5 & 6 \end{bmatrix}$$

위 두 결과를 보면 행렬의 곱셈에서는 교환법칙이 성립하지 않음을 알 수 있다.

이제 차원이 다른 두 행렬의 곱셈을 살펴보자.

$$\boldsymbol{A} = \begin{bmatrix} 1 & 2 & 3 \\ 3 & 2 & 1 \end{bmatrix}, \quad \boldsymbol{B} = \begin{bmatrix} 0 & 2 \\ 1 & -1 \\ 0 & 1 \end{bmatrix}$$

두 행렬의 곱셈은 다음과 같이 계산할 수 있다.

$$\boldsymbol{AB} = \begin{bmatrix} 1 & 2 & 3 \\ 3 & 2 & 1 \end{bmatrix} \begin{bmatrix} 0 & 2 \\ 1 & -1 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 2 & 3 \\ 2 & 5 \end{bmatrix}$$

두 행렬의 곱하는 순서를 바꾸면 차원이 전혀 다른 행렬이 얻어진다.

$$\boldsymbol{BA} = \begin{bmatrix} 0 & 2 \\ 1 & -1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 & 3 \\ 3 & 2 & 1 \end{bmatrix} = \begin{bmatrix} 6 & 4 & 2 \\ -2 & 0 & 2 \\ 3 & 2 & 1 \end{bmatrix}$$

행렬의 곱셈은 교환법칙이 성립하지 않는다.

$$\boldsymbol{AB} \neq \boldsymbol{BA} \tag{A.2}$$

Caution

교환법칙이 성립하지 않는다는 의미는 A.2 이 언제나 성립한다는 의미는 아니다. 아래와 같이 특별한 경우 교환법칙이 성립하는 경우도 있다.

$$\begin{bmatrix} 1 & 2 \\ 1 & 3 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 2 \\ 1 & 3 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 1 & 3 \end{bmatrix}$$

- 행렬의 곱셈은 결합법칙과 배분법칙은 성립한다.

$$(\boldsymbol{AB})\boldsymbol{C} = \boldsymbol{A}(\boldsymbol{BC})$$

$$(\boldsymbol{A} + \boldsymbol{B})\boldsymbol{C} = \boldsymbol{AC} + \boldsymbol{BC}$$

A.7. 단위벡터와 항등행렬

i번째 단위벡터 $\boldsymbol{e}_i$를 정의하자. 단위벡터 $\boldsymbol{e}_i$는 n- 차원 벡터로서 i번째 원소만 1이고 나머지는 0인 벡터이다.

$$\boldsymbol{e}_i = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

즉 n-차원 항등행렬 $\boldsymbol{I}$는 n개의 단위벡터들을 모아놓은 것이다. 단위행렬은 대각원소가 1이고 나머지는 0인 정방행렬이다.

$$\boldsymbol{I} = [\boldsymbol{e}_1 \;\; \boldsymbol{e}_2 \;\; \ldots \;\; \boldsymbol{e}_n]$$

A.8. 대각합

$\boldsymbol{A} = \{a_{ij}\}$를 $n \times n$ 정방행렬(square matrix)인 경우, 행렬의 대각 원소(diagonal element)들의 합(trace)을 $tr(\boldsymbol{A})$로 표시한다.

$$tr(\boldsymbol{A}) = \sum_{i=1}^{n} a_{ii}$$

두 행렬의 덧셈(뺄셈)에 대한 대각합에 대한 성질들은 다음과 같다.

$$tr(\boldsymbol{A} \pm \boldsymbol{B}) = tr(\boldsymbol{A}) \pm tr(\boldsymbol{B})$$

> **Caution**
>
> 행렬의 곱셈은 일반적으로 교환법칙이 성립하지 않지만 대각합의 연산은 교환법칙이 성립한다.
>
> $$tr(\boldsymbol{AB}) = tr(\boldsymbol{BA})$$

대각합은 교환법칙이 성립히기 때문에 다음과 같은 성질이 성립한다.

$$\mathrm{tr}(\boldsymbol{AKL}) = \mathrm{tr}(\boldsymbol{KLA})$$

벡터의 연산에서도 대각합의 교환법칙이 성립되어 다음과 같은 유용한 식이 성립한다.

$$\mathrm{tr}\left(\boldsymbol{x}\boldsymbol{y}^{t}\right) = \mathrm{tr}\left(\boldsymbol{y}^{t}\boldsymbol{x}\right) = \boldsymbol{y}^{t}\boldsymbol{x} \in \mathbb{R}.$$

대각합의 교환법칙때문에 어떤 행렬의 앞에 특정 행렬을 곱하고, 뒤에 역행렬을 곱해도 대각합은 변하지 않는다.

$$\mathrm{tr}\left(\boldsymbol{S}^{-1}\boldsymbol{AS}\right) = \mathrm{tr}\left(\boldsymbol{ASS}^{-1}\right) = \mathrm{tr}(\boldsymbol{A})$$

대각합에 대한 그 밖의 성질들은 다음과 같다.

- $\mathrm{tr}(\alpha\boldsymbol{A}) = \alpha\,\mathrm{tr}(\boldsymbol{A}), \alpha \in \mathbb{R}$ for $\boldsymbol{A} \in \mathbb{R}^{n\times n}$
- $\mathrm{tr}\left(\boldsymbol{I}_{n}\right) = n$

A.9. 행렬식

$\boldsymbol{A}$의 행렬식(determinant)을 $det(\boldsymbol{A}) = |\boldsymbol{A}|$로 표기한다.

이차원 행렬 $\boldsymbol{A}$ 의 행렬식은 다음과 같이 계산한다.

$$\det(\boldsymbol{A}) = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = a_{11}a_{22} - a_{12}a_{21}.$$

만약 행렬 $\boldsymbol{A}$가 대각행렬(diagonal matrix)이면 $|\boldsymbol{A}|$는 행렬의 대각원소의 곱이다 ($|\boldsymbol{A}| = \prod a_{ii}$).

두 행렬의 곱의 행렬식은 각 행렬의 행렬식의 곱이다.

$$|\boldsymbol{AB}| = |\boldsymbol{A}||\boldsymbol{B}|$$

행렬식에 대한 유용한 공식들은 다음과 같다.

- $|\boldsymbol{A}^t| = |\boldsymbol{A}|$
- $|c\boldsymbol{A}| = c^n|\boldsymbol{A}|$

만약 행렬 $\boldsymbol{A}$가 다음과 같은 분할행렬(partitioned matrix) 의 형태를 가지면

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{0} & \boldsymbol{A}_{22} \end{bmatrix}$$

행렬 $\boldsymbol{A}$의 행렬식은 다음과 같이 주어진다.

$$|\boldsymbol{A}| = |\boldsymbol{A}_{11}||\boldsymbol{A}_{22}|$$

A.10. 직교행렬

만약 정방행렬 $\boldsymbol{P}$가 다음과 같은 조건을 만족하면 직교행렬(orthogonal matrix)라고 부른다.

$$\boldsymbol{P}\boldsymbol{P}^t = \boldsymbol{P}^t\boldsymbol{P} = \boldsymbol{I}$$

직교행렬의 정의에서 주의할 점은 $\mathbf{P} \mathbf{P}^\mathrm{t} = \mathbf{I}$ 와 $\boldsymbol{P}^t\boldsymbol{P} = \boldsymbol{I}$ 이 모두 성립하해야 한다는 점이다.

행렬 $\boldsymbol{P}$ 의 역행렬은 $\boldsymbol{P}^t$ 이다.

$$\boldsymbol{P}^{-1} = \boldsymbol{P}^t$$

만약 $\boldsymbol{P}$가 직교행렬이면 다음과 같은 성질을 가진다.

- $|\boldsymbol{P}| = \pm 1$, 왜냐하면
$$|\boldsymbol{P}\boldsymbol{P}^t| = |\boldsymbol{P}||\boldsymbol{P}^t| = |\boldsymbol{P}|^2 = |\boldsymbol{I}| = 1$$
- 임의의 정방행렬 $\boldsymbol{A}$에 대하여 다음이 성립한다.
$$tr(\boldsymbol{P}\boldsymbol{A}\boldsymbol{P}^t) = tr(\boldsymbol{A}\boldsymbol{P}^t\boldsymbol{P}) = tr(\boldsymbol{A})$$

A.11. 우드베리 공식

다음은 우드베리공식(Woodbury formula) 과 파생된 유용한 공식들이다.

$$(\boldsymbol{A} + \boldsymbol{U}\boldsymbol{C}\boldsymbol{V})^{-1} = \boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{U}(\boldsymbol{C}^{-1} + \boldsymbol{V}\boldsymbol{A}^{-1}\boldsymbol{U})^{-1}\boldsymbol{V}\boldsymbol{A}^{-1}$$

$$(\boldsymbol{I} + \boldsymbol{U}\boldsymbol{C}\boldsymbol{V})^{-1} = \boldsymbol{I} - \boldsymbol{U}(\boldsymbol{C}^{-1} + \boldsymbol{V}\boldsymbol{U})^{-1}\boldsymbol{V}$$

$$(\boldsymbol{A} + \boldsymbol{u}\boldsymbol{v}^t)^{-1} = \boldsymbol{A}^{-1} - \frac{\boldsymbol{A}^{-1}\boldsymbol{u}\boldsymbol{v}^t\boldsymbol{A}^{-1}}{1 + \boldsymbol{v}^t\boldsymbol{A}^{-1}\boldsymbol{u}} \tag{A.3}$$

$$(a\boldsymbol{I}_n + b\mathbf{1}_n\mathbf{1}_n^t)^{-1} = \frac{1}{a}\left[\boldsymbol{I}_n - \frac{b}{a+nb}\mathbf{1}\mathbf{1}^t\right]$$

B. 벡터공간

B.1. 벡터공간의 정의와 의미

먼저 지금까지 우리가 배운 벡터의 개념을 일반화하여 다루기 위해서 벡터공간의 일반적 개념을 정의하고자 한다.

벡터는 숫자를 모아놓은 형태인 A.1 로 주로 나타내지만 이러한 벡터를 모아놓은 집합은 실벡터 공간(real vector space)이라고 한다. 즉, A.1 의 벡터는 p-차원 실벡터(real vector)라고 한다.

지금부터 논의할 추상적인 벡터 공간(abstract vector space)은 어떤 집합이든 원소에 대한 더하기와 스칼라곱이 정의되어 있는 공간을 말한다.

이제부터 $\mathbb{R}$ 을 실수 전체 집합이라고 하자. 또한 $\mathbb{R}^n$ 을 n-차원 실벡터(real vector)의 집합이라고 하자. 또한 $\mathbb{R}^{n\times p}$ 을 $n \times p$-차원 행렬의 집합이라고 하자.

벡터공간(vector space) 은 어떤 집합 S 에 다음과 같은 두 개의 연산이 정의된 공간을 말한다.

1. 두 개의 원소에 대한 더하기(addition, $+$) 연산의 정의되어 있다.

$$+ \ : S + S \to S \tag{B.1}$$

2. 하나의 실수와 한 개의 원소에 대한 스칼라곱(scalar product, $\cdot$) 연산이 정의되어 있다.

$$\cdot \ : \mathbb{R} \cdot S \to S \tag{B.2}$$

위에서 **더하기 연산이 정의되어 있다**는 의미는 다음에 주어진 규칙이 성립한다는 의미이다.

- 집합 S 가 연산에 대하여 닫혀있다 (closure).

$$s_1 + s_2 \in S \quad \forall s_1, s_2 \in S$$

- 결합법칙이 성립한다 (Associativity).

$$(s_1 + s_2) + s_3 = s_1 + (s_2 + s_3) \quad \forall s_1, s_2, s_3 \in S$$

- 항등원이 존재한다 (Neutral element).

$$s + e = e + s = s \quad \exists e \ \ \forall s \in S$$

- 역원이 존재한다 (Inverse element).

$$s + i = i + s = 0 \quad \exists i \quad \forall s \in S$$

일반적으로 항등원(e) 는 0 으로 표시하며 역원(i) 는 $-s$ 로 표시한다.

- 교환법칙이 성립한다 (Commutativity).

$$s_1 + s_2 = s_2 + s_1 \quad \forall s_1, s_2 \in S$$

또한 위에서 **스칼라곱 연산이 정의되어 있다**는 의미는 다음에 주어진 규칙이 성립한다는 의미이다.

- 스칼라곱 연산의 분배법칙이 성립한다 (Distributivity).

$$r_1(s_1 + s_2) = r_1 s_1 + r_2 s_2, \quad (r_1 + r_2)s = r_1 s + r_2 s \quad \forall s_1, s_2 \in S, \ \ \forall r_1, r_2 in \mathbb{R}$$

- 스칼라곱 연산의 결합법칙이 성립한다

$$r_1(r_2 s) = (r_1 r_2)s \quad \forall s \in S, \ \ \forall r_1, r_2 in \mathbb{R}$$

- 스칼라곱 연산의 항등원이 존재한다 (Neutral element).

$$1 \cdot s = s \quad \forall s \in S$$

> **Caution**
>
> 벡터 공간에서 주의할 점은 **두 벡터의 곱하기** 가 정의되어 있다는 것이 아니라 하나의 스칼라와 하나의 벡터에 대한 스칼라 곱하기가 정의되어 있다는 것이다.
>
> $$\begin{bmatrix}1\\2\end{bmatrix} \cdot \begin{bmatrix}3\\4\end{bmatrix} = ? \quad but \quad 3 \cdot \begin{bmatrix}1\\2\end{bmatrix} = \begin{bmatrix}3\\6\end{bmatrix}$$
>
> **두 벡터의 곱하기** 는 내적(inner product) 란 이름으로 따로 정의한다. 또한 두 벡터의 곱셈이 유일하게 정의되지 않는다는 점에 유의하자. 예를 들어 벡터의 곱셈은 외적(cross product) 이라는 이름으로 정의된다.

이 강의에서는 스칼라로 실수만 사용하고 벡터공간은 실벡터공간(real vector space)만 고려할 것이다. 하지만 벡터공간은 실벡터가 아닌 다른 일반적인 집합에 대해서도 정의할 수 있음을 유의하자. 예를 들어 n-차원 다항식들을 모두 모아 놓은 집합은 벡터공간이다. 또한 연속인 함수들을 모아 놓은 집합도 벡터공간이다.

B.2. 벡터의 선형독립

벡터공간에 속한 벡터 $\boldsymbol{v}_1,\ \boldsymbol{v}_2,\ \ldots\ ,\boldsymbol{v}_n$ 의 선형결합(또는 선형결합, linear combination)이란 각 벡터에 스칼라를 곱하여 더한 것들이다.

즉 다음과 같은 형태의 식을 벡터 $\boldsymbol{v}_1,\ \boldsymbol{v}_2,\ \ldots\ ,\boldsymbol{v}_n$ 의 선형결합(linear combination)이라고 한다:

$$r_1\boldsymbol{v}_1 + r_2\boldsymbol{v}_2 + \cdots + r_n\boldsymbol{v}_n, \quad r_1, r_2, \ldots, r_n \in \mathbb{R} \tag{B.3}$$

Definition B.1 (벡터의 선형독립과 선형종속)**.** 벡터공간에 속한 벡터 $\boldsymbol{v}_1,\ \boldsymbol{v}_2,\ \ldots\ ,\boldsymbol{v}_n$ 가 있다고 하자. 만약 다음 식이 만약 모두 0인 n개의 스칼라 $x_1, x_2, \ldots, x_n$에 대해서만 성립하면 n개 벡터 $\boldsymbol{v}_1,\ \boldsymbol{v}_2,\ \ldots\ ,\boldsymbol{v}_n$ 들은 선형독립(linearly independent)라고 한다.

$$x_1\boldsymbol{v}_1 + x_2\boldsymbol{v}_2 + \cdots + x_n\boldsymbol{v}_n = \mathbf{0} \quad \Longleftrightarrow x_1 = x_2 = \cdots = x_n = 0 \tag{B.4}$$

또한 벡터 $\boldsymbol{v}_1,\ \boldsymbol{v}_2,\ \ldots\ ,\boldsymbol{v}_n$ 가 선형독립이 아니면 선형종속(linear dependent)라고 한다. 벡터 $\boldsymbol{v}_1,\ \boldsymbol{v}_2,\ \ldots\ ,\boldsymbol{v}_n$ 가 선형종속이면 **모두 0이 아닌** $x_1, x_2, \ldots, x_n$ 이 존재하여 다음이 성립한다는 것이다.

$$\exists\, x_1, x_2, \ldots, x_n \in \mathbb{R} \text{ s.t. } (x_1, x_2, \ldots, x_n) \neq \mathbf{0}, \quad \boldsymbol{v}_1 + x_2\boldsymbol{v}_2 + \cdots + x_n\boldsymbol{v}_n = \mathbf{0} \tag{B.5}$$

■

예를 들어 다음과 같이 주어진 3개의 3-차원 벡터들은 선형종속이다.

$$\boldsymbol{v}_1 = \begin{bmatrix}1\\2\\3\end{bmatrix}, \quad \boldsymbol{v}_2 = \begin{bmatrix}1\\0\\1\end{bmatrix}, \quad \boldsymbol{v}_3 = \begin{bmatrix}3\\2\\5\end{bmatrix} \tag{B.6}$$

왜냐하면 다음과 같이 모두 0이 아닌 스칼라에 의해서 다음 식이 성립하기 때문이다. 즉 벡터 $\boldsymbol{v}_3$는 $\boldsymbol{v}_2$ 에 2를 곱하여 $\boldsymbol{v}_1$에 더한 값과 같다.

$$\boldsymbol{v}_3 = \boldsymbol{v}_1 + 2\boldsymbol{v}_2 \quad \Longleftrightarrow \quad \boldsymbol{v}_1 + 2\boldsymbol{v}_2 - \boldsymbol{v}_3 = 0$$

이제 다음과 같이 주어진 3개의 3-차원 벡터들은 선형독립이다. 즉 3개 벡터의 선형 조합이 0이 될 수 있도록 만드는 스칼라는 모두 0인 경우 밖에 없다.

$$\boldsymbol{v}_1 = \begin{bmatrix}1\\2\\3\end{bmatrix}, \quad \boldsymbol{v}_2 = \begin{bmatrix}1\\0\\1\end{bmatrix}, \quad \boldsymbol{v}_3 = \begin{bmatrix}3\\2\\4\end{bmatrix} \tag{B.7}$$

이제 다음과 같이 주어진 4개의 3-차원 벡터들은 선형종속이다.

$$\boldsymbol{v}_1 = \begin{bmatrix}1\\2\\3\end{bmatrix}, \quad \boldsymbol{v}_2 = \begin{bmatrix}1\\0\\1\end{bmatrix}, \quad \boldsymbol{v}_3 = \begin{bmatrix}3\\2\\4\end{bmatrix} \quad \boldsymbol{v}_4 = \begin{bmatrix}0\\0\\1\end{bmatrix} \tag{B.8}$$

$\boldsymbol{v}_3$ 가 다음과 같이 다른 벡터의 선형결합으로 나타난는 것을 보여준다.

$$\boldsymbol{v}_3 = \begin{bmatrix}3\\2\\4\end{bmatrix} = (1)\boldsymbol{v}_1 + (2)\boldsymbol{v}_2 + (-1)\boldsymbol{v}_4 = (1)\begin{bmatrix}1\\2\\3\end{bmatrix} + (2)\begin{bmatrix}1\\0\\1\end{bmatrix} + (-1)\begin{bmatrix}0\\0\\1\end{bmatrix}$$

B.8 와 같이 3차원 벡터가 4개인 경우 벡터의 값에 관계없이 선형종속으로 나타난다. 이러한 사실은 $\mathbb{R}^n$ 의 $n+1$ 개의 벡터는 항상 선형종속이라는 정리의 결과이다.

즉, $\mathbb{R}^n$ **에서** n**개보다 더 많은 벡터들은 항상 선형종속이다.**

B.3. 역행렬

정방행렬 $\boldsymbol{A}$ 의 역행렬(inverse matrix) $\boldsymbol{A}^{-1}$는 다음과 같은 성질을 만족하는 행렬이다.

$$\boldsymbol{A}\boldsymbol{A}^{-1} = \boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{I}$$

역행렬은 언제나 존재하는 것은 아니다. 만약 행렬 $\boldsymbol{A}$가 역행렬을 가지면 이를 $\boldsymbol{A}^{-1}$로 표시한다. 또한 역행렬이 존재하면 정칙행렬(non-singular matrix)이라고 한다.

역행렬은 존재하는 조건은 행렬식(determinant)이 0이 아니어야 한다.

B.4. 행렬의 계수

행렬의 계수(rank)란 일차 독립인 열들의 최대 수 또는 일차 독립인 행들의 최대 수로 정의된다

$$rank(\boldsymbol{A}) = rk(\boldsymbol{A}) = dim(Col(\boldsymbol{A})) = dim(Row(\boldsymbol{A}))$$

꼭 기억해야 할 것은 행렬의 계수는 열들을 이용하여 구한 계수와 행들을 이용하여 구한 계수가 같다는 것이다. 즉, 행렬의 계수는 열의 계수와 행의 계수 중 하나만 구해도 된다는 것이다.

예를 들어 B.6 에 주어진 3 개의 벡터를 열로 하는 행렬의 계수는 2이다. 왜냐하면 선형종속인 벡터가 하나 있기 때문이다.

$$\boldsymbol{A} = \begin{bmatrix}1 & 1 & 3\\2 & 0 & 2\\3 & 1 & 5\end{bmatrix} \quad \rightarrow \quad rank(\boldsymbol{A}) = 2$$

위에 주어진 행렬 $\boldsymbol{A}$의 행들을 고려하면 첫 번째 행과 두 번째 행의 합이 세 번째 행으로 나타난다. 즉, 서로 독립인 행의 최대 개수는 2 이며 이는 서로 독립인 열의 최대 개수와 같다. 따라서 행렬 $\boldsymbol{A}$의 계수는 2이다.

다음으로 B.7 에 주어진 3 개의 벡터를 열로 하는 행렬의 계수는 3이다. 3개의 열벡터와 3개의 행벡터들은 모두 선형독립이다.

$$\boldsymbol{A} = \begin{bmatrix} 1 & 1 & 3 \\ 2 & 0 & 2 \\ 3 & 1 & 4 \end{bmatrix} \quad \rightarrow \quad rank(\boldsymbol{A}) = 3$$

이제 B.8 주어진 4개의 벡터로 이루어진 행렬의 계수는 3이다. 왜냐하면 4개의 열벡터 중 3개의 열벡터는 선형독립이지만 4번째 열벡터는 선형종속이기 때문이다.

$$\boldsymbol{A} = \begin{bmatrix} 1 & 1 & 3 & 0 \\ 2 & 0 & 2 & 0 \\ 3 & 1 & 4 & 1 \end{bmatrix} \quad \rightarrow \quad rank(\boldsymbol{A}) = 3$$

행렬의 열과 행의 개수가 다를 때 행렬의 계수는 열의 개수와 행의 개수 중 작은 값보다 같거나 작다 예를 들어 $n \times p$ 행렬의 계수는 $min(n, p)$ 와 같거나 작다.

$$A \in \mathbb{R}^{n \times p} \quad \rightarrow \quad rank(\boldsymbol{A}) \leq min(n, p)$$

다음은 행렬의 계수에 관련된 주요 공식이다.

- $rank(\boldsymbol{A}) = rank(\boldsymbol{A}^t)$
- 행렬 $\boldsymbol{A}$ 가 정방행렬이고 계수가 n 이면 역행렬이 존재한다(정칙행렬).
- 또한 더 나아가 $\boldsymbol{A}$ 가 정칙행렬이라는 사실은 아래 나열된 조건들과 동치(equivalance)이다.

  $\Leftrightarrow \boldsymbol{A}$ 의 열들이 일차독립이다.

  $\Leftrightarrow \boldsymbol{A}$ 의 행들이 일차독립이다.

  $\Leftrightarrow \boldsymbol{A}$ 의 계수가 n 이다.

  $\Leftrightarrow \boldsymbol{A}$ 의 행렬식이 0이 아니다.

B.5. 생성집합과 기저

벡터공간 V 의 벡터 $\boldsymbol{v}_1, \boldsymbol{v}_n, \dots, \boldsymbol{v}_m$ 의 선형결합을 모두 모은 집합

$$W = span\{\boldsymbol{v}_1, \boldsymbol{v}_2, \dots, \boldsymbol{v}_m\} = \{r_1\boldsymbol{v}_1 + r_2\boldsymbol{v}_2 + \cdots + r_m\boldsymbol{v}_m : r_1, r_2, \dots, r_m \in \mathbb{R}\}$$

을 벡터 $\boldsymbol{v}_1, \boldsymbol{v}_n, \dots, \boldsymbol{v}_m$ 의 생성(span)이라고 하며 W 의 생성집합(generating set, spanning set) 이라고 한다.

또한 어떤 벡터공간의 **생성집합에 속한 벡터들이 선형독립일 때** 이 생성집합을 기저 (basis)라고 한다.

만약 주어진 벡터 공간의 부분집합이 다시 벡터공간의 정의를 만족한다면 이를 부분공간(subspace)이라고 한다. 위에서 정의한 생성집합 W는 벡터공간 V의 부분공간이다.

B.6. 벡터공간의 차원

- $\mathbb{R}^n$ 의 모든 기저는 n개의 원소를 갖는다.
- 임의의 벡터공간 V에 대해서 V의 부분집합 $B = \{\boldsymbol{b}_1, \dots, \boldsymbol{b}_n\}$ 가 V의 기저라고 하면 다음을 보일 수 있다.
  - V 의 모든 벡터들은 $\boldsymbol{b}_1, \dots, \boldsymbol{b}_n$ 의 선형결합으로 나타낼 수 있으며 유일하다.
  - V 의 부분집합이 n 개보다 많은 벡터를 포함하면 이 부분집합의 벡터들은 선형종속이다.
  - V 의 또 다른 기저 $C = \{\boldsymbol{c}_1, \dots, \boldsymbol{c}_m\}$ 가있다면 $m = n$ 이다.
- 벡터공간 V의 차원(dimension) 은 기저의 개수로 정의되며 $dim(V)$로 표시한다.

B.7. 행렬의 열공간과 영공간

$n \times p$ 행렬 $\boldsymbol{A}$ 에 의하여 생성되는 열공간(column space) $C(\boldsymbol{A})$는 행렬 $\boldsymbol{A}$를 구성하는 열벡터의 선형조합으로 나타낼 수 있는 모든 벡터들의 집합을 말한다.

$$C(\boldsymbol{A}) = \{\boldsymbol{y} | \boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}, \boldsymbol{x} \in \mathbb{R}^p\}$$

$n \times p$ 행렬 $\boldsymbol{A}$ 에 의하여 생성되는 영공간(null space) $N(\boldsymbol{A})$ 는 다음과 같이 정의되는 벡터들의 집합을 말한다.

$$N(\boldsymbol{A}) = \{\boldsymbol{x} | \boldsymbol{A}\boldsymbol{x} = \boldsymbol{0} \text{ for } \boldsymbol{x} \in \mathbb{R}^p\}$$

벡터공간과 영공간은 다음과 같은 성질을 가진다.

- $rank(\boldsymbol{A}) =$ dimension of $C(\boldsymbol{A}) = dim[C(\boldsymbol{A})]$
- $dim[C(\boldsymbol{A})] + dim[N(\boldsymbol{A})] = n$

B.8. 두 벡터의 사영

선형독립인 두 벡터 $\boldsymbol{a}_1$과 $\boldsymbol{a}_2$ 가 있다고 하자. 벡터 $\boldsymbol{a}_1$과 같은 방향을 가지면서 벡터 $\boldsymbol{a}_2$에 가장 가까운 벡터를 $proj_{\boldsymbol{a}_1}(\boldsymbol{a}_2)$ 라고 정의하고 이를 벡터 $\boldsymbol{a}_1$ 방향으로 벡터 $\boldsymbol{a}_2$ 의 사영(projection)이라고 부른다.

그러면 이러한 사영은 어떻게 구할 수 있나? 벡터 $\boldsymbol{a}_2$ 의 사영은 벡터 $\boldsymbol{a}_1$ 방향에 있으므로 어떤 실수 c 가 있어서 다음과 같이 표시할 수 있다.

$$proj_{\boldsymbol{a}_1}(\boldsymbol{a}_2) = c\boldsymbol{a}_1$$

이제 사영 $c\boldsymbol{a}_1$과 벡터 $\boldsymbol{a}_2$의 거리 $d(c)$ 를 생각하면 다음과 같다.

$$\begin{aligned} d^2(c) &= \|\boldsymbol{a}_2 - c\boldsymbol{a}_1\|^2 \\ &= (\boldsymbol{a}_2 - c\boldsymbol{a}_1)^t(\boldsymbol{a}_2 - c\boldsymbol{a}_1) \\ &= \boldsymbol{a}_2^t\boldsymbol{a}_2 - 2c\boldsymbol{a}_2^t\boldsymbol{a}_1 + c^2\boldsymbol{a}_1^t\boldsymbol{a}_1 \end{aligned}$$

위에서 $\|\boldsymbol{a}\|$ 는 벡터 $\boldsymbol{a}$의 길이를 나타낸다.

$$d(\boldsymbol{a}) = \|\boldsymbol{a}\| = \sqrt{\boldsymbol{a}^t\boldsymbol{a}}$$

상수 c 는 거리 $d(c)$를 최소로 만드는 수이다. $d^2(c)$은 c 에 대하여 미분 가능한 2차 함수이며 아래로 볼록한 함수이므로 이를 미분하여 c 를 구할 수 있다.

$$\frac{\partial d^2(c)}{\partial x}c = -2\boldsymbol{a}_2^t\boldsymbol{a}_1 + 2c\boldsymbol{a}_1^t\boldsymbol{a}_1 = 0$$

위의 방적식으로 부터 c를 얻고

$$c = \frac{\boldsymbol{a}_2^t\boldsymbol{a}_1}{\boldsymbol{a}_1^t\boldsymbol{a}_1}$$

다음과 같이 벡터 $\boldsymbol{a}_1$ 방향으로 벡터 $\boldsymbol{a}_2$ 의 사영을 나타낼 수 있다.

$$proj_{\boldsymbol{a}_1}(\boldsymbol{a}_2) = \frac{\boldsymbol{a}_2^t\boldsymbol{a}_1}{\boldsymbol{a}_1^t\boldsymbol{a}_1}\boldsymbol{a}_1 \tag{B.9}$$

이제 위의 두 벡터의 사영을 이용하면 벡터 $\boldsymbol{a}_1$ 과 직교하는 벡터 $\tilde{\boldsymbol{q}}_2$를 다음과 같이 찾을 수 있다.

$$\tilde{\boldsymbol{q}}_2 = \boldsymbol{a}_2 - proj_{\boldsymbol{a}_1}(\boldsymbol{a}_2) = \boldsymbol{a}_2 - \frac{\boldsymbol{a}_2^t\boldsymbol{a}_1}{\boldsymbol{a}_1^t\boldsymbol{a}_1}\boldsymbol{a}_1$$

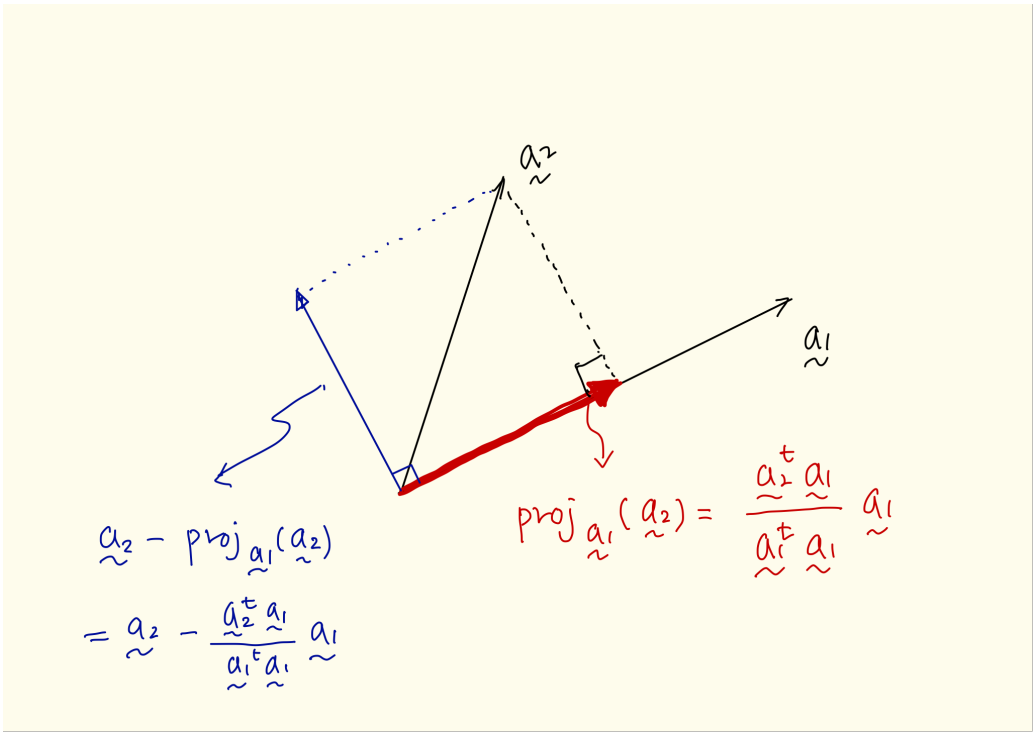


그림 B.1.: 벡터의 사영

두 벡터 $\boldsymbol{a}_1$와 $\tilde{\boldsymbol{q}}_2$의 직교성은 다음과 같이 보일 수 있다.

$$\begin{aligned}
\boldsymbol{a}_1^t\tilde{\boldsymbol{q}}_2 &= \boldsymbol{a}_1^t\left(\boldsymbol{a}_2 - \frac{\boldsymbol{a}_2^t\boldsymbol{a}_1}{\boldsymbol{a}_1^t\boldsymbol{a}_1}\boldsymbol{a}_1\right) \\
&= \boldsymbol{a}_1^t\boldsymbol{a}_2 - \frac{\boldsymbol{a}_2^t\boldsymbol{a}_1}{\boldsymbol{a}_1^t\boldsymbol{a}_1}\boldsymbol{a}_1^t\boldsymbol{a}_1 \\
&= \boldsymbol{a}_1^t\boldsymbol{a}_2 - \boldsymbol{a}_2^t\boldsymbol{a}_1 \\
&= 0
\end{aligned}$$

이제 두 벡터 $\boldsymbol{q}_1$과 $\boldsymbol{q}_2$를 다음과 같이 정규직교벡터로 만들 수 있다.

$$\begin{aligned}
\boldsymbol{q}_1 &= \boldsymbol{a}_1/\left\|\boldsymbol{a}_1\right\| \\
\boldsymbol{q}_2 &= \tilde{\boldsymbol{q}}_2/\left\|\tilde{\boldsymbol{q}}_2\right\|
\end{aligned}$$

B.9. 최소제곱법과 사영

회귀계수벡터의 값을 구하는 최소제곱법의 기준을 다시 살펴보자.

$$\min_{\boldsymbol{\beta}}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^t(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})$$

위에서 $\boldsymbol{X}\boldsymbol{\beta}$는 행렬 $\boldsymbol{X}$의 열벡터 $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_p$로 이루어진 선형조합이다.

$$\boldsymbol{X\beta} = [\boldsymbol{x}_1 \ \cdots \boldsymbol{x}_p]\boldsymbol{\beta} = \beta_1\boldsymbol{x}_1 + \cdots + \beta_p\boldsymbol{x}_p$$

행렬 $\boldsymbol{X}$의 열벡터로 생성된 공간을 $C(\boldsymbol{X})$ 라고 하자

$$C(\boldsymbol{X}) = span\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p\} = \{\boldsymbol{X\beta} \mid \boldsymbol{\beta} \in \mathbb{R}^p\}$$

따라서 최소제곱법으로 구한 회귀계수 벡터 $\hat{\boldsymbol{\beta}}$는 반응값 벡터 $\boldsymbol{y}$ 와 $\boldsymbol{X\beta}$의 거리가 최소가 되도록 만들어 준다.

$$\min_{\boldsymbol{\beta}}(\boldsymbol{y} - \boldsymbol{X\beta})^t(\boldsymbol{y} - \boldsymbol{X\beta}) = (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^t(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})$$

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{y}$$

따라서 예측값 벡터 $\hat{\boldsymbol{y}}$ 는 행렬 $\boldsymbol{X}$의 열벡터로 생성한 열공간 $C(\boldsymbol{X})$ 방향으로 반응값 벡터 $\boldsymbol{y}$를 사영한 벡터이다.

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{y}$$

위에서 행렬 $\boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t$를 열공간 $C(\boldsymbol{X})$의 사영행렬(projection matrix)라고 부른다. 사영행렬의 정의는 Appendix G 에서 공부한다.

$$\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t \tag{B.10}$$

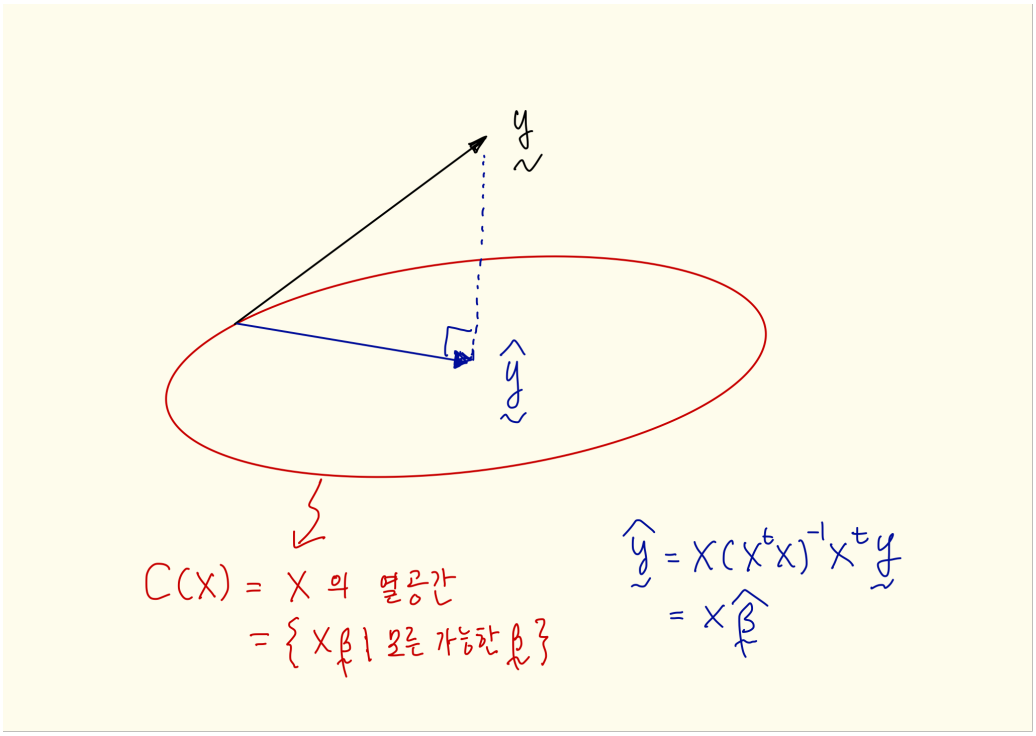


그림 B.2.: 최소제곱법을 설명한 그림

이제 계획행렬 $\boldsymbol{X}$ 두 개로 분할하는 경우를 생각해 보자.

$$\boldsymbol{X} = [\boldsymbol{X}_1, \boldsymbol{X}_2]$$

이 경우 계획행렬 $\boldsymbol{X}$ 과 계획행렬 $\boldsymbol{X}_1$ 을 사용한 사영행렬을 정의하면 다음과 같다.

$$\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t, \quad \boldsymbol{H}_1 = \boldsymbol{X}_1(\boldsymbol{X}_1^t\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^t$$

이제 두 개의 사영행렬 $\boldsymbol{H}$와 $\boldsymbol{H}_1$을 이용하여 반응값 벡터 $\boldsymbol{y}$를 사영하면 다음과 같다.

$$\hat{\boldsymbol{y}} = \boldsymbol{H}\boldsymbol{y}, \quad \hat{\boldsymbol{y}_1} = \boldsymbol{H}_1\boldsymbol{y}$$

두 개의 예측값 벡터 $\hat{\boldsymbol{y}}$와 $\hat{\boldsymbol{y}_1}$ 는 기하학적으로 어떤 관계에 있을까?

예를 들어서 Chapter 3 에서 설명한 부분-F 검정에서 사용한 두 개의 모형 3.13 과 3.14 을 생각해 보자. 두 개의 계획행렬 $\boldsymbol{X}$ 과 $\boldsymbol{X}_1$ 이 만드는 열공간은 다음과 같은 포함관계가 성립한다.

$$C(\boldsymbol{X}_1) \subset C(\boldsymbol{X})$$

따라서 두 개의 사영행렬 $\boldsymbol{H}$와 $\boldsymbol{H}_1$ 을 연속하여 적용한 다음과 같은 관계가 성립한다.

$$\boldsymbol{H}\boldsymbol{H}_1 = \boldsymbol{H}_1\boldsymbol{H} = \boldsymbol{H}_1$$

이 사실은 다음과 같은 그림으로 설명할 수 있다.

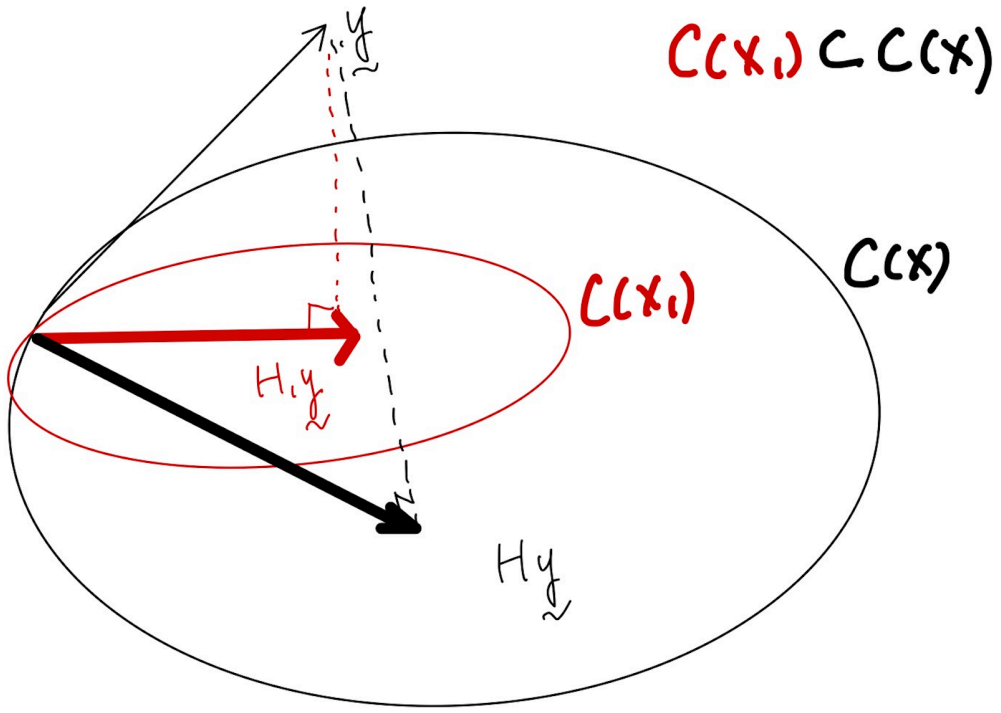


그림 B.3.: 포함관계에 있는 열공가에 대한 사영의 관계

이 경우 $\boldsymbol{H}\boldsymbol{X}_1 = \boldsymbol{X}_1$ 임도 알 수 있다. 왜냐하면 $\boldsymbol{H}$는 열공간 $C(\boldsymbol{X})$ 방향으로 사영하는데 $\boldsymbol{X}_1$는 $C(\boldsymbol{X})$의 부분집합이기 때문이다. 즉 $\boldsymbol{H}$는 $\boldsymbol{X}_1$을 그대로 유지한다.

C. 고유값과 고유벡터

> Caution
>
> 고유값과 고유벡터는 정방행렬(square matrix)에 대해서만 정의된다.

C.1. 특성다항식

특성다항식(Characteristic polynomial)은 다음과 같이 정의된다

실수 $\lambda \in \mathbb{R}$ 와 정방행렬(square matrix) $A \in \mathbb{R}^{n\times n}$ 에 대하여

$$\begin{aligned} p_A(\lambda) &:= \det(A - \lambda I) \\ &= c_0 + c_1\lambda + c_2\lambda^2 + \cdots + c_{n-1}\lambda^{n-1} + (-1)^n\lambda^n, \end{aligned} \tag{C.1}$$

C.2. 고유값과 고유벡터

C.2.1. 정의

n-차원 정방행렬 $\boldsymbol{A}$ 이 있을 때, 다음 식을 만족하는 λ 와 벡터 $\boldsymbol{x}$가 존재하면 λ 를 행렬 $\boldsymbol{A}$ 의 고유값(eigenvalue), $\boldsymbol{x}$ 를 행렬 $\boldsymbol{A}$ 의 고유벡터(eigenvector)라고 한다 (부교재 definition 4.6)

$$\boldsymbol{A}\boldsymbol{x} = \lambda\boldsymbol{x} \tag{C.2}$$

- 고유벡터는 유일하지 않다. 즉, 벡터 $\boldsymbol{x}$ 가 고유벡터이면 $c\boldsymbol{x}$ 도 고유벡터이다.

$$\boldsymbol{A}(c\boldsymbol{x}) = c\boldsymbol{A}\boldsymbol{x} = c\lambda\boldsymbol{x} = \lambda(c\boldsymbol{x})$$

C.2.2. 계산

다음 4개의 문장은 동치이다

- λ 는 행렬 $\boldsymbol{A}$ 의 고유값이다.
- 방정식 $(\boldsymbol{A}-\lambda\boldsymbol{I})\boldsymbol{x}=\boldsymbol{0}$ 은 영벡터이외의 해 $\boldsymbol{x}$ 를 가진다(nontrivial solution)
- 행렬 $\boldsymbol{A}-\lambda\boldsymbol{I}$ 의 행렬식이 0 이다.

$$\det(\boldsymbol{A}-\lambda\boldsymbol{I})=0 \tag{C.3}$$

- 행렬 $\boldsymbol{A}-\lambda\boldsymbol{I}$ 의 rank가 n 보다 작다.

위에서 행렬식이 0 인 방정식 C.3 을 푸는 것은 특성방정식 C.1 이 0 인 방정식을 을 푸는 것과 동일하다.

Exercise C.1. 다음과 같은 2×2 행렬의 고유값과 고유행렬을 구해보자.

$$\boldsymbol{A}=\begin{bmatrix}4 & 2\\ 1 & 3\end{bmatrix}$$

다음과 같이 행렬 $\boldsymbol{A}$ 에 대한 특성다항식을 이용하여 고유값을 구할 수 있다.

$$\begin{aligned}p_{\boldsymbol{A}}(\lambda)&=\det(\boldsymbol{A}-\lambda\boldsymbol{I})\\ &=\det\left(\begin{bmatrix}4 & 2\\ 1 & 3\end{bmatrix}-\begin{bmatrix}\lambda & 0\\ 0 & \lambda\end{bmatrix}\right)=\begin{vmatrix}4-\lambda & 2\\ 1 & 3-\lambda\end{vmatrix}\\ &=(4-\lambda)(3-\lambda)-(2)(1)\\ &=(2-\lambda)(5-\lambda)\end{aligned}$$

위의 방정식에서 λ에 대한 다항식 $p_{\boldsymbol{A}}(\lambda)=0$ 의 근 을 구하면 고유값을 구할 수 있다. 따라서 행렬 $\boldsymbol{A}$ 의 고유값은 $\lambda_1=2$ 와 $\lambda_2=5$ 이다.

이제 각각의 고유값에 대한 고유행렬을 다음과 같이 고유벡터의 정의 C.2 에 의하여 구해보자.

$$\begin{bmatrix}4 & 2\\ 1 & 3\end{bmatrix}\boldsymbol{x}=\lambda\boldsymbol{x}\quad\rightarrow\quad\begin{bmatrix}4-\lambda & 2\\ 1 & 3-\lambda\end{bmatrix}\boldsymbol{x}=\boldsymbol{0}$$

먼저, 고유값 $\lambda_2=5$ 고유행렬은 다음과 같이 정의된다.

$$\begin{bmatrix}4-5 & 2\\ 1 & 3-5\end{bmatrix}\begin{bmatrix}x_1\\ x_2\end{bmatrix}=\begin{bmatrix}-1 & 2\\ 1 & -2\end{bmatrix}\begin{bmatrix}x_1\\ x_2\end{bmatrix}=\boldsymbol{0}\quad\rightarrow x_1-2x_2=0$$

이제 위의 방정식을 만족하는 고유벡터는 다음과 같이 구할 수 있다.

$$\boldsymbol{x}_2=\begin{bmatrix}2\\ 1\end{bmatrix}$$

유의할 점은 고유벡터는 방정식을 만족하는 무수히 많은 벡터 중에 하나의 예일 뿐이다. 예를 들어 길이가 1 인 단위벡터(unit vector)인 고유벡터를 구하고 싶다면 위의 벡터를 길이가 1인 단위벡터로 바꾸면 된다.

$$\boldsymbol{x}_2 = \begin{bmatrix} 2/\sqrt{5} \\ 1/\sqrt{5} \end{bmatrix}$$

또한 $\lambda_1 = 2$ 에 대한 고유벡터는 다음과 같이 단위벡터로 구할 수 있다.

$$\begin{bmatrix} 4-2 & 2 \\ 1 & 3-2 \end{bmatrix} \boldsymbol{x} = \begin{bmatrix} 2 & 2 \\ 1 & 1 \end{bmatrix} \boldsymbol{x} = \boldsymbol{0} \quad \rightarrow x_1 + x_2 = 0$$

$$\boldsymbol{x}_1 = \begin{bmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{bmatrix}$$

■

C.2.3. 중복도와 고유공간

- 대수적 중복도(algebraic multiplicity) 는 특성다항식 C.1 이 0인 방정식을 푸는 경우 다항식에서 고유값이 중근(multiple root)의 해로 나타나는 차수를 의미한다.
- 기하적 중복도(geometric multiplicity) 는 고유값에 대응하는 고유벡터들 중 선형독립인 고유벡터들의 최대 개수를 의미한다.
- 고유 공간(eigenspace)은 고유값에 대응하는 고유벡터들이 생성하는 벡터공간을 의미한다.

Exercise C.2. 3차원 행렬 $\boldsymbol{A}$ 가 다음과 같을 때

$$\boldsymbol{A} = \begin{bmatrix} 0 & 0 & -2 \\ 1 & 2 & 1 \\ 1 & 0 & 3 \end{bmatrix}$$

행렬 $\boldsymbol{A}$의 특성다항식은 다음과 같다.

$$\det(\lambda \boldsymbol{I} - \boldsymbol{A}) = \begin{vmatrix} \lambda & 0 & 2 \\ -1 & \lambda - 2 & -1 \\ -1 & 0 & \lambda - 3 \end{vmatrix} = (\lambda - 1)(\lambda - 2)^2$$

참고로 특성방정식을 푸는 경우, 방정식 $\det(\boldsymbol{A} - \lambda \boldsymbol{I}) = 0$ 이나 $\det(\lambda \boldsymbol{I} - \boldsymbol{A}) = 0$ 중 어느 것을 사용해도 상관없다.

위의 식에서 3차원 행렬의 행렬식은 다음과 같이 구할 수 있다. 첫 번째 행을 기준으로 전개하면

$$\begin{aligned}\det(\lambda \boldsymbol{I} - \boldsymbol{A}) &= -\lambda \cdot \begin{vmatrix} 2-\lambda & 1 \\ 0 & 3-\lambda \end{vmatrix} + 0 \cdot \begin{vmatrix} 1 & 1 \\ 1 & 3-\lambda \end{vmatrix} + (-2) \cdot \begin{vmatrix} 1 & 2-\lambda \\ 1 & 0 \end{vmatrix} \\ &= -\lambda[(2-\lambda)(3-\lambda) - 0] - 2[(1)(0) - (1)(2-\lambda)] \\ &= (\lambda - 1)(\lambda - 2)^2\end{aligned}$$

첫번째 고유값은 $\lambda_1 = 1$ 이다. 고유벡터를 구하기 위해서는 다음과 같은 방정식을 풀면 된다.

$$(\lambda_1 \boldsymbol{I} - \boldsymbol{A})\boldsymbol{x} = \boldsymbol{0}$$

위의 방정식을 풀면

$$(\lambda_1 \boldsymbol{I} - \boldsymbol{A})\boldsymbol{x} = (\boldsymbol{I} - \boldsymbol{A})\boldsymbol{x} = \begin{bmatrix} 1 & 0 & 2 \\ -1 & -1 & -1 \\ -1 & 0 & -2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

아래와 같이 간단히 할 수 있으며

$$x_1 = -2x_3, \quad x_2 = x_3$$

다음과 같은 고유값과 고유벡터를 얻을 수 있다.

$$\lambda_1 = 1 \quad \rightarrow \quad \boldsymbol{x}_1 = \begin{bmatrix} -2 \\ 1 \\ 1 \end{bmatrix}$$

첫번째 고유값은 $\lambda_1 = 1$ 이며 대수적 중복도는 1이고 기하적 중복도도 1이다. 기하적 중복도도 1이란 의미는 고유벡터가 선형독립인 1개의 벡터로 이루어져 있다는 의미이다. 이 경우 고유공간 E_1 은 한 개의 고유벡터 $\boldsymbol{x}_1$ 이 생성하는 부분공간을 의미한다.

$$E_1 = \text{span}\left\{ \begin{bmatrix} -2 \\ 1 \\ 1 \end{bmatrix} \right\}$$

다음으로 두번째 고유값에 대한 방정식 $(\lambda_2 \boldsymbol{I} - \boldsymbol{A})\boldsymbol{x} = \boldsymbol{0}$ 을 풀면 다음과 같다.

$$(\lambda_2 \boldsymbol{I} - \boldsymbol{A})\boldsymbol{x} = (2\boldsymbol{I} - \boldsymbol{A})\boldsymbol{x} = \begin{bmatrix} 2 & 0 & 2 \\ -1 & 0 & -1 \\ -1 & 0 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

이 방정식은 아래와 같이 간단히 할 수 있으며

$$x_1 = -x_3$$

주어진 방정식이 하나이기 때문에 다음과 같이 서로 선형독립인 두 개의 고유벡터를 얻을 수 있다.

$$\lambda_2 = 2 \quad \rightarrow \quad \boldsymbol{x}_2 = \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix} \quad \boldsymbol{x}_3 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$$

위에서 **두번째 고유값은 $\lambda_2 = 2$ 이며 대수적 중복도는 2이다.** 또한 **선형독립인 2개의 고유벡터를 구할 수 있으므로 기하적 중복도는 2이다.**

이 경우 E_2 는 두 개의 고유벡터 $\boldsymbol{x}_2, \boldsymbol{x}_3$ 가 생성하는 부분공간을 의미한다.

$$E_2 = \text{span} \left\{ \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \right\}$$

■

이제 대수적 중복도와 기하적 중복도가 다른 경우에 대한 예제를 들어보자.

Exercise C.3. 3차원 행렬 $\boldsymbol{A}$ 가 다음과 같을 때

$$\boldsymbol{A} = \begin{bmatrix} 1 & 0 & 2 \\ -1 & 1 & 3 \\ 0 & 0 & 2 \end{bmatrix}$$

행렬 $\boldsymbol{A}$의 특성다항식은 다음과 같다.

$$\det(\lambda \boldsymbol{I} - \boldsymbol{A}) = \begin{vmatrix} \lambda - 1 & 0 & -2 \\ 1 & \lambda - 1 & -3 \\ 0 & 0 & \lambda - 2 \end{vmatrix} = (\lambda - 1)^2 (\lambda - 2)$$

첫번째 고유값은 $\lambda_1 = 1$ 이다. 고유벡터를 구하기 위해서는 다음과 같은 방정식을 풀면 된다.

$$(\lambda_1 \boldsymbol{I} - \boldsymbol{A})\boldsymbol{x} = \boldsymbol{0}$$

위의 방정식을 풀면

$$(\lambda_1 \boldsymbol{I} - \boldsymbol{A})\boldsymbol{x} = (\boldsymbol{I} - \boldsymbol{A})\boldsymbol{x} = \begin{bmatrix} 0 & 0 & -2 \\ 1 & 0 & -3 \\ 0 & 0 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

아래와 같이 간단히 할 수 있으며

$$x_1 = x_3 = 0$$

다음과 같은 하나의 고유벡터를 얻을 수 있다.

$$\lambda_1 = 1 \quad \rightarrow \quad x_1 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$$

첫번째 고유값은 $\lambda_1 = 1$ 이며 대수적 중복도는 2이지만 기하적 중복도는 1이다. 이 경우 고유공간 E_1 은 한 개의 고유벡터 $\boldsymbol{x}_1$ 이 생성하는 부분공간을 의미한다.

$$E_1 = \text{span} \left\{ \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \right\}$$

다음으로 두번째 고유값에 대한 방정식 $(\lambda_2 \boldsymbol{I} - \boldsymbol{A})\boldsymbol{x} = \boldsymbol{0}$ 을 풀면 다음과 같다.

$$(\lambda_2 \boldsymbol{I} - \boldsymbol{A})\boldsymbol{x} = (2\boldsymbol{I} - \boldsymbol{A})\boldsymbol{x} = \begin{bmatrix} 1 & 0 & 2 \\ 1 & 1 & -3 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

이 방정식은 아래와 같이 간단히 할 수 있으며

$$x_1 = -2x_3, \quad x_2 = 5x_3$$

다음과 같은 한 개의 고유벡터를 얻을 수 있다.

$$\lambda_2 = 2 \quad \rightarrow \quad x_2 = \begin{bmatrix} -2 \\ 5 \\ 1 \end{bmatrix}$$

위에서 **두번째 고유값은 $\lambda_2 = 2$ 이며 대수적 중복도는 1이다.** 또한 **선형독립인 1개의 고유벡터를 구할 수 있으므로 기하적 중복도는 1이다.**

이 경우 E_2 는 한 개의 고유벡터 $\boldsymbol{x}_2$ 가 생성하는 부분공간을 의미한다.

$$E_2 = \text{span}\left\{ \begin{bmatrix} -2 \\ 5 \\ 1 \end{bmatrix} \right\}$$

■

C.3. 대칭행렬의 대각화

Exercise C.4. 이제 대칭행렬에 대한 고유값과 고유행렬을 구해보자. 대칭행렬은 고유값이 실수이고 서로 다르며, 서로 직교하는 고유벡터를 가진다.

$$\boldsymbol{A} = \tfrac{1}{2} \begin{bmatrix} 5 & -2 \\ -2 & 5 \end{bmatrix}$$

행렬 $\boldsymbol{A}$ 의 특성다항식은 다음과 같이 구할 수 있다.

$$\begin{aligned} \det(\boldsymbol{A} - \lambda \boldsymbol{I}) &= \det\left(\begin{bmatrix} \frac{5}{2} - \lambda & -1 \\ -1 & \frac{5}{2} - \lambda \end{bmatrix} \right) \\ &= \left(\frac{5}{2} - \lambda \right)^2 - 1 = \lambda^2 - 5\lambda + \frac{21}{4} = \left(\lambda - \frac{7}{2} \right) \left(\lambda - \frac{3}{2} \right) \end{aligned}$$

따라서 행렬 $\boldsymbol{A}$ 의 고유값은 각각 $\lambda_1 = \frac{7}{2}$ 과 $\lambda_2 = \frac{3}{2}$ 이며 대응하는 고유벡터 $\boldsymbol{p}_1$ 과 $\boldsymbol{p}_2$ 는 다음과 같이 구할 수 있다.

$$\boldsymbol{A}\boldsymbol{p}_1 = \frac{7}{2}\boldsymbol{p}_1, \quad \boldsymbol{A}\boldsymbol{p}_2 = \frac{3}{2}\boldsymbol{p}_2$$

위의 고유빅터에 대한 방정식을 풀어서 길이가 1인 고유벡터를 구하면 다음과 같다.

$$\boldsymbol{p}_1 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ -1 \end{bmatrix}, \quad \boldsymbol{p}_2 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

이제 고유벡터는 서로 직교하는 단위벡터임을 알 수 있다.

$$\boldsymbol{p}_1^t \boldsymbol{p}_2 = 0$$

이제 고유값과 고유벡터를 구했으니 대칭행렬 $\boldsymbol{A}$ 를 대각화 해보자. 대칭행렬 $\boldsymbol{A}$ 는 다음과 같이 대각화 할 수 있다. 먼저 두 고유벡터를 열벡터로 하는 행렬 $\boldsymbol{P}$ 를 정의하자. 행렬 $\boldsymbol{P}$ 는 서로 직교하는 벡터로 구성되었으므로 직교행렬이다.

$$\boldsymbol{P} = [\boldsymbol{p}_1, \boldsymbol{p}_2] = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} \quad \rightarrow \quad \boldsymbol{P}\boldsymbol{P}^t = \boldsymbol{P}^t\boldsymbol{P} = \boldsymbol{I}$$

이제 $\boldsymbol{P}$ 의 역행렬은 $\boldsymbol{P}^t$ 이므로 다음과 같이 대칭행렬 $\boldsymbol{A}$의 대각화를 유도할 수 있다.

$$\boldsymbol{P}^{-t}\boldsymbol{A}\boldsymbol{P} = \begin{bmatrix} \frac{7}{2} & 0 \\ 0 & \frac{3}{2} \end{bmatrix} = \boldsymbol{D}.$$

위의 식은 다음과 같이 쓸수 있다.

$$\underbrace{\frac{1}{2}\begin{bmatrix} 5 & -2 \\ -2 & 5 \end{bmatrix}}_{\boldsymbol{A}} = \underbrace{\frac{1}{\sqrt{2}}\begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}}_{\boldsymbol{P}} \underbrace{\begin{bmatrix} \frac{7}{2} & 0 \\ 0 & \frac{3}{2} \end{bmatrix}}_{\boldsymbol{D}} \underbrace{\frac{1}{\sqrt{2}}\begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}}_{\boldsymbol{P}^t}$$

또한 위의 식은 다음과 같이 나타낼 수 있다.

$$\begin{aligned} \boldsymbol{A} &= \frac{1}{2}\begin{bmatrix} 5 & -2 \\ -2 & 5 \end{bmatrix} \\ &= \lambda_1 \boldsymbol{p}_1 \boldsymbol{p}_1^t + \lambda_2 \boldsymbol{p}_2 \boldsymbol{p}_2^t \\ &= \frac{7}{4}\begin{bmatrix} 1 \\ -1 \end{bmatrix}\begin{bmatrix} 1 & -1 \end{bmatrix} + \frac{3}{4}\begin{bmatrix} 1 \\ 1 \end{bmatrix}\begin{bmatrix} 1 & 1 \end{bmatrix} \\ &= \frac{7}{4}\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} + \frac{3}{4}\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \end{aligned}$$

■

D. 행렬의 분해

D.0.1. Gram–Schmidt 방법

서로 독립인 n차원의 벡터들이 p개 있을때

$$\boldsymbol{a}_1, \boldsymbol{a}_2, \dots, \boldsymbol{a}_p$$

이들이 만드는 열공간을 C 라고 하자.

$$\begin{aligned} C &= span\{\boldsymbol{a}_1, \boldsymbol{a}_2, \dots, \boldsymbol{a}_p\} \\ &= \{\ c_1\boldsymbol{a}_1 + c_2\boldsymbol{a}_2 + \cdots + c_p\boldsymbol{a}_p \mid \text{ all possible real values of } c_1, c_2, \dots, c_p\ \} \end{aligned}$$

이제 우리는 위와 동일한 열공간 C 만드는 정규직교 벡터들을 찾는 방법을 알아보고자 한다.

$$\boldsymbol{q}_1, \boldsymbol{q}_2, \dots, \boldsymbol{q}_p \quad \text{where } \boldsymbol{q}_i^t\boldsymbol{q}_j = 0,\ \ \boldsymbol{q}_i^t\boldsymbol{q}_i = 1$$

그리고

$$C = span\{\boldsymbol{q}_1, \boldsymbol{q}_2, \dots, \boldsymbol{q}_p\} = span\{\boldsymbol{a}_1, \boldsymbol{a}_2, \dots, \boldsymbol{a}_p\} \tag{D.1}$$

이제 앞 절의 벡터의 사영에 대한 결과를 사용하여 다음과 같은 직교하는 p 개의 벡터들을 축차적으로 만들어 보자.

$$\begin{aligned} \tilde{\boldsymbol{q}}_1 &= \boldsymbol{a}_1 \\ \tilde{\boldsymbol{q}}_2 &= \boldsymbol{a}_2 - proj_{\tilde{\boldsymbol{q}}_1}(\boldsymbol{a}_2) \\ \tilde{\boldsymbol{q}}_3 &= \boldsymbol{a}_3 - proj_{\tilde{\boldsymbol{q}}_1}(\boldsymbol{a}_3) - proj_{\tilde{\boldsymbol{q}}_2}(\boldsymbol{a}_3) \\ \tilde{\boldsymbol{q}}_4 &= \boldsymbol{a}_4 - proj_{\tilde{\boldsymbol{q}}_1}(\boldsymbol{a}_4) - proj_{\tilde{\boldsymbol{q}}_2}(\boldsymbol{a}_4) - proj_{\tilde{\boldsymbol{q}}_3}(\boldsymbol{a}_4) \\ &\dots \\ \tilde{\boldsymbol{q}}_p &= \boldsymbol{a}_p - \sum_{k=1}^{p} proj_{\tilde{\boldsymbol{q}}_k}(\boldsymbol{a}_p) \end{aligned}$$

축차적으로 만든 벡터들을 정규벡터로 만들면 원래의 벡터들 $\boldsymbol{a}_1, \boldsymbol{a}_2, \dots, \boldsymbol{a}_p$이 생성하는 동일한 열공간을 만드는 정규직교 벡터 $\boldsymbol{q}_1, \boldsymbol{q}_2, \dots, \boldsymbol{q}_p$를 만들 수 있다.

$$\boldsymbol{q}_i = \tilde{\boldsymbol{q}}_i / \left\|\tilde{\boldsymbol{q}}_i\right\|, \quad i = 1, 2, \dots, p \tag{D.2}$$

Gram–Schmidt 방법으로 만든 벡터들이 직교하는 것은 다음과 같이 증명할 수 있다. 먼저 B.10 에 의하여 $\tilde{\boldsymbol{q}}_1$ 과 $\tilde{\boldsymbol{q}}_2$는 직교한다. 이제 임의의 i에 대하여 $\tilde{\boldsymbol{q}}_1, \tilde{\boldsymbol{q}}_2, \cdots, \tilde{\boldsymbol{q}}_{i-1}$ 벡터들이 직교한다고 가정하자. 모든 $1 \le j \le i-1$ 에 대하여

$$\begin{aligned}
\tilde{\boldsymbol{q}}_j^t \tilde{\boldsymbol{q}}_i &= \tilde{\boldsymbol{q}}_j^t \left[\boldsymbol{a}_i - \sum_{k=1}^{i-1} proj_{\tilde{\boldsymbol{q}}_k}(\boldsymbol{a}_i) \right] \\
&= \tilde{\boldsymbol{q}}_j^t \left[\boldsymbol{a}_i - proj_{\tilde{\boldsymbol{q}}_j}(\boldsymbol{a}_i) \right] - \left[\sum_{\substack{1 \le k \le i-1 \\ k \ne j}} \tilde{\boldsymbol{q}}_j^t \, proj_{\tilde{\boldsymbol{q}}_k}(\boldsymbol{a}_i) \right] \\
&= 0 + 0
\end{aligned}$$

위에서 마지막 단계의 직교성은 다음과 같은 사실로 부터 유도된다.

- $\boldsymbol{a}_i - proj_{\tilde{\boldsymbol{q}}_j}(\boldsymbol{a}_i)$는 $\tilde{\boldsymbol{q}}_j^t$와 직교한다.
- 가정에 의하여 $\tilde{\boldsymbol{q}}_1, \tilde{\boldsymbol{q}}_2, \cdots, \tilde{\boldsymbol{q}}_{i-1}$ 는 직교하고 $proj_{\tilde{\boldsymbol{q}}_k}(\boldsymbol{a}_i)$ 는 $\tilde{\boldsymbol{q}}_k$ 와 같은 방향을 가진다.

$$\tilde{\boldsymbol{q}}_j^t proj_{\tilde{\boldsymbol{q}}_k}(\boldsymbol{a}_i) = 0 \quad \text{for } 1 \le j, k \le i-1, k \ne j$$

D.2 과 D.2 의 알고리즘을 Gram–Schmidt 방법이라고 부른다. 위의 두 식에 의한 알고리즘을 다음과 같은 사실을 이용하면 좀 더 간단한 방법의 알고리즘이 나온다.

$$proj_{\tilde{\boldsymbol{q}}_k}(\boldsymbol{a}_l) = \frac{\boldsymbol{a}_l^t \tilde{\boldsymbol{q}}_k}{\tilde{\boldsymbol{q}}_k^t \tilde{\boldsymbol{q}}_k} \tilde{\boldsymbol{q}}_k = \frac{\boldsymbol{a}_l^t \tilde{\boldsymbol{q}}_k}{\left\|\tilde{\boldsymbol{q}}_k\right\|^2} \tilde{\boldsymbol{q}}_k = (\boldsymbol{a}_l^t \boldsymbol{q}_k) \boldsymbol{q}_k$$

1. p개의 벡터 $\boldsymbol{a}_1, \boldsymbol{a}_2, \dots, \boldsymbol{a}_p$에 대하여
2. for $i = 1, 2, \dots, p$
    - $\tilde{\boldsymbol{q}}_i = \boldsymbol{a}_i - (\boldsymbol{q}_1^t \boldsymbol{a}_i)\boldsymbol{q}_1 - \cdots - (\boldsymbol{q}_{i-1}^t \boldsymbol{a}_i)\boldsymbol{q}_{i-1}$ (직교화)
    - $\boldsymbol{q}_i = \tilde{\boldsymbol{q}}_i / \left\|\boldsymbol{q}_i\right\|$ (정규화)

다음은 Gram–Schmidt 방법을 설명한 그림이다.

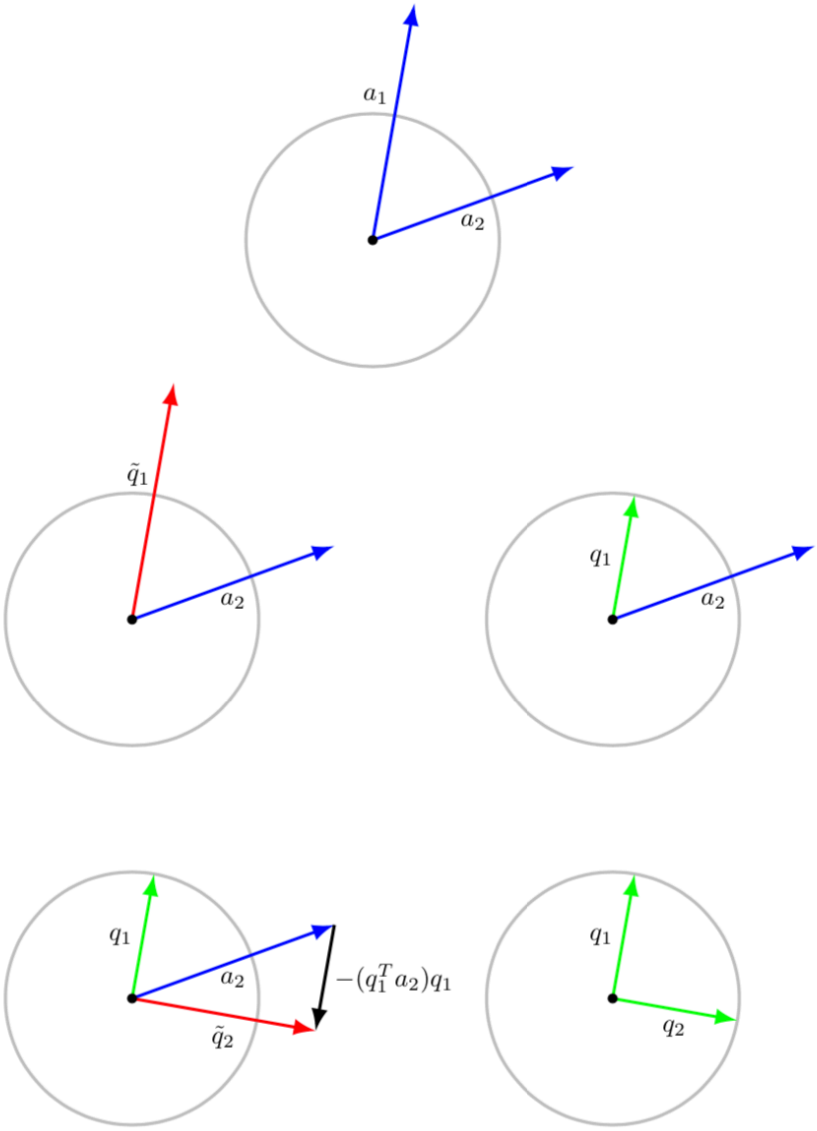


Figure 5.3 Gram–Schmidt algorithm applied to two 2-vectors a_1, a_2. *Top.* The original vectors a_1 and a_2. The gray circle shows the points with norm one. *Middle left.* The orthogonalization step in the first iteration yields $\tilde{q}_1 = a_1$. *Middle right.* The normalization step in the first iteration scales $\tilde{q}_1$ to have norm one, which yields q_1. *Bottom left.* The orthogonalization step in the second iteration subtracts a multiple of q_1 to yield the vector $\tilde{q}_2$, which is orthogonal to q_1. *Bottom right.* The normalization step in the second iteration scales $\tilde{q}_2$ to have norm one, which yields q_2.

그림 D.1.: Gram–Schmidt 방법(출처:Introduction to Applied Linear Algebra by Boyd and Vandenberghe, 2019)

D.0.2. LU 분해

정방행렬 $\boldsymbol{A}$를 다음과 같이 하삼각행렬 $\boldsymbol{L}$과 상삼각행렬 $\boldsymbol{U}$의 곱으로 나타내는 것을 LU 분해라고 한다.

$$\boldsymbol{A} = \boldsymbol{LU}$$

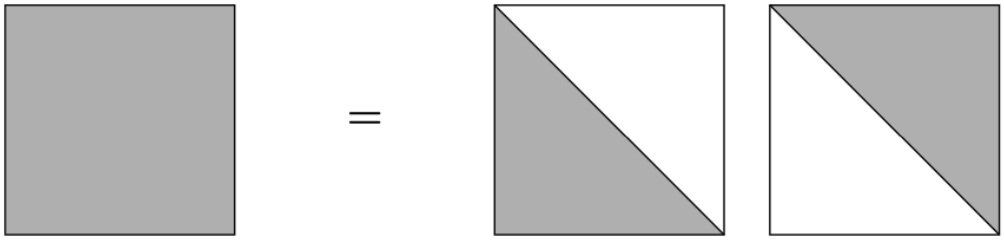


그림 D.2.: LU 분해

이러한 LU 분해는 행렬 $\boldsymbol{A}$에 행연산을 적용하여 쉽게 구할 수 있다. 예를 들어 위에서 고려한 2×2 행렬에 행연산을 적용하여 대각원소 아래를 0으로 만들면 LU 분해를 쉽게 유도할 수 있다.

$$\begin{bmatrix} 1 & 0 \\ -3 & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} = \begin{bmatrix} 1 & 2 \\ 0 & -2 \end{bmatrix}$$

따라서

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 3 & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 0 & -2 \end{bmatrix} = \boldsymbol{LU}$$

D.0.3. QR 분해

D.2 과 D.2 에 주어진 Gram–Schmidt 방법을 원래 벡터들 $\boldsymbol{a}_1, \boldsymbol{a}_2, \dots, \boldsymbol{a}_p$에 대하여 다시 다음과 같이 나타낼 수 있다.

$$\begin{aligned}
\boldsymbol{a}_1 &= \tilde{\boldsymbol{q}}_1 \\
&= \|\tilde{\boldsymbol{q}}_1\| \boldsymbol{q}_1 \\
\boldsymbol{a}_2 &= \tilde{\boldsymbol{q}}_2 + proj_{\tilde{\boldsymbol{q}}_1}(\boldsymbol{a}_2) \\
&= \tilde{\boldsymbol{q}}_2 + \frac{\boldsymbol{a}_2^t \tilde{\boldsymbol{q}}_1}{\tilde{\boldsymbol{q}}_1^t \tilde{\boldsymbol{q}}_1} \tilde{\boldsymbol{q}}_1 \\
&= (\boldsymbol{a}_2^t \boldsymbol{q}_1)\boldsymbol{q}_1 + \|\tilde{\boldsymbol{q}}_2\| \boldsymbol{q}_2 \\
\boldsymbol{a}_3 &= \tilde{\boldsymbol{q}}_3 + proj_{\tilde{\boldsymbol{q}}_1}(\boldsymbol{a}_3) + proj_{\tilde{\boldsymbol{q}}_2}(\boldsymbol{a}_3) \\
&= \tilde{\boldsymbol{q}}_3 + \frac{\boldsymbol{a}_3^t \tilde{\boldsymbol{q}}_1}{\tilde{\boldsymbol{q}}_1^t \tilde{\boldsymbol{q}}_1} \tilde{\boldsymbol{q}}_1 + \frac{\boldsymbol{a}_3^t \tilde{\boldsymbol{q}}_2}{\tilde{\boldsymbol{q}}_2^t \tilde{\boldsymbol{q}}_2} \tilde{\boldsymbol{q}}_2 \\
&= (\boldsymbol{a}_3^t \boldsymbol{q}_1)\boldsymbol{q}_1 + (\boldsymbol{a}_3^t \boldsymbol{q}_2)\boldsymbol{q}_2 + \|\tilde{\boldsymbol{q}}_3\| \boldsymbol{q}_3 \\
&\dots \\
\boldsymbol{a}_p &= (\boldsymbol{a}_p^t \boldsymbol{q}_1)\boldsymbol{q}_1 + (\boldsymbol{a}_p^t \boldsymbol{q}_2)\boldsymbol{q}_2 + \cdots + (\boldsymbol{a}_p^t \boldsymbol{q}_{p-1})\boldsymbol{q}_{p-1} + \left\|\tilde{\boldsymbol{q}}_p\right\| \boldsymbol{q}_p
\end{aligned}$$

즉 위의 축차식을 보면 원래 벡터 $\boldsymbol{a}_i$ 는 Gram–Schmidt 방법으로 구한 정규직교벡터 $\boldsymbol{q}_1, \boldsymbol{q}_2, \dots, \boldsymbol{q}_p$ 의 선형 조합으로 나타낼 수 있다.

이제 Gram–Schmidt 방법으로 구한 정규직교벡터들 $\boldsymbol{q}_1, \boldsymbol{q}_2, \cdots, \boldsymbol{q}_p$ 을 모아놓은 행렬을 $\boldsymbol{Q}$ 라고 하고 위에서 $\boldsymbol{a}_i$ 들이 직교행렬의 선형조합으로 표시될때 계수들을 모아놓는 상삼각행렬을 $\boldsymbol{R}$ 이라고 하자. 그려면 다음과 같은 QR 분해가 주어진다.

$$\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{R} \tag{D.3}$$

여기서

$$\boldsymbol{Q} = [\boldsymbol{q}_1 \;\; \boldsymbol{q}_2 \;\; \cdots \;\; \boldsymbol{q}_p], \quad \boldsymbol{Q}^t\boldsymbol{Q} = \boldsymbol{I}$$

$$\boldsymbol{R} = \begin{bmatrix} \|\tilde{\boldsymbol{q}}_1\| & \boldsymbol{a}_2^t\boldsymbol{q}_1 & \boldsymbol{a}_3^t\boldsymbol{q}_1 & \cdots & \boldsymbol{a}_p^t\boldsymbol{q}_1 \\ 0 & \|\tilde{\boldsymbol{q}}_2\| & \boldsymbol{a}_3^t\boldsymbol{q}_2 & \cdots & \boldsymbol{a}_p^t\boldsymbol{q}_2 \\ 0 & 0 & \|\tilde{\boldsymbol{q}}_3\| & \cdots & \boldsymbol{a}_p^t\boldsymbol{q}_3 \\ & & & \cdots & \\ 0 & 0 & 0 & \cdots & \|\tilde{\boldsymbol{q}}_p\| \end{bmatrix}$$

이제 Gram–Schmidt 방법과 QR 분해를 실제 예제를 통하여 구해보자

아래와 같이 4차원 벡터 3개가 있다.

$$\boldsymbol{a}_1 = \begin{bmatrix} -1 \\ 1 \\ -1 \\ 1 \end{bmatrix} \quad \boldsymbol{a}_2 = \begin{bmatrix} -1 \\ 3 \\ -1 \\ 3 \end{bmatrix} \quad \boldsymbol{a}_3 = \begin{bmatrix} 1 \\ 3 \\ 5 \\ 7 \end{bmatrix} \tag{D.4}$$

위의 벡터 $\boldsymbol{a}_1, \boldsymbol{a}_2, \boldsymbol{a}_3$에 대하여 Gram–Schmidt 방법을 적용해보자.

1. $i = 1$. 먼저 $\|\tilde{\boldsymbol{q}}_1\| = \|\boldsymbol{a}_1\| = 2$이므로 첫번째 벡터 $\boldsymbol{q}_1$를 만든다.

$$\boldsymbol{q}_1 = \tilde{\boldsymbol{q}}_1 / \|\tilde{\boldsymbol{q}}_1\| = \begin{bmatrix} -1/2 \\ 1/2 \\ -1/2 \\ 1/2 \end{bmatrix}$$

2. $i = 2$. 이제 두번째 직교벡터 $\boldsymbol{q}_2$를 만들자. $\boldsymbol{q}_1^t\boldsymbol{a}_2 = 4$이므로

$$\tilde{\boldsymbol{q}_2} = \boldsymbol{a}_2 - (\boldsymbol{q}_1^t\boldsymbol{a}_2)\boldsymbol{q}_1 = \begin{bmatrix} -1 \\ 3 \\ -1 \\ 3 \end{bmatrix} - 4 \begin{bmatrix} -1/2 \\ 1/2 \\ -1/2 \\ 1/2 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

그리고 $\|\tilde{\boldsymbol{q}}_2\| = 2$이므로

$$\boldsymbol{q}_2 = \tilde{\boldsymbol{q}}_2 / \|\tilde{\boldsymbol{q}}_2\| = \begin{bmatrix} 1/2 \\ 1/2 \\ 1/2 \\ 1/2 \end{bmatrix}$$

3. $i = 3$ 마지막으로 $\boldsymbol{q}_1^t \boldsymbol{a}_3 = 2, \boldsymbol{q}_2^t \boldsymbol{a}_3 = 8$ 이므로

$$\tilde{\boldsymbol{q}_3} = \boldsymbol{a}_3 - (\boldsymbol{q}_1^t \boldsymbol{a}_3)\boldsymbol{q}_1 - (\boldsymbol{q}_2^t \boldsymbol{a}_3)\boldsymbol{q}_2 = \begin{bmatrix} 1 \\ 3 \\ 5 \\ 7 \end{bmatrix} - 2 \begin{bmatrix} -1/2 \\ 1/2 \\ -1/2 \\ 1/2 \end{bmatrix} - 8 \begin{bmatrix} 1/2 \\ 1/2 \\ 1/2 \\ 1/2 \end{bmatrix} = \begin{bmatrix} -2 \\ -2 \\ 2 \\ 2 \end{bmatrix}$$

또한 $\|\tilde{\boldsymbol{q}_3}\| = 4$이므로

$$\boldsymbol{q}_3 = \tilde{\boldsymbol{q}}_3 / \|\tilde{\boldsymbol{q}}_3\| = \begin{bmatrix} -1/2 \\ -1/2 \\ 1/2 \\ 1/2 \end{bmatrix}$$

따라서 Gram–Schmidt 방법으로 만든 정규직교벡터는 다음과 같다.

$$\boldsymbol{q}_1 = \begin{bmatrix} -1/2 \\ 1/2 \\ -1/2 \\ 1/2 \end{bmatrix} \quad \boldsymbol{q}_2 = \begin{bmatrix} 1/2 \\ 1/2 \\ 1/2 \\ 1/2 \end{bmatrix} \quad \boldsymbol{q}_3 = \begin{bmatrix} -1/2 \\ -1/2 \\ 1/2 \\ 1/2 \end{bmatrix}$$

이제 위에서 구한 Gram-Schmidt 방법으로 얻은 결과를 이용하여 QR 분해를 구해보자.

D.4 에서 주어진 백터들을 열로 가진 행렬 $\boldsymbol{A}$의 QR분해를 구해보자.

$$\boldsymbol{A} = \begin{bmatrix} -1 & -1 & 1 \\ 1 & 3 & 3 \\ -1 & -1 & 5 \\ 1 & 3 & 7 \end{bmatrix}$$

앞의 예제에서 구한 직교벡터를 그대로 이용하면 $\boldsymbol{Q}$는 쉽게 구해진다.

$$\boldsymbol{Q} = \begin{bmatrix} -1/2 & 1/2 & -1/2 \\ 1/2 & 1/2 & -1/2 \\ -1/2 & 1/2 & 1/2 \\ 1/2 & 1/2 & 1/2 \end{bmatrix}$$

또한 D.3 에 주어진 공식을 이용하면 행렬 $\boldsymbol{R}$은 다음과 같이 구할 수 있다.

$$\boldsymbol{R} = \begin{bmatrix} \|\tilde{\boldsymbol{q}}_1\| & \boldsymbol{a}_2^t\boldsymbol{q}_1 & \boldsymbol{a}_3^t\boldsymbol{q}_1 \\ 0 & \|\tilde{\boldsymbol{q}}_2\| & \boldsymbol{a}_3^t\boldsymbol{q}_2 \\ 0 & 0 & \|\tilde{\boldsymbol{q}}_3\| \end{bmatrix} = \begin{bmatrix} 2 & 4 & 2 \\ 0 & 2 & 8 \\ 0 & 0 & 4 \end{bmatrix}$$

D.0.4. SVD 분해

D.0.4.1. 특이값과 특이벡터

고유값과 고유벡터는 정방행렬인 경우 정의되는 것으로서 행렬이 정방행렬이 아닌 경우에는 구할 수 없다. 이제 고유값과 유사한 성질을 가지는 특이값을 일반행렬에서 정의해보자.

$\boldsymbol{A}$가 $m \times n$ 일반행렬이라고 가정하고 그 계수 r이라고 하자 ($r(\boldsymbol{A}) = r$). 이제 서로 직교하는 n-차원의 벡터들의 집합 $\boldsymbol{v}_1, \boldsymbol{v}_2, \dots, \boldsymbol{v}_n$과 다른 직교하는 m-차원의 벡터들의 집합 $\boldsymbol{u}_1, \boldsymbol{u}_2, \dots, \boldsymbol{u}_m$을 생각하자.

행렬 $\boldsymbol{A}$의 특이값(singular values) $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_r > 0$과 왼쪽 특이벡터(left singular vectors) $\boldsymbol{u}_1, \boldsymbol{u}_2, \dots, \boldsymbol{u}_m$ 그리고 오른쪽 특이벡터(right singular vectors) $\boldsymbol{v}_1, \boldsymbol{v}_2, \dots, \boldsymbol{v}_n$ 는 다음과 같은 성질을 만족한다.

$$\boldsymbol{A}\boldsymbol{v}_1 = \sigma_1\boldsymbol{u}_1, \quad \boldsymbol{A}\boldsymbol{v}_2 = \sigma_2\boldsymbol{u}_2, \quad \dots \quad \boldsymbol{A}\boldsymbol{v}_r = \sigma_r\boldsymbol{u}_r, \quad \boldsymbol{A}\boldsymbol{v}_{r+1} = \boldsymbol{0}, \quad \dots, \quad \boldsymbol{A}\boldsymbol{v}_n = \boldsymbol{0} \tag{D.5}$$

$n \times n$ 정방행렬 $\boldsymbol{V}$와 $m \times m$ 정방행렬 $\boldsymbol{U}$ 를 각각 서로 직교하는 정규벡터 $\boldsymbol{v}_1, \boldsymbol{v}_2, \dots, \boldsymbol{v}_n$ 과 $\boldsymbol{u}_1, \boldsymbol{u}_2, \dots, \boldsymbol{u}_m$ 으로 구성되는 직교행렬이라고 하자.

$$\boldsymbol{V} = [\boldsymbol{v}_1\ \boldsymbol{v}_2\ \dots\ \boldsymbol{v}_n], \quad \boldsymbol{U} = [\boldsymbol{u}_1\ \boldsymbol{u}_2\ \dots\ \boldsymbol{u}_m]$$

D.5 에 나타난 관계를 행렬 $\boldsymbol{V}$와 $\boldsymbol{U}$로 나타내면 다음과 같이 표현할 수 있다.

$$\boldsymbol{A}\boldsymbol{V} = \boldsymbol{U}\boldsymbol{\Sigma} \tag{D.6}$$

위에서 $m \times n$ 행렬 $\boldsymbol{\Sigma}$는 다음과 같은 형태를 가진다.

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_r & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix}, \quad \boldsymbol{\Sigma}_r = \begin{bmatrix} \sigma_1 & & 0 \\ & \sigma_2 & \\ & & \ddots \\ 0 & & \sigma_r \end{bmatrix}$$

D.0.4.2. SVD 분해

이제 행렬 $\boldsymbol{V}$가 직교행렬을 이용하면 다음과 같은 SVD 분해(singular value decomposition; 특이값 분해)을 정의할 수 있다.

$$\underset{m\times n}{\boldsymbol{A}} = \underset{m\times m}{\boldsymbol{U}}\ \underset{m\times n}{\boldsymbol{\Sigma}}\ \underset{n\times n}{\boldsymbol{V}^t} \tag{D.7}$$

위의 D.7 을 전개하면 다음과 같이 계수가 1인 행렬 $\boldsymbol{u}_k\boldsymbol{v}_k^t$ 들의 선형조합으로 행렬 $\boldsymbol{A}$를 나타낼 수 있다.

$$\boldsymbol{A} = \sigma_1\boldsymbol{u}_1\boldsymbol{v}_1^t + \sigma_2\boldsymbol{u}_2\boldsymbol{v}_2^t + \dots \sigma_r\boldsymbol{u}_r\boldsymbol{v}_r^t \tag{D.8}$$

또한 D.6 에서 $\boldsymbol{\Sigma}$에서 0이 되는 값을 제외하면 처음 r개의 요소들만 이루어진 부분으로만 축소된 SVD 분해를 구할 수 있다.

$$\boldsymbol{A}\boldsymbol{V}_r = \boldsymbol{U}_r\boldsymbol{\Sigma}_r, \quad \boldsymbol{A}[\boldsymbol{v}_1\ \boldsymbol{v}_2\ \dots\ \boldsymbol{v}_r] = [\boldsymbol{u}_1\ \boldsymbol{u}_2\ \dots\ \boldsymbol{u}_r]\begin{bmatrix}\sigma_1 & & 0 \\ & \sigma_2 & \\ & & \ddots \\ 0 & & \sigma_r\end{bmatrix} \tag{D.9}$$

위의 D.9 에서 주의할 점은 행렬 $\boldsymbol{V}_r$과 $\boldsymbol{U}_r$은 정방행렬이 아니고 직교행렬도 아니다. $\boldsymbol{V}_r^t\boldsymbol{V}_r = \boldsymbol{I}$ 와 $\boldsymbol{U}_r^t\boldsymbol{U}_r = \boldsymbol{I}$이 성립하지만 일반적으로 $\boldsymbol{V}_r\boldsymbol{V}_r^t \neq \boldsymbol{I}, \boldsymbol{U}_r\boldsymbol{U}_r^t \neq \boldsymbol{I}$이다.

D.0.4.3. 특이값과 특이벡터의 계산

$m \times n$ 행렬 $\boldsymbol{A}$의 SVD 분해 D.7 로 부터 행렬 $\boldsymbol{A}^t\boldsymbol{A}$와 $\boldsymbol{A}\boldsymbol{A}^t$를 다음과 같이 나타낼 수 있다.

$$\begin{aligned}\boldsymbol{A}^t\boldsymbol{A} &= (\boldsymbol{V}\boldsymbol{\Sigma}^t\boldsymbol{U}^t)(\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^t) = \boldsymbol{V}\boldsymbol{\Sigma}^t\boldsymbol{\Sigma}\boldsymbol{V}^t \\ \boldsymbol{A}\boldsymbol{A}^t &= (\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^t)(\boldsymbol{V}\boldsymbol{\Sigma}^t\boldsymbol{U}^t) = \boldsymbol{U}\boldsymbol{\Sigma}\Sigma^t\boldsymbol{U}^t\end{aligned} \tag{D.10}$$

정방행렬의 역행렬은 다음과 같이 행렬식과 여인자 행렬을 이용하여 구한다. 예를 들어 다음과 같은 2×2 행렬 $\boldsymbol{A}$의 역행렬은

위에서 $\boldsymbol{A}^t\boldsymbol{A}$와 $\boldsymbol{A}\boldsymbol{A}^t$는 모두 대칭행렬이지만 서로 차원이 다르다. 또한 **?@eq-ata-mat** 과 D.10 을 보면 두 행렬이 모두 $\boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^t$ 의 형식으로 분해되는 것을 알 수 있다. 즉 다음과 같은 사실을 알 수 있다.

- $n \times n$ 비음정치행렬 $\boldsymbol{A}^t\boldsymbol{A}$의 고유벡터 행렬은 $\boldsymbol{V}$이다.
- $m \times m$ 비음정치행렬 $\boldsymbol{A}\boldsymbol{A}^t$의 고유벡터 행렬은 $\boldsymbol{U}$이다.
- 행렬 $\boldsymbol{A}^t\boldsymbol{A}$와 $\boldsymbol{A}\boldsymbol{A}^t$의 0이 아닌 고유값은 $\sigma_1^2, \sigma_2^2, \dots, \sigma_r^2$ 이다.

따라서 다음과 같은 방법으로 특이값과 특이벡터를 계산할 수 있다. 위의 방법은 두 행렬 $\boldsymbol{A}^t\boldsymbol{A}$와 $\boldsymbol{A}\boldsymbol{A}^t$를 모두 구하지 않고 $\boldsymbol{A}^t\boldsymbol{A}$의 고유값과 고유벡터만으로 SVD 분해를 구하는 방법이다 (만약 행렬 $\boldsymbol{A}$가 100000×5이라면 $\boldsymbol{A}\boldsymbol{A}^t$는 100000×100000이다!)

먼저 $\boldsymbol{A}^t\boldsymbol{A}$의 고유벡터 $\boldsymbol{v}_1, \dots, \boldsymbol{v}_r$을 다음과 같은 고유값과 고유벡터의 정의로 먼저 구한다.

$$\boldsymbol{A}^t\boldsymbol{A}\boldsymbol{v}_k = \lambda_k \boldsymbol{v}_k = \sigma_k^2 \boldsymbol{v}_k, \quad k = 1, 2, \dots, r \tag{D.11}$$

다음으로 다음의 식으로 $\boldsymbol{u}_1, \dots, \boldsymbol{u}_r$ 를 구한다.

$$\boldsymbol{u}_k = \frac{\boldsymbol{A}\boldsymbol{v}_k}{\sigma_k}, \quad k = 1, 2, \dots, r \tag{D.12}$$

D.12 에서 다음과 같이 $\boldsymbol{u}_k$가 행렬 $\boldsymbol{A}\boldsymbol{A}^t$의 고유벡터임을 확인할 수 있다.

$$\boldsymbol{A}\boldsymbol{A}^t\boldsymbol{u}_k = \boldsymbol{A}\boldsymbol{A}^t \left(\frac{\boldsymbol{A}\boldsymbol{v}_k}{\sigma_k} \right) = \boldsymbol{A} \left(\frac{\sigma_k^2 \boldsymbol{v}_k}{\sigma_k} \right) = \sigma_k^2 \boldsymbol{u}_k$$

또한 D.11 에서 $\boldsymbol{v}_k$는 정규직교벡터이므로 다음과 같이 $\boldsymbol{u}_k$도 정규직교행려임을 보일 수 있다.

$$\boldsymbol{u}_k^t \boldsymbol{u}_l = \left(\frac{\boldsymbol{A}\boldsymbol{v}_k}{\sigma_k} \right)^t \left(\frac{\boldsymbol{A}\boldsymbol{v}_l}{\sigma_l} \right) = \frac{\boldsymbol{v}_k^t (\boldsymbol{A}^t \boldsymbol{A} \boldsymbol{v}_l)}{\sigma_k \sigma_l} = \frac{\sigma_l}{\sigma_k} \boldsymbol{v}_k^t \boldsymbol{v}_l = \begin{cases} 1 & \text{if } k = l \\ 0 & \text{if } k \neq l \end{cases}$$

위에서 구한 r개의 $\boldsymbol{v}_k$와 $\boldsymbol{u}_k$외에 $n - r$과 $m - r$ 개의 서로 직교하는 나머지 $\boldsymbol{v}$와 $\boldsymbol{u}$도 구할 수 있다.

D.0.4.4. SVD 분해의 기하학적 의미

다음은 SVD 분해의 기하학적 의미를 설명한 그림이다.

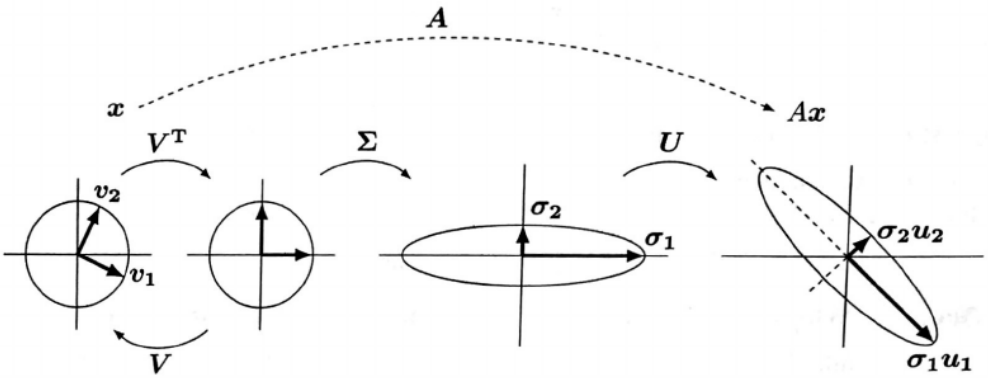


그림 D.3.: SVD 분해의 기하학적 의미

E. 벡터 미분

E.1. 용어

- vector differential: 벡터 미분
- partial derivative: 편미분
- gradient: 그레디언트
- Jacobian: 야코비안, 자코비안

E.2. 벡터 미분의 표기법

이제 다변량함수(multivariate function), $f : \mathbb{R}^n \to \mathbb{R}^m$에 대한 미분을 생각해보자.

먼저 간단한 예제를 고려해 보자. 두 열벡터

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \in \mathbb{R}_2, \quad \boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} \in \mathbb{R}^3$$

를 고려하고 다음과 같은 함수로 두 벡터의 관계가 정의된다고 하자.

$$y_1 = x_1^2 + x_2, \quad y_2 = \exp(x_1) + 3x_2, \quad y_3 = \sin(x_1) + x_2^3 \tag{E.1}$$

위의 관계를 함수 관계 $\boldsymbol{f} : \mathbb{R}^2 \to \mathbb{R}^3$ 로 나타내보면

$$\boldsymbol{f}(\boldsymbol{x}) = \begin{bmatrix} f_1(\boldsymbol{x}) \\ f_2(\boldsymbol{x}) \\ f_3(\boldsymbol{x}) \end{bmatrix} = \begin{bmatrix} x_1^2 + x_2 \\ \exp(x_1) + 3x_2 \\ \sin(x_1) + x_2^3 \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \boldsymbol{y}$$

이러한 경우 다변량 함수 $\boldsymbol{f}$를 벡터 $\boldsymbol{x}$로 미분하려면, 즉 미분 표기법을 이용하려면 편미분을 한 결과를 행렬의 형태를 정해야한다.

$$\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}} = (n \times m) - \text{matrix} \quad \text{or} \quad (m \times n) - \text{matrix?}$$

일단 각각의 편미분 $\frac{\partial f_i}{\partial x_j}$를 구해야 하며 이는 scalar 미분으로 쉽게 구해진다.

$$\begin{aligned} \frac{\partial f_1}{\partial x_1} &= 2x_1, & \frac{\partial f_2}{\partial x_1} &= \exp(x_1), & \frac{\partial f_3}{\partial x_1} &= \cos(x_1) \\ \frac{\partial f_1}{\partial x_2} &= 1, & \frac{\partial f_2}{\partial x_2} &= 3, & \frac{\partial f_3}{\partial x_2} &= 3x_2^2 \end{aligned} \tag{E.2}$$

이제 이제 편미분값들을 행렬의 형태로 정리해보자. 편미분을 행렬에 배치할 때 다음과 같은 규칙을 사용할 것이다.

- 행렬의 행은 $\boldsymbol{x}$의 차원 n 과 같다.
- 행렬의 열은 $\boldsymbol{f}$의 차원 m 과 같다.

위와 같이 편미분을 배치하는 벡타 미분 표기법을 분모 표기법 (denominator layout)이라고 한다.

> **분모 표기법**
>
> $$\boldsymbol{J} = \nabla_x \boldsymbol{x} = \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}} \equiv \frac{\partial \boldsymbol{f}^t}{\partial \boldsymbol{x}} \underset{def}{\equiv} \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_2}{\partial x_1} & \frac{\partial f_3}{\partial x_1} \\ \frac{\partial f_1}{\partial x_2} & \frac{\partial f_2}{\partial x_2} & \frac{\partial f_3}{\partial x_2} \end{bmatrix} = \begin{bmatrix} 2x_1 & \exp(x_1) & \cos(x_1) \\ 1 & 3 & 3x_2^2 \end{bmatrix}$$
>
> $\boldsymbol{J}$는 야코비안 행렬(Jacobian matrix)이라고 부른다.

이제 이러한 분자표기법의 특별한 결과를 알아보자

- $f : \mathbb{R}^n \to \mathbb{R}^1$ 인 경우

  $f : \mathbb{R}^n \to \mathbb{R}^1$ 인 경우 벡터미분 결과를 **그레디언트(gradient)**라고 부르며 다음과 같이 표기된다.

$$\nabla_x f = \frac{\partial f}{\partial \boldsymbol{x}} = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{bmatrix}$$

- $f : \mathbb{R}^1 \to \mathbb{R}^m$ 인 경우

$$\frac{\partial \boldsymbol{f}}{\partial x} = [\frac{\partial f_1}{\partial x}, \frac{\partial f_2}{\partial x}, \cdots, \frac{\partial f_m}{\partial x}]$$

참고로 E.1 에서 정의한 함수 관계를 두 벡터 $\boldsymbol{x}$ 와 $\boldsymbol{y}$ 의 사상관계로 보면

$$\boldsymbol{f} : \boldsymbol{x} \mapsto \boldsymbol{y}$$

다음과 같이 그레디언트 벡터를 표기할 수 있다.

$$\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}} = \frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}} = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_2}{\partial x_1} & \frac{\partial y_3}{\partial x_1} \\ \frac{\partial y_1}{\partial x_2} & \frac{\partial y_2}{\partial x_2} & \frac{\partial y_3}{\partial x_2} \end{bmatrix}$$

E.3. 함성함수의 미분법

이제 합성함수의 미분법(chain rule)에 대하여 알아보자.

두 개의 함수

$$\boldsymbol{g} : \mathbb{R}^n \mapsto \mathbb{R}^m, \quad \boldsymbol{f} : \mathbb{R}^m \mapsto \mathbb{R}^p$$

가 있을 때, $\boldsymbol{f}$와 $\boldsymbol{g}$의 합성함수 $\boldsymbol{h}$는 다음과 같이 정의된다.

$$\boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{f}(\boldsymbol{g}(\boldsymbol{x})) = \boldsymbol{f} \circ \boldsymbol{g}$$

즉,

$$\boldsymbol{h} : \mathbb{R}^n \mapsto \mathbb{R}^m \mapsto \mathbb{R}^p$$

이러한 합성함수의 미분은 다음과 같이 계산된다.

$$\frac{\partial \boldsymbol{h}}{\partial \boldsymbol{x}} = \frac{\partial \boldsymbol{f} \circ \boldsymbol{g}}{\partial \boldsymbol{x}} = \frac{\partial \boldsymbol{g}}{\partial \boldsymbol{x}} \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{g}} \tag{E.3}$$

E.3 에서 $\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{g}}$는 $m \times p$ Jacovian 벡터이고

$$\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{g}} \begin{bmatrix} \frac{\partial f_1}{\partial g_1} & \frac{\partial f_2}{\partial g_1} & \cdots & \frac{\partial f_p}{\partial g_1} \\ \frac{\partial f_1}{\partial g_2} & \frac{\partial f_2}{\partial g_2} & \cdots & \frac{\partial f_p}{\partial g_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_1}{\partial g_m} & \frac{\partial f_2}{\partial g_m} & \cdots & \frac{\partial f_p}{\partial g_m} \end{bmatrix} = (m \times p)$$

$\frac{\partial \boldsymbol{g}}{\partial \boldsymbol{x}}$는 $n \times m$ Jacovian 벡터이다

$$\frac{\partial \boldsymbol{g}}{\partial \boldsymbol{x}} = \begin{bmatrix} \frac{\partial g_1}{\partial x_1} & \frac{\partial g_2}{\partial x_1} & \cdots & \frac{\partial g_m}{\partial x_1} \\ \frac{\partial g_1}{\partial x_2} & \frac{\partial g_2}{\partial x_2} & \cdots & \frac{\partial g_m}{\partial x_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial g_1}{\partial x_n} & \frac{\partial g_2}{\partial x_n} & \cdots & \frac{\partial g_m}{\partial x_n} \end{bmatrix} = (n \times m)$$

함성함수의 미분 공식을 차원으로 나타내면 다음과 같다.

$$\underset{n \times p}{\frac{\partial \boldsymbol{h}}{\partial \boldsymbol{x}}} = \underset{n \times m}{\frac{\partial \boldsymbol{g}}{\partial \boldsymbol{x}}} \underset{m \times p}{\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{g}}}$$

E.4. 두 벡터 내적의 미분

E.4.1. 상수벡터와 변수벡터의 내적

먼저 상수 벡터 $\boldsymbol{a}$와 변수 벡터 $\boldsymbol{x}$의 내적의 미분을 생각해 보자.

참고로 다음과 같이 두 벡터의 내적은 스칼라이다.

$$\boldsymbol{a}^t\boldsymbol{x} = \boldsymbol{x}^t\boldsymbol{a} = a_1x_1 + a_2x_2 + \cdots + a_nx_n$$

따라서 그레이디언트를 구하는 방법과 같이 결과는 열벡터로 표기된다.

$$\frac{\partial \boldsymbol{a}^t\boldsymbol{x}}{\partial \boldsymbol{x}} = \frac{\partial \boldsymbol{x}^t\boldsymbol{a}}{\partial \boldsymbol{x}} = \boldsymbol{a} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix}$$

위의 식에서 상수벡터 $\boldsymbol{a}$는 가 전치로 앞에 나타나는 표현 $\boldsymbol{x}^t\boldsymbol{a}$ 를 사용하면 결과 벡터 $\boldsymbol{a}$가 열벡터로 그대로 나타나지므로 **내적의 미분 표기**로 사용할 것이다.

$$\frac{\partial \boldsymbol{x}^t\boldsymbol{a}}{\partial \boldsymbol{x}} = \frac{\partial \boldsymbol{x}^t}{\partial \boldsymbol{x}}\boldsymbol{a} = \boldsymbol{I}\boldsymbol{a} = \boldsymbol{a} \tag{E.4}$$

E.4.2. 상수벡터와 함수벡터의 내적

더 나아가서 상수 벡터 $\boldsymbol{a}$와 함수 벡터 $\boldsymbol{f}$의 내적의 미분도 E.4 을 표시하는 동일한 논리로 다음과 같이 표기할 수 있다.

$$\frac{\partial \boldsymbol{f}^t\boldsymbol{a}}{\partial \boldsymbol{x}} = \frac{\partial \boldsymbol{f}^t}{\partial \boldsymbol{x}}\boldsymbol{a} \tag{E.5}$$

참고로 E.5 에서 $\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}}$는 행벡터가 아닌 행렬로 나타날 수 있다.

E.4.3. 함수벡터와 함수벡터의 내적

이제 다음과 같이 같은 공간으로 사상되는 두 함수 $\boldsymbol{f}$ 와 $\boldsymbol{g}$ 의 내적을 생각해 보자.

$$\boldsymbol{f} : \mathbb{R}^n \mapsto \mathbb{R}^m, \quad \boldsymbol{g} : \mathbb{R}^n \mapsto \mathbb{R}^m$$

두 함수의 내적을 미분하는 경우 곱셉 법칙을 적용하여야 하는데 행렬의 곱셉에서는 교환법칙이 성립되지 않으므로 순서에 주의해야 한다.

내적 $\boldsymbol{f}^t\boldsymbol{g}$ 를 각각 따로 미분해야 하는데 각 벡터에 대해 따로 미분을 실행해 보자

- $\boldsymbol{f}$ 를 미분하는 경우 $\boldsymbol{g}$ 는 상수 벡터 $\boldsymbol{a}$ 로 취급한다. 그리고 E.5 를 적용한다.

$$\frac{\partial \boldsymbol{f}^t\boldsymbol{g}}{\partial \boldsymbol{x}} = \frac{\partial \boldsymbol{f}^t\boldsymbol{a}}{\partial \boldsymbol{x}} = \frac{\partial \boldsymbol{f}^t}{\partial \boldsymbol{x}}\boldsymbol{a} = \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}}\boldsymbol{g} \tag{E.6}$$

- $\boldsymbol{g}$ 를 미분하는 경우 $\boldsymbol{f}$ 는 상수 벡터 $\boldsymbol{a}$ 로 취급한다. 그리고 E.5 를 적용한다.

$$\frac{\partial \boldsymbol{f}^t\boldsymbol{g}}{\partial \boldsymbol{x}} = \frac{\partial \boldsymbol{a}^t\boldsymbol{g}}{\partial \boldsymbol{x}} = \frac{\partial \boldsymbol{g}^t\boldsymbol{a}}{\partial \boldsymbol{x}} = \frac{\partial \boldsymbol{g}^t}{\partial \boldsymbol{x}}\boldsymbol{a} = \frac{\partial \boldsymbol{g}}{\partial \boldsymbol{x}}\boldsymbol{f} \tag{E.7}$$

이제 위의 두 결과 E.6 과 E.7 를 합치면 다음과 같은 최종적인 결과를 얻을 수 있다.

$$\frac{\partial \boldsymbol{f}^t\boldsymbol{g}}{\partial \boldsymbol{x}} = \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}}\boldsymbol{g} + \frac{\partial \boldsymbol{g}}{\partial \boldsymbol{x}}\boldsymbol{f} \tag{E.8}$$

E.5. 벡터 미분의 응용

E.5.1. 선형사상의 미분상

이제 앞에서 배운 벡터의 미분을 이용하여 유용한 응용 공식을 유도해보자.

먼저 선형변환 $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$ 를 생각해 보자. 이때 $(M \times N)$-$\boldsymbol{A}$는 상수 행렬이다. 이때 $\boldsymbol{y}$를 $\boldsymbol{x}$로 미분하면 다음과 같다.

먼저 행렬 $\boldsymbol{A}$의 i 번째 행을 $\boldsymbol{a}_i^t$라고 하면

$$\boldsymbol{A} = \begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1N} \\ A_{21} & A_{22} & \cdots & A_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ A_{M1} & A_{M2} & \cdots & A_{MN} \end{bmatrix} = \begin{bmatrix} \boldsymbol{a}_1^t \\ \boldsymbol{a}_2^t \\ \vdots \\ \boldsymbol{a}_M^t \end{bmatrix}$$

선형변환 $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{A}\boldsymbol{x}$ 로 정의하면 다음과 같이 나타낼 수 있다.

$$\boldsymbol{A}\boldsymbol{x} = \begin{bmatrix} \boldsymbol{a}_1^t\boldsymbol{x} \\ \boldsymbol{a}_2^t\boldsymbol{x} \\ \vdots \\ \boldsymbol{a}_M^t\boldsymbol{x} \end{bmatrix} = \begin{bmatrix} f_1(\boldsymbol{x}) \\ f_2(\boldsymbol{x}) \\ \vdots \\ f_M(\boldsymbol{x}) \end{bmatrix} = \boldsymbol{f}(\boldsymbol{x})$$

따라서

$$\frac{\partial \boldsymbol{A}\boldsymbol{x}}{\partial \boldsymbol{x}} = \frac{\partial \boldsymbol{f}(\boldsymbol{x})}{\partial \boldsymbol{x}} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_2}{\partial x_1} & \cdots & \frac{\partial f_M}{\partial x_1} \\ \frac{\partial f_1}{\partial x_2} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_M}{\partial x_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_1}{\partial x_N} & \frac{\partial f_2}{\partial x_N} & \cdots & \frac{\partial f_M}{\partial x_N} \end{bmatrix} = \begin{bmatrix} A_{11} & A_{21} & \cdots & A_{M1} \\ A_{12} & A_{22} & \cdots & A_{M2} \\ \vdots & \vdots & \ddots & \vdots \\ A_{1N} & A_{2N} & \cdots & A_{MN} \end{bmatrix} = \boldsymbol{A}^t$$

따라서 선형사상의 미분은 선형변환 행렬의 전치이다.

$$\frac{\partial \boldsymbol{A}\boldsymbol{x}}{\partial \boldsymbol{x}} = \boldsymbol{A}^t \tag{E.9}$$

F. 다변량 확률변수의 성질

F.1. 일변량분포

일변량 확률변수 X가 확률밀도함수 $f(x)$를 가지는 분포를 따를때 기대값과 분산은 다음과 같이 정의된다.

$$E(X) = \int x f(x) dx = \mu, \quad V(X) = E[X - E(X)]^2 = \int (x-\mu)^2 f(x) dx = \sigma^2$$

새로운 확률변수 Y가 확률변수 X의 선형변환으로 표시된다면 (a와 b는 실수)

$$Y = aX + b$$

그 기대값(평균)과 분산은 다음과 같이 계산된다.

$$\begin{aligned}
E(Y) &= E(aX+b) \\
&= \int (ax+b) f(x) dx \\
&= a \int x f(x) dx + b \\
&= aE(X) + b \\
&= a\mu + b \\
V(Y) &= Var(aX+b) \\
&= E[aX + b - E(aX+b)]^2 \\
&= E[a(X-\mu)]^2 \\
&= a^2 E(X-\mu)^2 \\
&= a^2 \sigma^2
\end{aligned}$$

F.2. 확률벡터와 분포

확률벡터 $\boldsymbol{X}$가 p 차원의 다변량분포를 따른다고 하고 결합확률밀도함수 $f(\boldsymbol{x}) = f(x_1, x_2, \dots, x_p)$를 를 가진다고 하자.

$$\boldsymbol{X} = \begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ .. \\ X_p \end{bmatrix}$$

다변량 확률벡터의 기대값(평균벡터)과 공분산(행렬)은 다음과 같이 계산된다.

$$\boldsymbol{E}(\boldsymbol{X}) = \begin{bmatrix} E(X_1) \\ E(X_2) \\ E(X_3) \\ .. \\ E(X_p) \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ .. \\ \mu_p \end{bmatrix} = \boldsymbol{\mu}$$

$$V(\boldsymbol{X}) = Cov(\boldsymbol{X}) = E(\boldsymbol{X}-\boldsymbol{\mu})(\boldsymbol{X}-\boldsymbol{\mu})^t = \begin{bmatrix} \sigma_{11} & \sigma_{12} & ... & \sigma_{1p} \\ \sigma_{12} & \sigma_{22} & ... & \sigma_{2p} \\ & ... & ... & \\ \sigma_{1p} & \sigma_{2p} & ... & \sigma_{pp} \end{bmatrix} = \boldsymbol{\Sigma}$$

여기서 $\sigma_{ii} = V(X_i)$, $\sigma_{ij} = Cov(X_i, X_j) = Cov(X_j, X_i)$이다. 따라서 공분산 행렬 $\boldsymbol{\Sigma}$는 대칭행렬(symmetric matrix)이다. 다음 공식은 유용한 공식이다.

$$\boldsymbol{\Sigma} = E(\boldsymbol{X}-\boldsymbol{\mu})(\boldsymbol{X}-\boldsymbol{\mu})^t = E(\boldsymbol{X}\boldsymbol{X}^t) - \boldsymbol{\mu}\boldsymbol{\mu}^t$$

두 확률변수의 상관계수 ρ_{ij}는 다음과 같이 정의된다.

$$\rho_{ij} = \frac{Cov(X_i, X_j)}{\sqrt{V(X_i)V(X_j)}} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}}$$

새로운 확률벡터 $\boldsymbol{Y}$가 확률벡터 $\boldsymbol{X}$ 의 선형변환라고 하자.

$$\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X} + \boldsymbol{b}$$

단 여기서 $\boldsymbol{A} = \{a_{ij}\}$는 $p \times p$ 실수 행렬이고 $\boldsymbol{b} = (b_1 b_2 ... b_p)^t$는 $p \times 1$ 실수 벡터이다.

확률벡터 $\boldsymbol{Y}$의 기대값(평균벡터)과 공분산은 다음과 같이 계산된다.

$$
\begin{aligned}
E(\boldsymbol{Y}) &= E(\boldsymbol{AX}+\boldsymbol{b}) \\
&= \boldsymbol{A}E(\boldsymbol{X})+\boldsymbol{b} \\
&= \boldsymbol{A\mu}+\boldsymbol{b} \\
V(\boldsymbol{Y}) &= Var(\boldsymbol{AX}+\boldsymbol{b}) \\
&= E[\boldsymbol{AX}+\boldsymbol{b}-E(\boldsymbol{AX}+\boldsymbol{b})][\boldsymbol{AX}+\boldsymbol{b}-E(\boldsymbol{AX}+\boldsymbol{b})]^t \\
&= E[\boldsymbol{AX}-\boldsymbol{A\mu}][\boldsymbol{AX}-\boldsymbol{A\mu}]^t \\
&= E[\boldsymbol{A}(\boldsymbol{X}-\boldsymbol{\mu})][\boldsymbol{A}(\boldsymbol{X}-\boldsymbol{\mu})]^t \\
&= \boldsymbol{A}E[(\boldsymbol{X}-\boldsymbol{\mu})(\boldsymbol{X}-\boldsymbol{\mu})^t]\boldsymbol{A}^t \\
&= \boldsymbol{A\Sigma A}^t
\end{aligned}
$$

만약 표본 $\boldsymbol{X}_i, \boldsymbol{X}_2, \dots, \boldsymbol{X}_n$ 이 독립적으로 평균이 $\boldsymbol{\mu}$ 이고 공분산이 $\boldsymbol{\Sigma}$ 인 분포에서 추출되었다면 표본의 평균벡터 $\bar{\boldsymbol{X}}$ 는 평균이 $\boldsymbol{\mu}$ 이고 공분산이 $\frac{1}{n}\boldsymbol{\Sigma}$ 인 분포를 따른다.

$$
\bar{\boldsymbol{X}} = \begin{bmatrix} \sum_{i=1}^n X_{i1}/n \\ \sum_{i=1}^n X_{i2}/n \\ \sum_{i=1}^n X_{i3}/n \\ .. \\ \sum_{i=1}^n X_{ip}/n \end{bmatrix}
$$

여기서 X_{ij} 는 i 번째 표본벡터 $\boldsymbol{X}_i = (X_{i1}X_{i2} \dots X_{ip})^t$ 의 j 번째 확률변수이다.

F.3. 다변량 정규분포

일변량 확률변수 X 가 평균이 μ 이고 분산이 σ^2 인 정규분포를 따른다면 다음과 같이 나타내고

$$X \sim N(\mu, \sigma^2)$$

확률밀도함수 $f(x)$ 는 다음과 같이 주어진다.

$$f(x) = (2\pi\sigma^2)^{-1/2} \exp\left(-\frac{(x-\mu)^2}{2}\right)$$

p-차원 확률벡터 $\boldsymbol{X}$ 가 평균이 $\boldsymbol{\mu}$ 이고 공분산이 $\boldsymbol{\Sigma}$ 인 다변량 정규분포를 따른다면 다음과 같이 나타내고

$$\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

확률밀도함수 $f(\boldsymbol{x})$ 는 다음과 같이 주어진다.

$$f(\boldsymbol{x}) = (2\pi)^{-p/2}|\boldsymbol{\Sigma}|^{-1/2} \exp\left(-\frac{(\boldsymbol{x}-\boldsymbol{\mu})\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})^t}{2}\right)$$

다변량 정규분포 $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$를 따르는 확률벡터 $\boldsymbol{X}$를 다음과 같이 두 부분으로 나누면

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{bmatrix}, \quad \boldsymbol{X}_1 = \begin{bmatrix} \boldsymbol{X}_{11} \\ \boldsymbol{X}_{12} \\ \vdots \\ \boldsymbol{X}_{1p} \end{bmatrix}, \quad \boldsymbol{X}_2 = \begin{bmatrix} \boldsymbol{X}_{21} \\ \boldsymbol{X}_{22} \\ \vdots \\ \boldsymbol{X}_{2q} \end{bmatrix}$$

각각 다변량 정규분포를 따르고 다음과 같이 나타낼 수 있다.

$$\begin{bmatrix} E(\boldsymbol{X}_1) \\ E(\boldsymbol{X}_2) \end{bmatrix} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \quad \begin{bmatrix} V(\boldsymbol{X}_1) & Cov(\boldsymbol{X}_1, X_2) \\ Cov(\boldsymbol{X}_2 X_1) & V(\boldsymbol{X}_2) \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{12}^t & \boldsymbol{\Sigma}_{22} \end{bmatrix}$$

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{bmatrix} \sim N_{p+q} \left(\begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \Sigma_{12} \\ \boldsymbol{\Sigma}_{12}^t & \Sigma_{22} \end{bmatrix} \right)$$

확률벡터 $\boldsymbol{X}_2 = \boldsymbol{x}_2$가 주어진 경우 $\boldsymbol{X}_1$의 조건부 분포는 p-차원 다변량 정규분포를 따르고 평균과 공분산은 다음과 같다.

$$E(\boldsymbol{X}_1 | \boldsymbol{X}_2 = \boldsymbol{x}_2) = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{x}_2), \quad V(\boldsymbol{X}_1 | \boldsymbol{X}_2 = \boldsymbol{x}_2) = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{12}^t$$

예를 들어 2-차원 확률벡터 $\boldsymbol{X} = (X_1, X_2)^t$가 평균이 $\boldsymbol{\mu} = (\mu_1, \mu_2)^t$ 이고 공분산 $\boldsymbol{\Sigma}$가 다음과 같이 주어진

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix}$$

이변량 정규분포를 따른다면 확률밀도함수 $f(\boldsymbol{x})$에서 exp함수의 인자는 다음과 같이 주어진다.

$$(\boldsymbol{x} - \boldsymbol{\mu})\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})^t = \\ -\frac{1}{2(1-\rho^2)} \left[\left(\frac{(x_1 - \mu_1)^2}{\sigma_{11}} \right) + \left(\frac{(x_2 - \mu_2)^2}{\sigma_{22}} \right) - 2\rho \left(\frac{(x_1 - \mu_1)}{\sqrt{\sigma_{11}}} \right) \left(\frac{(x_2 - \mu_2)}{\sqrt{\sigma_{22}}} \right) \right]$$

그리고 $p = 2$인 경우 확률밀도함수의 상수부분은 다음과 같이 주어진다.

$$(2\pi)^{-p/2} |\boldsymbol{\Sigma}|^{-1/2} = \frac{1}{2\pi\sqrt{\sigma_{11}\sigma_{22}(1-\rho^2)}}$$

여기서 $\rho = \sigma_{12}/\sqrt{\sigma_{11}\sigma_{22}}$

만약 $X_2 = x_2$가 주어졌을 때 X_1의 조건부 분포는 정규분포이고 평균과 분산은 다음과 같이 주어진다.

$$E(X_1 | X_2 = x_2) = \mu_1 + \frac{\sigma_{12}}{\sigma_{22}}(\mu_2 - x_2) = \mu_1 + \rho\frac{\sqrt{\sigma_{11}}}{\sqrt{\sigma_{22}}}(\mu_2 - x_2)$$

$$V(X_1|X_2 = x_2) = \sigma_{11} - \frac{\sigma_{12}^2}{\sigma_{22}} = \sigma_{11}(1-\rho^2)$$

다변량 정규분포에서 공분산이 0인 두 확률 변수는 독립이다.

$$\sigma_{ij} = 0 \leftrightarrow X_i \text{ and } X_j \text{ are independent}$$

F.4. 표준정규분포로의 변환

일변량 확률변수 X가 평균이 μ 이고 분산이 σ^2인 경우 다음과 같은 선형변환을 고려하면.

$$Z = \frac{X-\mu}{\sigma} = (\sigma^2)^{-1/2}(X-\mu)$$

확률변수 Z 는 평균이 0 이고 분산이 1인 분포를 따른다.

p차원 확률벡터 $\boldsymbol{X}$ 가 평균이 $\boldsymbol{\mu}$ 이고 공분산이 $\boldsymbol{\Sigma}$인 분포를 가진다고 가정하자. 공분산 행렬 $\boldsymbol{\Sigma}$는 양정치 행렬(positive definite matrix)이며 다음과 같은 행렬의 분해가 가능하다.

$$\Sigma = \boldsymbol{C}\boldsymbol{C}^t$$

여기서 $\boldsymbol{C}$ 는 정칙행렬이며 역행렬 $\boldsymbol{C}^{-1}$가 존재한다. 위와 같은 행렬의 분해는 스펙트럴 분해(spectral decomposition)을 이용하여 구할 수 있다. 공분산 행렬 $\boldsymbol{\Sigma}$는 양정치 행렬이므로 고유치(eigen value) $(\lambda_1, \lambda_2, \dots, \lambda_p)$가 모두 양수이고 정규직교 고유벡터(orthonormal eigen vector)의 행렬 $\boldsymbol{P}$을 이용하여 다음과 같은 분해가 가능하다.

$$\Sigma = \boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}^t = \boldsymbol{P}\boldsymbol{\Lambda}^{1/2}\Lambda^{1/2}\boldsymbol{P}^t$$

여기서 $\boldsymbol{\Lambda}$는 고유치 $(\lambda_1, \lambda_2, \dots, \lambda_p)$를 대각원소로 가지는 대각행렬이며 $\boldsymbol{\Lambda}^{1/2}$는 고유치의 제곱근을 대각원소로 가지는 대각행렬이다. 따라서 $\boldsymbol{C} = \boldsymbol{P}\boldsymbol{\Lambda}^{1/2}$로 하면 위와 같은 행렬의 분해가 가능하다. 정규직교 고유벡터(orthonormal eigen vector)의 행렬 $\boldsymbol{P}$는 직교행렬이므로

$$\boldsymbol{C}^{-1} = (\boldsymbol{P}\boldsymbol{\Lambda}^{1/2})^{-1} = \boldsymbol{\Lambda}^{-1/2}\boldsymbol{P}^t$$

p차원 확률벡터 $\boldsymbol{X}$의 다음과 같은 선형변환을 고려하면.

$$\boldsymbol{Z} = \boldsymbol{C}^{-1}(\boldsymbol{X}-\boldsymbol{\mu}) = \boldsymbol{\Lambda}^{-1/2}\boldsymbol{P}^t(\boldsymbol{X}-\boldsymbol{\mu})$$

확률벡터 $\boldsymbol{Z}$ 는 평균이 $\boldsymbol{0}$ 이고 공분산이 $\boldsymbol{I}$인 분포를 따른다 (why?).

확률벡터 $\boldsymbol{X}$가 정규분포를 따른다면 선형변환한 확률벡터 $\boldsymbol{Z}$도 정규분포를 따른다.

F.5. 예제

예를 들어 이변량확률벡터 $\boldsymbol{X}$가 다음과 같은 평균벡터와 공분산을 가진 정규분포를 따른다고 하자

$$\boldsymbol{\mu} = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \quad \boldsymbol{\Sigma} = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$$

공분산행렬 $\boldsymbol{\sigma}$의 고유치는 $|\boldsymbol{\sigma} - \lambda \boldsymbol{I}| = 0$의 방정식을 풀어 구할 수 있다.

$$|\boldsymbol{\sigma} - \lambda \boldsymbol{I}| = \begin{bmatrix} 2-\lambda & 1 \\ 1 & 2-\lambda \end{bmatrix} = \lambda^2 - 4\lambda + 3 = 0$$

방정식을 풀면 고유치는 $(\lambda_1, \lambda_2) = (3, 1)$이다. 각 고유치에 대한 고유벡터 $\boldsymbol{p} = (p_1, p_2)^t$는 $\boldsymbol{\Sigma p} = \lambda \boldsymbol{p}$ 으로 구할 수 있다. 각 고유치에 대하여 방정식을 구하면 다음 두 개의 방정식을 얻을 수 있다.

$$p_1 - p_2 = 1 \text{ and } p_1 + p_2 = 0$$

정규직교 벡터의 조건을 만족 시키기 위해서 $p_1^2 + p_2^2 = 1$의 조건을 적용하면 다음과 같은 정규직교 고유행렬을 얻을 수 있다.

$$\boldsymbol{P} = \begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}$$

또한

$$\boldsymbol{\Lambda} = \begin{bmatrix} 3 & 0 \\ 0 & 1 \end{bmatrix} \quad \boldsymbol{\Lambda}^{1/2} = \begin{bmatrix} \sqrt{3} & 0 \\ 0 & 1 \end{bmatrix}$$

따라서 $C^{-1} = \Lambda^{-1/2} \boldsymbol{P}^t$ 이며

$$\boldsymbol{C}^{-1} = \boldsymbol{\Lambda}^{-1/2} \boldsymbol{P}^t = \begin{bmatrix} \frac{1}{\sqrt{3}} & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}$$

G. 이차형식과 제곱합의 분포

이 장에서는 선형 회귀모형의 추론과 검정에 자주 사용되는 제곱합의 분포와 관련이 있는 이차형식의 분포의 성질을 살펴본다.

G.1. 이차형식

n-차원 벡터 $\boldsymbol{x}^t = [x_1, x_2, \dots, x_n]$과 대칭행렬 $\boldsymbol{A}$에 대하여 이차형식(quadratic form)은 다음과 같이 정의된다.

$$Q_A(\boldsymbol{x}) = \boldsymbol{x}^t \boldsymbol{A} \boldsymbol{x} = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_i x_j \tag{G.1}$$

이차형식의 정의에서 반드시 행렬 $\boldsymbol{A}$를 대칭행렬로 정의하지 않아도 되지만 임의의 행렬에 대하여 이차형식의 값이 동일한 대칭행렬이 존재하기 때문에 정의에서 이차형식으로 국한하는 것이 일반적이다.

Definition G.1 (양정치 행렬)**.** 이차형식 $Q_A(\boldsymbol{x}) = \boldsymbol{x}^t \boldsymbol{A} \boldsymbol{x}$가 영벡터가 아닌 모든 벡터 $\boldsymbol{x}$에 대하여 0 보다 크면, 즉

$$\boldsymbol{x}^t \boldsymbol{A} \boldsymbol{x} > 0 \quad \text{for all} \quad \boldsymbol{x} \in \mathbb{R}^n$$

$\boldsymbol{A}$를 양정치(positive definite)라고 부른다.

만약 이차형식 $Q_A(\boldsymbol{x}) = \boldsymbol{x}^t \boldsymbol{A} \boldsymbol{x}$가 영벡터가 아닌 모든 벡터 $\boldsymbol{x}$에 대하여 0 보다 크거나 같다면

$$\boldsymbol{x}^t \boldsymbol{A} \boldsymbol{x} \geq 0 \quad \text{for all} \quad \boldsymbol{x} \in \mathbb{R}^n$$

$\boldsymbol{A}$를 양반정치(positive semi-definite)라고 부른다.

■

정칙행렬 $\boldsymbol{B}$에 대하여 다음과 같은 선형변환을 고려하자.

$$\boldsymbol{x} = \boldsymbol{B}\boldsymbol{y} \quad \text{or} \quad \boldsymbol{y} = \boldsymbol{B}^{-1}\boldsymbol{x}$$

벡터 $\boldsymbol{x}$로 정의된 이차형식은 벡터 $\boldsymbol{y}$의 형태로 다음과 같이 변환할 수 있다.

$$Q(\boldsymbol{x}) = \boldsymbol{x}^t\boldsymbol{A}\boldsymbol{x} = \boldsymbol{y}^t\boldsymbol{B}^t\boldsymbol{A}\boldsymbol{B}\boldsymbol{y} = Q^*(\boldsymbol{y})$$

이차형식의 성질은 정칙 선형변환에서 유지된다. 즉 행렬 $\boldsymbol{A}$가 양(반)정치 행렬이고 행렬 $\boldsymbol{B}$가 정칙행렬이면 행렬 $\boldsymbol{B}^t\boldsymbol{A}\boldsymbol{B}$도 양(반)정치 행렬이다.

G.2. 대칭행렬의 대각화

앞 절의 예제 Exercise C.4 에서 보았듯이 n차원 대칭행렬 $\boldsymbol{A}$ 에 대하여 직교행렬 $\boldsymbol{P}$가 존재하여 다음과 같은 분해가 가능하다.

$$\boldsymbol{P}^t\boldsymbol{A}\boldsymbol{P} = \boldsymbol{\Lambda} = diag(\lambda_1, \lambda_2, \dots, \lambda_n) \tag{G.2}$$

G.2 의 분해에서 λ_i는 행렬 $\boldsymbol{A}$의 고유치이며 행렬 $\boldsymbol{P}$의 i 번째 열은 대응하는 고유벡터 $\boldsymbol{p}_i$ 로 구성되어 있다.

$$\boldsymbol{P} = [\boldsymbol{p}_1 \;\; \boldsymbol{p}_2 \;\; \dots \;\; \boldsymbol{p}_n]$$

이제 위의 분해를 증명해 보자. 고유치 λ_i 와 대응하는 고유벡터 $\boldsymbol{p}_i$의 정의에 따라서 다음과 같은 n개의 식을 얻을 수 있고

$$\boldsymbol{A}\boldsymbol{p}_i = \lambda_i\boldsymbol{p}_i, \quad i = 1, 2, 3 \dots, n$$

위의 식을을 합쳐서 표기하면 다음과 같은 식을 얻으며 이는 G.2 를 의미한다.

$$\boldsymbol{A}\boldsymbol{P} = \boldsymbol{P}\boldsymbol{\Lambda}$$

G.2 를 다시 쓰면 다음과 같은 스펙트럴 분해(spectral decomposition)를 얻는다.

$$\boldsymbol{A} = \boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}^t = \sum_{i=1}^{n} \lambda_i\boldsymbol{p}_i\boldsymbol{p}_i^t \tag{G.3}$$

참고로 대각합과 행렬식에 대한 고유치의 관계를 나타내는 다음의 유용한 두 식을 반드시 기억하자.

$$tr(\boldsymbol{A}) = \sum_i \lambda_i, \quad |\boldsymbol{A}| = \prod_i \lambda_i$$

대칭행렬의 분해 G.2 를 이용하면 다음과 같은 이차형식의 분해를 얻을 수 있다.

$$Q(\boldsymbol{x}) = \boldsymbol{x}^t\boldsymbol{A}\boldsymbol{x} = \boldsymbol{x}^t\boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}^t\boldsymbol{x} = \boldsymbol{y}^t\boldsymbol{\Lambda}\boldsymbol{y} = \sum_{i=1}^{n} \lambda_i y_i^2 \tag{G.4}$$

위의 식에서

$$\boldsymbol{y} = \boldsymbol{P}^t\boldsymbol{x}$$

이차형식의 분해식 G.4 를 보면 행렬 $\boldsymbol{A}$의 모든 고유치가 0보다 크면 양정치 임을 알 수 있다. 또한 모든 고유치가 0보다 크거나 같으면 양반정치 임을 알 수 있다.

또한 $rank(\boldsymbol{A}) = rank(\boldsymbol{\Lambda})$이며 이는 0이 아닌 고유치의 개수가 행렬 $\boldsymbol{A}$의 계수(rank)임을 알 수 있다.

G.3. 멱등행렬

n-차원 행렬 $\boldsymbol{A}$ 가 다음과 같은 성질을 가지면 멱등행렬(idenpotent matrix)라고 부른다.

$$\boldsymbol{A}^2 = \boldsymbol{A}\boldsymbol{A} = \boldsymbol{A}$$

멱등행렬은 다음과 같은 성질을 가지고 있다.

- 멱등행렬의 고유치는 0 또는 1이다.
- 멱등행렬은 대각합이 계수와 같다.

$$tr(\boldsymbol{A}) = rank(\boldsymbol{A})$$

- 멱등행렬은 양반정치 행렬이다.
- $\boldsymbol{A}$ 멱등행렬이면 $\boldsymbol{I} - \boldsymbol{A}$도 멱등행렬이다.

특별히 대칭인 멱등행렬을 **사영행렬**(또는 투영행렬, projection matrix)라고 부른다.

최소제곱법에서 B.10 에서 나타난 행렬 $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t$는 멱등행렬이며 따라서 사영행렬이다.

G.4. 이차형식의 분포

G.4.1. 카이제곱 분포

만약 확률변수 x 이 표준 정규분포 $N(0,1)$ 을 따른다면 $y = x^2$ 은 자유도가 1인 카이제곱 분포 χ_1^2를 따른다. 더 나아가 n 개의 확률변수 $x_1, x_2, \dots, x_n$ 이 서로 독립이고 표준 정규분포 $N(0,1)$ 을 따른다면 제곱합 $v = x_1^2 + x_2^2 + \cdots + x_n^2$은 자유도가 n인 카이제곱 분포 χ_n^2를 따른다.

이렇게 카이제곱 분포는 표준 정규분포를 따르는 서로 독립인 확률변수들의 제곱값에 대한 분포이다.

G.4.2. 비중심 카이제곱 분포

만약 확률변수 x가 $N(\mu,1)$을 따른다면 $v=x^2$은 자유도가 1인 비중심 카이제곱 분포, $\chi_1^2(\lambda^2)$ 를 따른다. 여기서 비중심 카이제곱 분포의 자유도는 1이고 비중심모수 $\lambda^2=\mu^2$으로 주어진다.

이제 n개의 서로 독립인 확률 변수 $x_1,x_2,\cdots,x_n$이 각각 $N(\mu_i,1)$을 따른다면 $v=x_1^2+\cdots+x_n^2$은 자유도가 n이고 비중심 모수가 $\lambda^2=\sum_{i=1}^n \mu_i^2$인 비중심 카이제곱 분포, $\chi_n^2(\lambda^2)$ 를 따른다.

참고로 확률변수 x가 $N(0,1)$을 따른다면 $v=x^2$은 중심 카이제곱 분포, χ_1^2 를 따르며 이때는 비중심모수가 $\lambda^2=0$이다. 즉, 비중심모수가 0인 비중심 카이제곱 분포(non-central chi square distribution)를 중심 카이제곱 분포(central chi square distribution)라고 한다. 또한 중심 카이제곱 분포는 중심을 빼고 카이제곱 분포라고 부른다.

G.4.3. 이차형식의 분포

n개의 서로 독립인 확률 변수 $x_1,x_2,\cdots,x_n$이 각각 $N(\mu_i,\sigma^2)$를 따른다면 n-차원의 확률벡터 $\boldsymbol{x}$는 다음과 같은 다변량 정규분포를 따른다고 할 수 있다.

$$\boldsymbol{x}\sim N(\boldsymbol{\mu},\sigma^2\boldsymbol{I})$$

위에서 $\boldsymbol{\mu}^t=(\mu_1,\mu_2,\dots,\mu_n)$

이제 이차형식의 분포에 대하여 논의하자.

Theorem G.1 (이차형식의 분포)**.** n-차원의 확률벡터 $\boldsymbol{x}$가 $N(\boldsymbol{\mu},\sigma^2\boldsymbol{I})$를 따른다면 이차형식 $Q=\boldsymbol{x}^t\boldsymbol{A}\boldsymbol{x}$의 분포는 다음과 같다.

$$V=\frac{Q}{\sigma^2}=\frac{\boldsymbol{x}^t\boldsymbol{A}\boldsymbol{x}}{\sigma^2}\quad\equiv_d\quad\sum_{i=1}^n\lambda_i\frac{u_i^2}{\sigma^2} \tag{G.5}$$

위의 식에서 $x\equiv_d y$ 는 확률변수 x와 y가 동일한 분포를 가진다는 것을 의미한다.

G.5 에서 행렬 $\boldsymbol{A}$의 스펙트럴 분해는 $\boldsymbol{A}=\boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}^t$이며 λ_i는 행렬 $\boldsymbol{A}$의 고유치, 즉 행렬 $\boldsymbol{\Lambda}$의 대각원소이다.

$$\boldsymbol{x}^t\boldsymbol{A}\boldsymbol{x}=\boldsymbol{x}^t\boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}^t\boldsymbol{x}=\boldsymbol{u}^t\boldsymbol{\Lambda}\boldsymbol{u}=\sum_{i=1}^n\lambda_i u_i^2$$

또한 확률변수 u_i들은 서로 독립이며 정규분포 $N(\eta_i,\sigma^2)$를 따른다.

$$Cov(\boldsymbol{u})=Cov(\boldsymbol{P}^t\boldsymbol{x})=\boldsymbol{P}^tCov(\boldsymbol{x})\boldsymbol{P}=\boldsymbol{P}^t(\sigma^2\boldsymbol{I})\boldsymbol{P}=\sigma^2\boldsymbol{I}$$

여기서 $\eta_1,\eta_2,\dots,\eta_n$는 각각 u_i 들의 평균으로 다음과 같이 정의된다.

$$E(\boldsymbol{u}) = E(\boldsymbol{P}^t\boldsymbol{x}) = \boldsymbol{P}^t\boldsymbol{\mu} = \begin{bmatrix} \eta_1 \\ \eta_2 \\ \vdots \\ \eta_n \end{bmatrix} = \boldsymbol{\eta}$$

즉, *G.5* 에서 $u_1^2/\sigma^2, u_2^2/\sigma^2, \dots, u_n^2/\sigma^2$ 는 서로 독립이며 각각 자유도가 *1* 이고 비중심 모수가 $\eta_1^2/\sigma^2, \eta_2^2/\sigma^2, \dots, \eta_n^2/\sigma^2$ 인 비중심 카이제곱-분포를 따른다.

■

Theorem G.1 의 G.5 에서 나타난 이차형식의 분포는 비중심 카이제곱 분포를 따르는 서로 독립인 확률 변수들의 가중 평균과 같다는 것이다.

이는 이차형식의 분포가 비중심 카이제곱 분포를 따른다는 것이 아님을 주의해야 한다. 그러면 어느 경우에 이차형식의 분포가 비중심 카이제곱 분포를 따르는가 생각해 보자.

가장 쉽게 생각할 수 있는 경우가 G.5 에서 λ_i 들의 값들이 0 또는 1인 경우이다. 이러한 경우는 행렬
$\boldsymbol{A}$ 가 멱등행렬인 경우이다. 실제로 다음 정리는 이차형식의 분포가 비중심 카이제곱 분포를 따르는 필요충분 조건이 행렬
$\boldsymbol{A}$ 가 멱등행렬이라는 것을 말해준다.

Corollary G.1. n-차원의 확률벡터 $\boldsymbol{x}$ 가 $N(\boldsymbol{\mu}, \sigma^2\boldsymbol{I})$ 를 따른다면 이차형식 $Q = \boldsymbol{x}^t\boldsymbol{A}\boldsymbol{x}$ 의 분포가 자유도가 r 이며 다음과 같은 비중심 모수 λ^2 을 가지는 비중심 카이제곱 분포를 따르는 필요충분 조건은 $\boldsymbol{A}$ 가 멱등행렬이고 $rank(\boldsymbol{A}) = r$ 인 경우이다.

$$\lambda^2 = \frac{\boldsymbol{\mu}^t\boldsymbol{A}\boldsymbol{\mu}}{\sigma^2}$$

더 나아가 $\boldsymbol{A}\boldsymbol{\mu} = \boldsymbol{0}$ 이면 이차형식의 분포는 자유도가 r 인 (중심)카이제곱 분포를 따른다.

■

G.4.4. 이차형식의 독립

두 개의 이차형식이 독립일 조건은 다음 정리와 같다.

Theorem G.2 (이차형식의 독립)**.** n-차원의 확률벡터 $\boldsymbol{x}$ 가 $N(\boldsymbol{\mu}, \sigma^2\boldsymbol{I})$ 를 따른다고 하자. 두 이차형식 $Q_1 = \boldsymbol{x}^t\boldsymbol{A}\boldsymbol{x}$ 과 $Q_2 = \boldsymbol{x}^t\boldsymbol{B}\boldsymbol{x}$ 가 서로 독립일 필요충분 조건은 $\boldsymbol{A}\boldsymbol{B} = \boldsymbol{0}$ 이다.

■

G.4.5. 이차형식의 차이

만약 3 개의 이차형식 Q, Q_1, Q_2 가 있어서 다음과 같은 관계가 있다고 하자.

$$Q = \boldsymbol{x}^t \boldsymbol{A} \boldsymbol{x} = Q_1 + Q_2 = \boldsymbol{x}^t \boldsymbol{A}_1 \boldsymbol{x} + \boldsymbol{x}^t \boldsymbol{A}_2 \boldsymbol{x}$$

이러한 경우 두 이차형식 Q 과 Q_1 이 각각 카이제곱 분포를 따를 때 $Q_2 = Q - Q_1$ 이 카이제곱 분포를 따르는 조건이 중요하다. 다음 정리는 그 조건을 행렬 $\boldsymbol{A}_2$ 가 양반정치인 경우라는 것을 말해준다.

Theorem G.3 (이차형식의 차이)**.** n-차원의 확률벡터 $\boldsymbol{x}$ 가 $N(\boldsymbol{\mu}, \sigma^2 \boldsymbol{I})$ 를 따른다고 하자. 세 개의 이차형식 $Q = \boldsymbol{x}^t \boldsymbol{A} \boldsymbol{x}, Q_1 = \boldsymbol{x}^t \boldsymbol{A}_1 \boldsymbol{x}, Q_2 = \boldsymbol{x}^t \boldsymbol{A}_2 \boldsymbol{x}$ 가 있다고 하고 $Q = Q_1 + Q_2$ 인 관계를 가진다고 가정하자.

만약 Q/σ^2 이 $\chi_r^2(\lambda^2)$ 을 따르고 Q_1/σ^2 이 $\chi_{r_1}^2(\lambda_1^2)$ 을 따르며 행렬 $\boldsymbol{A}_2$ 가 양반정치 행렬이면 다음을 만족한다. 두 이차형식 Q_1 과 Q_2 는 서로 독립이다. 또한 이차형식 Q_2 는 자유도가 $r_2 = r - r_1$ 이고 비중심 모수가 $\lambda_2^2 = \lambda^2 - \lambda_1^2$ 인 비중심 카이제곱분포를 따른다.

■

G.5. 코크란의 정리

선형모형에서 자주 등장하는 제곱합들의 분해, 즉 이차형식의 분해를 생각할 때 각 제곱합들의 분포를 아는 것이 매우 중요하다. 다음에 제시된 코크란의 정리(Cochran's Theorem)는 총 제곱합을 분해했을 때 각 제곱합의 분포가 카이제곱 분포를 따를 조건을 말해준다.

$$\boldsymbol{x}^t \boldsymbol{x} = \sum_{j=1}^{k} \boldsymbol{x}^t \boldsymbol{A}_j \boldsymbol{x}$$

Theorem G.4 (COCHRAN'S THEOREM)**.** n-차원의 확률벡터 $\boldsymbol{x}$ 가 $N(\boldsymbol{\mu}, \sigma^2 \boldsymbol{I})$ 를 따른다고 하자. k 개의 이차형식 $Q_j = \boldsymbol{x}^t \boldsymbol{A}_j \boldsymbol{x}, j = 1, 2, \dots, k$ 를 생각하고 다음과 같은 관계를 가진다고 하자.

$$\boldsymbol{x}^t \boldsymbol{x} = \sum_{i=1}^{n} x_i^2 = \sum_{j=1}^{k} Q_j$$

즉, $\sum_{j=1}^{k} \boldsymbol{A}_j = \boldsymbol{I}$ 이다. 또한 $r_j = \operatorname{rank}(\boldsymbol{A}_j)$ 이고 $\lambda_j^2 = \boldsymbol{\mu}^t \boldsymbol{A}_j \boldsymbol{\mu}$ 라고 하자.

k 개의 이차형식 $Q_1, Q_2, \dots, Q_k$ 들이 모두 독립이고 각 이차형식 Q_j/σ^2 가 비중심 카이제곱 분포 $\chi_{r_j}^2(\lambda_j^2)$ 를 따를 필요충분 조건은 다음과 같다.

$$r_1 + r_2 + \cdots + r_k = n$$

■

이제 제곱합의 분포들에 대하여 지금까지 학습한 내용을 정리해보자. 만약 n-차원의 확률벡터 $\boldsymbol{x}$ 가 $N(\boldsymbol{\mu}, \sigma^2\boldsymbol{I})$ 를 따른다고 하고 위의 코크란의 정리와 같이 제곱합의 분해를 고려하자. 다음에 제시된 모든 문장은 서로 동치(equivalent)이다.

1. 이차형식 $Q_1, Q_2, \dots, Q_k$ 들이 모두 독립이다.
2. 모든 $j = 1, 2, \dots, k$ 에 대하여 이차형식 Q_j/σ^2 가 비중심 카이제곱 분포 $\chi^2_{r_j}\left(\lambda_j^2\right)$ 를 따른다.
3. $\boldsymbol{A}_1, \boldsymbol{A}_2, \dots, \boldsymbol{A}_k$ 가 모두 멱등행렬이다.
4. 모든 $j \neq k$ 에 대하여 $\boldsymbol{A}_j\boldsymbol{A}_k = \boldsymbol{0}$ 이다.
5. $r_1 + r_2 + \cdots + r_k = n$

H. 지수군 분포

> Caution
>
> 이 장에서는 벡터와 스칼라의 표기를 구분없이 모두 보통체로 사용한다.

H.1. 가능도함수와 그 성질

확률변수 y 가 확률밀도함수 $f(y;\theta)$를 따른다고 하자. 모수벡터 θ 에 대한 가능도함수 $L(\theta;y)$와 로그가능도함수 $\ell(\theta;y)$는 다음과 같이 정의한다.

$$L(\theta) = L(\theta;y) \equiv f(y;\theta), \quad \ell(\theta) = \ell(\theta;y) \equiv \log L(\theta;y)$$

로그가능도함수를 모수 θ로 한 번 미분한 도함수(gradient)를 스코어함수(score function) $s(\theta;y)$ 로 아래와 같이 정의한다. 또한 두 번 미분한 헤시안(hessian)의 음수를 관측피셔정보(observed Fisher information) $J(\theta;y)$ 라고 정의한다.

$$s(\theta) = s(\theta;y) \equiv \frac{\partial}{\partial\theta}\ell(\theta;y), \quad J(\theta) = J(\theta;y) \equiv -\frac{\partial^2}{\partial\theta^2}\ell(\theta;y)$$

위의 식에서 만약 모수벡터 θ의 차원이 p라면 $s(\theta)$는 $p\times 1$ 벡터이고 $J(\theta;y)$는 $p\times p$ 행렬이다.

로그가능도함수는 다음의 두 가지 중요한 방정식을 만족한다.

$$E\left\{\frac{\partial}{\partial\theta}\ell(\theta;y)\right\} = 0 \tag{H.1}$$

$$E\left\{\left[\frac{\partial}{\partial\theta}\ell(\theta;y)\right]\left[\frac{\partial}{\partial\theta}\ell(\theta;y)\right]^t\right\} + E\left\{\frac{\partial^2}{\partial\theta^2}\ell(\theta;y)\right\} = 0 \tag{H.2}$$

H.1 와 H.2 으로부터 다음과 같은 식이 유도되며

$$E[s(\theta;y)] = 0, \quad E[s(\theta;y)s^t(\theta;y)] = E[J(\theta;y)]$$

다음과 같은 공식이 주어진다.

$$
\begin{aligned}
Var[s(\theta;y)] &= E[s(\theta;y)s^t(\theta;y)] - \{E[s(\theta;y)]E[s^t(\theta;y)]\} \\
&= E[s(\theta;y)s^t(\theta;y)] - 0 \\
&= -E\left[\frac{\partial^2}{\partial\theta^2}\log f(y;\theta)\right] \\
&= E[J(\theta;y)] \\
&\equiv I(\theta)
\end{aligned}
$$

위의 식에서 스코어함수의 분산을 피셔정보(Fisher information)이라고 부르며 $I(\theta)$로 표기한다.

> **Caution**
>
> 관측피셔정보(observed Fisher information) $J(\theta;y)$ 는 모수와 확률변수로 정의되는 확률 이며 피셔정보(Fisher information) $I(\theta)$는 관측피셔정보의 기대값으로 모수만의 함수로서 더이상 확률변수가 아니다.

첫 번째 방정식 H.1 는 다음과 같이 적분과 미분의 교환에 의해 증명할 수 있다.

$$
\begin{aligned}
0 &= \frac{\partial}{\partial\theta}\int f(y;\theta)\,dy \\
&= \int \frac{\partial}{\partial\theta} f(y;\theta)\,dy \\
&= \int \frac{\frac{\partial}{\partial\theta}f(y;\theta)}{f(y;\theta)} f(y;\theta)\,dy \\
&= \int \frac{\partial}{\partial\theta}\ell(\theta;y) f(y;\theta)\,dy \\
&= E\left\{\frac{\partial}{\partial\theta}\ell(\theta;y)\right\}
\end{aligned}
$$

두 번째 방정식 H.2 는 아래와 같이 증명할 수 있다.

$$
\begin{aligned}
0 &= \frac{\partial}{\partial\theta}\int \frac{\partial}{\partial\theta}\ell(\theta;y) f(y;\theta)\,dy \\
&= \int \frac{\partial}{\partial\theta}\left\{ f(y;\theta)\left[\frac{\partial}{\partial\theta}\ell(\theta;y)\right]^t\right\} dy \\
&= \int \left\{ \left[\frac{\partial}{\partial\theta} f(y;\theta)\right]\left[\frac{\partial}{\partial\theta}\ell(\theta;y)\right]^t + f(y;\theta)\frac{\partial}{\partial\theta}\left[\frac{\partial}{\partial\theta}\ell(\theta;y)\right]^t\right\} dy \\
&= \int \left\{ \left[\frac{\partial}{\partial\theta}\ell(\theta;y)\right]\left[\frac{\partial}{\partial\theta}\ell(\theta;y)\right]^t f(y;\theta) + \left[\frac{\partial^2}{\partial\theta^2}\ell(\theta;y)\right] f(y;\theta)\right\} dy \\
&= E\left\{\left[\frac{\partial}{\partial\theta}\ell(\theta;y)\right]\left[\frac{\partial}{\partial\theta}\ell(\theta;y)\right]^t\right\} + E\left\{\frac{\partial^2}{\partial\theta^2}\ell(\theta;y)\right\}
\end{aligned}
$$

H.2. 독립표본

표본 $y_1, y_2, \dots, y_n$ 가 분포 $f(y;\theta)$에서 독립적으로 얻어졌다면 표본에 대한 가능도함수 $L_n(\theta)$ 은 다음과 같다.

$$L_n(\theta) = \prod_{i=1}^{n} f(y_i;\theta) \tag{H.3}$$

또한 표본에 대한 로그가능도함수 $\ell_n(\theta)$ 은 다음과 같다.

$$\ell_n(\theta) = \ell_n(\theta;y) = \log L_n(\theta) = \log \prod_{i=1}^{n} f(y_i;\theta) = \sum_{i=1}^{n} \log f(y_i;\theta) = \sum_{i=1}^{n} \ell(\theta;y_i) \tag{H.4}$$

표본에 의한 로그 가능도함수 $\ell_n(\theta)$를 미분한 값, 즉 표본에 의한 스코어 함수 $s_n(\theta)$는 다음과 같이 정의한다.

$$s_n(\theta) = \frac{\partial}{\partial\theta}\ell_n(\theta;y)$$

n개의 표본에 대한 관측피셔정보 $J_n(\theta)$와 피셔정보 $I_n(\theta)$도 한 개의 확률 변수 경우와 유사하게 다음과 같이 정의된다.

$$I_n(\theta) = E\left[J_n(\theta;y)\right] = E\left[-\frac{\partial^2}{\partial\theta\partial\theta^t}\ell_n(\theta;y)\right]$$

H.3. 지수군 분포

확률변수 y가 다음과 같은형태의 분포를 따른다면 y의 분포는 지수군(exponential family)에 속한다고 한다.

$$f(y;\theta,\phi) = \exp\left\{\frac{t(y)^t\xi(\theta) - b(\theta)}{a(\phi)} + c(y,\phi)\right\} \tag{H.5}$$

다시 쓰면

$$\log f(y;\theta,\phi) = \frac{t(y)^t\xi(\theta) - b(\theta)}{a(\phi)} + c(y,\phi)$$

지수군 분포 H.5 에서 $t(y)^t = [t_1(y), \dots, t_k(y)]$ 는 k 개의 충분통계량으로 구성된 벡터이고 k-차원 벡터 $\xi(\theta)^t = [\xi_1(\theta), \dots, \xi_k(\theta)]$를 **기본형 모수**(canonical parameter)라고 부른다. 또한 $a(\phi)$를 **스케일모수**(scale parameter)라고 부르며 많은 경우 $a(\phi) = a \times w$ 의 형태로 나타나며 여기서 w는 보통 가중치와 같은 역활을 한다. 또한 기본형 모수 $\xi(\theta)$ 는 y 의 평균 $\mu = E(y)$와 특별한 함수관계를 가진다.

지수군 분포는 일반적으로 H.5 과 같은 나타낼 수 있으며 많은 경우 모수 θ와 기본형 모수 $\xi(\theta)$는 일대일 대응관계를 가진다. 일대일 대응 관계가 아닌 경우 이를 곡선형 지수군(curved exponential family)라고 부른다. 따라서 지수군 분포의 성질을 간결하게 유도하고 설명하기 위해서 다음과 같은 단순화된 형태의 식을 사용하는 것이 편리하다. 이후 모든 성질의 유도는 아래 H.6 의 형태를 사용할 것이다.

$$\log f(y;\theta,\phi) = \frac{t^t\theta - b(\theta)}{a(\phi)} + c(y,\phi) \tag{H.6}$$

위의 H.6 은 충분통계량 벡터를 t 로 나타내고 대응하는 기본형 모수를 θ로 표시한 것이다. 또한 여기서 충분통계량 벡터 t의 평균을 μ라고 하자.

$$E(t) = \mu$$

H.3.1. 평균, 분산과 기본형 모수의 관계

이제 H.6 에 나오는 $b(\theta)$의 미분을 다음과 같이 표시하자.

$$b'(\theta) = \frac{\partial b(\theta)}{\partial \theta}, \quad b''(\theta) = \frac{\partial b(\theta)}{\partial \theta}$$

방정식 H.1 으로부터 다음과 같은 식이 유도된다..

$$\begin{aligned} 0 &= E\left(\frac{\partial \ell}{\partial \theta}\right) \\ &= E\left(t - b'(\theta)\right)/a(\phi)) \\ &= [E(t) - b'(\theta)]/a(\phi) \\ &= [\mu - b'(\theta)]/a(\phi) \end{aligned}$$

따라서 평균 μ와 함수 $b(\theta)$는 다음과 같은 관계가 성립한다.

$$E(t) = \mu = b'(\theta) \tag{H.7}$$

또한 방정식 H.2 으로부터 다음이 성립하고

$$E\left\{a^{-2}(\phi)[t - b'(\theta)][t - b'(\theta)]^t\right\} + E\left\{-a^{-1}(\phi)b''(\theta)\right\} = 0$$

따라서 t의 분산과 함수 $b(\theta)$는 다음과 같은 관계가 성립한다.

$$Var(t) = a(\phi)b''(\theta) \equiv a(\phi)v(\mu) \tag{H.8}$$

위의 식에서 $v(\mu) = b''(\theta)$ 으로 정의하고 분산함수(variance function)라고 부른다.

이제 분산함수의 정의로부터 μ, θ에 대하여 다음과 같은 관계가 얻어지고

$$\frac{\partial \mu}{\partial \theta} = \frac{\partial b'(\theta)}{\partial \theta} = b''(\theta) = v(\mu) \tag{H.9}$$

도함수의 역관계가 다음과 같이 주어진다.

$$\frac{\partial\theta}{\partial\mu} = \left[\frac{\partial\mu}{\partial\theta}\right]^{-1} = [b''(\theta)]^{-1} = v^{-1}(\mu) \tag{H.10}$$

H.3.2. 지수군 분포의 예제

Example H.1 (이항분포)**.** 확률변수 S 가 이항분포 $B(n,\mu)$를 따른다면(여기서 $\mu = p$ 성공확률) 표본 비율 $y = S/n$의 로그확률밀도함수는 다음과 같다.

$$\begin{aligned}\log f(y;\theta,\phi) &= \log\left\{\binom{n}{ny}\mu^{ny}(1-\mu)^{n-ny}\right\} \\ &= \frac{y\log\frac{\mu}{1-\mu} + \log(1-\mu)}{n^{-1}} + \log\binom{n}{ny} \\ &= \frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\end{aligned}$$

충분통계량 t는 표본비율 y이고 기본형 모수 θ와 평균 $E(y) = \mu = p$, 스케일 모수 $a(\phi)$은 다음과 같은 관계가 있다.

$$\theta = \log\frac{\mu}{1-\mu} = \log\frac{p}{1-p}, \quad b(\theta) = -\log(1-\mu), \quad a(\phi) = \frac{1}{n} \tag{H.11}$$

평균 μ를 기본형모수 θ의 함수로 역변환하면 θ의 로지스틱함수로 표현된다.

$$\mu = \frac{\exp(\theta)}{1+\exp(\theta)} = \frac{1}{1+\exp(-\theta)}, \quad 1-\mu = \frac{1}{1+\exp(\theta)}$$

또한 함수 $b(\theta)$를 기본형 모수로 나타내면

$$b(\theta) = -\log(1-\mu) = \log[1+\exp(\theta)]$$

특별히 $n = 1$인 경우는 베르누이 분포이다.

이항분포에서 평균 μ는 함수 b와 다음과 같은 관계가 있다.

$$b'(\theta) = \frac{\exp(\theta)}{1+\exp(\theta)} = \mu$$

또한

$$b''(\theta) = \frac{\exp(\theta)[1+\exp(\theta)] - [\exp(\theta)]^2}{[1+\exp(\theta)]^2} = \mu(1-\mu) = v(\mu)$$

따라서

$$Var(y) = a(\phi)v(\mu) = \frac{\mu(1-\mu)}{n}$$

Example H.2 (포아송분포)**.** 확률변수 y 가 포아송 분포 $poi(\mu)$를 따른다고 하자. 여기서 $E(y) = \mu$ 이다. 확률변수 y의 로그확률밀도함수는 다음과 같다.

$$\log f(y;\theta,\phi) = y\log\mu - \mu - log(y!)$$

충분통계량 t는 반응값 y 자체이고 기본형 모수 θ와 평균 $E(y) = \mu$, 스케일 모수 $a(\phi)$은 다음과 같은 관계가 있다. 기본형 모수 θ와 평균 μ, 스케일 모수 $a(\phi)$은 다음과 같은 관계가 있다.

$$\theta = \log\mu, \quad b(\theta) = \mu, \quad a(\phi) = 1$$

평균 μ를 기본형모수 θ의 함수로 역변환하면 θ의로그함수로 표현된다. 또한 함수 $b(\theta)$를 기본형 모수로 나타내면 다음과 같다.

$$\mu = \exp(\theta), \quad b(\theta) = \mu = \exp(\theta)$$

따라서 포아송분포에서는 평균 μ는 함수 b와 다음과 같은 관계가 있다.

$$b'(\theta) = \exp(\theta) = \mu, \quad b''(\theta) = \exp(\theta) = \mu = v(\mu)$$

따라서

$$Var(y) = a(\phi)v(\mu) = \mu$$

Example H.3 (정규분포)**.** 확률변수 y 가 정규분포 $N(\mu,\sigma^2)$를 따른다고 하자. 확률변수 y의 로그확률밀도함수는 다음과 같다. 이때 모수벡터를 $\theta^t = (\mu,\sigma^2)$이다.

$$\begin{aligned}
\log f(y;\theta,\phi) &= -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log\sigma^2 - \frac{(y-\mu)^2}{2\sigma^2} \\
&= -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log\sigma^2 - \frac{y^2 - 2y\mu + \mu^2}{2\sigma^2} \\
&= \left[y\frac{\mu}{\sigma^2} - y^2\frac{1}{2\sigma^2}\right] - \left[\frac{1}{2}\log\sigma^2 + \frac{\mu^2}{2\sigma^2}\right] - \frac{n}{2}\log(2\pi)
\end{aligned}$$

충분 통계량과 기본형 모수는 다음과 같다.

$$t(y)^t = (y, y^2), \quad \xi(\theta)^t = \left(\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2}\right), \quad a(\phi) = 1$$

또한

$$b(\theta) = \frac{1}{2}\log\sigma^2 + \frac{\mu^2}{2\sigma^2}$$

H.4. 최대가능도추정법

모수 θ 에 대한 최대가능도 추정량(Maximum Likelihood Estimator;MLE) $\hat{\theta}$는 가능도 함수를 최대로 하는 값으로 정의된다.

$$\hat{\theta}_{MLE} = \arg\max_{\theta} L_n(\theta)$$

많은 경우 가능도 함수를 최대화하는 값을 구하기 어려우므로 가능도 함수의 로그 함수, 즉 로그가능도함수를 최대로 하는 값으로 최대가능도 추정량을 구한다.

$$\hat{\theta}_{MLE} = \arg\max_{\theta} \ell_n(\theta)$$

만약 로그가능도 함수가 모수 θ에 대하며 미분가능한 함수이면 최대가능도 추정량은 다음과 같은 방정식에 의하여 구할 수 있다.

$$\frac{\partial}{\partial\theta}\ell_n(\theta; y) = s_n(\theta) = 0$$

최대가능도 추정량은 적당한 조건하에서 다음과 같은 점근적 성질(Asymptotical properties)을 가진다.

- $\hat{\theta}_{MLE}$는 모수의 참값 θ_0로 확률적 수렴한다.

$$\hat{\theta}_{MLE} \to_p \theta_0 \quad \text{as } n \to \infty$$

- 최대가능도추정량 $\hat{\theta}_{MLE}$는 점근적으로 정규분포를 따른다.

$$\hat{\theta}_{MLE} \sim_d N(\theta_0, I_n^{-1}(\theta_0))$$

I. 모형선택의 정보 기준

I.1. Kullback-Leibler 정보

앞에서 소개한 AIC 는 두 개의 분포에 대한 거리를 반양하는 정보 기준에 의하여 유도된 모형선택의 측도이다. 이제 AIC 가 어떻게 두 개의 분포에 대한 거리에서 유도되는지 알아보자.

Definition I.1 (Kullback-Leibler 정보)**.** 두 개의 분포 F 와 G 가 있다고 가정하고 각 분포에 대한 확률밀도함수가 f 와 g 로 주어졌다. **KL-정보(Kullback-Leibler information)** 는 두 개의 분포 F 와 G 의 거리를 다음과 같이 정의하는 정보기준이다.

$$\begin{aligned} I(g;f) &= E_G\left[\log\left\{\frac{g(y)}{f(y)}\right\}\right] \\ &= \int \log\left\{\frac{g(y)}{f(y)}\right\} g(y)dy \\ &= \int \left[\log g(y) - \log f(y)\right] g(y)dy \end{aligned}$$

■

두 분포의 거리를 나타내는 KL-정보는 다음과 같은 성질을 가진다.

- $I(g;f) \geq 0$
- 만약 $I(g;f) = 0$ 이면 $g(y) = f(y)$ a.e.

Example I.1 (정규분포의 거리)**.** 두 개의 정규분포 $F \equiv N(\mu, \sigma^2)$ 와 $G \equiv N(\xi, \tau^2)$ 을 고려하자.

분포 G 에서 $(y-\mu)^2$ 의 기대값은 다음과 같이 주어지므로

$$E_G(y-\mu)^2 = E_G(y-\xi+\xi-\mu)^2 = \tau^2 + (\xi-\mu)^2$$

분포 G 를 가정하고 확률밀도함수 f 의 로그에 대한 기대값은 다음과 같다.

$$E_G[\log f(y)] = E_G\left[-\frac{1}{2}\log(2\pi\sigma^2) - (y-\mu)^2/(2\sigma^2)\right]$$
$$= -\frac{1}{2}\log(2\pi\sigma^2) - [\tau^2 + (\xi-\mu)^2]/(2\sigma^2)$$

유사한 방법으로 분포 G 에서 확률밀도함수 g 의 로그에 대한 기대값도 다음과 같이 구할 수 있다.

$$E_G[\log g(y)] = -\frac{1}{2}\log(2\pi\tau^2) - \frac{1}{2}$$

위에서 구한 제곱합의 기대값을 이용하면 두 개의 정규분포 F 와 G 의 KL-정보는 다음과 같이 주어진다.

$$I(g;f) = E_G[\log g(y)] - E_G[\log f(y)] = \frac{1}{2}\left\{\log\frac{\sigma^2}{\tau^2} + \frac{\tau^2 + (\xi-\mu)^2}{\sigma^2} - 1\right\}$$

먼저 $\boldsymbol{y} = \{y_1, y_2, \cdots, y_n\}$ 가 참분포(true distribution) $G(y)$ or $g(y)$ 에서 독립적으로 얻은 확률변수라고 하자.

이제 $F(y)$ or $f(y)$ 는 얻어진 자료에 대한 모형으로 사용하고자 하는 분포이며 이를 후보 모형(candidate distribution)이라고 하다. 일반적으로 고려하는 분포들을 모아놓은 집합을 후보 모형군(family of candidate distributions) 이라고 하며 $\{f(y|\theta)|\theta \in \Theta\}$ 라고 표기한다. 이런 후보 모형군은 분포를 결정하는 모수들을 모아놓은 모수 집합 Θ으로 표시하기도 한다.

이제 후보 모형군에 속하는 임의의 분포 $f(y) = f(y|\theta)$ 와 참모형 $g(y)$ 의 KL-정보는 다음과 같다.

$$I(g;f) = E_G[\log g(y)] - E_G[\log f(y)]$$

위에서 정의한 $I(g;f)$ 의 값이 작을수록 좋은 것이며 이는 고려한 후보 분포 $f(y)$ 가 참모형 $g(y)$ 에 더 가깝다는 의미이기 때문이다.

I.2. 가능도 함수

이제 자료에 대한 후보 분포를 $F(y)$ 또는 $f(y)$ 라고 하고 로그 가능도 함수의 기대값을 고려하자. 이 때 기대값은 참분포에서 계산된 기대값이다.

$$E_G[\log f(y)] = \int \log f(y) g(y) dy \tag{I.1}$$

실제 자료를 분석하는 경우 참분포를 알 수 없기 때문에 로그 가능도 함수의 기대값 I.1 을 구하는 것은 불가능하다. 따라서 참분포에 대한 추정을 하여 구해야 하는데 참분포 G에 대한 추정량은 자료를 이용하여 구할 수 있는 경험분포함수(emprical distribution)를 이용할 수 있다.

표본 자료 $\boldsymbol{y} = \{y_1, y_2, \cdots, y_n\}$ 를 이용하여 얻은 참분포 G에 대한 경험적 추정 분포는 다음과 같다.

$$\hat{G}(y) = \frac{1}{n}\sum_{i=1}^{n} I(y \le y_i) \tag{I.2}$$

이제 자료에서 얻은 경험분포를 사용하여 로그 가능도 함수의 기대값 I.1 에 대한 추정량을 구하면 다음과 같다.

$$E_{\hat{G}}[\log f(y)] = \int \log f(y) d\hat{G}(y) = \frac{1}{n}\sum_{i=1}^{n} \log f(y_i) \tag{I.3}$$

대수의 법칙(the law of large numbers)에 의하여 표본의 개수 n 이 커지만 다음이 성립한다.

$$\frac{1}{n}\sum_{i=1}^{n} \log f(y_i) \to_{a.e.} E_G[\log f(y)]$$

표본에 대한 모수적 확률 모형들을 모아놓은 집합, 모형공간 $\{f(y|\boldsymbol{\theta}) | \boldsymbol{\theta} \in \boldsymbol{\Theta} \subset R^p\}$. 을 고려하자. 표본자료 $\boldsymbol{y} = \{y_1, y_2, \cdots, y_n\}$로 부터 얻은 로그가능도함수는 다음과 같이 주어진다.

일단 여기서는 참분포 $g(y)$ 가 모수적 확률 모형 집합에 속하는 분포라고 가정하자.

$$g(x) = f(y|\boldsymbol{\theta}_0) \text{ for some } \boldsymbol{\theta}_0 \in \boldsymbol{\Theta}$$

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log f(x_i|\boldsymbol{\theta})$$

최대가능도 추정량 $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(\boldsymbol{y})$은 다음과 같이 로그가능도함수를 최대로 하는 추정량이다.

$$\hat{\boldsymbol{\theta}} = arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \ell(\boldsymbol{\theta})$$

이제 $\boldsymbol{\theta}_0$ 을 다음에 주어진 방정식의 근이라고 하자.

$$\int \frac{\partial \log f(y|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} f(y|\boldsymbol{\theta}) dy = 0$$

위의 식에서 다음과 같이 로그 확률함수를 모수벡터로 미분한 양을 스코어함수(score function) $u(\boldsymbol{\theta};\boldsymbol{y})$이라고 부른다.

$$u(\boldsymbol{\theta};\boldsymbol{y}) = \frac{\partial \log f(y|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$$

따라서 $\boldsymbol{\theta}_0$ 을 다음에 주어진 방정식의 근이다.

$$E_\theta[u(\boldsymbol{\theta};\boldsymbol{y})] = 0$$

만약 정상적인 조건들(regularity conditions)이 만족하면 다음과 같은 결과를 얻을 수 있다.

- 가능도 방정식 $\partial \ell(\boldsymbol{\theta})/\partial \boldsymbol{\theta} = 0$ 은 점근적으로(with probability 1) 방정식의 해 $\hat{\boldsymbol{\theta}}$ 를 가진다.
- 최대가능도추정량(MLE) $\hat{\boldsymbol{\theta}}$ 는 점근적으로 $\boldsymbol{\theta}_0$ 에 수렴한다.
- 최대가능도추정량은 점근적으로 다음과 같은 정규분포를 따른다.

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \to_d N(0, I(\boldsymbol{\theta}_0)) \tag{I.4}$$

위의 식에서 $I(\boldsymbol{\theta})$ 는 피셔정보(Fisher information matrix) 이라고 부르며 다음과 같이 정의된다.

$$I(\boldsymbol{\theta}) = \int f(y|\boldsymbol{\theta}) \frac{\partial \log f(y|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial \log f(y|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^t} dy$$

위에서 최대가능도 추정량에 대한 모든 성질은 참분포 $g(y)$ 가 모수적 확률 모형들의 집합 $\{f(y|\boldsymbol{\theta}) | \boldsymbol{\theta} \in \boldsymbol{\Theta} \subset R^p\}$ 에 속한다고 가정하였다

$$g(x) = f(y|\boldsymbol{\theta}_0)) \text{ for some } \boldsymbol{\theta}_0$$

만약 $g(y)$ 가 우리가 고려하고 있는 모수적 모형들의 집합에 속해있지 않다면 앞에서 구한 최대가능도 추정량의 점근적 성질들은 어떻게 될까?

$$g(x) \neq f(y|\boldsymbol{\theta}) \text{ for all } \boldsymbol{\theta}$$

이제 참분포 $g(y)$ 가 모수적 확률 모형들의 집합에 속하지 않을 수도 있다고 가정하자.

또한 $\boldsymbol{\theta}_0$ 를 다음 방정식의 근이라고 하자.

$$\int g(y) \frac{\partial \log f(y|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} dy = 0$$

이러한 가정하에서는 다음이 성립한다.

- 최대가능도추정량(MLE) $\hat{\boldsymbol{\theta}}$ 는 점근적으로 $\boldsymbol{\theta}_0$ 에 수렴한다.
- 최대가능도추정량은 점근적으로 다음과 같은 정규분포를 따른다.

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \to_d N(0, J^{-1}(\boldsymbol{\theta}_0) I(\boldsymbol{\theta}_0) J^{-1}(\boldsymbol{\theta}_0))$$

위의 식에서 $I(\boldsymbol{\theta})$ 와 $J(\boldsymbol{\theta})$ 는 다음과 같이 정의되는 양이다.

$$I(\boldsymbol{\theta}) = \int g(y) \frac{\partial \log f(y|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial \log f(y|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^t} dy$$
$$J(\boldsymbol{\theta}) = -\int g(y) \frac{\partial^2 \log f(y|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\boldsymbol{\theta}^t} dy$$

주목할 점은 만약 참분포가 모수적 분포의 집합에 속하면, 즉 $g(y) = f(y|\boldsymbol{\theta}_0)$ 이면 $I(\boldsymbol{\theta}_0) = J(\boldsymbol{\theta}_0)$ 이 성립하고 I.4 이 성립힌다.

I.3. AIC

이제 $\boldsymbol{y} = \{y_1, y_2, \cdots, y_n\}$ 는 참모형 $g(y)$에서 얻어진 독립표본이라고 하자. 모수적 확률 분포의 집합 $\{f(y|\boldsymbol{\theta})|\boldsymbol{\theta} \in \boldsymbol{\theta} \subset R^p\}$을 고려한다. 또한 모르는 모수 $\boldsymbol{\theta}$ 는 최대가능도 추정량 $\hat{\boldsymbol{\theta}}$ 에 의하여 추정된다고 하자.

이제 추정된 모수로 구한 확률분포 $f(y|\hat{\boldsymbol{\theta}})$ 와 참분포 $g(y)$ 가 얼마나 차이가 있는지 관심이 있으며 이 거리를 K-L 정보 를 이용하여 구하면 다음과 같다.

$$I(g(z);\ f(z|\hat{\boldsymbol{\theta}})) = E_G[\log g(z)] - E_G[\log f(z|\hat{\boldsymbol{\theta}})] \tag{I.5}$$

위의 I.5 에서 기대값 $E_G()$ 는 참분포 $g(z)$에 추출한 새로운 확률변수 z에 대한 기대값이며 표본으로 부터 구한 $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(\boldsymbol{y}_n)$ 는 표본 $\boldsymbol{y}$의 함수로서 기대값 $E_G()$과 관계없이 고정된 양이다.

위의 I.5 에 주어진 K-L 정보에서 앞의 기대값 $E_G[\log g(z)]$ 은 언제나 주어진 상수이므로 참분포와 모수적 분포의 거리를 나타내는 양으로 K-L 정보에서 뒤의 기대값이 모수적 분포의 적함도를 반영하는 중요한 측도이다.

$$E_G[\log f(z|\hat{\boldsymbol{\theta}})] = \int \log f(z|\hat{\boldsymbol{\theta}}) g(z) dz \tag{I.6}$$

$I(g(z);\ f(z|\hat{\boldsymbol{\theta}})) \geq 0$ 이므로 위의 I.6 에 주어진 양이 크면 클수록 참분포와 거리가 작아지므로 더 좋은 분포의 추정량이라고 말할 수 있다.

여기서 중요한 점은 실제 문제에서는 참분포 $g(y)$ 를 알 수 없으며 I.6 의 값을 추정하려면 참분포 $g(y)$ 에 대한 추정량이 필요하다. 가장 간단한 추정량은 참분포의 분포함수 G 를 경험적 표본 분포함수로 추정하는 것이다. 아래는 참분포의 분포함수에 대한 단순 추정량 $\hat{G}$ 이다.

$$E_{\hat{G}}[\log f(z|\hat{\boldsymbol{\theta}})] = \int \log f(z|\hat{\boldsymbol{\theta}}) d\hat{G}(z) = \frac{1}{n} \sum_{i=1}^{n} \log f(y_i|\hat{\boldsymbol{\theta}}) \tag{I.7}$$

사실 I.7 는 최대값을 가지는 로그 가능도함수를 n 으로 나눈 양이다

이제 I.7 으로 주어진 추정량으로 I.6 를 추정해야 하는데 사실 두 양이 모두 표본 $y_1, y_2, \cdots, y_n$에 의해 얻어진 $\hat{\boldsymbol{\theta}}$의 함수이다. 따라서 두 통계량 모두 참분포 $g(y)$ 에서 얻어진 표본 $y_1, y_2, \cdots, y_n$ 도 고려해야 한다.

$$E_{G(y)}\left[\frac{1}{n}\sum_{i=1}^{n}\log f(y_i|\hat{\boldsymbol{\theta}})\right] =_? E_{G(y)}\left[E_{G(z)}[\log f(z|\hat{\boldsymbol{\theta}})]\right] \tag{I.8}$$

불행하게도 위의 두 기대값의 값이 다르기 때문에 I.7 에 나타난 추정량은 I.6 의 불편추정량이 아니다. 따라서 I.7 에 나타난 추정량은 다음과 같이 주어진 편이(bias)를 구해서 보종할 수 있다.

$$b(G) = E_{G(\boldsymbol{y})}\left[\log f(\boldsymbol{y}|\hat{\boldsymbol{\theta}}) - E_G(\log f(z|\hat{\boldsymbol{\theta}}(\boldsymbol{y})))\right] \tag{I.9}$$

I.9 에 나타난 편이에 대한 해석은 다음과 같이 말할 수 있다.

- 실제로 추정해야 하는 측도는 $E_{G(y)}[E_{G(z)}[\log f(z|\hat{\boldsymbol{\theta}})]]$ 이며 이는 표본 $\boldsymbol{y}$ 에서 추정량을 이용하여 새로운 반응변수 z 를 예측할 때의 측도이다.
- 하지만 표본 추정량에 근거한 $E_{G(y)}[\sum_{i=1}^{n}\log f(y_i|\hat{\boldsymbol{\theta}})/n]$은 표본 $\boldsymbol{y}$ 에서 추정량을 이용하여 다시 표본에서 얻은 반응값을 예측하는 측도이다.
- 따라서 두 측도 사이에는 차이가 존재하며 표본 추정량에 근거한 $E_{G(y)}[\sum_{i=1}^{n}\log f(y_i|\hat{\boldsymbol{\theta}})]$는 실제로 과대 추정된다(과적합 발생).
- 이러한 이유로 가능도함수로 나타난 측도 $-2\sum_{i=1}^{n}\log f(y_i|\hat{\boldsymbol{\theta}})$ 를 추정된 편이로 보정해주어야 올바른 추론이다.

만약 우리가 I.9 에 나타난 편이 $b(G)$을 추정할 수 있다면 우리가 찾은 모수적 최적 모형과 참모형의 K-L거리에 근거한 모형의 적합도 $IC(\boldsymbol{y};\hat{\boldsymbol{\theta}})$ 를 다음과 같이 정의할 수 있다.

$$\begin{aligned} IC(\boldsymbol{X};\hat{G}) &= -2(\text{log-likelihood of the model} - \text{bias estimator}) \\ &= -2\sum_{i=1}^{n}\log f(y_i|\hat{\boldsymbol{\theta}}) + 2(\text{bias estimator } b(G)) \end{aligned}$$

본 강의에서는 구하지 않겠지만 I.9 에 나타난 편이 $b(G)$을 는 다음과 같이 두 행렬의 곱에 대각원소의 합으로 나타난다.

$$b(G) = tr[I(\boldsymbol{\theta}_0)J(\boldsymbol{\theta}_0)^{-1}] \tag{I.10}$$

위의 식에서 $I(\boldsymbol{\theta})$ 와 $J(\boldsymbol{\theta})$는 다음과 같이 정의된 양이다.

$$\begin{aligned} I(\boldsymbol{\theta}) &= \int g(y)\frac{\partial \log f(y|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\frac{\partial \log f(y|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^t}dy \\ J(\boldsymbol{\theta}) &= -\int g(y)\frac{\partial^2 \log f(y|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\boldsymbol{\theta}^t}dy \end{aligned}$$

만약 참모형 $g(z)$ 이 다음과 같이 고려한 모수적 모형의 집합 $\{f(y|\boldsymbol{\theta})|\boldsymbol{\theta}\in\boldsymbol{\theta}\subset R^p\}$ 에 속한다면

$$\text{If } g(y) = f(y|\boldsymbol{\theta}_0) \quad \text{for some } \theta_0, \quad \text{then} \quad I(\boldsymbol{\theta}) = J(\boldsymbol{\theta})$$

이런 조건에서는 I.10 에 주어진 편이가 다음과 같이 모수의 개수로 나타난다.

$$b(G) = tr[I(\boldsymbol{\theta}_0)J(\boldsymbol{\theta}_0)^{-1}] = tr(\boldsymbol{I}_p) = p$$

따라서 K-L 정보기준으로 유도된 참모형과 최대가능도 추정법으로 선택된 분포의 거리로 표현되는 AIC(Akaike Information Criteria)는 다음과 같이 정의된다.

$$AIC = -2\sum_{i=1}^{n} \log f(y_i|\hat{\boldsymbol{\theta}}) + 2p$$

I.4. BIC

베이지안 정보 기준(BIC) 또는 슈바르츠 정보 기준(SIC, SBC, SBIC)은 모형공간에서 최적의 모형을 선택하는 기준으로, 일반적으로 BIC가 낮은 모델이 선호된다. 이는 부분적으로 (AIC)과 밀접한 관련이 있다.

먼저 BIC 를 유도하는 과정을 살펴보려면 베이지안 통계에서 나타나는 모수에 대한 사전분포(prior distribution) $g(\theta) = p(\theta)$ 을 고려해야 한다.

또한 모형 공간을 $\mathcal{M} = \{m_1, \ldots, m_M\}$ 이라고 하자. 또한 $m \in \mathcal{M}$ 을 모형공간에 속하는 하나의 모형을 나타낸다.

$\ell(\theta)$ 을 로그 가능도 함수라고 하자. 아래 식에서 $f(y \mid \theta, m)$ 은 모형 m 에서 주어진 모수 θ 에 대한 반응변수 y 의 조건부 확률밀도함수이다.

$$\ell(\theta) = \log f(y \mid \theta, m)$$

다음으로 로그 가능도 함수와 사전분포를 이용하여 함수 g 와 h 를 다음과 같이 정의하자.

$$\begin{aligned} g(\theta) &= p(\theta \mid m) \\ h(\theta) &= \frac{1}{n}\ell(\theta). \end{aligned} \tag{I.11}$$

베이지안 통계에 의하면 반응변수 y에 대한 주변분포(marginal distribution) $p(y \mid m)$ 는 다음과 같이 주어진다.

$$\begin{aligned} p(y \mid m) &= \int_{\Theta} f(y \mid \theta, m)p(\theta \mid m)\mathrm{d}\theta \\ &= \int_{\Theta} \exp[nh(\theta)]g(\theta)\mathrm{d}\theta \end{aligned} \tag{I.12}$$

This is an integral suitable for Laplace approximation which states that

이제 위의 적분에 대한 라플라스 근사를 적용하면 다음과 같이 주어진다.

$$\int_{\Theta} \exp[nh(\theta)]g(\theta)\mathrm{d}\theta = \left(\sqrt{\frac{2\pi}{n}}\right)^{p} \exp\left[nh\left(\theta_0\right)\right]\left(g\left(\theta_0\right)\left|H\left(\theta_0\right)\right|^{-1/2} + O(1/n)\right) \tag{I.13}$$

위의 식에서

θ_0 는 $h(\theta)$ 를 최대화하는 값이고 $H\left(\theta_0\right)$ 는 θ_0 에서 계산된 $h(\theta)$수의 헤시안 행렬(Hessian matrix) 이다.

우리는 지금 최대가능도 추정법을 다루고 있으므로 위의 식에서 나타난 θ_0 는 최대가능도 추정량 $\hat{\theta}$ 이다.

$$\hat{\theta} = \arg\max_{\theta} \ell(\theta).$$

위의 결과에서 I.13 를 I.11 와 I.12 에 적용하면 다음 결과를 얻을 수 있다.

$$p(y \mid m) \approx \left(\sqrt{\frac{2\pi}{n}}\right)^{p} f(y \mid \hat{\theta}, m)p(\hat{\theta} \mid m)|H(\hat{\theta})|^{-1/2}. \tag{I.14}$$

I.14 에 로그를 취하고 -2 다음과 같은 결과를 얻는다.

$$-2\log p(y \mid m) \approx -2\ell(\hat{\theta}) + p\log n - p\log(2\pi) - 2\log p(\hat{\theta} \mid m) + \log|J(\hat{\theta})|. \tag{I.15}$$

표본의 크기가 커지면$(n \to \infty)$, I.15 의 마지막 3개의 항은 $O_p(1)$ 으로 나머지 항에 비교하여 무시할 수 있다.

이제 모형 $\mathcal{M} = \{m_1, \ldots, m_M\}$ 에서 최적의 모형을 선택하는 기준은 모형에 대한 사후분포 $p\left(m_j \mid y\right)$ 를 최대로 하는 기준을 사용하는데 이는 우리가 근사한 주변분포 $p\left(y \mid m_j\right)$ 에 비례하는 것을 이용하여 다음과 같이 모형의 선택 기준으로 BIC 를 정의할 수 있다.

$$\mathrm{BIC}(m) = -2\log f(y \mid \hat{\theta}, m) + p\log n.$$

J. R-실습: 중회귀 모형 적합

아래 자료는 강근석 and 유형조 (2016) 의 예제 3.3에 나온 중고차 가격 자료를 이용한 R 실습입니다.

`regbook` 패키지를 참고서적 사이트에 따라 설치할 수 있습니다.

J.1. 예제 3.3 자료

중고차 가격에는 어떠한 변수들이 영향을 주는 걸까? 국내의 중고차 사이트에는 기본적으로 연식, 차종, 색, 도어, 엔진, 배기량, 연료, 주행거리 등의 정보가 올라와 있다. 자료 `usedcars`는 국내 유명 중고차 사이트에서 2007년에 수집한 것으로 가솔린 엔진의 중형차를 대상으로 각 중고차에 대한 가격, 연식, 주행거리, 배기량, 변속기종류를 조사하였다. 예제 3.3에 나온 중고차 가격자료를 이용한 R 실습입니다.

이 자료는 다음과 같은 열로 구성되어 있다.

- `price` : 가격(만원)
- `year` : 연식(개월)은 현재까지 몇 개월이 지났는지를 계산하였다.
- `mileage` : 주행거리(km)
- `cc`: 배기량(cc)
- `automatic` : 변속기종류. 수동(0), 자동(1).

먼저 자료와 변수의 개수를 알아보고 자료가 어떻게 구성되어 있는지 살펴봅시다.

```
dim(usedcars)
```

```
[1] 30  5
```

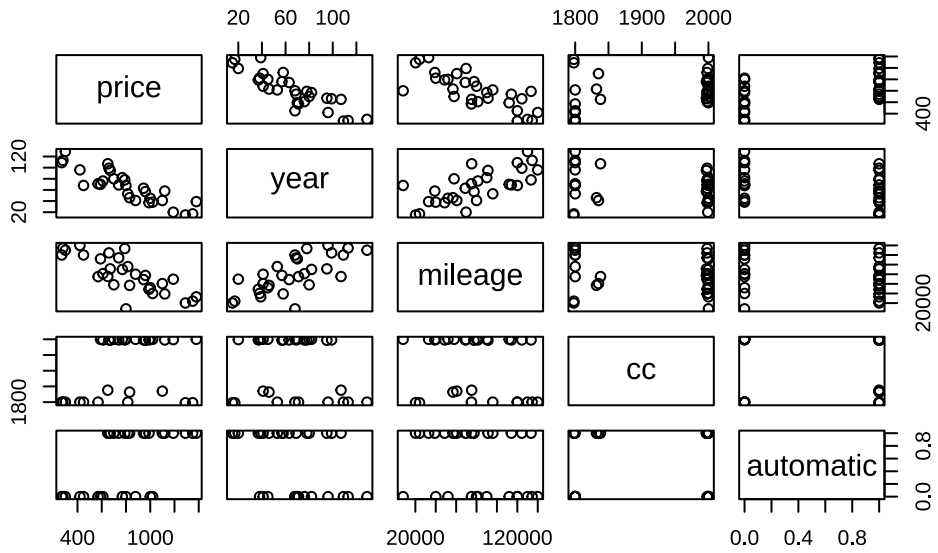
```
head(usedcars)
```

```
  price year mileage   cc automatic
1   790   78  133462 1998         1
2  1380   39   33000 2000         1
3   270  109  120000 1800         0
4  1190   20   69727 1999         1
5   590   70  112000 2000         0
6  1120   58   39106 1998         1
```

J.1.1. 산점도 행렬

각 변수간의 산점도를 모두 그려보면 변수들간의 관계를 쉽게 파악할 수 있다.

```
pairs(usedcars)
```



J.2. 중회귀 모형의 적합

중고차의 가격을 예측하기 위한 중회귀 모형을 적합해보자. 반응변수는 price(y)이고 설명변수는 year(x_1), mileage(x_2), cc(x_3), automatic(x_4)이다.

$$\text{price} = \beta_0 + \beta_1 \text{year} + \beta_2 \text{mileage} + \beta_3 \text{cc} + \beta_4 \text{automatic} + \epsilon$$

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \epsilon_i$$

중회귀 모형을 적합하느 함수 lm 의 결과는 fit0에 저장한다.

```
fit0 <- lm(price ~ year + mileage + cc + automatic, usedcars)
```

계획행렬 $\boldsymbol{X}$ 은 다음과 같이 구할 수 있다.

```
model.matrix(fit0)
```

```
   (Intercept) year mileage   cc automatic
1            1   78  133462 1998         1
2            1   39   33000 2000         1
3            1  109  120000 1800         0
4            1   20   69727 1999         1
5            1   70  112000 2000         0
6            1   58   39106 1998         1
7            1   53   95935 1800         1
8            1   68  120000 1800         0
9            1   15   20215 1798         1
10           1   96  140000 1800         0
11           1   63   68924 1998         1
12           1   82   90000 2000         0
13           1   76   81279 1998         0
14           1   17   24070 1798         1
15           1   38   40000 2000         0
16           1   46   56887 1832         1
17           1   95   91216 1997         1
18           1   37   48680 1998         1
19           1   68    8000 2000         0
20           1   41   60634 1835         1
21           1   69  114131 1998         1
22           1   71   75000 1800         0
23           1   99  124417 1998         1
24           1  129  130000 1800         0
25           1   57   77559 1997         1
26           1  107   75216 1838         1
27           1   45   52000 2000         0
28           1   80   58000 2000         1
29           1  113  134500 1800         0
30           1   41   80000 2000         0
attr(,"assign")
[1] 0 1 2 3 4
```

fit0 에 저장된 결과를 다음과 같이 함수 str을 이용하여 볼 수 있다.

```
str(fit0)
```

```
List of 12
 $ coefficients : Named num [1:5] 525.28696 -5.79964 -0.00226 0.38879 165.31263
  ..- attr(*, "names")= chr [1:5] "(Intercept)" "year" "mileage" "cc" ...
 $ residuals    : Named num [1:30] 76.98 212.69 -51.4 -4.01 -53.45 ...
```

```
 ..- attr(*, "names")= chr [1:30] "1" "2" "3" "4" ...
$ effects      : Named num [1:30] -4407 -1434 -369 -229 419 ...
 ..- attr(*, "names")= chr [1:30] "(Intercept)" "year" "mileage" "cc" ...
$ rank         : int 5
$ fitted.values: Named num [1:30] 713 1167 321 1194 643 ...
 ..- attr(*, "names")= chr [1:30] "1" "2" "3" "4" ...
$ assign       : int [1:5] 0 1 2 3 4
$ qr           :List of 5
 ..$ qr   : num [1:30, 1:5] -5.477 0.183 0.183 0.183 0.183 ...
 .. ..- attr(*, "dimnames")=List of 2
 .. .. ..$ : chr [1:30] "1" "2" "3" "4" ...
 .. .. ..$ : chr [1:5] "(Intercept)" "year" "mileage" "cc" ...
 .. ..- attr(*, "assign")= int [1:5] 0 1 2 3 4
 ..$ qraux: num [1:5] 1.18 1.18 1.08 1.03 1.26
 ..$ pivot: int [1:5] 1 2 3 4 5
 ..$ tol  : num 1e-07
 ..$ rank : int 5
 ..- attr(*, "class")= chr "qr"
$ df.residual  : int 25
$ xlevels      : Named list()
$ call         : language lm(formula = price ~ year + mileage + cc + automatic, data = usedca
$ terms        :Classes 'terms', 'formula'  language price ~ year + mileage + cc + automatic
 .. ..- attr(*, "variables")= language list(price, year, mileage, cc, automatic)
 .. ..- attr(*, "factors")= int [1:5, 1:4] 0 1 0 0 0 0 0 1 0 0 ...
 .. .. ..- attr(*, "dimnames")=List of 2
 .. .. .. ..$ : chr [1:5] "price" "year" "mileage" "cc" ...
 .. .. .. ..$ : chr [1:4] "year" "mileage" "cc" "automatic"
 .. ..- attr(*, "term.labels")= chr [1:4] "year" "mileage" "cc" "automatic"
 .. ..- attr(*, "order")= int [1:4] 1 1 1 1
 .. ..- attr(*, "intercept")= int 1
 .. ..- attr(*, "response")= int 1
 .. ..- attr(*, ".Environment")=<environment: R_GlobalEnv>
 .. ..- attr(*, "predvars")= language list(price, year, mileage, cc, automatic)
 .. ..- attr(*, "dataClasses")= Named chr [1:5] "numeric" "numeric" "numeric" "numeric" ...
 .. .. ..- attr(*, "names")= chr [1:5] "price" "year" "mileage" "cc" ...
$ model        :'data.frame':  30 obs. of  5 variables:
 ..$ price    : int [1:30] 790 1380 270 1190 590 1120 815 450 1290 420 ...
 ..$ year     : int [1:30] 78 39 109 20 70 58 53 68 15 96 ...
 ..$ mileage  : int [1:30] 133462 33000 120000 69727 112000 39106 95935 120000 20215 140000 .
 ..$ cc       : int [1:30] 1998 2000 1800 1999 2000 1998 1800 1800 1798 1800 ...
 ..$ automatic: int [1:30] 1 1 0 1 0 1 1 0 1 0 ...
 ..- attr(*, "terms")=Classes 'terms', 'formula'  language price ~ year + mileage + cc + auto
```

```
  .. .. ..- attr(*, "variables")= language list(price, year, mileage, cc, automatic)
  .. .. ..- attr(*, "factors")= int [1:5, 1:4] 0 1 0 0 0 0 0 1 0 0 ...
  .. .. .. ..- attr(*, "dimnames")=List of 2
  .. .. .. .. ..$ : chr [1:5] "price" "year" "mileage" "cc" ...
  .. .. .. .. ..$ : chr [1:4] "year" "mileage" "cc" "automatic"
  .. .. ..- attr(*, "term.labels")= chr [1:4] "year" "mileage" "cc" "automatic"
  .. .. ..- attr(*, "order")= int [1:4] 1 1 1 1
  .. .. ..- attr(*, "intercept")= int 1
  .. .. ..- attr(*, "response")= int 1
  .. .. ..- attr(*, ".Environment")=<environment: R_GlobalEnv>
  .. .. ..- attr(*, "predvars")= language list(price, year, mileage, cc, automatic)
  .. .. ..- attr(*, "dataClasses")= Named chr [1:5] "numeric" "numeric" "numeric" "numeric" ..
  .. .. .. ..- attr(*, "names")= chr [1:5] "price" "year" "mileage" "cc" ...
 - attr(*, "class")= chr "lm"
```

예를 들어 다음과 같이 회귀계수의 값을 구할 수 있다.

```
fit0$coefficients
```

```
  (Intercept)          year       mileage            cc     automatic
525.286960604  -5.799637101  -0.002262844   0.388787346 165.312632517
```

함수 lm 의 결과인 fit0 은 class 함수를 사용하면 lm 객체(object) 라는 것을 알 수 있으며 lm 객체에 적용할 수 있는 함수를 찾는방법은 methods 함수를 사용하면 된다.

```
class(fit0)
```

```
[1] "lm"
```

```
methods(class=class(fit0))
```

```
 [1] add1                addterm             alias
 [4] anova               Anova               avPlot
 [7] avPlot3d            Boot                bootCase
[10] boxcox              boxCox              brief
[13] case.names          ceresPlot           coerce
[16] confidenceEllipse   confint            Confint
[19] cooks.distance      crPlot              crPlot3d
[22] deltaMethod         deviance            dfbeta
[25] dfbetaPlots         dfbetas             dfbetasPlots
[28] drop1               dropterm            dummy.coef
```

```
[31] durbinWatsonTest    effects             emm_basis
[34] extractAIC          family              formula
[37] fortify             hatvalues           hccm
[40] infIndexPlot        influence           influencePlot
[43] initialize          inverseResponsePlot kappa
[46] labels              leveneTest          leveragePlot
[49] linearHypothesis    lmbeta              logLik
[52] logtrans            mcPlot              mmp
[55] model.frame         model.matrix        ncvTest
[58] nextBoot            nobs                ols_regress
[61] ols_test_normality  outlierTest         plot
[64] powerTransform      predict             Predict
[67] press               print               proj
[70] qqnorm              qqPlot              qr
[73] recover_data        residualPlot        residualPlots
[76] residuals           rpredict            rstandard
[79] rstudent            S                   show
[82] sigmaHat            simulate            slotsFromS3
[85] spreadLevelPlot     stdcoef             sumary
[88] summary             symbox              variable.names
[91] vcov                vif
see '?methods' for accessing help and source code
```

lm 객체에 적용할 수 있는 함수는 summary, anova, predict, plot 등이 있다.

J.2.1. 회귀계수의 추정과 결정계수

함수 summary 는 각 회귀계수의 추정값과 가설 $H_0 : \beta_k = 0$에 대한 t-검정 결과를 보여준다. 또한 결정계수 R^2도 구해준다.

```
summary(fit0)
```

```
Call:
lm(formula = price ~ year + mileage + cc + automatic, data = usedcars)

Residuals:
    Min      1Q  Median      3Q     Max
-177.35  -63.91   -0.99   70.34  212.69

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)  5.253e+02  3.998e+02   1.314 0.200823
year        -5.800e+00  9.283e-01  -6.247 1.55e-06 ***
mileage     -2.263e-03  7.211e-04  -3.138 0.004324 **
cc           3.888e-01  2.022e-01   1.923 0.065958 .
automatic    1.653e+02  3.986e+01   4.147 0.000339 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 101.1 on 25 degrees of freedom
Multiple R-squared:  0.9045,    Adjusted R-squared:  0.8892
F-statistic: 59.21 on 4 and 25 DF,  p-value: 2.184e-12
```

위의 결과로 다음과 같은 추정값과 통계량을 알 수 있다.

- `Estimate` : 회귀계수의 추정값
- `Std. Error` : 회귀계수의 표준오차
- `t value` : t-검정 통계량
- `Pr(>|t|)` : p-value
- `Multiple R-squared` : 결정계수 R^2
- `Adjusted R-squared` : 수정된 결정계수 R^2_{adj}
- `F-statistic` : F-검정 통계량
- `p-value` : F-검정의 p-value
- `Residual standard error` : 오차항의 표준편차에 대한 추정값 ($\sqrt{MSE} = \hat{\sigma}$)
- `degrees of freedom` : 잔차 제곱합의 자유도 ($n - p$)

각 회귀 계수에 대한 95% 신뢰구간은 함수 `confint`로 구할 수 있다.

```
confint(fit0)
```

```
                    2.5 %        97.5 %
(Intercept) -2.981256e+02  1.348699e+03
year        -7.711605e+00 -3.887669e+00
mileage     -3.748021e-03 -7.776672e-04
cc          -2.763072e-02  8.052054e-01
automatic    8.322275e+01  2.474025e+02
```

예를 들어 설명변수 `cc` 에 대한 회귀계수의 추정값은 `0.3888` 이며 배기량이 `1cc` 증가할 때 가격은 약 0.4만원, 즉 4000원이 평균적으로 증가한다고 해석할 수 있다.

설명변수 `cc` 에 대한 회귀계수가 0인지 검정하기 위해서는 t-검정 통계량을 구하고 p-value를 구하면 된다. t-검정 통계량은 다음과 같이 주어지며

$$t_0 = \frac{\hat{\beta}_3}{SE(\hat{\beta}_3)} = \frac{0.3888}{0.2022} = 1.923$$

p-값은 0.0659852 이다. 즉 귀무가설 $H_0 : \beta_k = 0$ 을 기각할 수 없다.

참고로 p-value 를 이용하지 않고 t-검정 통계량을 이용하여 유의수준 $\alpha = 0.05$에서 검정할 수 있다. 즉, 자유도가 25인 것을 이용하면 기각역 $t_{0.975,25} = 2.059539$ 이므로 $|t_0| < t_{0.975,25}$ 이므로 귀무가설을 기각할 수 없다.

```
qt(0.975, 25)
```

```
[1] 2.059539
```

회귀계수에 대한 95% 신뢰구간은 (-2.763072e-02, 8.052054e-01) 으로 0을 포함하고 있다.

이제 모든 설명변수에 대한 F-검정을 해보자. 즉, 모든 설명변수의 계수가 0인지 검정해보자. 즉 귀무가설은 다음과 같다.

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$$

함수 summary 의 결과에서 F-statistic의 값을 보면 $F_0 = 59.21$ 이고 p-value는 2.184e-12 이다. 즉 귀무가설을 기각할 수 있다. 즉, 적어도 하나의 설명변수는 반응변수에 영향을 준다고 해석할 수 있다.

참고로 p-value 를 사용하지 않고 F-검정 통계량을 이용하여 유의수준 $\alpha = 0.05$에서 검정할 수 있다. 즉, 자유도가 4, 25인 것을 이용하면 기각역 $F_{0.95,4,25} = 2.75871$ 이므로 $F_0 > F_{0.95,4,25}$ 이므로 귀무가설을 기각할 수 있다.

```
qf(0.95, 4, 25)
```

```
[1] 2.75871
```

J.2.2. 예측값

자료에 주어진 각 반응변수에 대한 예측값 $\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$는 함수 predict를 이용한다.

```
predict(fit0)
```

```
       1        2        3        4        5        6        7        8
713.0214 1167.3146  321.4025 1194.0114  643.4485 1042.5270  865.9501  559.1876
       9       10       11       12       13       14       15       16
1256.9013  351.5409  946.0553  623.6355  677.3900 1236.5788  991.9617 1007.3483
      17       18       19       20       21       22       23       24
 709.6348 1142.6549  890.3836 1029.0340  808.9611  643.6167  611.6964  182.7813
      25       26       27       28       29       30
 960.9247  614.4275  924.2101  872.9584  265.3927  884.0490
```

새로운 자료에 대한 예측값 $\widehat{E(y|x)}$은 다음과 같이 데이터프레임을 만들고 예측한다.

```
nw <- data.frame(year=60, mileage=10000, cc=200, automatic=1)
nw
```

```
  year mileage  cc automatic
1   60   10000 200         1
```

```
predict(fit0, newdata=nw, interval="confidence")
```

```
       fit       lwr      upr
1 397.7504 -342.6272 1138.128
```

새로운 자료에 대한 예측값(**반응변수의평균에 대한 예측값**)은 다음과 같이 신뢰구간을 구할 수 있다. 주어진 새로운 설병변수 벡터를 $\boldsymbol{x}_0$ 라고 하면 예측값은 다음과 같이 구할 수 있다.

$$\widehat{E(y|x_0)} = \boldsymbol{x}_0^T \hat{\boldsymbol{\beta}}$$

새로운 예측값에 대한 분산은 다음과 같이 구할 수 있다.

$$\widehat{Var(E(y|x_0))} = \widehat{Var}(\boldsymbol{x}_0^T \hat{\boldsymbol{\beta}}) = \hat{\sigma}^2 \boldsymbol{x}_0^T (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{x}_0$$

따라서 95% 신뢰구간은 다음과 같이 구할 수 있다.

$$\widehat{E(y|x_0)} \pm t_{0.975,n-p} \sqrt{\widehat{Var}(E(y|x_0))}$$

새로운 관측값에 대한 예측은 다음과 같이 한다.

```
predict(fit0, newdata=nw, interval="prediction")
```

```
       fit       lwr      upr
1 397.7504 -371.3501 1166.851
```

새로운 관측값은 평균이 아닌 특정한 확률변수의 예측값이다. 즉, 예측값에 대한 불확실성에 대하여 오차항의 분산을 더해주어 95% 신뢰구간은 다음과 같이 구할 수 있다.

$$\widehat{E(y|x_0)} \pm t_{0.975,n-p} \sqrt{\widehat{Var}(E(y|x_0)) + \hat{\sigma}^2}$$

J.3. 제곱합의 종류

각 설명변수가 반응변수에 미치는 영향을 알아보기 위해서 가각에 대한 제곱합을 구할 수 있다. 각 설명변수에 대한 제곱합은 순차제곱한(sequential sum of squares)과 편제곱합(partial sum of squares)로 구분할 수 있다.

J.3.1. 순차제곱합

순차제곱합은 모형에 들어가는 변수의 순서에 따라서 제곱합이 틀려진다.

다음의 예를 보면 두 모형이 같은 변수들로 적합되지만 순서가 달라지면 순차제곱합이 다르다.

```
model1 <- price ~ year + mileage + cc + automatic
model2 <- price ~ mileage + automatic + cc + year
fit1 <- lm(model1, usedcars)
fit2 <- lm(model2, usedcars)
anova(fit1)
```

```
Analysis of Variance Table

Response: price
          Df  Sum Sq Mean Sq  F value    Pr(>F)
year       1 2056608 2056608 201.2036 1.841e-13 ***
mileage    1  135864  135864  13.2919 0.0012228 **
cc         1   52409   52409   5.1273 0.0324794 *
automatic  1  175828  175828  17.2018 0.0003389 ***
Residuals 25  255538   10222
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(fit2)
```

```
Analysis of Variance Table

Response: price
          Df  Sum Sq Mean Sq  F value    Pr(>F)
mileage    1 1637355 1637355 160.1870 2.274e-12 ***
automatic  1  341741  341741  33.4335 5.006e-06 ***
cc         1   42683   42683   4.1758   0.05168 .
year       1  398929  398929  39.0283 1.552e-06 ***
Residuals 25  255538   10222
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

하지만 회귀계수의 추정량은 동일하다.

```
summary(fit1)
```

```
Call:
lm(formula = model1, data = usedcars)

Residuals:
    Min      1Q  Median      3Q     Max
-177.35  -63.91   -0.99   70.34  212.69

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.253e+02  3.998e+02   1.314 0.200823
year        -5.800e+00  9.283e-01  -6.247 1.55e-06 ***
mileage     -2.263e-03  7.211e-04  -3.138 0.004324 **
cc           3.888e-01  2.022e-01   1.923 0.065958 .
automatic    1.653e+02  3.986e+01   4.147 0.000339 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 101.1 on 25 degrees of freedom
Multiple R-squared:  0.9045,    Adjusted R-squared:  0.8892
F-statistic: 59.21 on 4 and 25 DF,  p-value: 2.184e-12
```

```
summary(fit2)
```

```
Call:
lm(formula = model2, data = usedcars)

Residuals:
    Min      1Q  Median      3Q     Max
-177.35  -63.91   -0.99   70.34  212.69

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.253e+02  3.998e+02   1.314 0.200823
mileage     -2.263e-03  7.211e-04  -3.138 0.004324 **
automatic    1.653e+02  3.986e+01   4.147 0.000339 ***
cc           3.888e-01  2.022e-01   1.923 0.065958 .
year        -5.800e+00  9.283e-01  -6.247 1.55e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 101.1 on 25 degrees of freedom
Multiple R-squared:  0.9045,    Adjusted R-squared:  0.8892
F-statistic: 59.21 on 4 and 25 DF,  p-value: 2.184e-12
```

J.3.1.1. 편제곱합

편제곱합은 주어진 설명변수가 설명하는 부분을 다른 변수들로 보정된 제곱합으로 순서에 관계없이 일정하다. 패키지 car 에 있는 함수 Anova 를 사용하면 편제곱합을 구할 수 있다.

```
Anova(fit1, type="III")
```

```
Anova Table (Type III tests)

Response: price
             Sum Sq Df F value    Pr(>F)
(Intercept)  17645  1  1.7262 0.2008228
year        398929  1 39.0283 1.552e-06 ***
mileage     100649  1  9.8467 0.0043244 **
cc           37794  1  3.6975 0.0659577 .
automatic   175828  1 17.2018 0.0003389 ***
Residuals   255538 25
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Anova(fit2, type="III")
```

```
Anova Table (Type III tests)

Response: price
             Sum Sq Df F value    Pr(>F)
(Intercept)  17645  1  1.7262 0.2008228
mileage     100649  1  9.8467 0.0043244 **
automatic   175828  1 17.2018 0.0003389 ***
cc           37794  1  3.6975 0.0659577 .
year        398929  1 39.0283 1.552e-06 ***
Residuals   255538 25
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

J.4. 부분 F 검정

먼저, 하나의 독립변수 배기량(cc)에 대한 계수가 0인지 검정해보자.

$$H_0 : \beta_3 = 0$$

배기량이 0일 때의 모형을 reducemodel1로 정의하고, 배기량이 0이 아닐 때의 모형을 fullmodel로 정의한다. 두 모형의 결과를 함수 anova를 이용하여 비교한다.

```
fullmodel <- price ~ year + mileage + cc + automatic
reducemodel1 <- price ~ year + mileage + automatic
fitfull <- lm(fullmodel, data=usedcars)
fitreduce1 <- lm(reducemodel1, data=usedcars)
anova(fitreduce1, fitfull)
```

```
Analysis of Variance Table

Model 1: price ~ year + mileage + automatic
Model 2: price ~ year + mileage + cc + automatic
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1     26 293332
2     25 255538  1     37794 3.6975 0.06596 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

두 모형의 잔차제곱합은 함수 deviance 를 이용할 수 있으며 구하면 다음과 같다.

```
deviance(fitfull)
```

```
[1] 255538.1
```

```
deviance(fitreduce1)
```

```
[1] 293331.9
```

따라서 부분 F-검정 통계량은 다음과 같이 구할 수 있다.

$$F_0 = \frac{(SSE(R) - SSE(F))/(df_R - df_F)}{SSE(F)/(df_F)} = \frac{(293331.9 - 255538.1)/(26 - 25)}{255538.1/(25)} = 3.6975$$

F_0의 p-value는 0.06596 이다. 즉 유의수준 $\alpha = 0.05$에서 귀무가설을 기각할 수 없다. 즉 배기량이 0이라는 가설을 기각할 수 없다. 만약 기각역을 사용한다면 $F_{0.95,1,25} = 4.241699$ 이므로 $F_0 < F_{0.95,1,25}$ 이므로 귀무가설을 기각할 수 없다.

```
qf(0.95, 1, 25)
```

```
[1] 4.241699
```

하나의 계수에 대한 부분 F-검정은 앞에서 설명한 t-검정으로도 가능하며 부분 F-검정의 결과와 동일하다.

> **Note**
>
> 참고로 모형의 잔차제곱합은 `summary` 함수에서 나오는 `Residual standard error`의 제곱에 자유도를 곱한값과 같다.
>
> $$SSE = \hat{\sigma}^2(n-p) = MSE(n-p)$$

이제 두 개 이상 의 변수에 대하여 부분 F-검정을 해보자. 설명변수 `cc` 와 `automatic`에 대한 계수가 0인지 검정해보자.

$$H_0 : \beta_3 = \beta_4 = 0$$

```
reducemodel2 <- price ~ year + mileage
fitreduce2 <- lm(reducemodel2, data=usedcars)
anova(fitreduce2, fitfull)
```

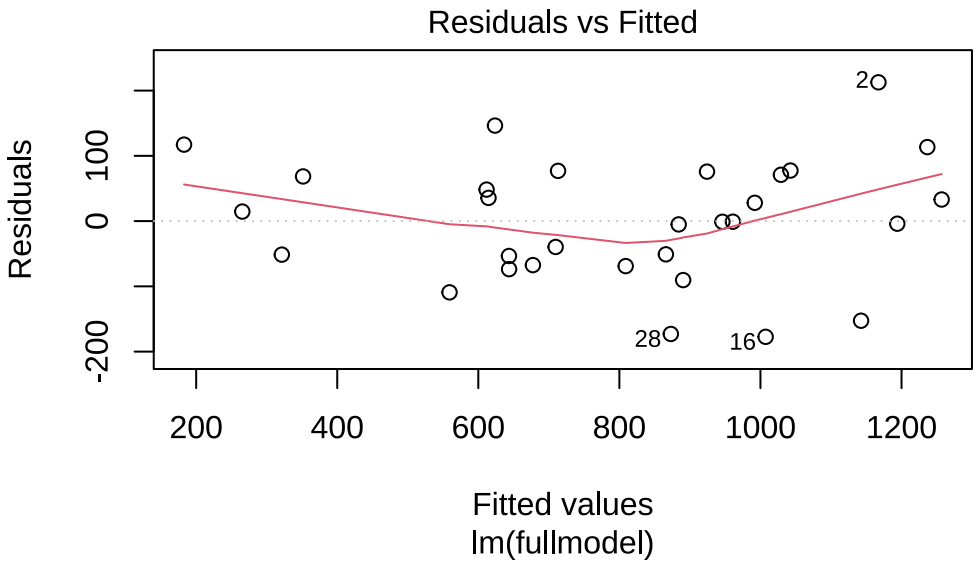
```
Analysis of Variance Table

Model 1: price ~ year + mileage
Model 2: price ~ year + mileage + cc + automatic
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     27 483775
2     25 255538  2    228237 11.165 0.0003429 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

J.5. 모형의 진단

이제 잔차를 이용하여 등분산성에 대한 진단을 실시해 보자. 아래 그림은 x축에 적합값 $\hat{y}$를 두고 y축에 잔차를 두고 그린 그림이다. 즉, 잔차는 적합값에 대한 함수로 그린 그림이다.

잔차가 0을 기준으로 하여 고르게 분포되어 있는지 확인한다. 즉, 잔차가 적합값에 대하여 독립적이고 등분산성을 만족하는지 확인한다. 아래 그림에서는 0을 기준으로 특별한 패턴이 보이지 않으므로 등분산성을 만족한다고 볼 수 있다.

```
plot(fitfull, which=1)
```



K. R-실습: 중회귀 모형 진단

K.1. 변수변환

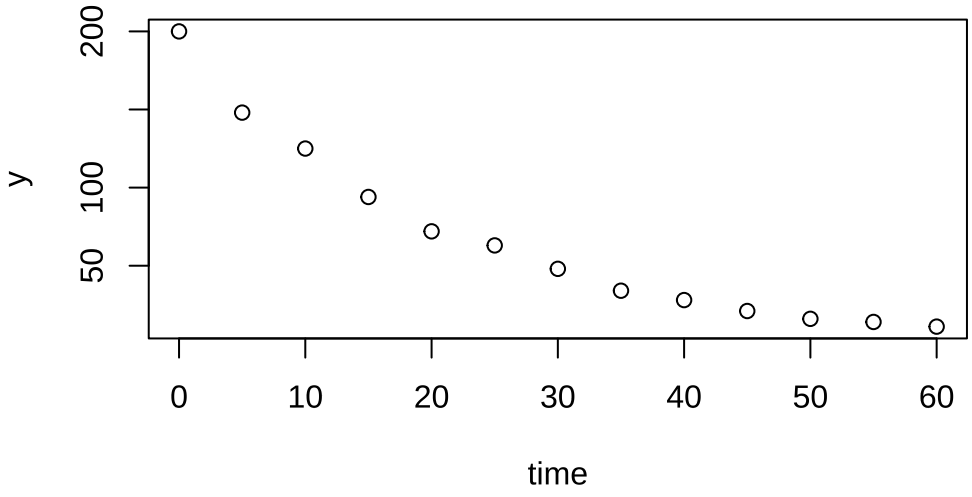
K.1.1. 예제 4.8

여기서 이용한 자료 bug 는 살충제의 독성실험에서 살충제에 노출된 벌레들의 생존개체수를 시간대별로 관측한 것이다.

- y :생존벌레의 수
- time :시간(분)

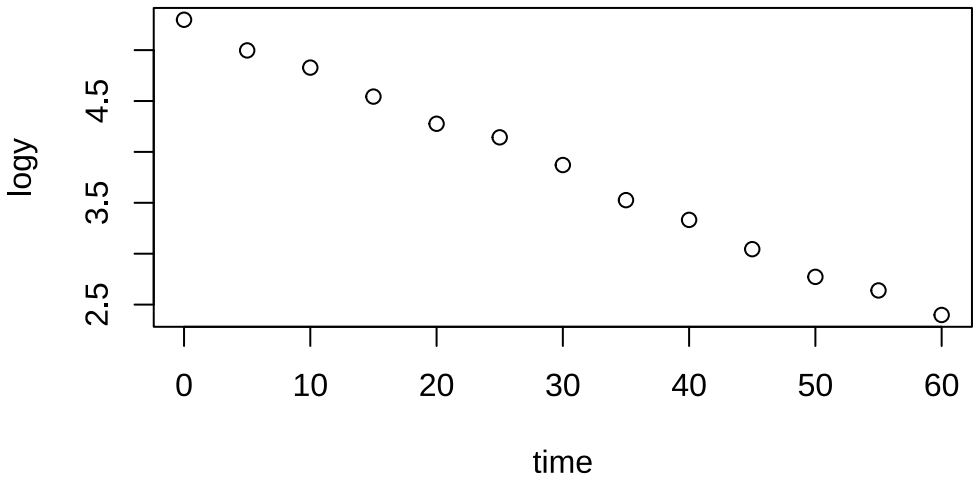
$$y = \beta_0 + \beta_1 time + e$$

```
plot(y~time, regbook::bug)
```



이제 로그변화을 고려해 보자.

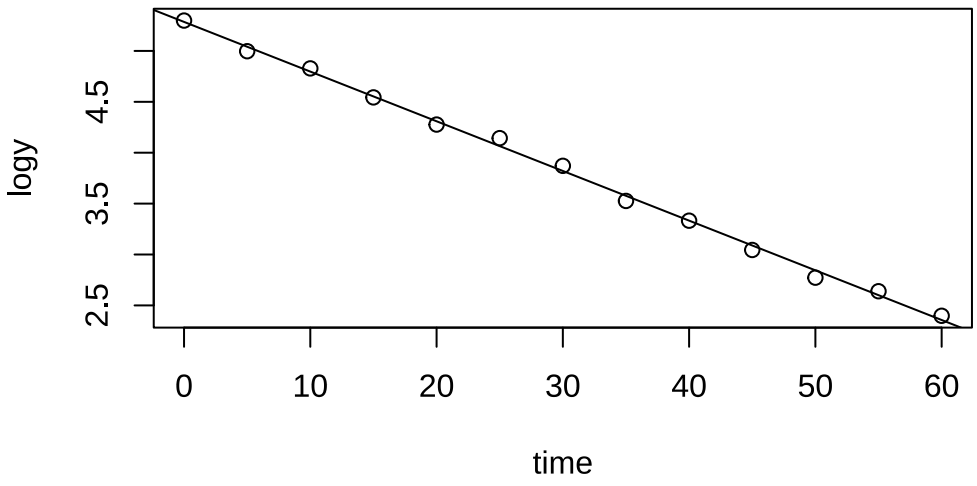
```
bug2 <-regbook::bug
bug2$logy <- log(bug2$y)
plot(logy~time, bug2)
```



변환된 자료에 대한 회귀분석을 수행해 보자.

$$\log(y) = \beta_0 + \beta_1 time + e$$

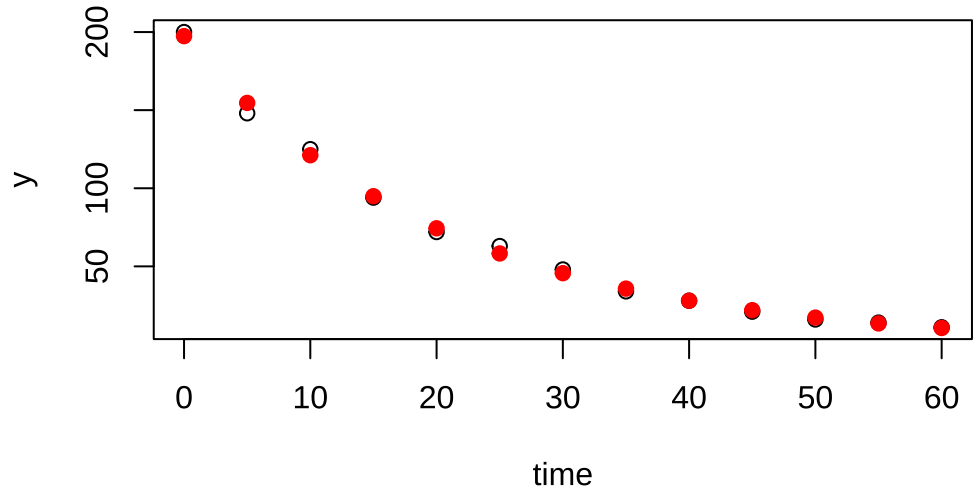
```
fitlog <- lm(logy~time, bug2)
plot(logy~time, bug2)
abline(fitlog)
```



만약 반응변수에 대한 예측값을 구하고 싶다면 다음과 같이 역변환을 이용하면 된다.

$$\hat{E}(y|x) = e^{\hat{\beta}_0 + \hat{\beta}_1 x}$$

```
bug2$pred <- exp(fitlog$fitted.values)
plot(y~time, bug2)
points(bug2$pred~bug2$time, col="red", pch=19)
```



K.2. Box-Cox 변환

Box-Cox 변환은 다음과 같이 수행한다. 패키지 `MASS` 의 함수 `boxcox` 를 이용한다.

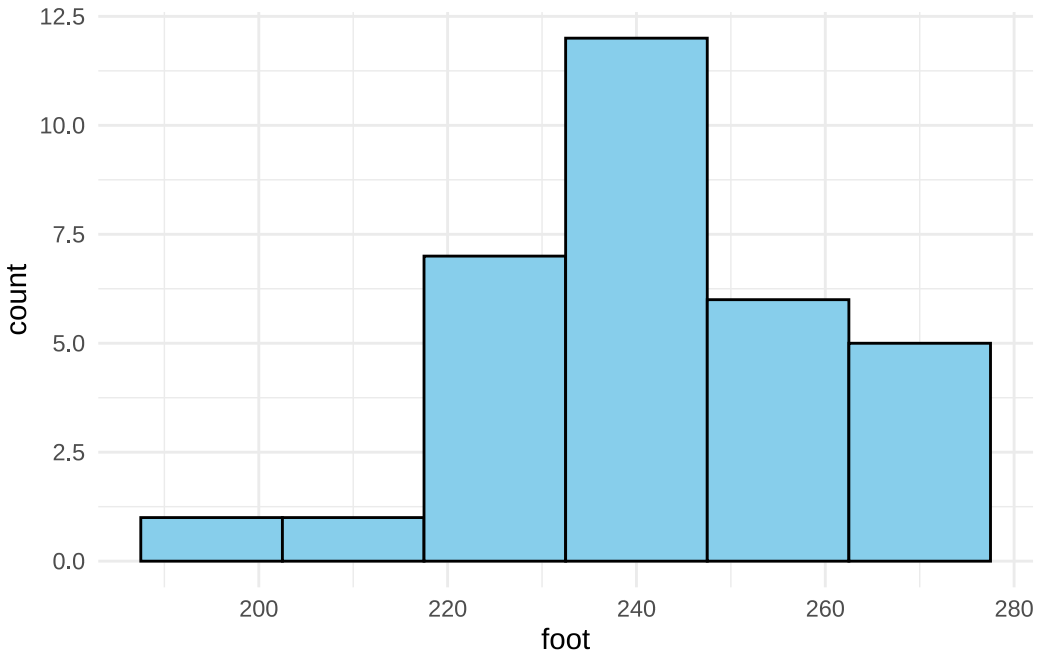
K.2.1. 예제 4.10

의학적으로 정확하지는 않지만 인체의 특성 중에는 양팔을 벌린 길이는 키와 같다든지, 목둘레는 허리둘레 또는 팔목굵기와 관계가 있다든지 하는 흥미로운 사실들이 많이 있다. 자료 `aflength`는 S대학교 학생들 중 무작위로 남녀 각각 16명을 추출하여 조사한 자료로 발길이와 팔안쪽길이이다.

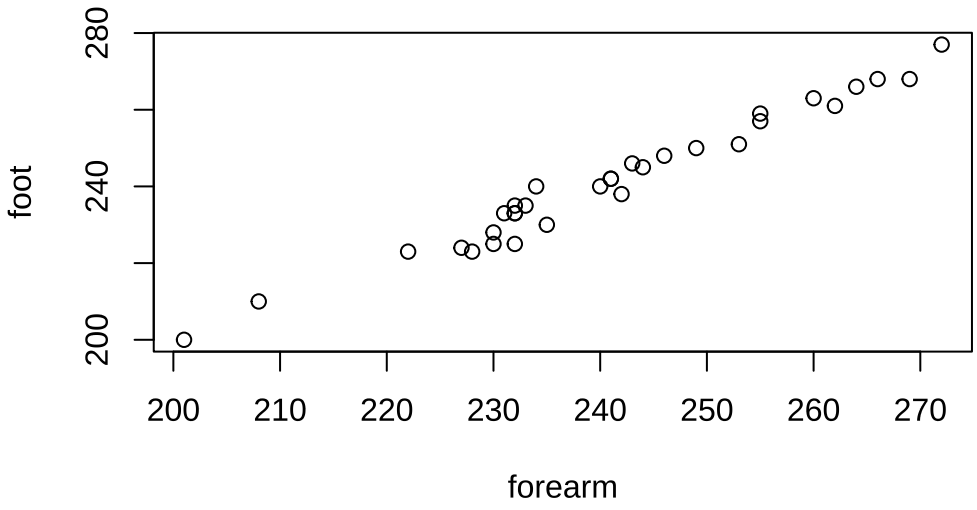
- `foot`:발길이(mm), 양말을 벗은 상태로 측정하였고 오른쪽 발만 측정하였다.
- `forearm`: 팔안쪽길이(mm), 손목부터 팔꿈치가 접히는 부분까지의 길이이다. 오른쪽 팔만 측정하였다.

변환이 필요없는 경우에 대한 예제이다.

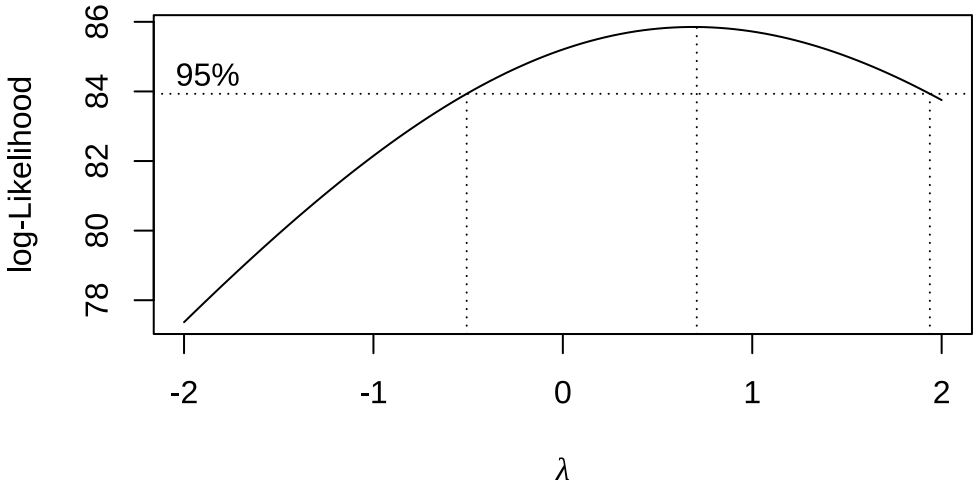
```
# plot histogram of foot by ggplot2
aflength %>% ggplot(aes(x=foot)) + geom_histogram(binwidth=15, fill="skyblue", color="black")
```




```
plot(foot ~ forearm, data=aflength)
```




```
ex411 <- boxcox(lm(foot ~ forearm, data=aflength))
```



K.2.2. 예제 4.11

예제 4.11 자료 `wool` 는 Box & Cox의 1964년 논문에서 사용한 예제로, 양모의 강력을 알아보기 위해 3^3 요인실험을 수행한 결과이다.

- `cycle` :반응변수. 시편이 끊어질 때까지의 측정 횟수.
- `length` :시편의 길이

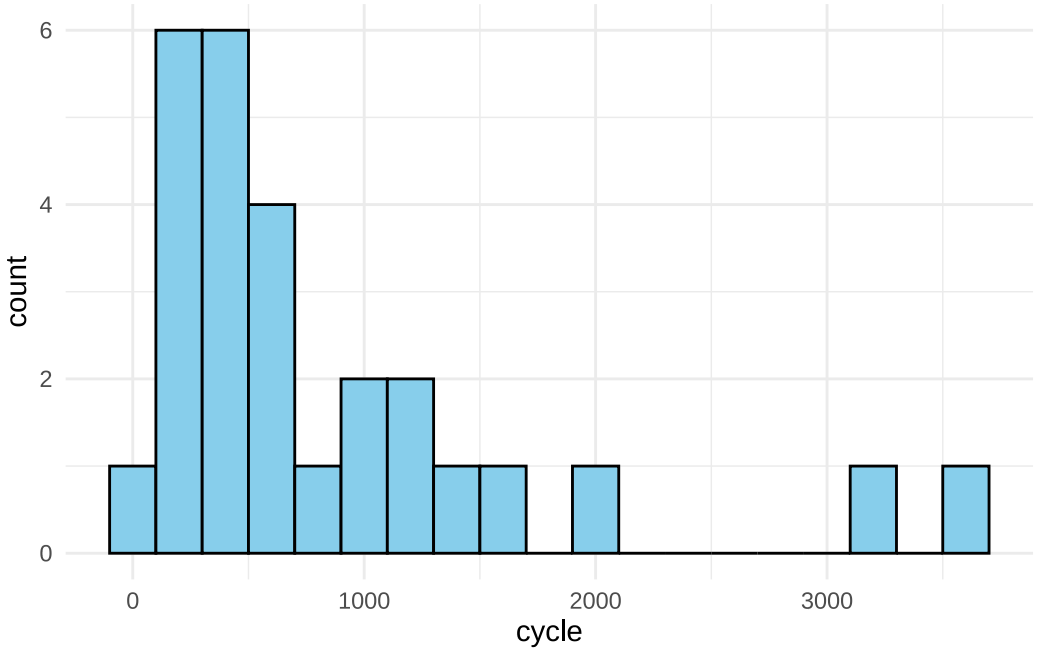
- load : 시편에 가한 하중
- amplitude :하중을 가한 폭

```
head(wool)
```

```
  cycle length amplitude load
1   674    250         8   40
2   370    250         8   45
3   292    250         8   50
4   338    250         9   40
5   266    250         9   45
6   210    250         9   50
```

반응변수 cycle 의 히스토그램을 보면 오른쪽으로 꼬리가 긴 분포로서 정규분포와 매우 다른 모양을 보인다.

```
# plot histogram of foot by ggplot2
wool %>% ggplot(aes(x=cycle)) + geom_histogram(binwidth=200, fill="skyblue", color="black") +
```



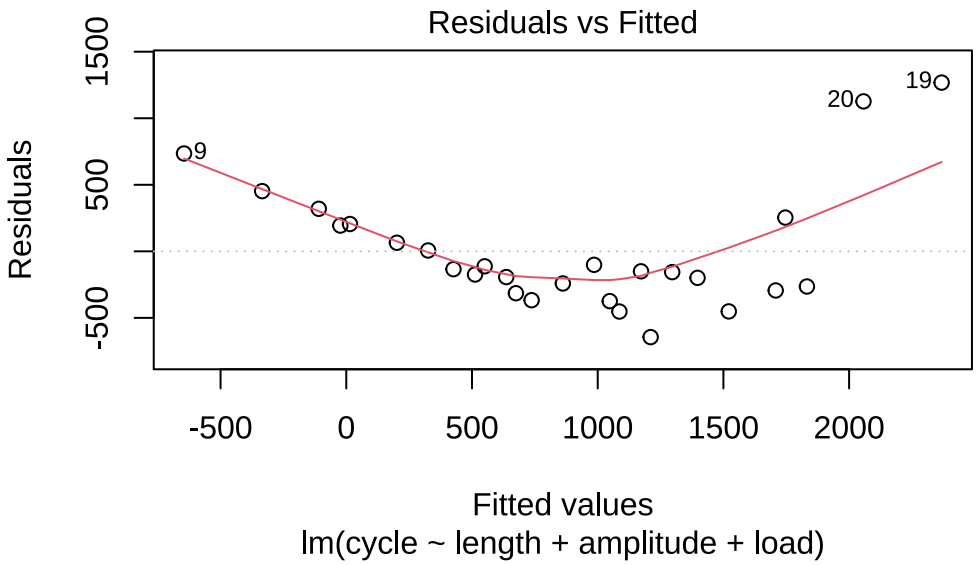
잔차 분석의 결과를 보면 잔차가 0을 중심으로 고르게 분포하지 않고, 정규성 가정이 만족되지 않음을 알 수 있다.

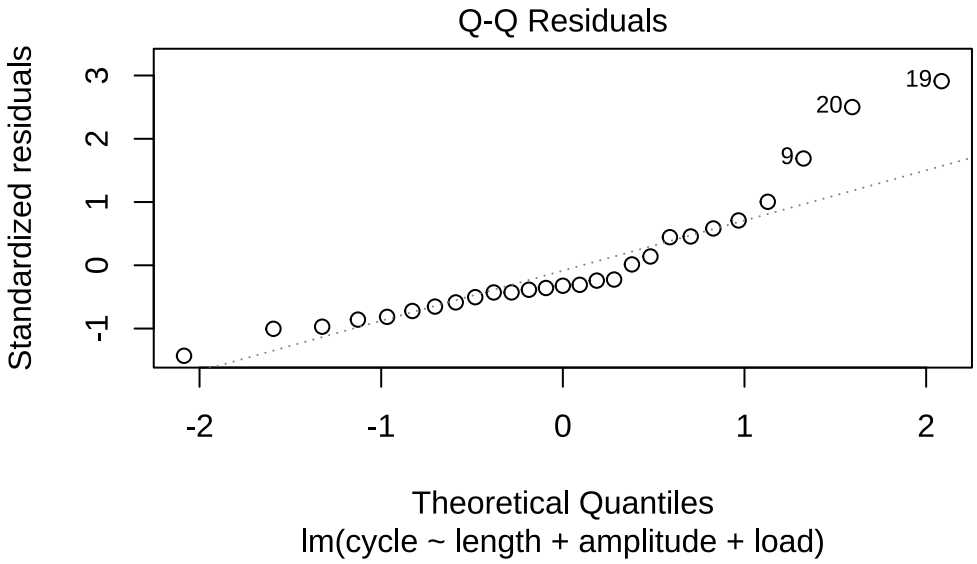
```
woolfm1 <- lm(cycle~length + amplitude + load, data=wool)
summary(woolfm1)
```

```
Call:
lm(formula = cycle ~ length + amplitude + load, data = wool)
```

```
Residuals:
   Min     1Q Median     3Q    Max
-644.5 -279.1 -150.2  199.5 1268.0

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 4521.370   1621.721   2.788 0.010454 *
length        13.200      2.301   5.736 7.66e-06 ***
amplitude   -535.833    115.057  -4.657 0.000109 ***
load         -62.167     23.011  -2.702 0.012734 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 488.1 on 23 degrees of freedom
Multiple R-squared:  0.7291,    Adjusted R-squared:  0.6937
F-statistic: 20.63 on 3 and 23 DF,  p-value: 1.028e-06
```

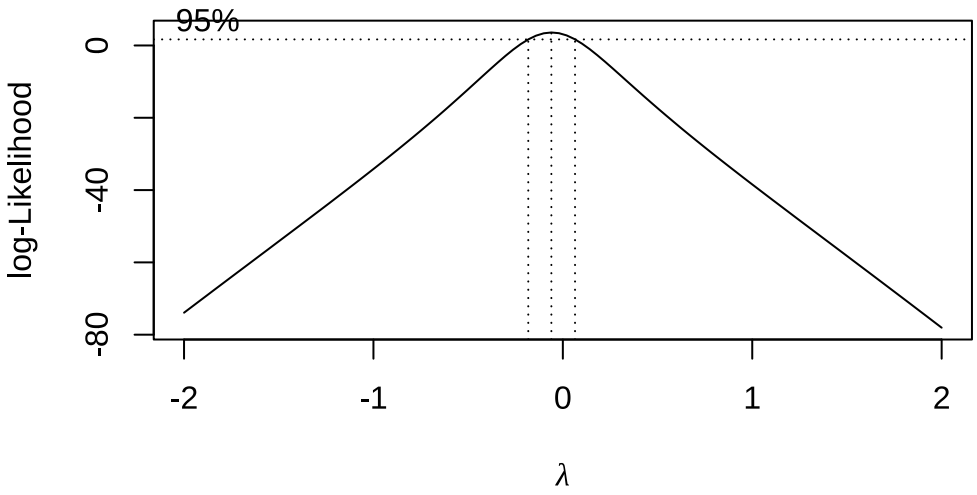
```
plot(woolfm1, which=c(1,2))
```





이제 Box-Cox 변환을 적용해 보자.

```
boxcox(woolfm1)
```

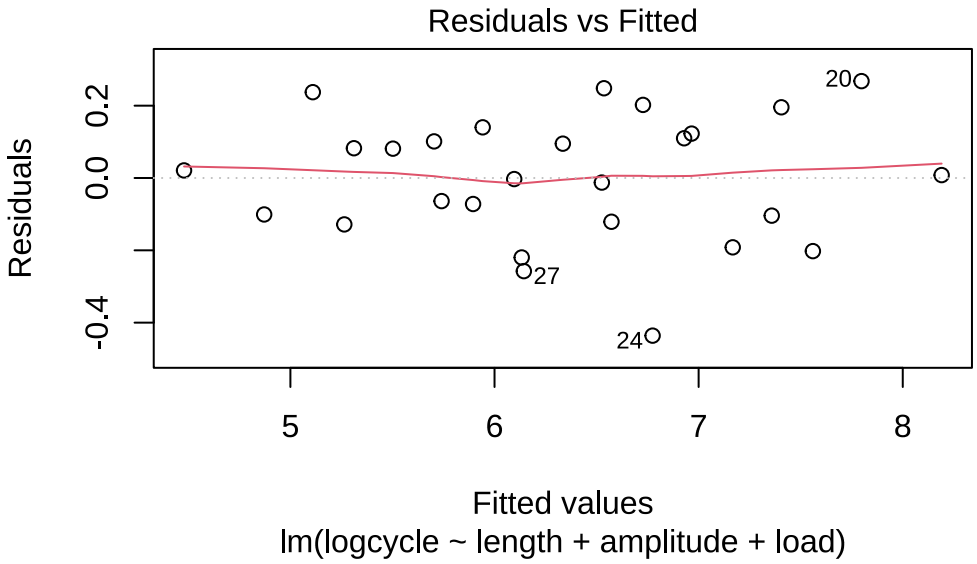


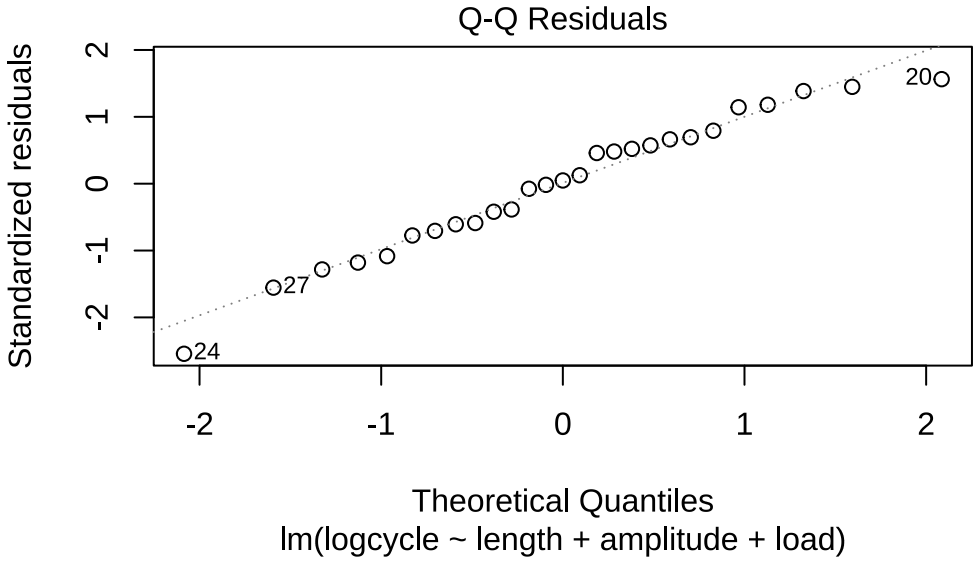
위의 결과에서 $\lambda = 0$이 가장 좋은 변환으로 나타났다. 이는 로그 변환이 가장 적절하다는 의미이다. 이제 이 변환을 적용해 보자.

```
wool$logcycle <- log(wool$cycle)
woolfm2 <- lm(logcycle~length + amplitude + load, data=wool)
summary(woolfm2)
```

```
Call:
lm(formula = logcycle ~ length + amplitude + load, data = wool)

Residuals:
     Min       1Q   Median       3Q      Max
-0.43592 -0.11250  0.00802  0.11635  0.26790

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 10.551813   0.616683  17.111 1.41e-14 ***
length       0.016648   0.000875  19.025 1.43e-15 ***
amplitude   -0.630866   0.043752 -14.419 5.22e-13 ***
load        -0.078524   0.008750  -8.974 5.66e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1856 on 23 degrees of freedom
Multiple R-squared:  0.9658,    Adjusted R-squared:  0.9614
F-statistic: 216.8 on 3 and 23 DF,  p-value: < 2.2e-16
```

```
plot(woolfm2, which=c(1,2))
```





K.3. 다중공선성

K.3.1. 고유값과 고유벡터에 대한 이론

선형모형 $E(\boldsymbol{y}|\boldsymbol{X}) = \boldsymbol{X}\boldsymbol{\beta}$ 에서 계획행렬 $\boldsymbol{X}$의 열들이 선형독립이 아닌 경우 다중공선성이 발생한다. 다중공선성은 계획행렬 $\boldsymbol{X}$의 열들이 선형종속인 경우에 발생한다.

대칭행렬 $\boldsymbol{X}^t\boldsymbol{X}$의 고유값 λ_i와 그에 대응하는 고유벡터 $\boldsymbol{p}_i$는 다음을 만족하는 실수와 벡터이다.

$$(\boldsymbol{X}^t\boldsymbol{X})\boldsymbol{p}_i = \lambda_i\boldsymbol{p}_i$$

고유값 λ_i을 구하는 방법은 다음의 방정식을 만족하는 해를 구하는 것이다.

$$det\left(\boldsymbol{X}^t\boldsymbol{X} - \lambda_i\boldsymbol{I}\right) = 0$$

여기서 $det(\boldsymbol{A})$는 행렬 $\boldsymbol{A}$의 행렬식을 의미한다.

$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$를 $\boldsymbol{X}^t\boldsymbol{X}$의 고유값이라고 하자. $\boldsymbol{X}^t\boldsymbol{X}$의 각 고유값에 대한 정규직교 고유벡터(orthonormal eigenvector)를 $\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_p$라고 하자, 즉

$$\boldsymbol{p}_i^t\boldsymbol{p}_i = 1, \quad \boldsymbol{p}_i^t\boldsymbol{p}_j = 0 \quad (i \neq j)$$

더 나아가 행렬 $\boldsymbol{P}$를 고유벡터를 모아놓은 행렬로 정의하자.

$$\boldsymbol{P} = [\boldsymbol{p}_1\ \boldsymbol{p}_2\ \cdots\ \boldsymbol{p}_p]$$

이때 $p \times p$ - 차원의 행렬 $\boldsymbol{P}$는 직교행렬이다.

$$\boldsymbol{P}^t\boldsymbol{P} = \boldsymbol{P}\boldsymbol{P}^t = \boldsymbol{I}$$

이제 다음과 같이 $\boldsymbol{X}^t\boldsymbol{X}$를 나타낼 수 있다.

$$\boldsymbol{P}^t(\boldsymbol{X}^t\boldsymbol{X})\boldsymbol{P} = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_p) = \boldsymbol{\Lambda}$$

또한

$$\boldsymbol{P}^t(\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{P} = \text{diag}\left(\frac{1}{\lambda_1}, \frac{1}{\lambda_2}, \dots, \frac{1}{\lambda_p}\right) = \boldsymbol{\Lambda}^{-1}$$

위의 식에서 알 수 있듯이 $1/\lambda_i$는 $(\boldsymbol{X}^t\boldsymbol{X})^{-1}$의 고유값이다.

행렬 $\boldsymbol{P}$가 직교행렬이기 때문에 다음과 같은 표현도 가능하다.

$$(\boldsymbol{X}^t\boldsymbol{X}) = \boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}^t, \quad (\boldsymbol{X}^t\boldsymbol{X})^{-1} = \boldsymbol{P}\boldsymbol{\Lambda}^{-1}\boldsymbol{P}^t$$

고유벡터와 고유값의 정의에 의하여 고유값 λ_k이 매우 0에 가까우면 다음이 성립하고

$$\lambda_i = \boldsymbol{p}_k^t(\boldsymbol{X}^t\boldsymbol{X})\boldsymbol{p}_k = (\boldsymbol{X}\boldsymbol{p}_k)^t(\boldsymbol{X}\boldsymbol{p}_k) \approx 0$$

위의 식은 다음과 같이 행렬 $\boldsymbol{X}$의 열들간에 선형관계 $\boldsymbol{X}\boldsymbol{p}_k = \boldsymbol{0}$ 이 있다는 것을 의미한다.

$$p_{1k}\boldsymbol{x}_1 + p_{2k}\boldsymbol{x}_2 + \dots p_{p,k}\boldsymbol{x}_p \approx 0$$

위에서 $\boldsymbol{p}_k$와 $\boldsymbol{X}$는 다음과 같이 표시한다.

$$\boldsymbol{X} = [\boldsymbol{x}_1\ \boldsymbol{x}_2\ \dots\ \boldsymbol{x}_p], \quad \boldsymbol{p}_k = \begin{bmatrix} p_{1k} \\ p_{2k} \\ \vdots \\ p_{p,k} \end{bmatrix}$$

또한 회귀계수 벡터 $\hat{\beta}$의 공분산 행렬이 다음과 같이 주어지므로

$$Cov(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}^t\boldsymbol{X})^{-1} = \sigma^2\boldsymbol{P}\boldsymbol{\Lambda}^{-1}\boldsymbol{P}^t \tag{K.1}$$

다음과 같은 식이 성립한다.

$$var(\hat{\beta}_k)/\sigma^2 = \frac{p_{k1}^2}{\lambda_1} + \frac{p_{k2}^2}{\lambda_2} + ... \frac{p_{k,p}^2}{\lambda_p} \tag{K.2}$$

K.3.2. 고유값과 고유벡터에 대한 예제: 두 개의 독립변수

이제 다음과 두 개의 독립변수가 있는 회귀 모형을 고려해 보자.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + e_i, i = 1, 2, \cdots, n$$

절편을 제외한 두 개의 표준화된 독립변수들로 이루어진 행렬을 $\boldsymbol{X}$로 표시하자.

$$\boldsymbol{X} = [\boldsymbol{x}_1 \, \boldsymbol{x}_2]$$

위에서 계획 행렬 $\boldsymbol{X}$는 원래 독립 변수의 계획 행렬 X의 열들을 표준화한 변수로 구성된 것이다..

$$\sum_{i=1}^{n} x_{i1} = 0, \quad \sum_{i=1}^{n} x_{i2} = 0, \quad \sum_{i=1}^{n} x_{i1}^2 = 1, \quad \sum_{i=1}^{n} x_{i2}^2 = 1, \quad \sum_{i=1}^{n} x_{i1} x_{i2} = \rho$$

이제 $\boldsymbol{X}^t\boldsymbol{X}$는 두 독립변수의 상관계수 행렬임을 알 수 있다.

$$\boldsymbol{X}^t\boldsymbol{X} = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} = \boldsymbol{R}, \quad 0 < \rho < 1$$

여기서 두 독립변수 x_1과 x_2의 상관계수 ρ는 0보다 크다고 가정하자.

이제 $\boldsymbol{X}^t\boldsymbol{X}$의 고유값($\lambda_i$)과 고유벡터($\boldsymbol{p}_i$)는 다음과 같은 방정식을 만족하는 수 λ_i와 벡터 $\boldsymbol{p}_i$ 이다.

$$(\boldsymbol{X}^t\boldsymbol{X})\boldsymbol{p}_i = \lambda_i \boldsymbol{p}_i, \quad \boldsymbol{p}_i^t\boldsymbol{p}_i = 1$$

일단 먼저 고유값을 구하는 방법은 $det(\boldsymbol{X}^t\boldsymbol{X} - \lambda_i\boldsymbol{I}) = 0$을 만족하는 값을 찾는 것이다. 여기서 $det(\boldsymbol{A})$는 $\boldsymbol{A}$의 행렬식을 의미한다.

$$det(\boldsymbol{X}^t\boldsymbol{X} - \lambda_i\boldsymbol{I}) = det\left(\begin{bmatrix} 1-\lambda_i & \rho \\ \rho & 1-\lambda_i \end{bmatrix}\right) = 0$$

위의 방정식은 다음과 같이 요약할 수 있고

$$\lambda_i^2 - 2\lambda_i + (1 - \rho^2) = 0$$

해는 다음과 같이 주어진다.

$$\lambda_1 = 1+\rho, \quad \lambda_2 = 1-\rho \quad (\lambda_1 \geq \lambda_2)$$

이제 각 고유값에 대한 고유벡터를 구해보자. 각 고유값 λ_i에 대한 고유벡터를 $\boldsymbol{p}_i$ 라고 하면

$$\boldsymbol{p}_1 = \begin{bmatrix} p_{11} \\ p_{21} \end{bmatrix}, \; p_{11}^2 + p_{21}^2 = 1 \qquad \boldsymbol{p}_2 = \begin{bmatrix} p_{12} \\ p_{22} \end{bmatrix}, \; p_{12}^2 + p_{11}^2 = 1$$

다음과 같은 방정식을 만족해야 한다.

$$(\boldsymbol{X}^t\boldsymbol{X})\boldsymbol{p}_1 = \lambda_1\boldsymbol{p}_1, \quad (\boldsymbol{X}^t\boldsymbol{X})\boldsymbol{p}_2 = \lambda_2\boldsymbol{p}_2$$

즉,

$$\begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \begin{bmatrix} p_{11} \\ p_{21} \end{bmatrix} = (1+\rho) \begin{bmatrix} p_{11} \\ p_{21} \end{bmatrix}, \quad \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \begin{bmatrix} p_{12} \\ p_{22} \end{bmatrix} = (1-\rho) \begin{bmatrix} p_{12} \\ p_{22} \end{bmatrix}$$

위의 두 방정식은 정리하면 다음과 더 단순한 방정식을 얻는다.

$$p_{11} - p_{21} = 0, \quad p_{12} + p_{22} = 0$$

이제 위의 식을 만족하고 길이가 1인 두 벡터를 찾으면 다음과 같은 두 개의 직교하고 길이가 1인 고유벡터 $\boldsymbol{p}_1$과 $\boldsymbol{p}_2$를 찾을 수 있다.

$$\boldsymbol{p}_1 = \begin{bmatrix} p_{11} \\ p_{21} \end{bmatrix} = \begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix}, \qquad \boldsymbol{p}_2 = \begin{bmatrix} p_{12} \\ p_{22} \end{bmatrix} = \begin{bmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{bmatrix}$$

따라서 앞 절의 이론에서 나온 고유벡터로 구성된 행렬 $\boldsymbol{P}$와 고유값을 대각원소로 하는 행렬 $\boldsymbol{\Lambda}$는 다음과 같다.

$$\boldsymbol{P} = [\boldsymbol{p}_1 \, \boldsymbol{p}_2] = \begin{bmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{bmatrix} = \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{bmatrix}, \qquad \boldsymbol{\Lambda} = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} = \begin{bmatrix} 1+\rho & 0 \\ 0 & 1-\rho \end{bmatrix}$$

이제 다음이 성립함을 확인할 수 있다.

$$\boldsymbol{P}^t(\boldsymbol{X}^t\boldsymbol{X})\boldsymbol{P} = \boldsymbol{\Lambda}, \quad (\boldsymbol{X}^t\boldsymbol{X})^{-1} = \boldsymbol{P}\boldsymbol{\Lambda}^{-1}\boldsymbol{P}^t$$

즉,

$$\begin{aligned}\boldsymbol{P}^t(\boldsymbol{X}^t\boldsymbol{X})\boldsymbol{P} &= \begin{bmatrix}1/\sqrt{2} & 1/\sqrt{2}\\ 1/\sqrt{2} & -1/\sqrt{2}\end{bmatrix}\begin{bmatrix}1 & \rho\\ \rho & 1\end{bmatrix}\begin{bmatrix}1/\sqrt{2} & 1/\sqrt{2}\\ 1/\sqrt{2} & -1/\sqrt{2}\end{bmatrix}\\ &= \begin{bmatrix}1+\rho & 0\\ 0 & 1-\rho\end{bmatrix}\\ &= \boldsymbol{\Lambda}\end{aligned}$$

또한 다음도 성립함을 확인할 수 있다.

$$(\boldsymbol{X}^t\boldsymbol{X})^{-1} = \boldsymbol{P}\boldsymbol{\Lambda}^{-1}\boldsymbol{P}^t$$

즉,

$$\begin{aligned}(\boldsymbol{X}^t\boldsymbol{X})^{-1} &= \boldsymbol{P}\boldsymbol{\Lambda}^{-1}\boldsymbol{P}^t\\ &= \begin{bmatrix}p_{11} & p_{12}\\ p_{21} & p_{22}\end{bmatrix}\begin{bmatrix}\frac{1}{\lambda_1} & 0\\ 0 & \frac{1}{\lambda_2}\end{bmatrix}\begin{bmatrix}p_{11} & p_{21}\\ p_{12} & p_{22}\end{bmatrix}\\ &= \begin{bmatrix}1/\sqrt{2} & 1/\sqrt{2}\\ 1/\sqrt{2} & -1/\sqrt{2}\end{bmatrix}\begin{bmatrix}\frac{1}{1+\rho} & 0\\ 0 & \frac{1}{1-\rho}\end{bmatrix}\begin{bmatrix}1/\sqrt{2} & 1/\sqrt{2}\\ 1/\sqrt{2} & -1/\sqrt{2}\end{bmatrix}\\ &= \begin{bmatrix}p_{11}^2\frac{1}{\lambda_1}+p_{12}^2\frac{1}{\lambda_2} & p_{11}p_{21}\frac{1}{\lambda_1}+p_{12}p_{22}\frac{1}{\lambda_2}\\ p_{11}p_{21}\frac{1}{\lambda_1}+p_{12}p_{22}\frac{1}{\lambda_2} & p_{21}^2\frac{1}{\lambda_1}+p_{22}^2\frac{1}{\lambda_2}\end{bmatrix}\\ &= \begin{bmatrix}(\frac{1}{\sqrt{2}})^2\frac{1}{1+\rho}+(\frac{1}{\sqrt{2}})^2\frac{1}{1-\rho} & (\frac{1}{\sqrt{2}})^2\frac{1}{1+\rho}+(\frac{1}{\sqrt{2}})(-\frac{1}{\sqrt{2}})\frac{1}{1-\rho}\\ (\frac{1}{\sqrt{2}})^2\frac{1}{1+\rho}+(\frac{1}{\sqrt{2}})(-\frac{1}{\sqrt{2}})\frac{1}{1-\rho} & (\frac{1}{\sqrt{2}})^2\frac{1}{1+\rho}+(-\frac{1}{\sqrt{2}})^2\frac{1}{1-\rho}\end{bmatrix}\\ &= \frac{1}{1-\rho^2}\begin{bmatrix}1 & -\rho\\ -\rho & 1\end{bmatrix}\end{aligned}$$

앞 절에서 나온 회귀계수 추정량의 분산 공식 K.1 과 K.2 를 적용하면 다음과 같은 식을 얻을 수 있다.

$$\begin{aligned}Var(\hat{\beta}_k)/\sigma^2 &= \frac{p_{k1}^2}{\lambda_1}+\frac{p_{k2}^2}{\lambda_2}\\ &= \frac{1}{2}\left(\frac{1}{1+\rho}+\frac{1}{1-\rho}\right)\\ &= \frac{1}{1-\rho^2}\end{aligned}$$

위의 분산 공식에서 제일 작은 두 번째 고유값 $\lambda_2 = 1-\rho$가 0에 가까우면 분산이 매우 커지는 것을 알 수 있다. 이 고유값은 상관계수 ρ가 1에 가까울 수록 0에 가까워 진다.

K.3.3. 예제 4.13

중고차 예제에서 가상의 변수를 만들어 적합할 때 완벽한 선형관계가 존재하면 적합 시 변수를 제거하는 것을 알 수 있다.

```
usedcars2 <- usedcars %>%  mutate(ccmile = cc + mileage)
fitcoll1 <- lm(price ~ year + mileage + cc + automatic + ccmile, usedcars2)
summary(fitcoll1)
```

```
Call:
lm(formula = price ~ year + mileage + cc + automatic + ccmile,
    data = usedcars2)

Residuals:
    Min      1Q  Median      3Q     Max
-177.35  -63.91   -0.99   70.34  212.69

Coefficients: (1 not defined because of singularities)
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.253e+02  3.998e+02   1.314 0.200823
year        -5.800e+00  9.283e-01  -6.247 1.55e-06 ***
mileage     -2.263e-03  7.211e-04  -3.138 0.004324 **
cc           3.888e-01  2.022e-01   1.923 0.065958 .
automatic    1.653e+02  3.986e+01   4.147 0.000339 ***
ccmile              NA         NA      NA       NA
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 101.1 on 25 degrees of freedom
Multiple R-squared:  0.9045,    Adjusted R-squared:  0.8892
F-statistic: 59.21 on 4 and 25 DF,  p-value: 2.184e-12
```

K.3.4. 예제 4.14

자료 hald는 1932년 미국 화학공학 학회지에 발표된 것으로 시멘트가 굳을 때 발생하는 열량에 시멘트의 네 가지 성분들이 미치는 영향을 분석한 것이다.

이 데이터프레임은 다음과 같은 열로 구성되어 있다.

- y : 시멘트 1g당 발생하는 열량(cal)
- x1 : 성분1의 무게의 구성비(%)
- x2 : 성분2의 무게의 구성비(%)

- x3 : 성분3의 무게의 구성비(%)
- x4 : 성분4의 무게의 구성비(%)

```
head(hald)
```

```
      y x1 x2 x3 x4
1  78.5  7 26  6 60
2  74.3  1 29 15 52
3 104.3 11 56  8 20
4  87.6 11 31  8 47
5  95.9  7 52  6 33
6 109.2 11 55  9 22
```

모든 변수가 포함된 모형을 적합해 보자.

```
hald.lm <- lm(y~ ., data=hald)
summary(hald.lm)
```

```

Call:
lm(formula = y ~ ., data = hald)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.1750 -1.6709  0.2508  1.3783  3.9254 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  62.4054    70.0710   0.891   0.3991
x1            1.5511     0.7448   2.083   0.0708 .
x2            0.5102     0.7238   0.705   0.5009
x3            0.1019     0.7547   0.135   0.8959
x4           -0.1441     0.7091  -0.203   0.8441
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.446 on 8 degrees of freedom
Multiple R-squared:  0.9824,    Adjusted R-squared:  0.9736 
F-statistic: 111.5 on 4 and 8 DF,  p-value: 4.756e-07
```

상관계수 행렬의 고유값을 계산해 보자.

```
R <- cor(hald[2:5])
R
```

```
           x1         x2         x3         x4
x1  1.0000000  0.2285795 -0.8241338 -0.2454451
x2  0.2285795  1.0000000 -0.1392424 -0.9729550
x3 -0.8241338 -0.1392424  1.0000000  0.0295370
x4 -0.2454451 -0.9729550  0.0295370  1.0000000
```

```
solve(R)
```

```
         x1        x2        x3       x4
x1 38.49621  94.11969  41.88410  99.7858
x2 94.11969 254.42317 105.09139 267.5394
x3 41.88410 105.09139  46.86839 111.1451
x4 99.78580 267.53942 111.14509 282.5129
```

```
diag(solve(R))
```

```
       x1        x2        x3        x4
 38.49621 254.42317  46.86839 282.51286
```

```
eigenval <- eigen(R)$values
eigenval
```

```
[1] 2.235704035 1.576066070 0.186606149 0.001623746
```

```
sqrt(max(eigenval)/eigenval)
```

```
[1]  1.000000  1.191022  3.461339 37.106342
```

VIF를 구해보자.

```
car::vif(hald.lm)
```

```
       x1        x2        x3        x4
 38.49621 254.42317  46.86839 282.51286
```

```
summary(regbook::vif(hald.lm))
```

```
VIF:
    x1     x2     x3     x4
 38.50 254.42  46.87 282.51

Variance Proportion:
  Eigenvalues Cond.Index          x1           x2          x3           x4
1 2.235704035   1.000000 0.002632084 0.0005589686 0.001481988 0.0004753347
2 1.576066070   1.191022 0.004269804 0.0004272931 0.004954638 0.0004572915
3 0.186606149   3.461339 0.063519491 0.0020822791 0.046495910 0.0007243995
4 0.001623746  37.106342 0.929578621 0.9969314592 0.947067464 0.9983429744
```

x_2를 제외하고 분석해 보자.

```
hald.lm2 <- lm(y~ x1 + x3 + x4, data=hald)
summary(hald.lm2)
```

```
Call:
lm(formula = y ~ x1 + x3 + x4, data = hald)

Residuals:
    Min      1Q  Median      3Q     Max
-2.9323 -1.8090  0.4806  1.1398  3.7771

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 111.68441    4.56248  24.479 1.52e-09 ***
x1            1.05185    0.22368   4.702  0.00112 **
x3           -0.41004    0.19923  -2.058  0.06969 .
x4           -0.64280    0.04454 -14.431 1.58e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.377 on 9 degrees of freedom
Multiple R-squared:  0.9813,    Adjusted R-squared:  0.975
F-statistic: 157.3 on 3 and 9 DF,  p-value: 4.312e-08
```

```
summary(regbook::vif(hald.lm2))
```

```
VIF:
   x1    x3    x4
3.678 3.460 1.181

Variance Proportion:
  Eigenvalues Cond.Index           x1         x3         x4
1   1.8683737   1.000000 0.0720157120 0.07053018 0.02229687
2   0.9838532   1.378056 0.0002285765 0.02382939 0.79011946
3   0.1477731   3.555775 0.9277557115 0.90564042 0.18758367
```

L. R-실습: 관측값에 대한 진단

L.1. 교과서 예제 3.8

자료 usedcars 에 대한 잔차 분석 (예제 3.8, 예제 5.3) 입니다.

먼저 자료 usedcars 에서 주어진 모든 설명변수를 사용하여 중회귀모형을 적합해 봅니다.

```
usedcars.lm <- lm(price ~ year + mileage + cc + automatic, usedcars)
summary(usedcars.lm)
```

```
Call:
lm(formula = price ~ year + mileage + cc + automatic, data = usedcars)

Residuals:
    Min      1Q  Median      3Q     Max
-177.35  -63.91   -0.99   70.34  212.69

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.253e+02  3.998e+02   1.314 0.200823
year        -5.800e+00  9.283e-01  -6.247 1.55e-06 ***
mileage     -2.263e-03  7.211e-04  -3.138 0.004324 **
cc           3.888e-01  2.022e-01   1.923 0.065958 .
automatic    1.653e+02  3.986e+01   4.147 0.000339 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 101.1 on 25 degrees of freedom
Multiple R-squared:  0.9045,    Adjusted R-squared:  0.8892
F-statistic: 59.21 on 4 and 25 DF,  p-value: 2.184e-12
```

L.1.1. 잔차그림

R 에서는 plot 함수를 이용하여 기본적인 잔차그림을 그릴 수 있습니다.

- 그림 `Reidual vs Fitted` 는
- 그림 `Normal Q-Q` 는 잔차가 정규분포를 따르는지를 확인하는 그림이다.
- 그림 `Scale-Location` 은 예측값 $\hat{y}_i$ 에 대한 표준화된 잔차 $r_i/\sqrt{1-h_{ii}}$ 를 그린 것이다.

1. 잔차 대 예측값(Residuals vs Fitted)

- 이 그래프는 등분산성(homoscedasticity)과 선형성 가정을 확인하는 데 도움이 된다.
- 각 예측값 $\hat{y}_i$ 에 대한 잔차 r_i 를 그린 것이다.
- 이상적으로는, 잔차들이 수평축(0-라인) 주변에 무작위로 흩어져 있어야 하며, 이는 관계가 선형이고 오류 항의 분산이 일정함을 나타낸다.

2. 정규 Q-Q(Quantile-Quantile)

- 정규 Q-Q 그래프는 잔차의 정규성을 확인하는 데 사용된다.
- 점들이 제공된 직선을 따라 배치되어야 이상적이며, 이는 잔차가 정규 분포에 가깝다는 것을 나타낸다.
- 주어진 선으로부터의 이탈은 정규성으로부터의 벗어남을 나타낸다.

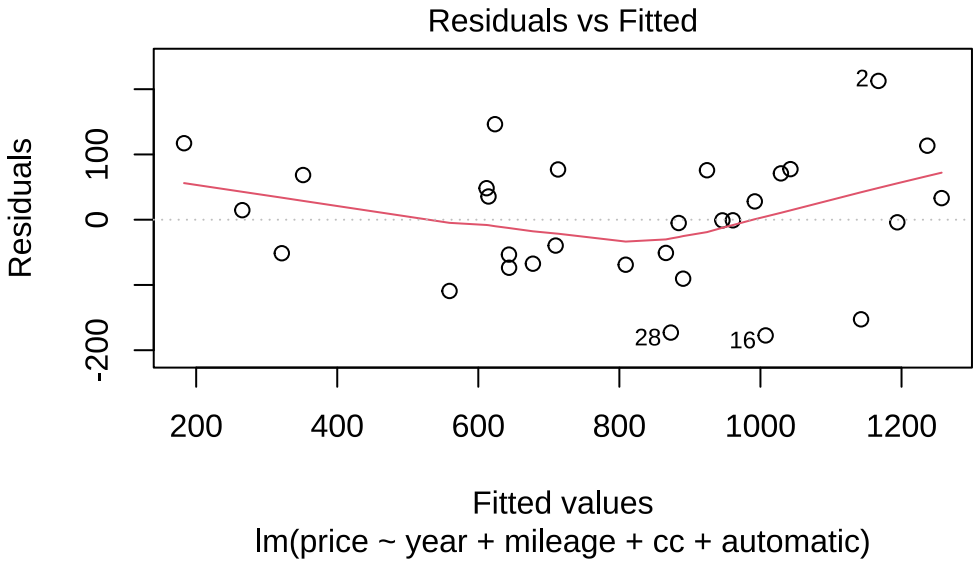
3. 스케일-위치(Scale-Location) 또는 스프레드-위치(Spread-Location)

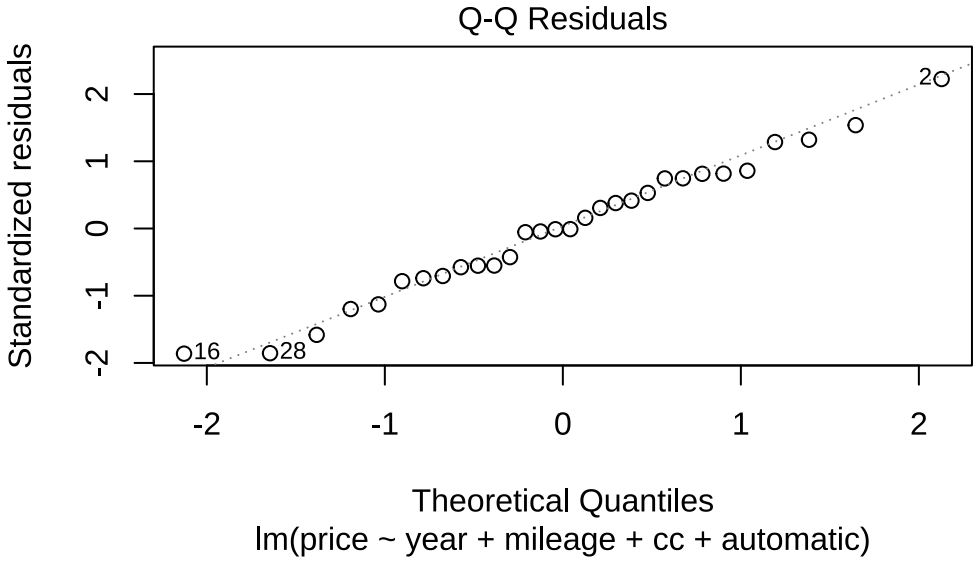
- 잔차 대 예측값 그래프와 유사하게, 스케일-위치 그래프는 잔차의 퍼진 정도를 보여주므로 등분산성을 확인하는 데 사용된다.
- 스케일-위치 그래프(또는 스프레드-위치 그래프)의 y축은 일반적으로 표준화된 잔차 절대값 의 제곱근을 나타낸다.
- 잔차 절대값 의 제곱근을 보여주는 것은 잔차의 분산을 안정화하는 데 도움을 주어, 이질적 분산(등분산성이 아닌)의 패턴을 시각적으로 더 쉽게 식별할 수 있게 한다.
- 점들이 대략적으로 수평선을 이루고 균등하게 퍼져 있어야 등분산성의 가정이 적합하다고 판단된다.

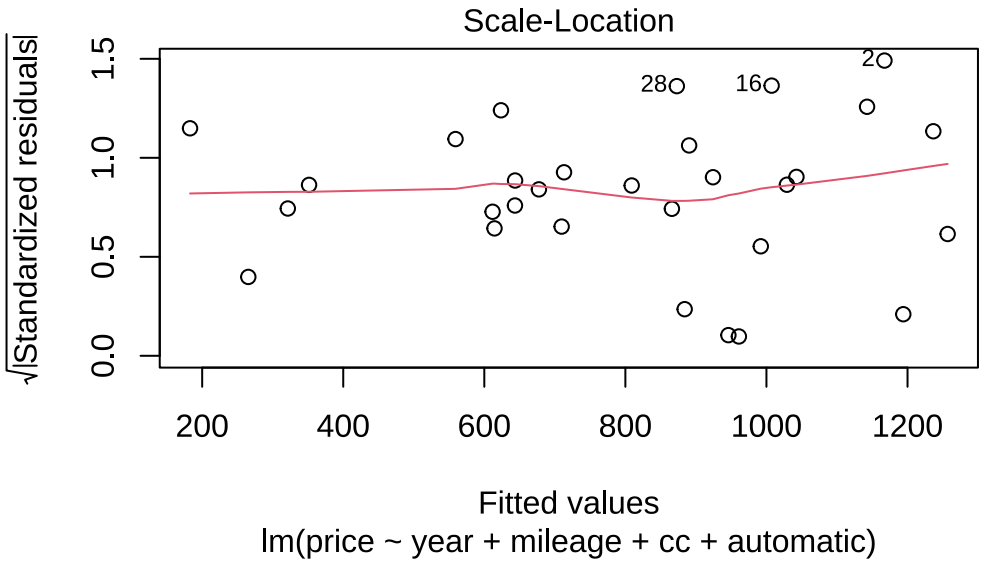
4. 잔차 대 지렛값(Residuals vs Leverage)

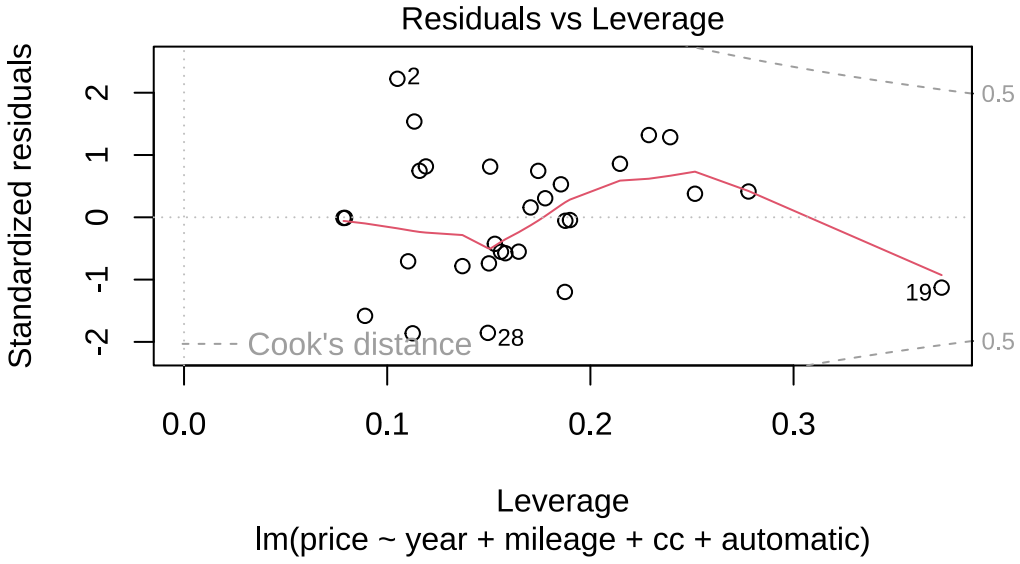
- 이 그래프는 회귀선에 영향을 줄 수 있는 영향력 있는 관찰값을 식별하는 데 도움이 된다.
- 높은 지렛값와 큰 잔차를 가진 관찰값(이상치)을 잘 파악할 수 있도록 만들어진 그림이다.
- 이 그래프에서는 다음과 같은 통계량들이 제시되는 것에 유의하자.
  - x축: 지렛값(leverage Values, h_{ii})
  - y축: 표준화된 잔차(Standardized Residuals)
  - 등고선: Cook's Distance

```
plot(usedcars.lm)
```


Residuals vs Fitted
Residuals
-200
0
100
200
400
600
800
1000
1200
2
28
16
Fitted values
lm(price ~ year + mileage + cc + automatic)


Q-Q Residuals
Standardized residuals
-2
-1
0
1
2
16
28
2
Theoretical Quantiles
lm(price ~ year + mileage + cc + automatic)





L.1.2. 잔차

다음과 같은 함수를 통하여 다양한 잔차들와 지렛값을 구할 수 있다.

```
resid_inter <- rstandard(usedcars.lm)  # internally studentized residual - 내표준화 잔차
resid_exter <- rstudent(usedcars.lm)   # externally studentized residual - 외표준화 잔차
hatval <- hatvalues(usedcars.lm)       # leverage value - 지렛값
data.frame(resid_inter , resid_exter, hatval)
```

```
    resid_inter  resid_exter     hatval
```

```
1   0.859118727  0.854468915 0.21455013
2   2.223678962  2.432560486 0.10501551
3  -0.553417436 -0.545588388 0.15599165
4  -0.044084943 -0.043195925 0.18996253
5  -0.576180936 -0.568325835 0.15814304
6   0.816421059  0.810807796 0.11903880
7  -0.551399974 -0.543574935 0.16470220
8  -1.198080391 -1.209098031 0.18743332
9   0.378399054  0.371820157 0.25147525
10  0.745189938  0.738380670 0.17431785
11 -0.010873657 -0.010653990 0.07847932
12  1.537452127  1.583088042 0.11334881
13 -0.706665346 -0.699408397 0.11029244
14  1.286252868  1.304157036 0.23928783
15  0.305838319  0.300221293 0.17774944
16 -1.862061057 -1.965848316 0.11253640
17 -0.425962246 -0.418878889 0.15297508
18 -1.582023043 -1.634007936 0.08908012
19 -1.128902196 -1.135412108 0.37287989
20  0.746526192  0.739734875 0.11591279
21 -0.739861832 -0.732982630 0.15005336
22 -0.783820912 -0.777598695 0.13701582
23  0.529387292  0.521623446 0.18549015
24  1.320232140  1.341155689 0.22878139
25 -0.009531759 -0.009339195 0.07924745
26  0.414031097  0.407063963 0.27781733
27  0.813419294  0.807745473 0.15066704
28 -1.855054061 -1.957267375 0.14953912
29  0.158642701  0.155515766 0.17056129
30 -0.055408868 -0.054292716 0.18765466
```

L.1.3. 영향점 측도

`influence.measures` 함수를 통하여 영향점을 파악하는 진단값을 구할 수 있다.

가장 중요한 통계량은 다음과 같다.

- `difft` : DFFITS
- `cook.d` : Cook's distance
- `hat` : leverage value

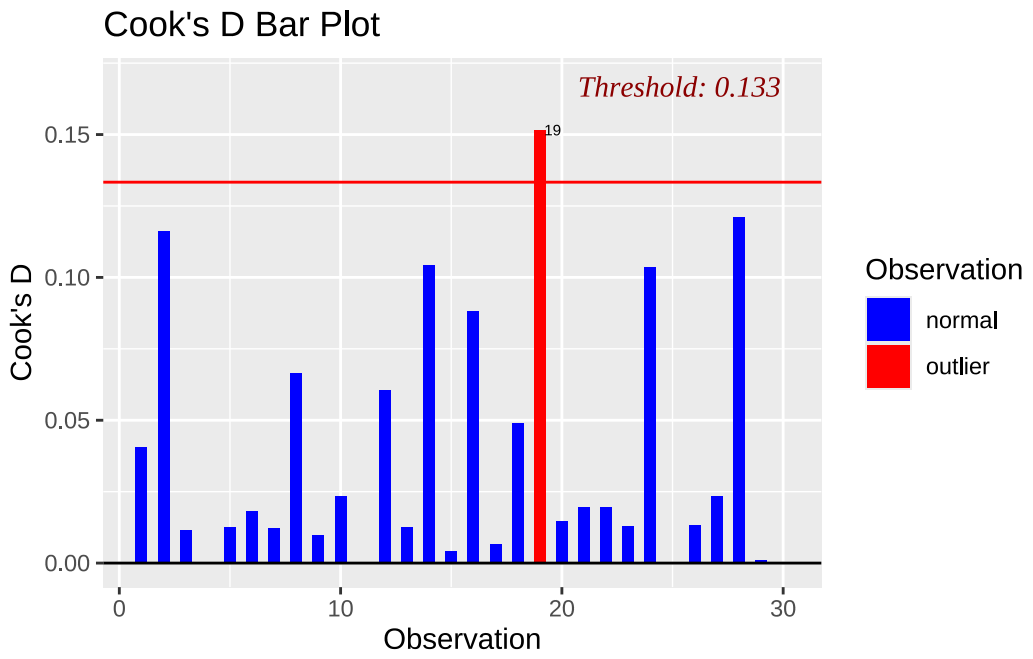
```
# DFBETAS for each model variable, DFFITS, covariance ratios,
# Cook's distances and the diagonal elements of the hat matrix
# Cases which are influential with respect to any of these measures
# are marked with an asterisk.
influence.measures(usedcars.lm)
```

```
Influence measures of
     lm(formula = price ~ year + mileage + cc + automatic, data = usedcars) :

     dfb.1_  dfb.year  dfb.milg   dfb.cc dfb.atmt    dffit cov.r   cook.d
1  -0.20716 -0.108544  0.327125  0.17932  0.19256  0.44658 1.344 4.03e-02
2  -0.18773  0.033416 -0.347354  0.24746  0.23435  0.83326 0.455 1.16e-01
3  -0.10302 -0.086990  0.008925  0.10950  0.06372 -0.23455 1.366 1.13e-02
4   0.00469  0.016237 -0.011729 -0.00589 -0.00320 -0.02092 1.513 9.12e-05
5   0.11228  0.102465 -0.135083 -0.12498  0.13462 -0.24632 1.363 1.25e-02
6  -0.07885  0.136480 -0.181209  0.08714  0.11239  0.29805 1.216 1.80e-02
7  -0.14228  0.100579 -0.106201  0.14682 -0.10174 -0.24137 1.381 1.20e-02
8  -0.27687  0.308498 -0.320654  0.25704  0.24993 -0.58070 1.123 6.62e-02
9   0.15471 -0.066016 -0.054370 -0.13883  0.03146  0.21551 1.592 9.62e-03
10  0.13093 -0.056206  0.169173 -0.13765 -0.10220  0.33927 1.328 2.34e-02
11  0.00143 -0.000670  0.000312 -0.00142 -0.00166 -0.00311 1.331 2.01e-06
12 -0.27881  0.085189 -0.022194  0.31082 -0.33867  0.56603 0.842 6.04e-02
13  0.10909 -0.027792  0.028700 -0.12774  0.15983 -0.24625 1.246 1.24e-02
14  0.52930 -0.229975 -0.166587 -0.47750  0.11713  0.73144 1.145 1.04e-01
15 -0.02434 -0.037299 -0.032166  0.04432 -0.10023  0.13959 1.464 4.04e-03
16 -0.47183  0.092199  0.101866  0.45460 -0.29124 -0.70004 0.655 8.79e-02
17  0.08120 -0.112348  0.030779 -0.06848 -0.09956 -0.17801 1.396 6.55e-03
18  0.14607  0.138551  0.019412 -0.18052 -0.15882 -0.51098 0.795 4.90e-02
19  0.08767 -0.454297  0.733107 -0.15988  0.36621 -0.87551 1.505 1.52e-01
20  0.17302 -0.084980  0.007572 -0.16482  0.10281  0.26785 1.239 1.46e-02
21  0.14757  0.081228 -0.204480 -0.13332 -0.13993 -0.30798 1.292 1.93e-02
22 -0.21369 -0.011866  0.084126  0.19253  0.16353 -0.30984 1.255 1.95e-02
23 -0.12798  0.071546  0.083597  0.10512  0.13711  0.24893 1.423 1.28e-02
24  0.22299  0.430646 -0.103308 -0.26300 -0.09994  0.73047 1.108 1.03e-01
25  0.00129  0.000354 -0.000686 -0.00128 -0.00137 -0.00274 1.332 1.56e-06
26  0.06915  0.204983 -0.141142 -0.08585  0.12561  0.25248 1.641 1.32e-02
27 -0.08134 -0.093035 -0.048156  0.12751 -0.25004  0.34021 1.263 2.35e-02
28  0.28532 -0.571184  0.447521 -0.26551 -0.37866 -0.82073 0.688 1.21e-01
29  0.02625  0.019365  0.010861 -0.02922 -0.01619  0.07052 1.471 1.04e-03
30  0.00732  0.016726 -0.010214 -0.01018  0.01709 -0.02609 1.509 1.42e-04
      hat inf
1  0.2146
```

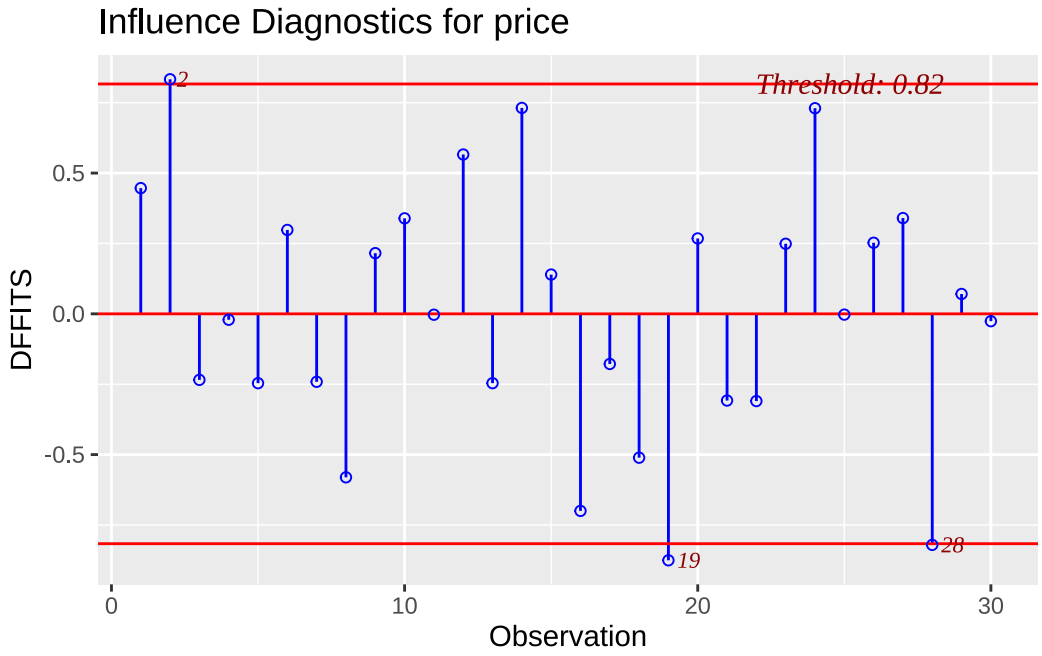
```
2  0.1050
3  0.1560
4  0.1900
5  0.1581
6  0.1190
7  0.1647
8  0.1874
9  0.2515
10 0.1743
11 0.0785
12 0.1133
13 0.1103
14 0.2393
15 0.1777
16 0.1125
17 0.1530
18 0.0891
19 0.3729
20 0.1159
21 0.1501
22 0.1370
23 0.1855
24 0.2288
25 0.0792
26 0.2778   *
27 0.1507
28 0.1495
29 0.1706
30 0.1877
```

패키지 `olsrr` 의 함수 `ols_plot_cooksd_bar` 과 `ols_plot_dffits` 를 이용하여 각각 Cook's distance 와 DFFIT 를 시각화할 수 있다.

```
ols_plot_cooksd_bar(usedcars.lm)
```




```
ols_plot_dffits(usedcars.lm)
```



L.2. 교과서 연습문제 5.14

```
head(MLB1)
```

```
  hit85 hit86
```

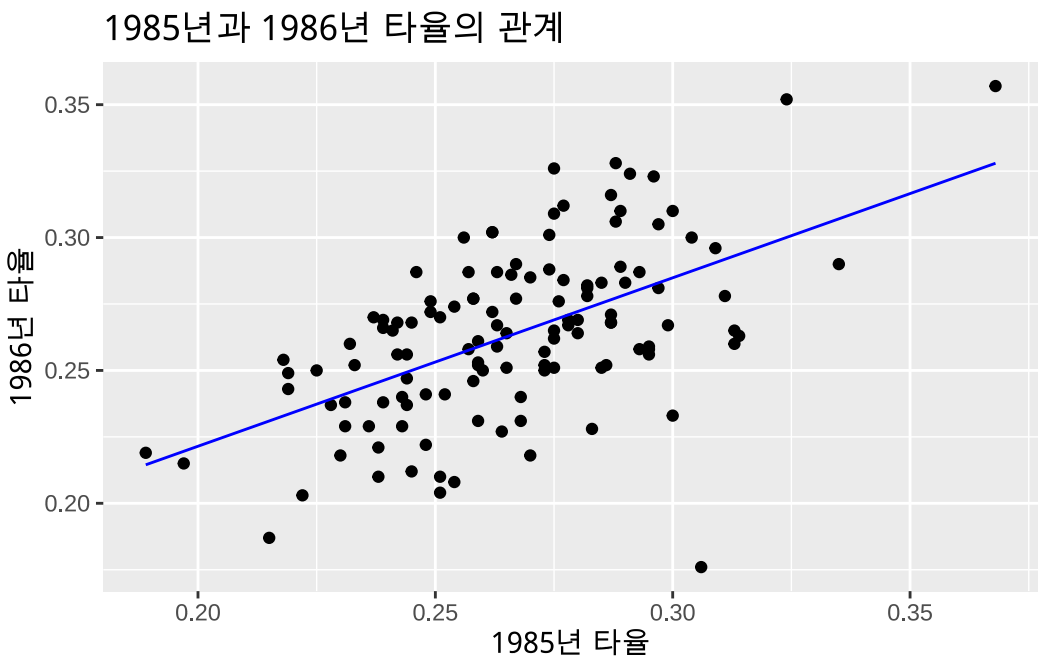
```
1 0.265 0.264
2 0.309 0.296
3 0.268 0.240
4 0.243 0.229
5 0.289 0.289
6 0.266 0.286
```

L.2.1. 회귀모형 적합

```
mlb1.lm <- lm(hit86 ~ hit85, data=MLB1)
summary(mlb1.lm)
```

```

Call:
lm(formula = hit86 ~ hit85, data = MLB1)

Residuals:
     Min       1Q   Median       3Q      Max
-0.11265 -0.01708 -0.00075  0.01887  0.05700

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.09470    0.02310   4.100 7.49e-05 ***
hit85        0.63383    0.08622   7.351 2.47e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.02679 on 122 degrees of freedom
Multiple R-squared:  0.307, Adjusted R-squared:  0.3013
F-statistic: 54.04 on 1 and 122 DF,  p-value: 2.471e-11
```

```
ggplot(MLB1, aes(x=hit85, y=hit86)) + geom_point() + labs(x = "1985년 타율", y = "1986년 타율") +
  labs(title="1985년과 1986년 타율의 관계") +
  geom_line(aes(y=mlb1.lm$fitted.values), color="blue")
```

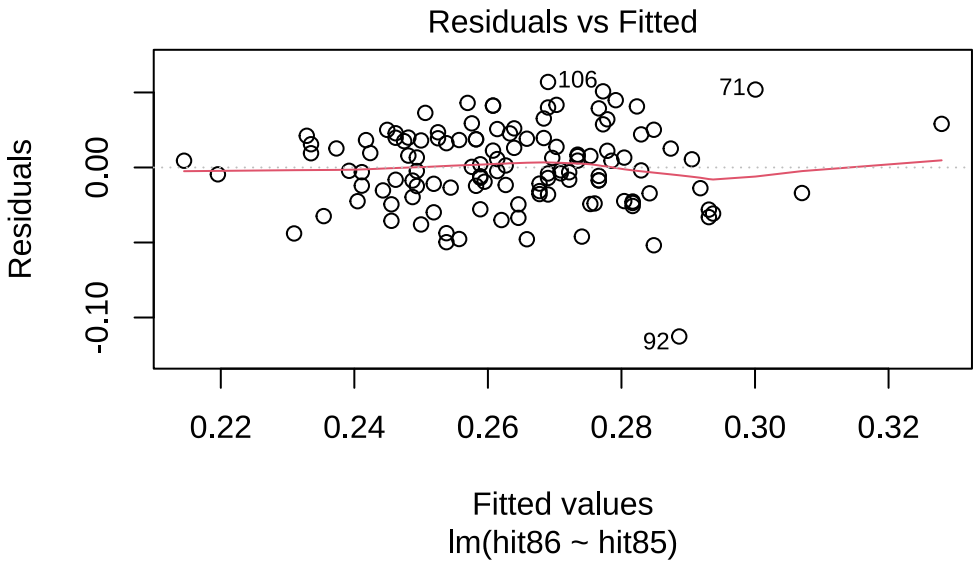


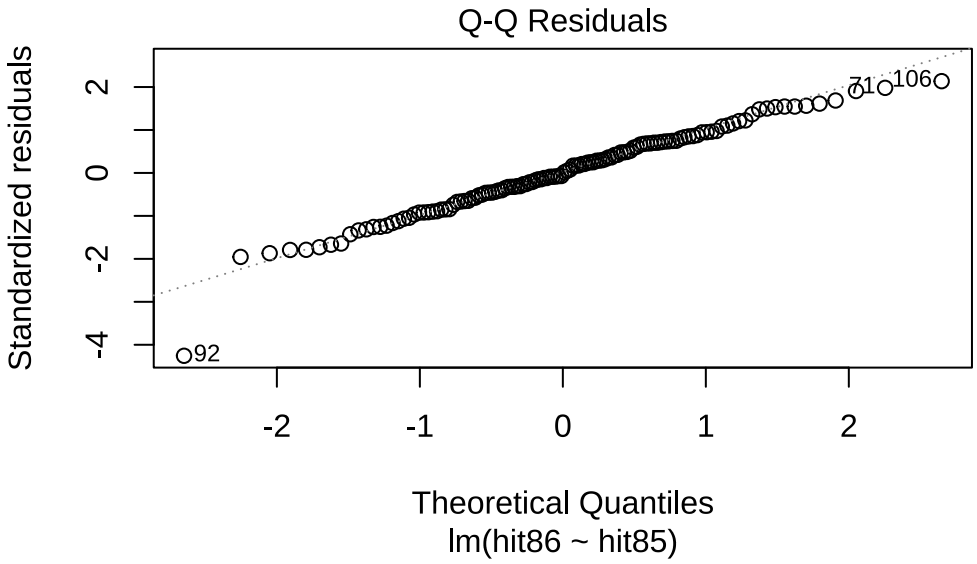
L.2.2. 잔차그림

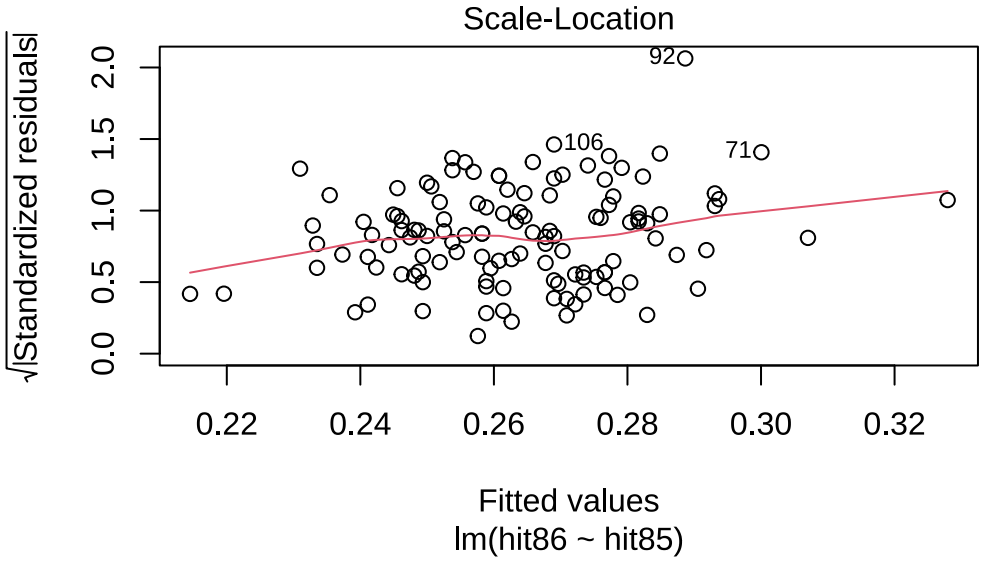
다음잔차 그림을 보면 다음과 같은 관측값이 이상점 또는 영향점일 가능성이 크다.

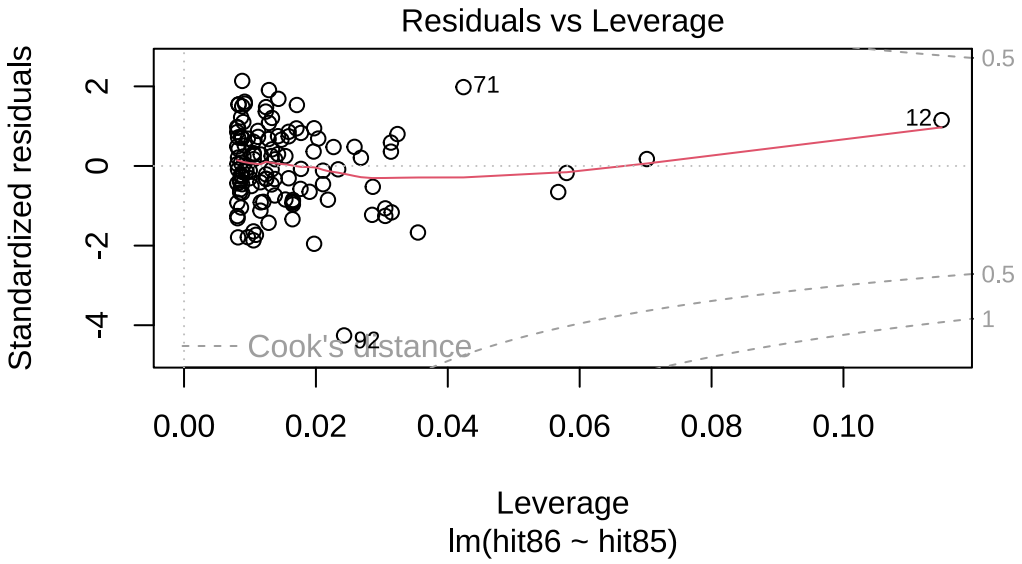
- 잔차가 큰 관측값 번호: 71, 92, 106
- 지렛값이 큰 관측값 번호 : 12

```
mlb1.lm <- lm(hit86 ~ hit85, data=MLB1)
plot(mlb1.lm)
```




Q-Q Residuals
Standardized residuals
-4
-2
0
2
92
71
106
-2
-1
0
1
2
Theoretical Quantiles
lm(hit86 ~ hit85)


Scale-Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
92
106
71
0.22
0.24
0.26
0.28
0.30
0.32
Fitted values
lm(hit86 ~ hit85)



L.2.3. 잔차와 지렛값 분석

아래 잔차와 지렛값을 보면 위에서 그림을 그려서 파악한 것과 일치하는 것을 확인할 수 있다.

```
resid_inter <- rstandard(mlb1.lm)  # internal studentized residual
resid_exter <- rstudent(mlb1.lm)   # external studentized residual
hatval <- hatvalues(mlb1.lm)       # leverage
resid_df <- data.frame(resid_inter , resid_exter, hatval)
resid_df
```

```
    resid_inter resid_exter      hatval
1    0.05000699  0.04980213 0.008086100
2    0.20604678  0.20523630 0.026823219
3   -0.92071669 -0.92013786 0.008089609
4   -0.74122875 -0.73985250 0.013757221
5    0.41794462  0.41652651 0.013334552
6    0.85074812  0.84977870 0.008066554
7   -0.11752895 -0.11705291 0.021076586
8    1.49950517  1.50730174 0.008822848
9    0.47628862  0.47477421 0.022674209
10   1.22323441  1.22575081 0.008655952
11  -0.96655846 -0.96629590 0.016511088
12   1.15252304  1.15408980 0.114892991
13  -1.12220362 -1.12340818 0.011587908
14   0.94814587  0.94775033 0.019727886
```

```
15  -0.65464742 -0.65310705 0.056746543
16   0.94426865  0.94384613 0.017044019
17   0.29746497  0.29635083 0.014253231
18  -0.64880397 -0.64725708 0.019043095
19   0.35993460  0.35864690 0.031379027
20   0.85838160  0.85744961 0.015865556
21  -0.52346010 -0.52189678 0.028627834
22   1.54555587  1.55450230 0.008269034
23   0.60818462  0.60660721 0.010534910
24  -0.84308211 -0.84207635 0.015369379
25  -0.17556386 -0.17486494 0.058014551
26  -0.21071097 -0.20988382 0.012441433
27  -0.32339097 -0.32220100 0.012441433
28   0.24968479  0.24872294 0.013281927
29  -1.78885907 -1.80534673 0.009668354
30   0.20995888  0.20913441 0.008187340
31   1.54555587  1.55450230 0.008269034
32  -0.40887179 -0.40747191 0.011587908
33  -0.08987603 -0.08950989 0.008187340
34   1.20706156  1.20934748 0.013334552
35  -0.92405584 -0.92349840 0.016444430
36   1.61388664  1.62469547 0.009194229
37   1.08032096  1.08106766 0.012877634
38   1.36629192  1.37121192 0.012393486
39   0.47981611  0.47829711 0.025854922
40   0.68662786  0.68513312 0.009668354
41   0.28365424  0.28258253 0.010571163
42  -0.89128320 -0.89052689 0.016511088
43  -0.32339097 -0.32220100 0.012441433
44   0.74859901  0.74724285 0.014253231
45   0.71830103  0.71686860 0.008195527
46  -0.66485460 -0.66332696 0.008509772
47  -1.79282669 -1.80945914 0.008195527
48   1.53177579  1.54036942 0.017113016
49   0.17108078  0.17039863 0.010571163
50   0.65981706  0.65828292 0.014769957
51  -1.16526569 -1.16699260 0.031490124
52  -0.08849917 -0.08813855 0.013281927
53   0.28750459  0.28642092 0.011631177
54  -0.14646855 -0.14587986 0.009447831
55  -1.25801434 -1.26105391 0.008089609
56  -1.25437758 -1.25736066 0.030515312
```

```
57  -0.50371052 -0.50216434 0.010225342
58   0.73588754  0.73449735 0.008655952
59  -0.35597858 -0.35470091 0.008494569
60  -0.45705663 -0.45556979 0.021076586
61  -0.32779322 -0.32659089 0.013757221
62  -1.33805577 -1.34244750 0.016444430
63  -0.84844614 -0.84746568 0.021821186
64   0.24777624  0.24682078 0.015369379
65   0.51479711  0.51324069 0.009218787
66   0.08012340  0.07979645 0.008638411
67  -0.43719987 -0.43574587 0.008086100
68   0.95954615  0.95923199 0.008187340
69  -1.31293093 -1.31687542 0.008126362
70   0.80156331  0.80038181 0.032372217
71   1.98105733  2.00544148 0.042377621
72   0.74550574  0.74414103 0.015865556
73  -0.25726760 -0.25628058 0.008638411
74  -0.07337522 -0.07307550 0.017735660
75   0.68826166  0.68676971 0.020352702
76  -0.21977971 -0.21892046 0.008638411
77  -1.22738844 -1.22996533 0.028523754
78  -0.85364557 -0.85269021 0.016511088
79   0.36253805  0.36124383 0.019649534
80   0.83049085  0.82942806 0.017735660
81   0.97690284  0.97671857 0.008067723
82   0.16870598  0.16803274 0.013812185
83   0.48970050  0.48816941 0.008067723
84  -1.04451327 -1.04490630 0.008638411
85  -0.30800228 -0.30685670 0.015865556
86  -0.40245640 -0.40106992 0.008509772
87   0.88119841  0.88038574 0.011216193
88   0.01518447  0.01512212 0.008988241
89  -0.15011541 -0.14951273 0.008822848
90   1.90682355  1.92793919 0.012877634
91   0.23854907  0.23762483 0.009010460
92  -4.25681260 -4.59422430 0.024271666
93   1.68633876  1.69933527 0.014310534
94  -0.90035488 -0.89965120 0.012025947
95   0.32117873  0.31999502 0.010571163
96  -1.64324327 -1.65491152 0.010534910
97  -1.95470304 -1.97789425 0.019727886
98   0.58748696  0.58590363 0.031379027
```

```
99   0.42112866  0.41970434 0.008269034
100 -1.72921841 -1.74361729 0.010903785
101 -0.08361786 -0.08327684 0.023372533
102 -0.30660015 -0.30545871 0.009968065
103 -0.45847196 -0.45698295 0.008802968
104  1.10252512  1.10350851 0.008988241
105 -1.42714716 -1.43330075 0.012827348
106  2.13685857  2.16906138 0.008822848
107 -1.86838606 -1.88791915 0.010534910
108 -1.67127243 -1.68379544 0.035476081
109 -0.57561180 -0.57402789 0.017664324
110 -1.06483510 -1.06542465 0.030515312
111  0.70374907  0.70228584 0.008802968
112 -0.67499469 -0.67348138 0.008822848
113  0.70374907  0.70228584 0.008802968
114  1.47948910  1.48681148 0.012441433
115 -0.11903487 -0.11855291 0.009968065
116  0.73105151  0.72964915 0.011216193
117  0.67662401  0.67511317 0.012827348
118 -0.91392269 -0.91330116 0.011631177
119  0.17440445  0.17370986 0.070186094
120 -0.46425912 -0.46276146 0.013281927
121 -0.07146215 -0.07117016 0.009447831
122 -0.26258954 -0.26158508 0.008822848
123 -0.58988369 -0.58830072 0.008509772
124  1.56476539  1.57421623 0.009218787
```

```
resid_df %>% dplyr::arrange(desc(abs(resid_inter))) %>% head(10)
```

```
    resid_inter resid_exter      hatval
92    -4.256813   -4.594224 0.024271666
106    2.136859    2.169061 0.008822848
71     1.981057    2.005441 0.042377621
97    -1.954703   -1.977894 0.019727886
90     1.906824    1.927939 0.012877634
107   -1.868386   -1.887919 0.010534910
47    -1.792827   -1.809459 0.008195527
29    -1.788859   -1.805347 0.009668354
100   -1.729218   -1.743617 0.010903785
93     1.686339    1.699335 0.014310534
```

```
resid_df %>% dplyr::arrange(desc(hatval)) %>% head(10)
```

```
    resid_inter resid_exter     hatval
12    1.1525230   1.1540898 0.11489299
119   0.1744045   0.1737099 0.07018609
25   -0.1755639  -0.1748649 0.05801455
15   -0.6546474  -0.6531071 0.05674654
71    1.9810573   2.0054415 0.04237762
108  -1.6712724  -1.6837954 0.03547608
70    0.8015633   0.8003818 0.03237222
51   -1.1652657  -1.1669926 0.03149012
19    0.3599346   0.3586469 0.03137903
98    0.5874870   0.5859036 0.03137903
```

L.2.4. 영향점 측도

아래 COOK 거리와 DFFIT 값을 보면 다음 관측값등이 회귀 적합에 큰 영향을 미치는 것으로 나타난다.

- 92, 71, 12, 108, 97

위에서 잔차와 지렛값만으로 파악한 것과 거의 일치하는 것을 확인할 수 있다.

다만 차이가 나는 점은 다음과 같다.

- 106 번 관측값이 잔차는 매우 크지만 지렛값이 상대적으로 작아서 COOK 거리와 DFFIT 값이 크게 나타나지 않는다는 것이다.
- 97, 108 번 관측값은 잔차는 작지만 지렛값이 크기 때문에 COOK 거리와 DFFIT 값이 크게 나타난다는 것이다.

92, 71, 12 관측값은 잔차와 지렛값이 모두 크기 때문에 COOK 거리와 DFFIT 값이 크게 나타난다.

```
influence.measures(mlb1.lm)
```

```
Influence measures of
     lm(formula = hit86 ~ hit85, data = MLB1) :

       dfb.1_  dfb.ht85    dffit cov.r   cook.d     hat inf
1    0.000699 -0.000232  0.00450 1.025 1.02e-05 0.00809
2   -0.026394  0.028494  0.03407 1.044 5.85e-04 0.02682
3   -0.004039 -0.004628 -0.08310 1.011 3.46e-03 0.00809
4   -0.062872  0.056210 -0.08738 1.022 3.83e-03 0.01376
5   -0.026354  0.030441  0.04842 1.027 1.18e-03 0.01333
6    0.009192 -0.001218  0.07663 1.013 2.94e-03 0.00807
```

```
7   -0.014528  0.013495 -0.01718 1.038 1.49e-04 0.02108
8   -0.027304  0.041692  0.14221 0.988 1.00e-02 0.00882
9   -0.053240  0.058048  0.07232 1.036 2.63e-03 0.02267
10  -0.018262  0.029939  0.11454 1.000 6.53e-03 0.00866
11   0.079949 -0.089550 -0.12520 1.018 7.84e-03 0.01651
12  -0.387291  0.400945  0.41580 1.124 8.62e-02 0.11489   *
13  -0.077277  0.067073 -0.12164 1.007 7.38e-03 0.01159
14  -0.093863  0.103379  0.13445 1.022 9.05e-03 0.01973
15   0.141276 -0.148373 -0.16019 1.070 1.29e-02 0.05675   *
16   0.098625 -0.090211  0.12429 1.019 7.73e-03 0.01704
17   0.026146 -0.023481  0.03564 1.030 6.40e-04 0.01425
18   0.061989 -0.068474 -0.09018 1.029 4.09e-03 0.01904
19   0.058748 -0.055642  0.06455 1.047 2.10e-03 0.03138
20   0.084010 -0.076341  0.10887 1.021 5.94e-03 0.01587
21   0.070568 -0.075934 -0.08960 1.042 4.04e-03 0.02863
22   0.036802 -0.022323  0.14195 0.985 9.96e-03 0.00827
23   0.035849 -0.030310  0.06259 1.021 1.97e-03 0.01053
24   0.064199 -0.072531 -0.10521 1.020 5.55e-03 0.01537
25  -0.041733  0.040267 -0.04340 1.079 9.49e-04 0.05801   *
26   0.011921 -0.013973 -0.02356 1.029 2.80e-04 0.01244
27   0.018301 -0.021450 -0.03616 1.028 6.59e-04 0.01244
28   0.020330 -0.018086  0.02886 1.029 4.20e-04 0.01328
29  -0.089226  0.072652 -0.17838 0.973 1.56e-02 0.00967
30   0.004279 -0.002327  0.01900 1.024 1.82e-04 0.00819
31   0.036802 -0.022323  0.14195 0.985 9.96e-03 0.00827
32  -0.028029  0.024328 -0.04412 1.026 9.80e-04 0.01159
33  -0.001831  0.000996 -0.00813 1.025 3.33e-05 0.00819
34  -0.076515  0.088384  0.14059 1.006 9.85e-03 0.01333
35  -0.093489  0.085243 -0.11941 1.019 7.14e-03 0.01644
36   0.069829 -0.054861  0.15651 0.983 1.21e-02 0.00919
37  -0.064901  0.075488  0.12348 1.010 7.61e-03 0.01288
38   0.103195 -0.090783  0.15361 0.998 1.17e-02 0.01239
39   0.068818 -0.064637  0.07792 1.040 3.06e-03 0.02585
40   0.033861 -0.027572  0.06770 1.019 2.30e-03 0.00967
41  -0.011489  0.014223  0.02921 1.026 4.30e-04 0.01057
42   0.073680 -0.082528 -0.11539 1.020 6.67e-03 0.01651
43   0.018301 -0.021450 -0.03616 1.028 6.59e-04 0.01244
44   0.065925 -0.059208  0.08985 1.022 4.05e-03 0.01425
45  -0.001461  0.008239  0.06517 1.016 2.13e-03 0.00820
46   0.007749 -0.014057 -0.06145 1.018 1.90e-03 0.00851
47   0.003689 -0.020796 -0.16448 0.972 1.33e-02 0.00820
48  -0.132459  0.147796  0.20325 0.995 2.04e-02 0.01711
```

```
49  -0.006928  0.008577  0.01761 1.027 1.56e-04 0.01057
50   0.060215 -0.054307  0.08060 1.024 3.26e-03 0.01477
51   0.169415 -0.181494 -0.21043 1.026 2.21e-02 0.03149
52  -0.007204  0.006409 -0.01023 1.030 5.27e-05 0.01328
53  -0.014418  0.017206  0.03107 1.027 4.86e-04 0.01163
54   0.004051 -0.005452 -0.01425 1.026 1.02e-04 0.00945
55  -0.005535 -0.006343 -0.11388 0.998 6.45e-03 0.00809
56   0.178355 -0.191340 -0.22307 1.022 2.48e-02 0.03052
57  -0.028057  0.023463 -0.05104 1.023 1.31e-03 0.01023
58  -0.010943  0.017940  0.06863 1.016 2.36e-03 0.00866
59  -0.010679  0.007387 -0.03283 1.023 5.43e-04 0.00849
60  -0.056545  0.052524 -0.06685 1.035 2.25e-03 0.02108
61  -0.027754  0.024813 -0.03857 1.029 7.49e-04 0.01376
62  -0.135900  0.123913 -0.17358 1.003 1.50e-02 0.01644
63  -0.107969  0.100501 -0.12658 1.027 8.03e-03 0.02182
64  -0.018817  0.021259  0.03084 1.031 4.79e-04 0.01537
65  -0.012600  0.017518  0.04951 1.022 1.23e-03 0.00922
66   0.002659 -0.001920  0.00745 1.025 2.80e-05 0.00864
67  -0.006114  0.002033 -0.03934 1.022 7.79e-04 0.00809
68   0.019626 -0.010675  0.08715 1.010 3.80e-03 0.00819
69  -0.022710  0.010399 -0.11920 0.996 7.06e-03 0.00813
70   0.133778 -0.126857  0.14640 1.040 1.07e-02 0.03237
71  -0.358382  0.379615  0.42187 0.994 8.68e-02 0.04238   *
72   0.072909 -0.066253  0.09448 1.024 4.48e-03 0.01587
73  -0.008540  0.006166 -0.02392 1.024 2.88e-04 0.00864
74   0.006522 -0.007251 -0.00982 1.035 4.86e-05 0.01774
75   0.082988 -0.076917  0.09899 1.030 4.92e-03 0.02035
76  -0.007295  0.005267 -0.02044 1.025 2.10e-04 0.00864
77  -0.189194  0.178493 -0.21076 1.021 2.21e-02 0.02852
78   0.070550 -0.079022 -0.11048 1.021 6.12e-03 0.01651
79   0.042469 -0.039270  0.05114 1.035 1.32e-03 0.01965
80  -0.074025  0.082301  0.11145 1.023 6.23e-03 0.01774
81   0.007426  0.001756  0.08809 1.009 3.88e-03 0.00807
82  -0.011176  0.012828  0.01989 1.030 1.99e-04 0.01381
83   0.003712  0.000878  0.04403 1.021 9.75e-04 0.00807
84  -0.034820  0.025141 -0.09754 1.007 4.75e-03 0.00864
85  -0.030065  0.027320 -0.03896 1.031 7.65e-04 0.01587
86   0.004686 -0.008499 -0.03716 1.023 6.95e-04 0.00851
87   0.057715 -0.049704  0.09377 1.015 4.40e-03 0.01122
88   0.000601 -0.000462  0.00144 1.026 1.05e-06 0.00899
89   0.002708 -0.004136 -0.01411 1.025 1.00e-04 0.00882
90  -0.115741  0.134624  0.22020 0.969 2.37e-02 0.01288
```

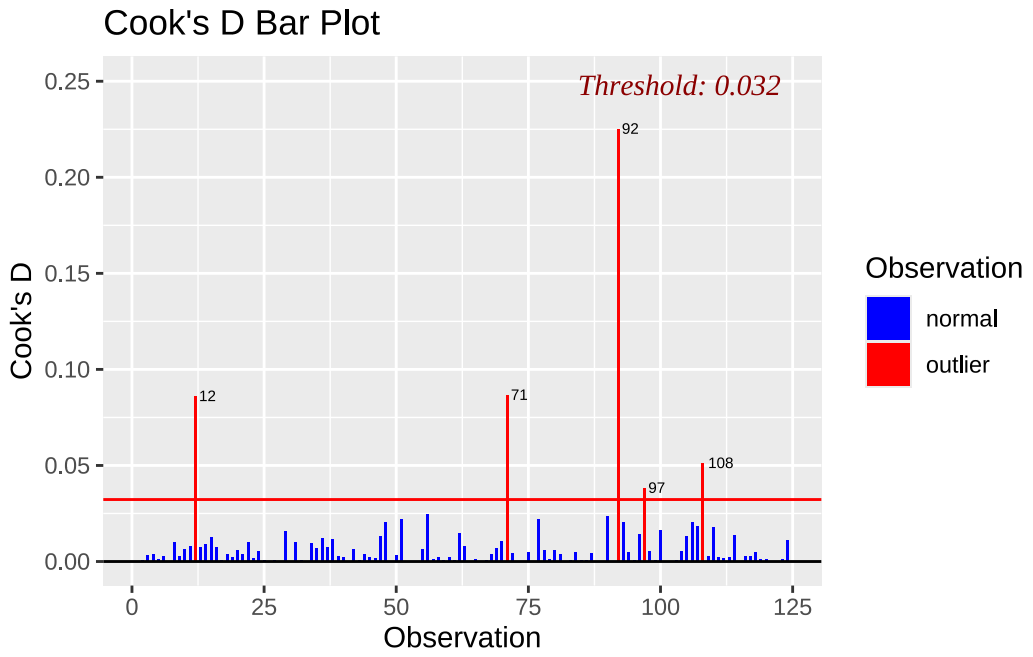
```
91  -0.005069  0.007342  0.02266 1.025 2.59e-04 0.00901
92   0.545385 -0.592108 -0.72460 0.755 2.25e-01 0.02427   *
93  -0.118528  0.135273  0.20476 0.984 2.06e-02 0.01431
94   0.048192 -0.056968 -0.09926 1.015 4.93e-03 0.01203
95  -0.013010  0.016106  0.03308 1.026 5.51e-04 0.01057
96  -0.097802  0.082691 -0.17076 0.982 1.44e-02 0.01053
97   0.195887 -0.215746 -0.28059 0.973 3.84e-02 0.01973
98   0.095974 -0.090900  0.10546 1.044 5.59e-03 0.03138
99   0.009936 -0.006027  0.03832 1.022 7.39e-04 0.00827
100  0.076513 -0.093419 -0.18307 0.978 1.65e-02 0.01090
101 -0.011158  0.010426 -0.01288 1.041 8.37e-05 0.02337
102  0.010450 -0.013394 -0.03065 1.025 4.73e-04 0.00997
103 -0.016699  0.012473 -0.04307 1.022 9.33e-04 0.00880
104  0.043875 -0.033691  0.10509 1.005 5.51e-03 0.00899
105 -0.112509  0.099557 -0.16338 0.996 1.32e-02 0.01283
106 -0.039292  0.059996  0.20464 0.950 2.03e-02 0.00882   *
107 -0.111573  0.094334 -0.19480 0.969 1.86e-02 0.01053
108 -0.298349  0.283857 -0.32292 1.006 5.14e-02 0.03548
109 -0.061854  0.056746 -0.07698 1.029 2.98e-03 0.01766
110  0.151129 -0.162132 -0.18902 1.029 1.78e-02 0.03052
111  0.025662 -0.019169  0.06618 1.017 2.20e-03 0.00880
112  0.012200 -0.018629 -0.06354 1.018 2.03e-03 0.00882
113  0.025662 -0.019169  0.06618 1.017 2.20e-03 0.00880
114 -0.084450  0.098983  0.16688 0.993 1.38e-02 0.01244
115  0.004056 -0.005198 -0.01190 1.027 7.13e-05 0.00997
116  0.047833 -0.041194  0.07771 1.019 3.03e-03 0.01122
117  0.052994 -0.046894  0.07696 1.022 2.97e-03 0.01283
118  0.045973 -0.054864 -0.09908 1.015 4.91e-03 0.01163
119  0.046341 -0.044900  0.04773 1.093 1.15e-03 0.07019   *
120 -0.037825  0.033650 -0.05369 1.027 1.45e-03 0.01328
121  0.001976 -0.002660 -0.00695 1.026 2.44e-05 0.00945
122  0.004739 -0.007235 -0.02468 1.024 3.07e-04 0.00882
123  0.006873 -0.012467 -0.05450 1.019 1.49e-03 0.00851
124 -0.038647  0.053732  0.15185 0.985 1.14e-02 0.00922
```

```
data.frame(influence.measures(mlb1.lm)$infmat) %>% arrange(desc(cook.d)) %>% head(10)
```

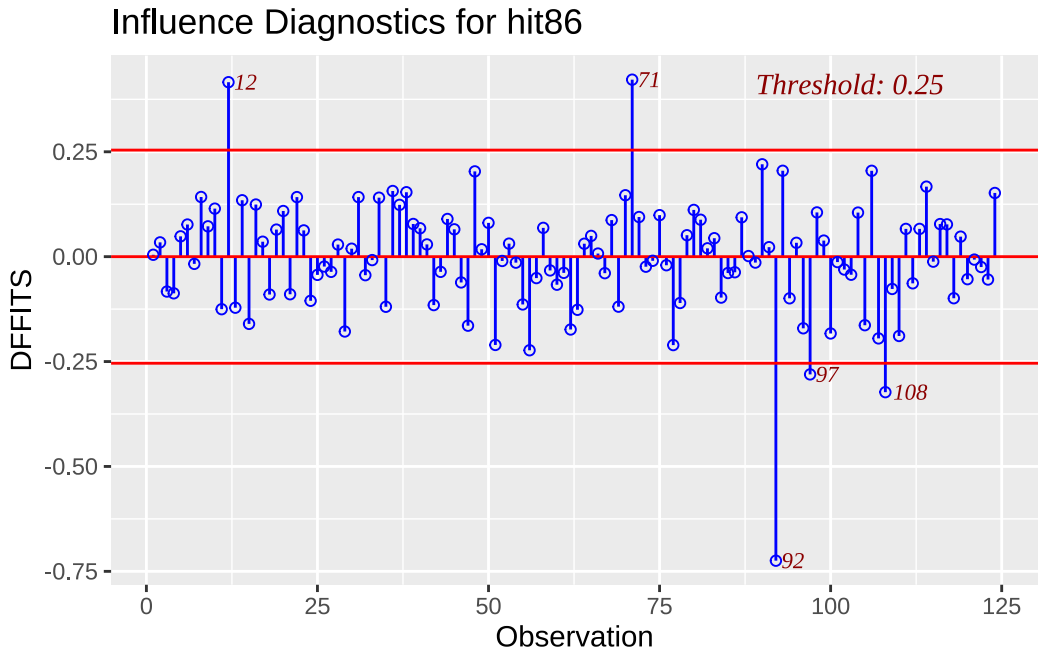
```
        dfb.1_   dfb.ht85      dffit     cov.r     cook.d        hat
92   0.5453850 -0.5921082 -0.7245987 0.7553712 0.22537707 0.02427167
71  -0.3583817  0.3796146  0.4218724 0.9943835 0.08683731 0.04237762
12  -0.3872908  0.4009454  0.4158038 1.1236842 0.08621185 0.11489299
108 -0.2983494  0.2838571 -0.3229242 1.0062796 0.05136735 0.03547608
```

```
97   0.1958869 -0.2157455 -0.2805886 0.9731151 0.03844727 0.01972789
56   0.1783549 -0.1913398 -0.2230737 1.0217219 0.02476301 0.03051531
90  -0.1157414  0.1346236  0.2202043 0.9693882 0.02371680 0.01287763
77  -0.1891943  0.1784931 -0.2107561 1.0207619 0.02211610 0.02852375
51   0.1694152 -0.1814937 -0.2104279 1.0264159 0.02207447 0.03149012
93  -0.1185276  0.1352728  0.2047561 0.9838364 0.02064312 0.01431053
```

```
ols_plot_cooksd_bar(mlb1.lm)
```



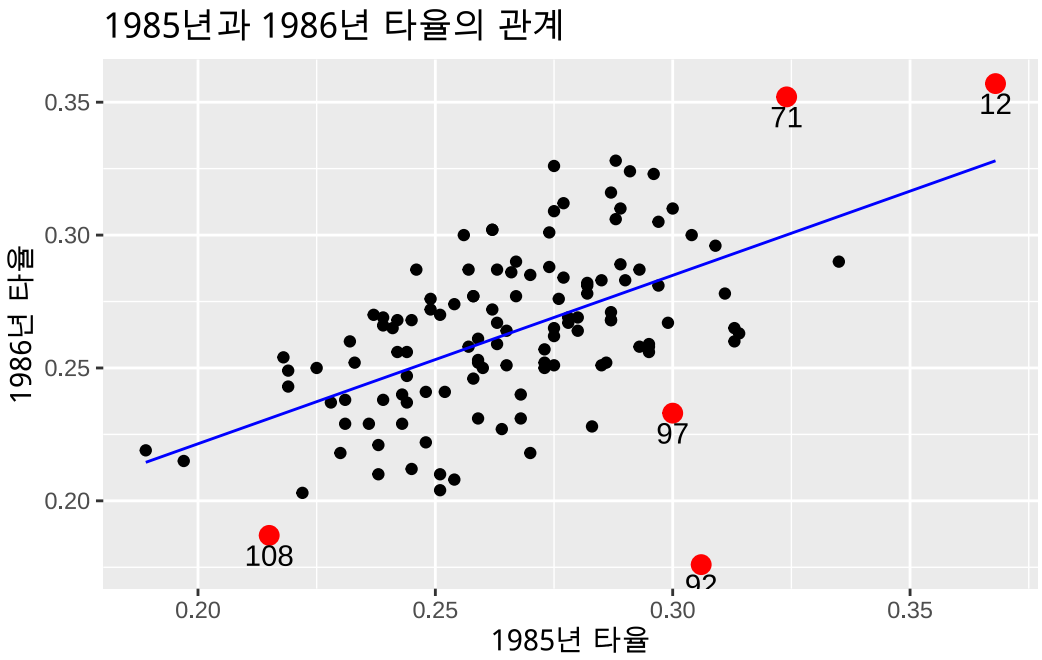

```
ols_plot_dffits(mlb1.lm)
```



L.2.5. 결론

위에서 나타난 관측값들을 나타내는 산점도를 다시 그려보자.

```
influnce_obs <- c(92, 71, 12, 108, 97)
MLB1$row_number <- seq_len(nrow(MLB1))

ggplot(MLB1, aes(x=hit85, y=hit86)) + geom_point() + labs(x = "1985년 타율", y = "1986년 타율") +
  labs(title="1985년과 1986년 타율의 관계") +
  geom_line(aes(y=mlb1.lm$fitted.values), color="blue") +
  geom_point(data = MLB1[influnce_obs, ], aes(x=hit85, y=hit86),
                    color = "red", size = 3) +
  geom_text(data = MLB1[influnce_obs, ], aes(x=hit85, y=hit86, label = row_number),
                   color = "black", vjust = 1.5, hjust = 0.5)
```



이제 영향점과 이상점을 제거한 후 회귀모형을 다시 적합해보자.

```
MLB1_clean <- MLB1 %>% filter(!row_number %in% influnce_obs)

mlb1_clean.lm <- lm(hit86 ~ hit85, data=MLB1_clean)
summary(mlb1_clean.lm)
```

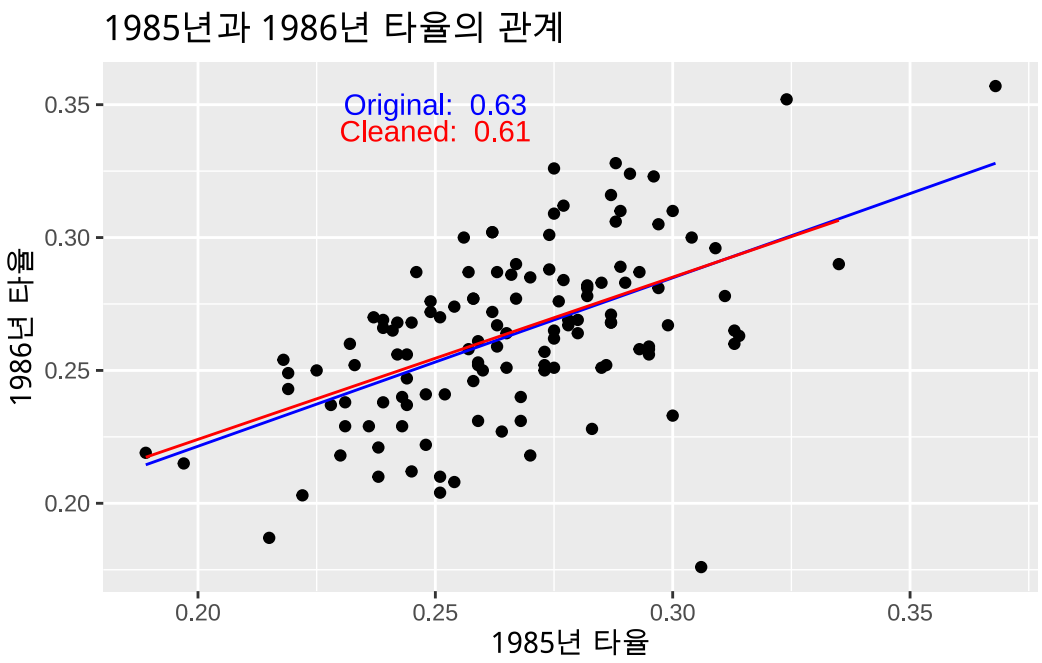
```
Call:
lm(formula = hit86 ~ hit85, data = MLB1_clean)
```

```
Residuals:
      Min        1Q    Median        3Q       Max
-0.051190 -0.016541 -0.000852  0.017538  0.056161

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.10198    0.02283   4.468 1.84e-05 ***
hit85        0.61039    0.08577   7.117 9.55e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.02385 on 117 degrees of freedom
Multiple R-squared:  0.3021,    Adjusted R-squared:  0.2961
F-statistic: 50.65 on 1 and 117 DF,  p-value: 9.554e-11
```

영향점과 이상점을 제거하기 전과 후의 회귀선을 그려서 비교해 보자.

- 영향점과 이상점을 제거하기 전의 기울기의 추정치는 0.633832 이다.
- 영향점과 이상점을 제거한 후의 기울기의 추정치는 0.6103926 이다.

```
p <- ggplot(MLB1, aes(x=hit85, y=hit86)) + geom_point() + labs(x = "1985년 타율", y = "1986년 타율
  labs(title="1985년과 1986년 타율의 관계") +
  geom_line(aes(y=mlb1.lm$fitted.values), color="blue") +
  geom_line(data = MLB1_clean, aes(y=mlb1_clean.lm$fitted.values), color="red") +
  #add label for two regression lines
  annotate("text", x = 0.25, y = 0.35, label = paste("Original: ", round(mlb1.lm$coef[2], 2)),
  annotate("text", x = 0.25, y = 0.34, label = paste("Cleaned: ", round(mlb1_clean.lm$coef[2],

p
```


1985년과 1986년 타율의 관계
Original: 0.63
Cleaned: 0.61
1986년 타율
0.35
0.30
0.25
0.20
0.20
0.25
0.30
0.35
1985년 타율

M. R-실습: 모형의 선택

M.1. AIC 와 BIC 계산

통계모형을 선택하는 척도로서 가능도함수이론에 근거한 AIC(Akaike information criteria, 6.5)와 베이지안 검정이론에 기초한 BIC(bayesian or schwartz information criteria, 6.4)가 있으며 정의는 모형의 선택에 대한 절에 나타니 있다.

AIC 와 BIC 는 함수 `AIC` 와 `BIC` 를 이용하면 구할 수 있다.

```
fit1 <- lm(price ~ ., data=houseprice)
fit1
```

```
Call:
lm(formula = price ~ ., data = houseprice)

Coefficients:
(Intercept)          tax       ground        floor         year
    1.21874      0.05195      0.01159      0.34941     -0.21894
```

```
AIC(fit1)
```

```
[1] 121.5611
```

```
BIC(fit1)
```

```
[1] 129.3361
```

위의 AIC 와 BIC 에서 로그 가능도 함수 $\ell(\hat{\boldsymbol{\theta}})$ 는 함수 `logLik` 로 구할 수 있다.

참고로 선형모형에 대한 AIC 와 BIC 는 6.6 에 의하여 다음과 같이 계산할 수 있다.

$$AIC = n\log(2\pi) + n + n\log\frac{SSE_p}{n} + 2(p+1)$$
$$BIC = n\log(2\pi) + n + n\log\frac{SSE_p}{n} + (\log n)(p+1)$$

이제 회귀모형을 적합한 결과과 6.6 를 이용하여 직접 AIC 와 BIC 를 계산해보자.

```
n <- dim(houseprice)[1]
p <- length(coef(fit1))
sse <- deviance(fit1)
c(n,p,sse)
```

```
[1] 27.00000  5.00000 91.44876
```

```
aic1 <- -2*logLik(fit1) + 2*(p+1)
bic1 <- -2*logLik(fit1) + log(n)*(p+1)
c(aic1, bic1)
```

```
[1] 121.5611 129.3361
```

```
aic2 <- n*log(2*pi) + n + n*log(deviance(fit1)/n)  + 2*(p+1)
bic2 <- n*log(2*pi) + n + n*log(deviance(fit1)/n)  + log(n)*(p+1)
c(aic2, bic2)
```

```
[1] 121.5611 129.3361
```

M.2. 예제 6.1

M.2.1. 변수선택의 통계량

패키지 `leaps` 에 수록된 `regsubsets()` 함수를 이용하면 가능한 모든 모형에 대한 중요한 선택 기준들이 계산된다.

- `rss`: 잔차 제곱합
- `rsq`: R^2
- `adjr2` : 수정된 R^2
- `cp` : 맬로우의 C_p
- `bic` : BIC

교재 예제에 나타난 자료 `houseprice`는 독립변수의 수가 4개이므로 가능한 회귀식의 개수가 $2^4 = 16$개이므로 이러한 방법을 적용할 수 있다. 하지만 독립변수의 수가 10개만 되면 가능한 회귀식의 수가 1024개나 되고 20개가 모형의 수가 100만개가 넘으므로 이들 모두에 대한 통계량을 계산하는 것은 실제로 쉽지않다.

이제 예제 자료 `houseprice`에 대하여 `regsubsets`함수를 사용하여 가능한 회귀식을 모두 적합하고 각 식에 대한 모형선택의 기준값을 계산해보자. `regsubsets`함수에서 `nbest=6`는 독립변수의 수가 같은 모형들 중에서 가장 좋은 6개의 모형만을 보여주라는 명령문이다.

모든 가능한 회귀식에 대한 통계량은 `summaryf`함수를 통하여 볼 수 있다.

```
houseprice.rgs <- regsubsets(price ~ . , data=houseprice, nbest=6)
summaryf(houseprice.rgs)
```

```
         tax ground floor year        rss        rsq      adjr2         cp
1  ( 1 )               *       182.66315 0.86271916 0.85722792  20.943616
1  ( 2 )   *                   215.97764 0.83768158 0.83118885  28.958146
1  ( 3 )         *             628.61671 0.52756188 0.50866435 128.227503
1  ( 4 )                     * 1202.47186 0.09627993 0.06013112 266.280912
2  ( 1 )   *           *        95.19550 0.92845564 0.92249361   1.901360
2  ( 2 )               *    *  153.80917 0.88440442 0.87477146  16.002160
2  ( 3 )         *     *       168.58785 0.87329747 0.86273892  19.557496
2  ( 4 )   *     *             191.93607 0.85575007 0.84372925  25.174420
2  ( 5 )   *                 *  214.64151 0.83868576 0.82524290  30.636710
2  ( 6 )         *           *  626.80968 0.52891996 0.48966329 129.792782
3  ( 1 )   *           *    *   92.31936 0.93061720 0.92156727   3.209442
3  ( 2 )   *     *     *        93.26197 0.92990878 0.92076645   3.436208
3  ( 3 )         *     *    *  150.10029 0.88719183 0.87247773  17.109908
3  ( 4 )   *     *           *  187.51305 0.85907420 0.84069258  26.110366
4  ( 1 )   *     *     *    *   91.44876 0.93127151 0.91877542   5.000000
                bic
1  ( 1 ) -47.022941
1  ( 2 ) -42.499601
1  ( 3 ) -13.654236
1  ( 4 )   3.858312
2  ( 1 ) -61.323304
2  ( 2 ) -48.369244
2  ( 3 ) -45.892146
2  ( 4 ) -42.390102
2  ( 5 ) -39.371316
2  ( 6 ) -10.436125
3  ( 1 ) -58.855794
3  ( 2 ) -58.581513
3  ( 3 ) -45.732450
3  ( 4 ) -39.723741
4  ( 1 ) -55.815783
```

위의 결과를 보면 두 개의 변수 `tax`와 `floor`가 포함된 회귀식이 다음과 같은 통계량으로 가장 좋은 모형으로 나타난다.

- `rss`: 잔차제곱합(residual sum of square), $SSE = 95.20$
- `rsq`: 결정계수(R_p^2), $R^2 = 0.92846$
- `adjr2`: 수정된 결정계수(R_{ap}^2), $R_a^2 = 0.92249$

- `cp`: 맬로우즈 C_p, $C_p = 1.901$
- `bic`: BIC(Bayesian Information Criteria), $BIC = -61.323$

패키지 `olsrr`에 있는 함수 `ols_step_all_possible()`을 이용하면 가능한 모든 회귀모형에 대한 통계량을 구하고 여러 가지 통계량에 대한 그림을 `plot()` 함수를 이용하여 쉽게 그릴 수 있다.

```
fit1 <- lm(price ~ ., data=houseprice)
houseprice.rgs2 <- ols_step_all_possible(fit1)
houseprice.rgs2
```

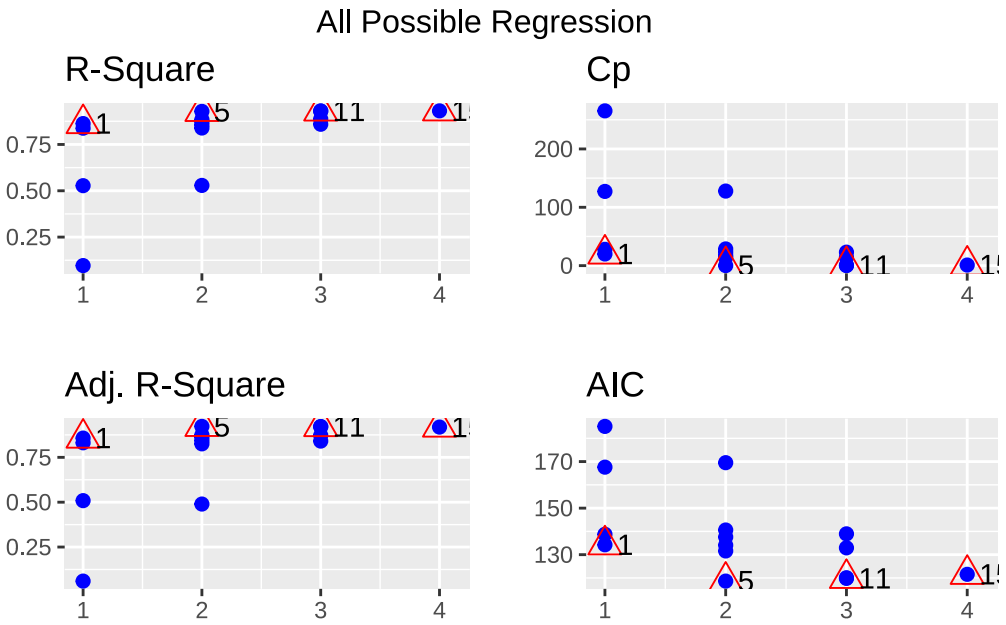
```
   Index N            Predictors   R-Square Adj. R-Square Mallow's Cp
3      1 1                 floor 0.86271916    0.85722792   20.943616
1      2 1                   tax 0.83768158    0.83118885   28.958146
2      3 1                ground 0.52756188    0.50866435  128.227503
4      4 1                  year 0.09627993    0.06013112  266.280912
6      5 2             tax floor 0.92845564    0.92249361    1.901360
10     6 2            floor year 0.88440442    0.87477146   16.002160
8      7 2          ground floor 0.87329747    0.86273892   19.557496
5      8 2            tax ground 0.85575007    0.84372925   25.174420
7      9 2              tax year 0.83868576    0.82524290   30.636710
9     10 2           ground year 0.52891996    0.48966329  129.792782
13    11 3        tax floor year 0.93061720    0.92156727    3.209442
11    12 3      tax ground floor 0.92990878    0.92076645    3.436208
14    13 3     ground floor year 0.88719183    0.87247773   17.109908
12    14 3       tax ground year 0.85907420    0.84069258   26.110366
15    15 4 tax ground floor year 0.93127151    0.91877542    5.000000
```

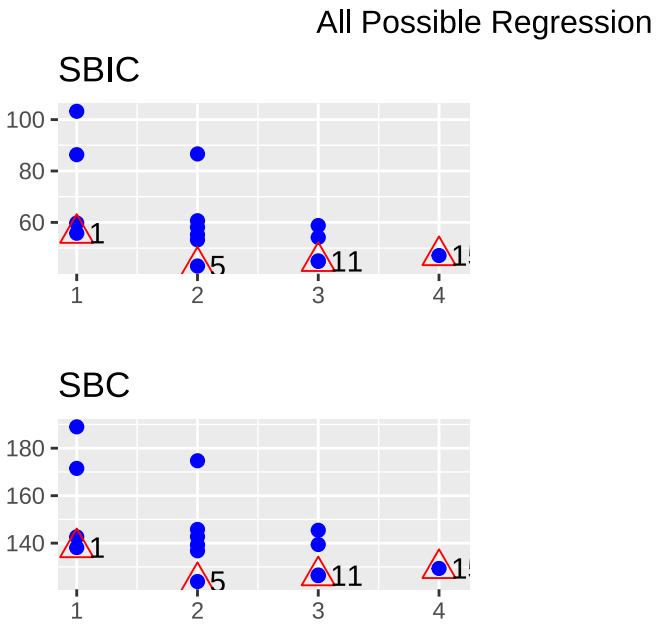
```
houseprice.rgs2$result %>% dplyr::select(mindex, n, predictors, rsquare, adjr, cp, aic, sbc) %
```

```
   mindex n            predictors    rsquare       adjr         cp      aic
1       5 2             tax floor 0.92845564 0.92249361   1.901360 118.6453
2      11 3        tax floor year 0.93061720 0.92156727   3.209442 119.8169
3      12 3      tax ground floor 0.92990878 0.92076645   3.436208 120.0912
4      15 4 tax ground floor year 0.93127151 0.91877542   5.000000 121.5611
5       6 2            floor year 0.88440442 0.87477146  16.002160 131.5993
6       1 1                 floor 0.86271916 0.85722792  20.943616 134.2415
7       7 2          ground floor 0.87329747 0.86273892  19.557496 134.0764
8      13 3     ground floor year 0.88719183 0.87247773  17.109908 132.9403
9       2 1                   tax 0.83768158 0.83118885  28.958146 138.7648
10      8 2            tax ground 0.85575007 0.84372925  25.174420 137.5785
11     14 3       tax ground year 0.85907420 0.84069258  26.110366 138.9490
12      9 2              tax year 0.83868576 0.82524290  30.636710 140.5973
```

```
13      3 1                ground 0.52756188 0.50866435 128.227503 167.6102
14     10 2           ground year 0.52891996 0.48966329 129.792782 169.5324
15      4 1                  year 0.09627993 0.06013112 266.280912 185.1227
        sbc
1  123.8286
2  126.2961
3  126.5704
4  129.3361
5  136.7827
6  138.1290
7  139.2598
8  139.4195
9  142.6523
10 142.7618
11 145.4282
12 145.7806
13 171.4977
14 174.7158
15 189.0102
```

```
plot(houseprice.rgs2)
```





PRESS 잔차는 함수 `press()` 로 구할 수 있다.

$$\text{PRESS}_p = \sum_{i=1}^{n} (y_i - \hat{y}_{i(i)})^2$$

또한 교과서 264 페이지에 나온 교차확인(cross validation)에 의거한 R^2_{pred} 은 다음과 같이 계산되며 `press()` 함수에 의해 주어진다.

$$R^2_{pred} = 1 - \frac{\text{PRESS}_p}{\text{SST}}$$

```
press(houseprice.rgs)
```

```
         tax ground floor year     PRESS pred.r.squared
1  ( 1 )                 *       220.8853      0.8339932
1  ( 2 )   *                     316.0868      0.7624443
1  ( 3 )         *               838.8121      0.3695891
1  ( 4 )                       * 1430.6084      0.0000000
2  ( 1 )   *             *       144.3991      0.8914766
2  ( 2 )                 *     * 207.6594      0.8439331
2  ( 3 )         *       *       233.7615      0.8243161
2  ( 4 )   *     *               358.2145      0.7307832
2  ( 5 )   *                   * 341.4294      0.7433981
2  ( 6 )         *             * 911.3553      0.3150691
3  ( 1 )   *             *     * 153.5941      0.8845661
3  ( 2 )   *     *       *       159.1752      0.8803716
3  ( 3 )         *       *     * 233.2240      0.8247200
```

```
3  ( 4 )   *      *          *   380.1744      0.7142792
4  ( 1 )   *      *     *    *   170.4810      0.8718747
```

M.2.2. 이상치를 제거한 경우

교과서에서 분석하였듯이 9,10,27 번째 자료가 이상점 또는 영향점일 가능성이 높으므로 이를 제거하고 모형의 선택 기준을 다시 계산해 보자.

```
housepriceEX <- houseprice[-c(9,10,27),]
housepriceEX.lm <- lm(price ~ . , data=housepriceEX)
summary(housepriceEX.lm)
```

```
Call:
lm(formula = price ~ ., data = housepriceEX)

Residuals:
    Min      1Q  Median      3Q     Max
-2.4130 -1.0459 -0.2558  0.8593  2.8197

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.32538    2.14531   2.948  0.00825 **
tax          0.06589    0.01874   3.516  0.00231 **
ground       0.00974    0.02168   0.449  0.65835
floor        0.08578    0.09057   0.947  0.35547
year        -0.11142    0.26781  -0.416  0.68205
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.543 on 19 degrees of freedom
Multiple R-squared:  0.7819,    Adjusted R-squared:  0.7359
F-statistic: 17.02 on 4 and 19 DF,  p-value: 4.404e-06
```

이상치를 제거하면 tax 만 포함된 모형이 최적모형으로 나타난다. 이상치와 영향점이 모형의 선택에도 영향을 주는 점에 꼭 유의하자.

```
housepriceEX.rgs <- regsubsets(price ~ . , data=housepriceEX, nbest=6)
summaryf(housepriceEX.rgs)
```

```
        tax ground floor year       rss       rsq     adjr2        cp
```

```
1  ( 1 )   *                    48.17790 0.7675502 0.7569843  0.2458362
1  ( 2 )                *      102.73566 0.5043188 0.4817878 23.1726808
1  ( 3 )          *            125.87070 0.3926964 0.3650917 32.8947363
1  ( 4 )                      * 174.34503 0.1588164 0.1205808 53.2651405
2  ( 1 )   *            *       46.18667 0.7771576 0.7559345  1.4090560
2  ( 2 )   *      *             47.39011 0.7713512 0.7495751  1.9147822
2  ( 3 )   *                   *  48.11577 0.7678500 0.7457405  2.2197243
2  ( 4 )                *     *  83.74651 0.5959380 0.5574559 17.1928590
2  ( 5 )          *     *       85.27968 0.5885408 0.5493542 17.8371405
2  ( 6 )          *           * 116.96847 0.4356480 0.3819002 31.1537464
3  ( 1 )   *      *     *       45.62514 0.7798669 0.7468469  3.1730851
3  ( 2 )   *            *     *  45.69348 0.7795371 0.7464677  3.2018056
3  ( 3 )   *      *           *  47.34784 0.7715551 0.7372884  3.8970174
3  ( 4 )          *     *     *  74.63053 0.6399210 0.5859092 15.3620431
4  ( 1 )   *      *     *     *  45.21326 0.7818541 0.7359287  5.0000000
                bic
1  ( 1 ) -28.661838
1  ( 2 ) -10.487628
1  ( 3 )  -5.613326
1  ( 4 )   2.205420
2  ( 1 ) -26.496810
2  ( 2 ) -25.879469
2  ( 3 ) -25.514758
2  ( 4 ) -12.214327
2  ( 5 ) -11.778929
2  ( 6 )  -4.195690
3  ( 1 ) -23.612331
3  ( 2 ) -23.576407
3  ( 3 ) -22.722834
3  ( 4 ) -11.802150
4  ( 1 ) -20.651921
```

```
housepriceEX.rgs2 <- ols_step_all_possible(lm(price ~ ., data=housepriceEX))
housepriceEX.rgs2$result %>% as.data.frame() %>% dplyr::select(mindex, n, predictors, rsquare,
```

```
   mindex n            predictors   rsquare      adjr         cp       aic
1       1 1                   tax 0.7675502 0.7569843  0.2458362  90.83337
2       5 2             tax floor 0.7771576 0.7559345  1.4090560  91.82034
3       6 2            tax ground 0.7713512 0.7495751  1.9147822  92.43768
4       7 2              tax year 0.7678500 0.7457405  2.2197243  92.80240
5      11 3      tax ground floor 0.7798669 0.7468469  3.1730851  93.52677
6      12 3        tax floor year 0.7795371 0.7464677  3.2018056  93.56269
```

```
7      13 3       tax ground year 0.7715551 0.7372884  3.8970174  94.41627
8      15 4 tax ground floor year 0.7818541 0.7359287  5.0000000  95.30913
9       8 2            floor year 0.5959380 0.5574559 17.1928590 106.10283
10     14 3     ground floor year 0.6399210 0.5859092 15.3620431 105.33695
11      9 2          ground floor 0.5885408 0.5493542 17.8371405 106.53823
12      2 1                floor 0.5043188 0.4817878 23.1726808 109.00758
13      3 1               ground 0.3926964 0.3650917 32.8947363 113.88188
14     10 2           ground year 0.4356480 0.3819002 31.1537464 114.12146
15      4 1                 year 0.1588164 0.1205808 53.2651405 121.70063
         sbc
1   94.36753
2   96.53256
3   97.14990
4   97.51461
5   99.41704
6   99.45296
7  100.30654
8  102.37745
9  110.81504
10 111.22722
11 111.25044
12 112.54174
13 117.41604
14 118.83368
15 125.23479
```

```
press(housepriceEX.rgs)
```

```
         tax ground floor year     PRESS pred.r.squared
1   ( 1 )   *                    56.10399     0.72930829
1   ( 2 )                *      117.36430     0.43373824
1   ( 3 )          *            157.83378     0.23848023
1   ( 4 )                     * 202.40162     0.02344834
2   ( 1 )   *            *       57.66354     0.72178375
2   ( 2 )   *      *             64.24819     0.69001399
2   ( 3 )   *                 *  63.42708     0.69397570
2   ( 4 )                *    * 108.09369     0.47846724
2   ( 5 )          *     *      114.27185     0.44865875
2   ( 6 )          *          * 160.37982     0.22619605
3   ( 1 )   *      *     *       68.98320     0.66716837
3   ( 2 )   *            *    *  65.44694     0.68423019
3   ( 3 )   *      *          *  72.26745     0.65132247
```

```
3  ( 4 )          *     *     * 109.76504     0.47040332
4  ( 1 )    *     *     *     *  79.24660     0.61764930
```

M.3. 변수선택 방법

회귀식에서 변수를을 선택할 때 가장 큰 모형을 적합시키면 각 변수에 대한 중요도를 각 회귀계수에 대한 가설 검정 $H_o : \beta_i = 0$에 대한 t-통계량을 보고 판단할 수 있다.

다시 모든 자료를 고려한 houseprice에 대하여 가장 큰 모형을 적합시킨 후 각 변수에 대한 t-검정을 실시해보자.

```
summary(fit1)
```

```
Call:
lm(formula = price ~ ., data = houseprice)

Residuals:
    Min      1Q  Median      3Q     Max
-3.4891 -1.3574  0.1337  1.0686  3.4938

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.21874    2.04661   0.595  0.55759
tax          0.05195    0.01383   3.756  0.00109 **
ground       0.01159    0.02534   0.458  0.65169
floor        0.34941    0.07268   4.807 8.41e-05 ***
year        -0.21894    0.33149  -0.660  0.51582
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.039 on 22 degrees of freedom
Multiple R-squared:  0.9313,    Adjusted R-squared:  0.9188
F-statistic: 74.53 on 4 and 22 DF,  p-value: 1.817e-12
```

위는 4개의 독립변수를 포함한 가장 큰 모형에 대하여 각 계수의 유의성 검정에 대한 결과이다. 변수 floor가 가장 유의한 변수이고 다음으로 tax가 유의함을 알 수 있다. 또한 2개의 변수 year와 ground는 유의하지 않음을 알 수 있다.

이렇게 독립변수들은 반응변수를 설명하는 정도가 다르므로 모든 가능한 회귀식을 적합하여 모형을 선택하는 것보다 변수의 중요도를 고려하여 변수들을 유의한 정도에 따라 차례로 모형에 포함시키거나 제거하는 절차가 더 효율적이다. 이렇게 가장 단순한 모형(평균모형)에서 시작하여 설명력이 높은 변수들을 순차적으로 포함시키거나

(forward selection; 전진선택) 가장 큰 모형(full model)에서 시작하여 설명력이 낮은 변수들을 차례로 제거하는 방법(backward elimination; 후방제거)을 단계별 회귀(stepwise regression)이라고 한다.

단계별 회귀는 다음과 같은 세 가지 방법이 있다.

- 전진선택 (forward selection)
- 후방제거 (backward elimination)
- 단계별 선택 (stepwise selection)

단계별 선택은 전진선택과 후방제거를 결합한 형태로서 새로운 변수가 추가되는 경우마다(전진선택) 제거할 변수를 있는지 판단하여 유의하지 않다면 제거하는(후방제거) 방법이다.

M.3.1. 단계별 선택의 적용

단계별 선택에서 전진선택과 후방제거는 `add1()`와 `drop1()`함수를 사용한다.

1. 평균모형

$$y = \beta_0 + \epsilon$$

```
model0 <- lm(price~1, houseprice)
summary(model0)
```

```
Call:
lm(formula = price ~ 1, data = houseprice)

Residuals:
   Min     1Q Median     3Q    Max
-6.300 -4.275 -0.800  1.125 23.200

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   19.250      1.377   13.98 1.32e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.154 on 26 degrees of freedom
```

2. 첫번째 변수의 추가

이제 가장 설명력있는 변수를 추가하는데 다음과 같은 2개의 모형을 비교하는 부분 F 검정을 이용하여 가장 유의한 변수를 추가한다.

$$H_0 : y = \beta_0 + \epsilon \quad \text{vs} \quad H_1 : y = \beta_0 + \beta_1 x_i + \epsilon$$

```
add1(model0, scope=~tax+ ground+ floor+ year, test="F")
```

```
Single term additions

Model:
price ~ 1
       Df Sum of Sq     RSS     AIC  F value    Pr(>F)
<none>              1330.58 107.233
tax     1   1114.60  215.98  60.142 129.0183 2.310e-11 ***
ground  1    701.96  628.62  88.987  27.9170 1.792e-05 ***
floor   1   1147.92  182.66  55.619 157.1084 2.807e-12 ***
year    1    128.11 1202.47 106.500   2.6634    0.1152
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

위의 결과에서 가장 유의한 변수가 floor이므로 회귀식에 추가한다. 변수를 추가하는 경우는 고려하는 변수를 포함한 모형과 포함하지 않는 모형이 가장 유의한 차이를 보이는 변수를 선택한다.

```
model1 <- update(model0, . ~ . + floor)
summary(model1)
```

```
Call:
lm(formula = price ~ floor, data = houseprice)

Residuals:
    Min      1Q  Median      3Q     Max
-4.6682 -2.1916 -0.1459  2.0517  4.9644

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.25296    1.52716    0.82     0.42
floor        0.59511    0.04748   12.53 2.81e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 2.703 on 25 degrees of freedom
Multiple R-squared:  0.8627,    Adjusted R-squared:  0.8572
F-statistic: 157.1 on 1 and 25 DF,  p-value: 2.807e-12
```

위에서 update() 함수는 앞에서 적합한 모형 model0에 변수 floor를 추가한다.

3. 두 번째 변수의 추가

```
add1(model1, scope=~tax+ ground+ floor+ year, test="F")
```

```
Single term additions

Model:
price ~ floor
       Df Sum of Sq     RSS    AIC F value    Pr(>F)
<none>              182.663 55.619
tax     1    87.468  95.195 40.023 22.0517 8.997e-05 ***
ground  1    14.075 168.588 55.454  2.0037   0.16976
year    1    28.854 153.809 52.977  4.5023   0.04437 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
model2 <- update(model1, . ~ . + tax)
summary(model2)
```

```
Call:
lm(formula = price ~ floor + tax, data = houseprice)

Residuals:
   Min     1Q Median     3Q    Max
-3.009 -1.658  0.048  1.220  3.548

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.39497    1.13994   0.346    0.732
floor        0.34835    0.06313   5.518 1.13e-05 ***
tax          0.05742    0.01223   4.696 9.00e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.992 on 24 degrees of freedom
```

```
Multiple R-squared:  0.9285,    Adjusted R-squared:  0.9225
F-statistic: 155.7 on 2 and 24 DF,  p-value: 1.798e-14
```

위의 결과로부터 변수 tax가 가장 유의한 변수임을 알 수 있고 이를 모형에 추가한다.

4. 변수의 제거

이제 독립변수가 두 개가(floor와 tax) 모형에 포함되었고 가장 최근에 포함된 변수 tax를 제외한 나머지 변수를 제거할 수 있는지 감정한다. 후방제거하는 경우는 제거할 변수가 포함된 모형과 포함하지 않는 모형의 설명력에 유의한 차이가 없어야 한다. 즉 F-검정이 유의하니 않으면 제거한다.

```
drop1(model2, test="F")
```

```
Single term deletions

Model:
price ~ floor + tax
       Df Sum of Sq     RSS    AIC F value    Pr(>F)
<none>               95.195 40.023
floor   1   120.782 215.978 60.142  30.451 1.126e-05 ***
tax     1    87.468 182.663 55.619  22.052 8.997e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

위의 함수 drop1()의 결과에서 모든 변수에 대한 F-통계량이 유의하므로 변수를 제거하지 않는다.

5. 세 번째 변수의 추가

독립변수가 두 개가(floor와 tax) 모형에 새로운 변수를 추가하는 검정을 실시해 보자

```
add1(model2, scope=~tax+ ground+ floor+ year, test="F")
```

```
Single term additions

Model:
price ~ floor + tax
       Df Sum of Sq    RSS    AIC F value Pr(>F)
<none>              95.195 40.023
ground  1    1.9335 93.262 41.469  0.4768 0.4968
year    1    2.8761 92.319 41.194  0.7165 0.4060
```

위의 결과에서 유의한 변수가 없으므로 더 이상 변수를 추가하지 않는다.

6. 최종 모형

더 이상 추가할 변수와 제거할 변수가 없으면 단계별 선택을 중단한다. 따라서 단계별 선택법에 의한 최종 모형은 floor와 tax, 두 개의 독립변수를 포함하는 모형이다.

이러한 단계별 회귀의 결과는 모든 가능한 회귀에 의한 방법과 모형이 일치한다. 독립변수의 수가 많는 경우 이러한 단계별 회귀 방법은 유용하게 사용된다.

M.3.2. 함수 step()

위에서 논의한 세 종류의 변수선택법은 함수 step()을 이용하여 한 번에 결과를 얻을 수 있다.

Caution

주의할 점은 함수 step()은 변수의 선택 기준이 F-검정이 아니 AIC 를 이용한다는 점이다.

- 전진 선택

```
model0 <- lm(price~1, houseprice)
step(model0,  scope=~tax+ ground+ floor+ year, direction="forward")
```

```
Start:  AIC=107.23
price ~ 1

         Df Sum of Sq     RSS     AIC
+ floor   1   1147.92  182.66  55.619
+ tax     1   1114.60  215.98  60.142
+ ground  1    701.96  628.62  88.987
+ year    1    128.11 1202.47 106.500
<none>                1330.58 107.233

Step:  AIC=55.62
price ~ floor

         Df Sum of Sq     RSS    AIC
+ tax     1    87.468  95.195 40.023
+ year    1    28.854 153.809 52.977
+ ground  1    14.075 168.588 55.454
<none>                182.663 55.619

Step:  AIC=40.02
price ~ floor + tax

         Df Sum of Sq    RSS    AIC
```

```
<none>                 95.195 40.023
+ year    1    2.8761 92.319 41.194
+ ground  1    1.9335 93.262 41.469


Call:
lm(formula = price ~ floor + tax, data = houseprice)

Coefficients:
(Intercept)        floor          tax
    0.39497      0.34835      0.05742
```

- 후방 제거

```
fit <- lm(price~tax+ ground+ floor+ year, houseprice)
step(fit, direction="backward")
```

```
Start:  AIC=42.94
price ~ tax + ground + floor + year

         Df Sum of Sq     RSS    AIC
- ground  1     0.871  92.319 41.194
- year    1     1.813  93.262 41.469
<none>                 91.449 42.938
- tax     1    58.652 150.100 54.318
- floor   1    96.064 187.513 60.326

Step:  AIC=41.19
price ~ tax + floor + year

        Df Sum of Sq     RSS    AIC
- year   1     2.876  95.195 40.023
<none>                92.319 41.194
- tax    1    61.490 153.809 52.977
- floor  1   122.322 214.642 61.975

Step:  AIC=40.02
price ~ tax + floor

        Df Sum of Sq     RSS    AIC
<none>                95.195 40.023
- tax    1    87.468 182.663 55.619
- floor  1   120.782 215.978 60.142
```

```
Call:
lm(formula = price ~ tax + floor, data = houseprice)

Coefficients:
(Intercept)          tax        floor
    0.39497      0.05742      0.34835
```

- 단계별 선택

```
model0 <- lm(price~1, houseprice)
step(model0,  scope=~tax+ ground+ floor+ year, direction="both")
```

```
Start:  AIC=107.23
price ~ 1

         Df Sum of Sq     RSS     AIC
+ floor   1   1147.92  182.66  55.619
+ tax     1   1114.60  215.98  60.142
+ ground  1    701.96  628.62  88.987
+ year    1    128.11 1202.47 106.500
<none>                1330.58 107.233

Step:  AIC=55.62
price ~ floor

         Df Sum of Sq     RSS     AIC
+ tax     1     87.47   95.20  40.023
+ year    1     28.85  153.81  52.977
+ ground  1     14.08  168.59  55.454
<none>                182.66  55.619
- floor   1   1147.92 1330.58 107.233

Step:  AIC=40.02
price ~ floor + tax

         Df Sum of Sq     RSS    AIC
<none>                95.195 40.023
+ year    1     2.876  92.319 41.194
+ ground  1     1.934  93.262 41.469
- tax     1    87.468 182.663 55.619
- floor   1   120.782 215.978 60.142
```

```
Call:
lm(formula = price ~ floor + tax, data = houseprice)

Coefficients:
(Intercept)        floor          tax
    0.39497      0.34835      0.05742
```

M.3.3. 패키지 olsrr 의 이용

패키지 olsrr 의 함수 ols_step_both_p() 이용하여 F-검정을 이용한 단계별 선택을 할 수 있다.

> 변수선택법에서의 유의수준
>
> 함수 ols_step_both_p() 이용하여 F-검정에서는 유의수준을 지정해주지 않으면 유의 수준을 0.3 으로 사용한다.
> 참고로 SAS 의 모형 선택에서 자동적으로 사용되는 유의수준은 전진선택에서는 0.5, 후방제거에서는 0.1 이다.
>
> ```
> ols_step_forward_p(model, penter = 0.3,
>   progress = FALSE, details = FALSE, ...)
> ols_step_backward_p(model, prem = 0.3,
>   progress = FALSE, details = FALSE, ...)
>
> Forward Selection (FORWARD)
>
> The p-values for these F statistics are compared to the SLENTRY= value that is specified in
>
> Backward Elimination (BACKWARD)
>
> F statistics significant at the SLSTAY= level specified in the MODEL statement (or at the
> ```

```
res <- ols_step_both_p(fit1, details = TRUE)

Stepwise Selection Method
-------------------------

Candidate Terms:

1. tax
2. ground
3. floor
4. year
```

```
Step    => 0
Model   => price ~ 1
R2      => 0

Initiating stepwise selection...

Step      => 1
Selected  => floor
Model     => price ~ floor
R2        => 0.863

Step      => 2
Selected  => tax
Model     => price ~ floor + tax
R2        => 0.928


No more variables to be added or removed.
```

```
res
```

```
                           Stepwise Summary
--------------------------------------------------------------------------
Step    Variable        AIC        SBC       SBIC        R2       Adj. R2
--------------------------------------------------------------------------
 0      Base Model    185.856    188.448    105.725    0.00000    0.00000
 1      floor (+)     134.241    138.129     55.779    0.86272    0.85723
 2      tax (+)       118.645    123.829     43.032    0.92846    0.92249
--------------------------------------------------------------------------

Final Model Output
------------------

                        Model Summary
---------------------------------------------------------------
R                       0.964       RMSE                 1.878
R-Squared               0.928       MSE                  3.526
Adj. R-Squared          0.922       Coef. Var           10.346
Pred R-Squared          0.891       AIC                118.645
MAE                     1.572       SBC                123.829
---------------------------------------------------------------
```

```
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error
 AIC: Akaike Information Criteria
 SBC: Schwarz Bayesian Criteria
```

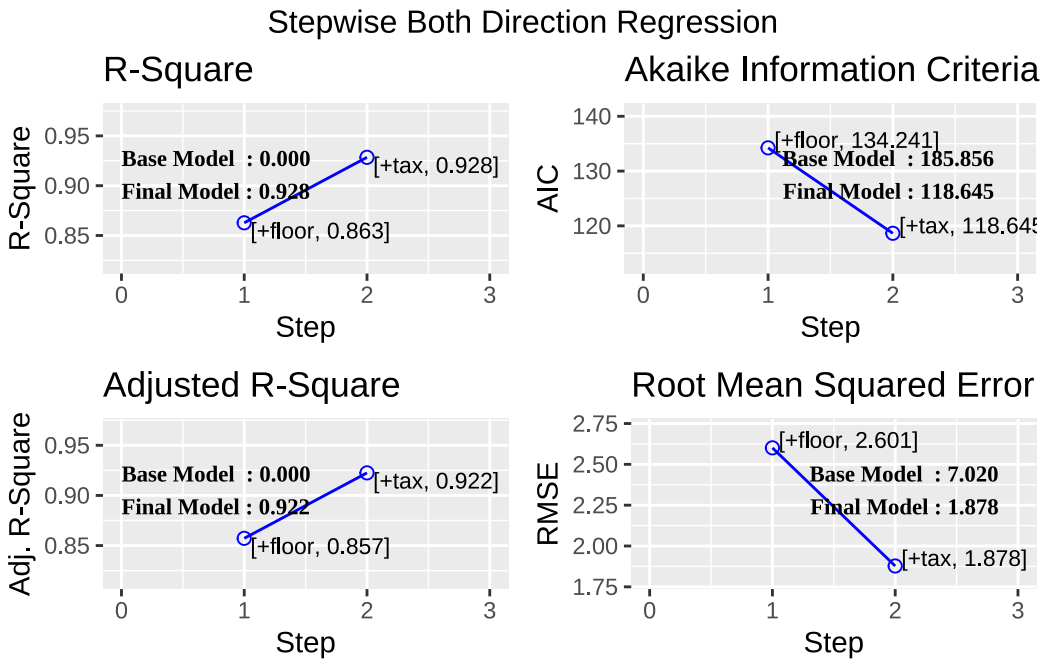
ANOVA

	Sum of Squares	DF	Mean Square	F	Sig.
Regression	1235.385	2	617.692	155.728	0.0000
Residual	95.195	24	3.966		
Total	1330.580	26			

Parameter Estimates

model	Beta	Std. Error	Std. Beta	t	Sig	lower	upper
(Intercept)	0.395	1.140		0.346	0.732	-1.958	2.748
floor	0.348	0.063	0.544	5.518	0.000	0.218	0.479
tax	0.057	0.012	0.463	4.696	0.000	0.032	0.083

```
plot(res)
```



함수 ols_step_both_aic() 는 함수 step() 과 동일하게 AIC 를 이용하여 변수선택을 한다.

```
ols_step_forward_aic(fit1)
```

```
                            Stepwise Summary
-------------------------------------------------------------------------
Step    Variable        AIC        SBC       SBIC        R2       Adj. R2
-------------------------------------------------------------------------
 0      Base Model    185.856    188.448    105.725    0.00000    0.00000
 1      floor         134.241    138.129     55.779    0.86272    0.85723
 2      tax           118.645    123.829     43.032    0.92846    0.92249
-------------------------------------------------------------------------

Final Model Output
------------------

                         Model Summary
---------------------------------------------------------------
R                       0.964       RMSE                 1.878
R-Squared               0.928       MSE                  3.526
Adj. R-Squared          0.922       Coef. Var           10.346
Pred R-Squared          0.891       AIC                118.645
MAE                     1.572       SBC                123.829
---------------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error
 AIC: Akaike Information Criteria
 SBC: Schwarz Bayesian Criteria

                                ANOVA
-------------------------------------------------------------------------
                Sum of
               Squares        DF    Mean Square       F          Sig.
-------------------------------------------------------------------------
Regression    1235.385         2        617.692    155.728    0.0000
Residual        95.195        24          3.966
Total         1330.580        26
-------------------------------------------------------------------------

                            Parameter Estimates
```

```
-------------------------------------------------------------------------------
      model     Beta     Std. Error    Std. Beta      t       Sig      lower    upper
-------------------------------------------------------------------------------
(Intercept)    0.395         1.140                  0.346    0.732    -1.958    2.748
      floor    0.348         0.063        0.544     5.518    0.000     0.218    0.479
        tax    0.057         0.012        0.463     4.696    0.000     0.032    0.083
-------------------------------------------------------------------------------
```

M.4. 연습문제 6.10

자료 cars93은 1993년 미국에서 판매된 93가지 종류의 저동차에 대한 자료이다. 변수 MPG.highway를 반응변수로 하고 EngineSize, Weight, Price, Width, Length, Horsepower, Wheelbae 7개의 변수를 고려하여 가장 적합한 모형을 선택해보자.

```
dat93 <- Cars93[c("MPG.highway","EngineSize", "Weight", "Price", "Width", "Length", "Horsepo
head(dat93)
```

```
  MPG.highway EngineSize Weight Price Width Length Horsepower Wheelbase
1          31        1.8   2705  15.9    68    177        140       102
2          25        3.2   3560  33.9    71    195        200       115
3          26        2.8   3375  29.1    67    180        172       102
4          26        2.8   3405  37.7    70    193        172       106
5          30        3.5   3640  30.0    69    186        208       109
6          31        2.2   2880  15.7    69    189        110       105
```

M.4.1. 완전모형

```
carfit <- lm(MPG.highway~ . , data=dat93)
summary(carfit)
```

```

Call:
lm(formula = MPG.highway ~ ., data = dat93)

Residuals:
    Min      1Q  Median      3Q     Max 
-6.3952 -1.9678 -0.1942  1.4516 10.1313 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 21.119770  12.809426   1.649   0.1029
EngineSize   0.483792   0.698257   0.693   0.4903
```

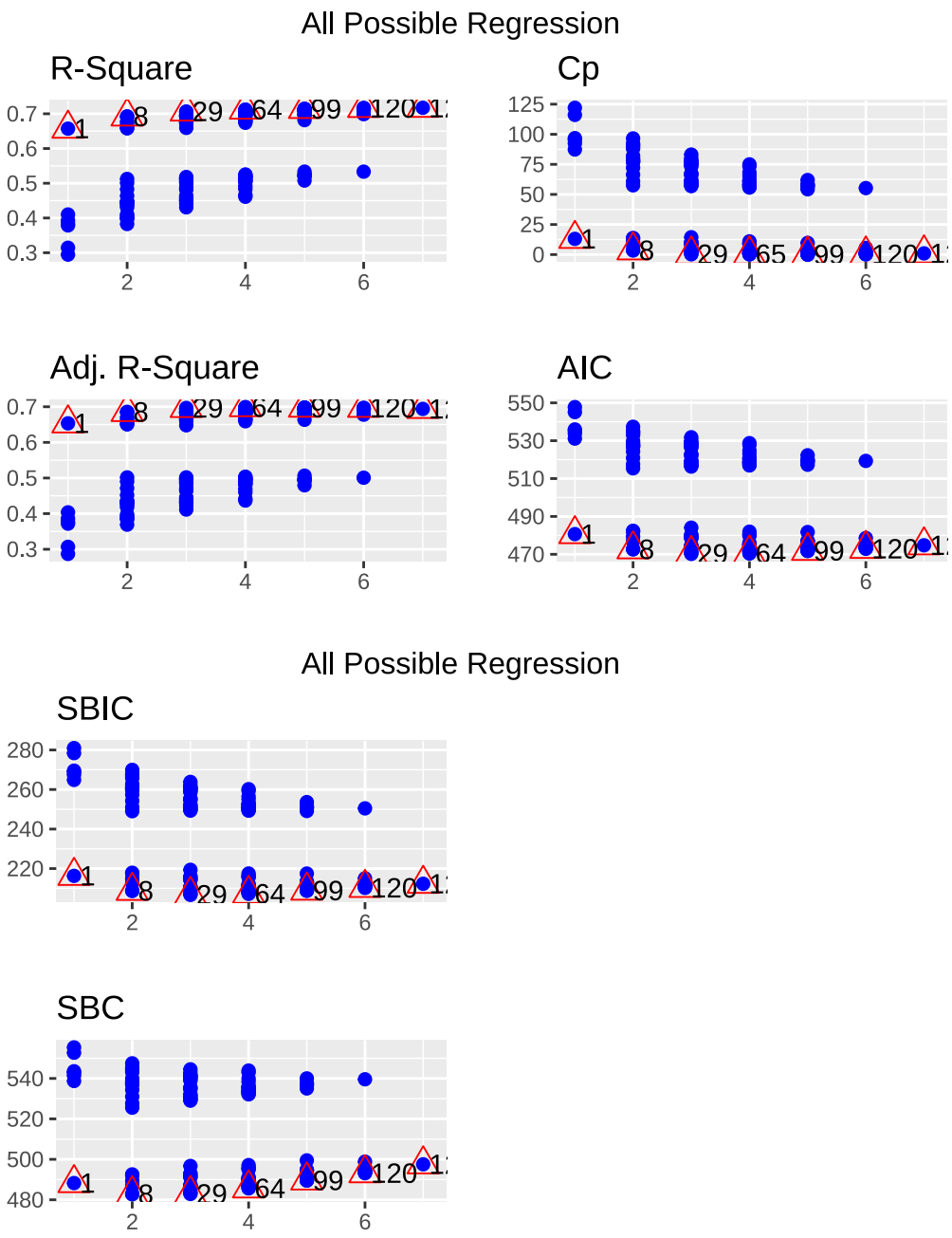
```
Weight      -0.012833   0.001726  -7.434 7.69e-11 ***
Price       -0.047157   0.059498  -0.793   0.4302
Width        0.079770   0.228192   0.350   0.7275
Length       0.058766   0.043322   1.356   0.1785
Horsepower   0.013296   0.013443   0.989   0.3254
Wheelbase    0.277231   0.118027   2.349   0.0212 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.95 on 85 degrees of freedom
Multiple R-squared:  0.7172,    Adjusted R-squared:  0.694
F-statistic:  30.8 on 7 and 85 DF,  p-value: < 2.2e-16
```

M.4.2. 모든 가능한 회귀

```
carfit.rgs2 <- ols_step_all_possible(carfit)
carfit.rgs2$result %>% dplyr::select(mindex, n, predictors, rsquare, adjr, cp, aic, sbc) %>%
    dplyr::arrange(sbc) %>%
    head(10)
```

```
   mindex n                         predictors   rsquare      adjr       cp
1       8 2                     Weight Length 0.6921969 0.6853568 5.527269
2       9 2                  Weight Wheelbase 0.6920227 0.6851788 5.579611
3      29 3            Weight Length Wheelbase 0.7065024 0.6966092 3.226956
4      30 3        EngineSize Weight Wheelbase 0.7045887 0.6946310 3.802229
5      31 3             Weight Width Wheelbase 0.7039881 0.6940102 3.982755
6      64 4 EngineSize Weight Length Wheelbase 0.7120588 0.6989706 3.556655
7      32 3                Weight Width Length 0.6974384 0.6872397 5.951620
8      65 4      Weight Width Length Wheelbase 0.7113915 0.6982729 3.757264
9      33 3        Weight Horsepower Wheelbase 0.6965448 0.6863160 6.220248
10     34 3           EngineSize Weight Length 0.6950183 0.6847380 6.679132
         aic      sbc
1  472.6393 482.7697
2  472.6919 482.8223
3  470.2133 482.8763
4  470.8178 483.4808
5  471.0066 483.6696
6  470.4358 485.6314
7  473.0419 485.7049
8  470.6511 485.8467
9  473.3162 485.9792
10 473.7829 486.4459
```

```
plot(carfit.rgs2)
```



M.4.3. 전진선택법

```
model0 <- lm(MPG.highway~1, data=dat93)
step(model0,  scope=~EngineSize+ Weight+ Price+ Width+ Length+ Horsepower+ Wheelbase, direct
```

```
Start:  AIC=312.3
MPG.highway ~ 1
```

```
             Df Sum of Sq     RSS    AIC
+ Weight      1   1718.70  896.62 214.74
+ Width       1   1072.43 1542.88 265.22
+ EngineSize  1   1027.48 1587.83 267.89
+ Horsepower  1   1002.23 1613.08 269.36
+ Wheelbase   1    990.41 1624.90 270.04
+ Price       1    822.16 1793.16 279.20
+ Length      1    770.83 1844.48 281.82
<none>                    2615.31 312.30

Step:  AIC=214.74
MPG.highway ~ Weight

             Df Sum of Sq    RSS    AIC
+ Length      1    91.615 805.00 206.72
+ Wheelbase   1    91.160 805.46 206.77
+ Width       1    53.010 843.61 211.07
+ EngineSize  1    31.068 865.55 213.46
<none>                    896.62 214.74
+ Price       1     5.845 890.77 216.13
+ Horsepower  1     2.334 894.28 216.50

Step:  AIC=206.72
MPG.highway ~ Weight + Length

             Df Sum of Sq    RSS    AIC
+ Wheelbase   1    37.413 767.59 204.29
<none>                    805.00 206.72
+ Width       1    13.708 791.29 207.12
+ EngineSize  1     7.379 797.62 207.86
+ Price       1     4.135 800.87 208.24
+ Horsepower  1     0.207 804.79 208.69

Step:  AIC=204.29
MPG.highway ~ Weight + Length + Wheelbase

             Df Sum of Sq    RSS    AIC
<none>                    767.59 204.29
+ EngineSize  1   14.5319 753.06 204.51
+ Width       1   12.7865 754.80 204.73
+ Horsepower  1    7.7950 759.79 205.34
+ Price       1    1.0351 766.55 206.16
```

```
Call:
lm(formula = MPG.highway ~ Weight + Length + Wheelbase, data = dat93)

Coefficients:
(Intercept)       Weight       Length    Wheelbase
   26.31687     -0.01107      0.08170      0.21005
```

M.4.4. 후방제거

```
step(carfit, direction="backward")
```

```
Start:  AIC=208.83
MPG.highway ~ EngineSize + Weight + Price + Width + Length +
    Horsepower + Wheelbase

             Df Sum of Sq     RSS    AIC
- Width       1      1.06  740.58 206.96
- EngineSize  1      4.18  743.69 207.35
- Price       1      5.47  744.98 207.51
- Horsepower  1      8.51  748.02 207.89
- Length      1     16.01  755.52 208.82
<none>                     739.51 208.82
- Wheelbase   1     48.00  787.51 212.67
- Weight      1    480.82 1220.33 253.41

Step:  AIC=206.96
MPG.highway ~ EngineSize + Weight + Price + Length + Horsepower +
    Wheelbase

             Df Sum of Sq     RSS    AIC
- EngineSize  1      7.73  748.31 205.93
- Price       1      9.67  750.25 206.17
- Horsepower  1     10.29  750.87 206.24
<none>                     740.58 206.96
- Length      1     19.54  760.12 207.38
- Wheelbase   1     51.04  791.62 211.16
- Weight      1    514.25 1254.83 254.00

Step:  AIC=205.93
MPG.highway ~ Weight + Price + Length + Horsepower + Wheelbase
```

```
             Df Sum of Sq     RSS    AIC
- Price       1     11.48  759.79 205.34
<none>                     748.31 205.93
- Horsepower  1     18.24  766.55 206.16
- Length      1     32.26  780.57 207.85
- Wheelbase   1     51.62  799.93 210.13
- Weight      1    516.55 1264.86 252.74

Step:  AIC=205.34
MPG.highway ~ Weight + Length + Horsepower + Wheelbase

             Df Sum of Sq     RSS    AIC
- Horsepower  1      7.79  767.59 204.29
<none>                     759.79 205.34
- Length      1     33.84  793.63 207.39
- Wheelbase   1     45.00  804.79 208.69
- Weight      1    508.94 1268.73 251.03

Step:  AIC=204.29
MPG.highway ~ Weight + Length + Wheelbase

            Df Sum of Sq     RSS    AIC
<none>                    767.59 204.29
- Wheelbase  1     37.41  805.00 206.72
- Length     1     37.87  805.46 206.77
- Weight     1    846.75 1614.34 271.43


Call:
lm(formula = MPG.highway ~ Weight + Length + Wheelbase, data = dat93)

Coefficients:
(Intercept)       Weight       Length    Wheelbase
   26.31687     -0.01107      0.08170      0.21005
```

M.4.5. 단계별 선택

```
model0 <- lm(MPG.highway~1, data=dat93)
step(model0,  scope=~EngineSize+ Weight+ Price+ Width+ Length+ Horsepower+ Wheelbase, direct
```

```
Start:  AIC=312.3
MPG.highway ~ 1
```

```
             Df Sum of Sq     RSS    AIC
+ Weight      1   1718.70  896.62 214.74
+ Width       1   1072.43 1542.88 265.22
+ EngineSize  1   1027.48 1587.83 267.89
+ Horsepower  1   1002.23 1613.08 269.36
+ Wheelbase   1    990.41 1624.90 270.04
+ Price       1    822.16 1793.16 279.20
+ Length      1    770.83 1844.48 281.82
<none>                   2615.31 312.30

Step:  AIC=214.74
MPG.highway ~ Weight

             Df Sum of Sq     RSS    AIC
+ Length      1     91.62  805.00 206.72
+ Wheelbase   1     91.16  805.46 206.77
+ Width       1     53.01  843.61 211.07
+ EngineSize  1     31.07  865.55 213.46
<none>                   896.62 214.74
+ Price       1      5.85  890.77 216.13
+ Horsepower  1      2.33  894.28 216.50
- Weight      1   1718.70 2615.31 312.30

Step:  AIC=206.72
MPG.highway ~ Weight + Length

             Df Sum of Sq     RSS    AIC
+ Wheelbase   1     37.41  767.59 204.29
<none>                   805.00 206.72
+ Width       1     13.71  791.29 207.12
+ EngineSize  1      7.38  797.62 207.86
+ Price       1      4.14  800.87 208.24
+ Horsepower  1      0.21  804.79 208.69
- Length      1     91.62  896.62 214.74
- Weight      1   1039.48 1844.48 281.82

Step:  AIC=204.29
MPG.highway ~ Weight + Length + Wheelbase

             Df Sum of Sq     RSS    AIC
<none>                   767.59 204.29
```

```
+ EngineSize  1     14.53  753.06 204.51
+ Width       1     12.79  754.80 204.73
+ Horsepower  1      7.79  759.79 205.34
+ Price       1      1.04  766.55 206.16
- Wheelbase   1     37.41  805.00 206.72
- Length      1     37.87  805.46 206.77
- Weight      1    846.75 1614.34 271.43


Call:
lm(formula = MPG.highway ~ Weight + Length + Wheelbase, data = dat93)

Coefficients:
(Intercept)       Weight       Length    Wheelbase
   26.31687     -0.01107      0.08170      0.21005
```

```
ols_step_forward_aic(carfit, details = TRUE)
```

```
Forward Selection Method
------------------------

Candidate Terms:

1. EngineSize
2. Weight
3. Price
4. Width
5. Length
6. Horsepower
7. Wheelbase


Step     => 0
Model    => MPG.highway ~ 1
AIC      => 578.2207

Initiating stepwise selection...

                       Table: Adding New Variables
-----------------------------------------------------------------------------
Predictor     DF      AIC        SBC       SBIC        R2       Adj. R2
-----------------------------------------------------------------------------
```

```
Weight          1    480.663    488.261    216.331    0.65717    0.65340
Width           1    531.141    538.739    264.864    0.41006    0.40358
EngineSize      1    533.812    541.410    267.447    0.39287    0.38620
Horsepower      1    535.280    542.878    268.867    0.38322    0.37644
Wheelbase       1    535.959    543.556    269.524    0.37870    0.37187
Price           1    545.122    552.720    278.402    0.31436    0.30683
Length          1    547.746    555.344    280.948    0.29474    0.28699
---------------------------------------------------------------------------

Step     => 1
Added    => Weight
Model    => MPG.highway ~ Weight
AIC      => 480.6632

                     Table: Adding New Variables
---------------------------------------------------------------------------
Predictor     DF      AIC        SBC       SBIC        R2       Adj. R2
---------------------------------------------------------------------------
Length          1    472.639    482.770    208.747    0.69220    0.68536
Wheelbase       1    472.692    482.822    208.797    0.69202    0.68518
Width           1    476.996    487.126    212.824    0.67744    0.67027
EngineSize      1    479.384    489.514    215.061    0.66905    0.66169
Price           1    482.055    492.185    217.566    0.65940    0.65183
Horsepower      1    482.421    492.551    217.909    0.65806    0.65046
---------------------------------------------------------------------------

Step     => 2
Added    => Length
Model    => MPG.highway ~ Weight + Length
AIC      => 472.6393

                     Table: Adding New Variables
---------------------------------------------------------------------------
Predictor     DF      AIC        SBC       SBIC        R2       Adj. R2
---------------------------------------------------------------------------
Wheelbase       1    470.213    482.876    206.718    0.70650    0.69661
Width           1    473.042    485.705    209.299    0.69744    0.68724
EngineSize      1    473.783    486.446    209.975    0.69502    0.68474
Price           1    474.160    486.823    210.320    0.69378    0.68346
Horsepower      1    474.615    487.278    210.736    0.69228    0.68190
---------------------------------------------------------------------------
```

```
Step     => 3
Added    => Wheelbase
Model    => MPG.highway ~ Weight + Length + Wheelbase
AIC      => 470.2133
```

Table: Adding New Variables

Predictor	DF	AIC	SBC	SBIC	R2	Adj. R2
EngineSize	1	470.436	485.631	207.247	0.71206	0.69897
Width	1	470.651	485.847	207.438	0.71139	0.69827
Horsepower	1	471.264	486.460	207.982	0.70948	0.69628
Price	1	472.088	487.283	208.714	0.70690	0.69358

```
No more variables to be added.

Variables Selected:

=> Weight
=> Length
=> Wheelbase
```

Stepwise Summary

Step	Variable	AIC	SBC	SBIC	R2	Adj. R2
0	Base Model	578.221	583.286	311.963	0.00000	0.00000
1	Weight	480.663	488.261	216.331	0.65717	0.65340
2	Length	472.639	482.770	208.747	0.69220	0.68536
3	Wheelbase	470.213	482.876	206.718	0.70650	0.69661

```
Final Model Output
------------------
```

Model Summary

R	0.841	RMSE	2.873
R-Squared	0.707	MSE	8.254

```
Adj. R-Squared          0.697       Coef. Var           10.097
Pred R-Squared          0.665       AIC                470.213
MAE                     2.139       SBC                482.876
-----------------------------------------------------------------

 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error
 AIC: Akaike Information Criteria
 SBC: Schwarz Bayesian Criteria

                              ANOVA
-------------------------------------------------------------------
               Sum of
              Squares        DF    Mean Square      F         Sig.
-------------------------------------------------------------------
Regression    1847.724         3        615.908    71.413    0.0000
Residual       767.588        89          8.625
Total         2615.312        92
-------------------------------------------------------------------

                                Parameter Estimates
-----------------------------------------------------------------------------------------------
      model      Beta     Std. Error    Std. Beta      t        Sig      lower     upper
-----------------------------------------------------------------------------------------------
(Intercept)    26.317          7.088                3.713     0.000    12.233    40.401
     Weight    -0.011          0.001       -1.225   -9.909     0.000    -0.013    -0.009
     Length     0.082          0.039        0.224    2.095     0.039     0.004     0.159
  Wheelbase     0.210          0.101        0.269    2.083     0.040     0.010     0.410
-----------------------------------------------------------------------------------------------
```

```
ols_step_forward_p(carfit, details = TRUE)
```

```
Forward Selection Method
------------------------

Candidate Terms:

1. EngineSize
2. Weight
3. Price
4. Width
5. Length
```

```
6. Horsepower
7. Wheelbase


Step    => 0
Model   => MPG.highway ~ 1
R2      => 0

Initiating stepwise selection...

                    Selection Metrics Table
-----------------------------------------------------------------
Predictor     Pr(>|t|)    R-Squared    Adj. R-Squared      AIC
-----------------------------------------------------------------
Weight         0.00000        0.657             0.653    480.663
Width          0.00000        0.410             0.404    531.141
EngineSize     0.00000        0.393             0.386    533.812
Horsepower     0.00000        0.383             0.376    535.280
Wheelbase      0.00000        0.379             0.372    535.959
Price          0.00000        0.314             0.307    545.122
Length         0.00000        0.295             0.287    547.746
-----------------------------------------------------------------

Step      => 1
Selected  => Weight
Model     => MPG.highway ~ Weight
R2        => 0.657

                    Selection Metrics Table
-----------------------------------------------------------------
Predictor     Pr(>|t|)    R-Squared    Adj. R-Squared      AIC
-----------------------------------------------------------------
Length         0.00190        0.692             0.685    472.639
Wheelbase      0.00195        0.692             0.685    472.692
Width          0.01952        0.677             0.670    476.996
EngineSize     0.07563        0.669             0.662    479.384
Price          0.44422        0.659             0.652    482.055
Horsepower     0.62912        0.658             0.650    482.421
-----------------------------------------------------------------

Step      => 2
```

```
Selected  => Length
Model     => MPG.highway ~ Weight + Length
R2        => 0.692

                    Selection Metrics Table
-----------------------------------------------------------------
Predictor     Pr(>|t|)    R-Squared    Adj. R-Squared      AIC
-----------------------------------------------------------------
Wheelbase      0.04014        0.707             0.697    470.213
Width          0.21761        0.697             0.687    473.042
EngineSize     0.36665        0.695             0.685    473.783
Price          0.49960        0.694             0.683    474.160
Horsepower     0.88015        0.692             0.682    474.615
-----------------------------------------------------------------

Step      => 3
Selected  => Wheelbase
Model     => MPG.highway ~ Weight + Length + Wheelbase
R2        => 0.707

                    Selection Metrics Table
-----------------------------------------------------------------
Predictor     Pr(>|t|)    R-Squared    Adj. R-Squared      AIC
-----------------------------------------------------------------
EngineSize     0.19593        0.712             0.699    470.436
Width          0.22536        0.711             0.698    470.651
Horsepower     0.34463        0.709             0.696    471.264
Price          0.73113        0.707             0.694    472.088
-----------------------------------------------------------------

Step      => 4
Selected  => EngineSize
Model     => MPG.highway ~ Weight + Length + Wheelbase + EngineSize
R2        => 0.712

                    Selection Metrics Table
-----------------------------------------------------------------
Predictor     Pr(>|t|)    R-Squared    Adj. R-Squared      AIC
-----------------------------------------------------------------
Width          0.46764        0.714             0.697    471.869
Horsepower     0.56977        0.713             0.697    472.088
Price          0.61570        0.713             0.696    472.165
```

```
---------------------------------------------------------------


No more variables to be added.

Variables Selected:

=> Weight
=> Length
=> Wheelbase
=> EngineSize


                          Stepwise Summary
-------------------------------------------------------------------------
Step    Variable        AIC        SBC       SBIC        R2       Adj. R2
-------------------------------------------------------------------------
 0      Base Model    578.221    583.286    311.963    0.00000    0.00000
 1      Weight        480.663    488.261    216.331    0.65717    0.65340
 2      Length        472.639    482.770    208.747    0.69220    0.68536
 3      Wheelbase     470.213    482.876    206.718    0.70650    0.69661
 4      EngineSize    470.436    485.631    207.247    0.71206    0.69897
-------------------------------------------------------------------------

Final Model Output
------------------

                       Model Summary
---------------------------------------------------------------
R                       0.844       RMSE                 2.846
R-Squared               0.712       MSE                  8.097
Adj. R-Squared          0.699       Coef. Var           10.057
Pred R-Squared          0.664       AIC                470.436
MAE                     2.081       SBC                485.631
---------------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error
 AIC: Akaike Information Criteria
 SBC: Schwarz Bayesian Criteria

                              ANOVA
```

```
---------------------------------------------------------------------
               Sum of
              Squares        DF    Mean Square      F         Sig.
---------------------------------------------------------------------
Regression    1862.256        4        465.564    54.404    0.0000
Residual       753.056       88          8.557
Total         2615.312       92
---------------------------------------------------------------------

                                Parameter Estimates
-------------------------------------------------------------------------------------------
      model      Beta     Std. Error    Std. Beta      t        Sig      lower     upper
-------------------------------------------------------------------------------------------
(Intercept)    28.439         7.246                   3.925    0.000    14.040    42.839
     Weight    -0.012         0.001       -1.334     -8.960    0.000    -0.015    -0.009
     Length     0.063         0.041        0.172      1.511    0.134    -0.020     0.145
  Wheelbase     0.233         0.102        0.298      2.282    0.025     0.030     0.435
 EngineSize     0.766         0.587        0.149      1.303    0.196    -0.402     1.933
-------------------------------------------------------------------------------------------
```

M.5. 연습문제 6.12

```
set.seed(13111)

Nsim <- 10
num_of_ind_var <- 50
n <- 100
N <- (num_of_ind_var+1)*n


randata <- matrix(rnorm(N), ncol=num_of_ind_var+1)
randata <- data.frame(randata)
indvarname <- names(randata)[1:50]

variables.set <- NULL
r2.set <- NULL
r2adj.set <- NULL
p <- NULL

for (i in 1:Nsim) {

  randata <- matrix(rnorm(N), ncol=num_of_ind_var+1)
  randata <- data.frame(randata)

  # === 함수 step(..direction="backward")는  AIC 를 이용
  fit.lm <- lm(X51 ~ ., data=randata)
  res <- step(fit.lm, direction="backward", trace=0)
  variable.entered <- as.character(attr(res$terms, "variables"))[-c(1,2)]

  # ==== 함수 step(...direction = "forward"..) 은 AIC 를 이용
  # fit.null <- lm(X51 ~ 1, data=randata)
  # res <- step(fit.null, scope~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8 + X9 + X10 + X11 + X1
  #  X13 + X14 + X15 + X16 + X17 + X18 + X19 + X20 + X21 + X22 +
  #  X23 + X24 + X25 + X26 + X27 + X28 + X29 + X30 + X31 + X32 +
  #  X33 + X34 + X35 + X36 + X37 + X38 + X39 + X40 + X41 + X42 +
  #  X43 + X44 + X45 + X46 + X47 + X48 + X49 + X50  , direction = "forward", trace=0)
  # ariable.entered <- as.character(attr(res$terms, "variables"))[-c(1,2)]


  # === 함수 ols_step_backward_p 는 F-검정을 이용
  # fit.lm <- lm(X51 ~ ., data=randata)
  # aa <- ols_step_backward_p(fit.lm)
  # variable.entered <- names(aa$model$coefficients)[-1]

  # ===  함수 ols_step_forward_p 는 F-검정을 이용
  # fit.lm <- lm(X51 ~ ., data=randata)
  # aa <- ols_step_forward_p(fit.lm)
  # variable.entered <- names(aa$model$coefficients)[-1]
```

```
variables.set
 X1  X2  X3  X4  X5  X6  X7  X8  X9 X10 X11 X12 X13 X14 X15 X16 X17 X18 X19 X20
  2   3   3   3   2   5   5   1   5   0   3   4   0   4   2   4   3   2   1   2
X21 X22 X23 X24 X25 X26 X27 X28 X29 X30 X31 X32 X33 X34 X35 X36 X37 X38 X39 X40
  6   3   4   2   4   4   3   1   4   3   2   3   2   1   3   4   2   1   2   5
X41 X42 X43 X44 X45 X46 X47 X48 X49 X50
  2   2   3   2   2   3   1   2   4   0
```

```
summary(p)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    7.0     9.0    14.0    13.4    17.0    20.0
```

```
summary(r2.set)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.2228  0.2909  0.3612  0.3570  0.4050  0.5094
```

```
summary(r2adj.set)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.1545  0.2094  0.2635  0.2595  0.2938  0.3852
```

N. R-실습: 분산분석 모형

N.1. 일원배치: 예제와 R 프로그램

N.1.1. 예제: 실험계획

4개의 서로 다른 원단업체에서 직물을 공급받고 있다. 공급한 직물의 긁힘에 대한 저항력을 알아보기 위하여 각 업체마다 4개의 제품을 랜덤하게 선택하여 ($a = 4, r = 4$) 일원배치법에 의하여 마모도 검사을 실시하였다.

N.1.2. 자료의 생성

```
company<- as.factor(rep(c(1:4), each=4))
response<- c(1.93, 2.38, 2.20, 2.25,
             2.55, 2.72, 2.75, 2.70,
             2.40, 2.68, 2.32, 2.28,
             2.33, 2.38, 2.28, 2.25)
df31<- data.frame(company=company, response= response)
df31
```

```
   company response
1        1     1.93
2        1     2.38
3        1     2.20
4        1     2.25
5        2     2.55
6        2     2.72
7        2     2.75
8        2     2.70
9        3     2.40
10       3     2.68
11       3     2.32
12       3     2.28
13       4     2.33
14       4     2.38
15       4     2.28
16       4     2.25
```

각 수준에 대한 표보 평균을 구해보자.

```
df31s <- df31 %>% group_by(company)  %>%  summarise(mean=mean(response), median= median(resp
df31s
```

```
# A tibble: 4 x 6
  company  mean median     sd   min   max
  <fct>   <dbl>  <dbl>  <dbl> <dbl> <dbl>
1 1        2.19   2.22 0.189   1.93  2.38
2 2        2.68   2.71 0.0891  2.55  2.75
3 3        2.42   2.36 0.180   2.28  2.68
4 4        2.31   2.30 0.0572  2.25  2.38
```

N.1.3. 선형모형의 적합(set-to-zero)

이제 자료를 다음과 같은 선형 모형으로 적합해 보자. 선형 모형의 적합은 `lm()` 함수를 사용한다.

$$y_{ij} = \mu + \alpha_i + e_{ij}$$

여기서 선형식의 모수와 R의 변수는 다음과 같은 관계를 가진다,

선형식의 모수	R의 변수
μ	(Intercept)
α_1	company1
α_2	company2
α_3	company3
α_4	company4

```
fit1 <- lm(response~company,data=df31)
summary(fit1)
```

```
Call:
lm(formula = response ~ company, data = df31)

Residuals:
    Min      1Q  Median      3Q     Max
-0.2600 -0.0700  0.0150  0.0625  0.2600

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.19000    0.07050  31.062 7.79e-13 ***
company2     0.49000    0.09971   4.914 0.000357 ***
company3     0.23000    0.09971   2.307 0.039710 *
company4     0.12000    0.09971   1.204 0.251982
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.141 on 12 degrees of freedom
Multiple R-squared:  0.6871,    Adjusted R-squared:  0.6089
F-statistic: 8.785 on 3 and 12 DF,  p-value: 0.002353
```

위에서 적합한 결과를 보면 평균 μ와 4개의 처리 α_1, α_2, α_3, α_4 가 모형에 있지만 모수의 추정량은 평균(intercept)과 3개의 모수(company2, company3, company4)만 추정량이 주어진다.
R 에서 옵션을 지정하지 않고 함수 lm()으로 선형모형을 적합하는 경우 set-to-zero 조건을 적용하며 자료에 나타난 처리의 수준들 중 순위가 가장 낮은 수준의 효과를 0으로 지정한다 (company1=0). set-to-zero 조건을 강제로 지정하려면 다음과 같은 명령문을 먼저 실행한다.

```
options(contrasts=c("contr.treatment", "contr.poly"))
```

위의 결과를 보면 (Intercept)에 대한 추정량이 첫 번째 처리 company1의 평균과 같은 것을 알 수 있다. set-to-zero 조건에서의 계획행렬은 다음과 같이 볼 수 있다.

```
model.matrix(fit1)
```

```
   (Intercept) company2 company3 company4
1            1        0        0        0
2            1        0        0        0
3            1        0        0        0
4            1        0        0        0
5            1        1        0        0
```

```
6            1        1        0        0
7            1        1        0        0
8            1        1        0        0
9            1        0        1        0
10           1        0        1        0
11           1        0        1        0
12           1        0        1        0
13           1        0        0        1
14           1        0        0        1
15           1        0        0        1
16           1        0        0        1
attr(,"assign")
[1] 0 1 1 1
attr(,"contrasts")
attr(,"contrasts")$company
[1] "contr.treatment"
```

이제 각 처리 평균에 대한 추정값 $\widehat{\mu+\alpha_i}$을 구해보자.

```
emmeans(fit1, "company")
```

```
 company emmean     SE df lower.CL upper.CL
 1          2.19 0.0705 12     2.04     2.34
 2          2.68 0.0705 12     2.53     2.83
 3          2.42 0.0705 12     2.27     2.57
 4          2.31 0.0705 12     2.16     2.46

Confidence level used: 0.95
```

이 경우 처리 평균에 대한 추정값은 산술 평균과 동일하게 나온다.

N.1.4. 선형모형의 적합 (sum-to-zero)

이제 일원배치 모형에서 sum-to-zero 조건을 적용하여 모수를 추정해 보자. sum-to-zero 조건을 적용하려면 다음과 같은 명령어를 실행해야 한다.

```
options(contrasts=c("contr.sum", "contr.poly"))
```

이제 다시 선형모형을 적합하고 추정결과를 보자.

```
fit2 <- lm(response~company,data=df31)
summary(fit2)
```

```
Call:
lm(formula = response ~ company, data = df31)

Residuals:
    Min      1Q  Median      3Q     Max
-0.2600 -0.0700  0.0150  0.0625  0.2600

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.40000    0.03525  68.081  < 2e-16 ***
company1    -0.21000    0.06106  -3.439 0.004901 **
company2     0.28000    0.06106   4.586 0.000626 ***
company3     0.02000    0.06106   0.328 0.748892
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.141 on 12 degrees of freedom
Multiple R-squared:  0.6871,    Adjusted R-squared:  0.6089
F-statistic: 8.785 on 3 and 12 DF,  p-value: 0.002353
```

이제 sum-to-zero 조건에 따라서 위의 set-to-zero 결과와 모수의 추정값이 다르게 나타나는 것을 알 수 있다. 마지막 모수 company4(α_4)는 sum-to-zero 조건을 이용하여 다음과 같은 관계를 이용하여 구할 수 있다.

$$\alpha_4 = -(\alpha_1 + \alpha_2 + \alpha_3)$$

sum-to-zero 조건에서의 계획행렬은 다음과 같이 볼 수 있다.

```
model.matrix(fit2)
```

```
   (Intercept) company1 company2 company3
1            1        1        0        0
2            1        1        0        0
3            1        1        0        0
4            1        1        0        0
5            1        0        1        0
6            1        0        1        0
7            1        0        1        0
8            1        0        1        0
9            1        0        0        1
10           1        0        0        1
11           1        0        0        1
12           1        0        0        1
13           1       -1       -1       -1
```

```
14           1       -1       -1       -1
15           1       -1       -1       -1
16           1       -1       -1       -1
attr(,"assign")
[1] 0 1 1 1
attr(,"contrasts")
attr(,"contrasts")$company
[1] "contr.sum"
```

이제 각 처리 평균에 대한 추정값 $\widehat{\mu+\alpha_i}$을 구해보면 set-to-zero 조건에서의 추정값과 동일함을 알 수 있다.

```
emmeans(fit2, "company")
```

```
 company emmean     SE df lower.CL upper.CL
 1          2.19 0.0705 12     2.04     2.34
 2          2.68 0.0705 12     2.53     2.83
 3          2.42 0.0705 12     2.27     2.57
 4          2.31 0.0705 12     2.16     2.46

Confidence level used: 0.95
```

N.1.5. 분산분석

분산분석의 결과는 어떠한 제약 조건에서도 동일하다.

```
res1 <- anova(fit1)
res1
```

```
Analysis of Variance Table

Response: response
          Df Sum Sq  Mean Sq F value   Pr(>F)
company    3 0.5240 0.174667  8.7846 0.002353 **
Residuals 12 0.2386 0.019883
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
res2<- anova(fit2)
res2
```

```
Analysis of Variance Table

Response: response
```

```
            Df Sum Sq  Mean Sq F value   Pr(>F)
company      3 0.5240 0.174667  8.7846 0.002353 **
Residuals 12 0.2386 0.019883
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

N.1.6. 다중비교 예제

앞에서 살펴본 일원배치법 예제은 4개의 처리가 있다. 따라서 $\binom{4}{2} = 6$ 개의 가설 검정(또는 신뢰구간)을 수행해야 한다.
4개의 company가 처리 수준이며 각 처리수준 은 1, 2, 3, 4로 표시된다.

```
df31s
```

```
# A tibble: 4 x 6
  company  mean median     sd   min   max
  <fct>   <dbl>  <dbl>  <dbl> <dbl> <dbl>
1 1        2.19   2.22 0.189   1.93  2.38
2 2        2.68   2.71 0.0891  2.55  2.75
3 3        2.42   2.36 0.180   2.28  2.68
4 4        2.31   2.30 0.0572  2.25  2.38
```

N.1.6.1. 다중비교 방법을 적용하지 않는 경우

먼저 다중비교 방법을 적용하지 않는 경우 결과를 보자. 함수 LSD.test 에서 p.adj=c("none")를 지정하면 다중 비교를 적용하지 않는다. 명령문 p.adj 를 지정하지 않으면 수정을 하지 않는 LSD 방법에 의한 신뢰구간 7.24 와 검정 방법 7.25 로 구한 결과를 준다.

```
anova.res <- aov(response~company,data=df31) #일원배치
test1 <- LSD.test(anova.res, "company", alpha = 0.05, group = FALSE, console = FALSE, p.adj=
test1$comparison
```

```
      difference pvalue signif.         LCL         UCL
1 - 2      -0.49 0.0004     *** -0.70724487 -0.27275513
1 - 3      -0.23 0.0397       * -0.44724487 -0.01275513
1 - 4      -0.12 0.2520         -0.33724487  0.09724487
2 - 3       0.26 0.0229       *  0.04275513  0.47724487
2 - 4       0.37 0.0030      **  0.15275513  0.58724487
3 - 4       0.11 0.2916         -0.10724487  0.32724487
```

N.1.6.2. 본페로니 수정(Bonferroni correction)

이제 다중비교 방법 중에 가장 보수적인 본페로니 수정(Bonferroni correction)을 적용해 보자. 함수 LSD.test 에서 p.adj=c("bonferroni")를 이용한다.

아래의 결과는 본페로니 수정 방법에 의한 신뢰 구간 와 검정 방법으로 구한 결과이다.
본페로니 수정이 적용된 신뢰구간은 LSD 방법의 신뢰구간보다 길며 수정된 p-값은 LSD 방법으로 구한 값의 6배이다. LSD 방법을 적용하는 경우 유의한 차이를 보이는 조합이 4개로 나타났는데(1-2,1-3,2-3,2-4) 본페로니 수정을 적용한 경우에는 2개로 줄어 들었다(1-2,2-4)
수정한 p-값이 1이 초과하면 확률이기 때문에 1로 주어진다.

```
test2 <- LSD.test(anova.res, "company", alpha = 0.05, group = FALSE, console = FALSE, p.adj=
test2$comparison
```

```
      difference pvalue signif.         LCL         UCL
1 - 2      -0.49 0.0021      ** -0.80434725 -0.17565275
1 - 3      -0.23 0.2383         -0.54434725  0.08434725
1 - 4      -0.12 1.0000         -0.43434725  0.19434725
2 - 3       0.26 0.1374         -0.05434725  0.57434725
2 - 4       0.37 0.0179       *  0.05565275  0.68434725
3 - 4       0.11 1.0000         -0.20434725  0.42434725
```

N.1.6.3. Tukey의 HSD

함수TukeyHSD는 분산분석을 실행한 결과를 이용하여 다중비교 방법 중 가장 많이 이용되는 Tukey's Honest Significant Difference (HSD) 방법으로 다중비교를 제공한다.
Tukey의 HSD는 너무 보수적인 결과를 주는 본페로니 수정을 개선한 것이다. 따라서 Tukey의 HSD 에서 얻은 결과는 수정하지 않는 LDS 의 결과와 Bonferoni 방법의 중간에 있다고 할 수 있다.
Tukey의 HSD 에서는 본페로니와 유사하게 2개의 조합(1-2,2-4)만이 유의한 차이가 있다고 나타난다.

```
anova.res <- aov(response~company,data=df31) #일원배치
test3 <- TukeyHSD(anova.res, conf.level = 0.95, ordered=FALSE)
test3
```

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = response ~ company, data = df31)

$company
     diff         lwr         upr     p adj
2-1  0.49  0.19397708  0.78602292 0.0017496
3-1  0.23 -0.06602292  0.52602292 0.1509207
4-1  0.12 -0.17602292  0.41602292 0.6362891
3-2 -0.26 -0.55602292  0.03602292 0.0924227
4-2 -0.37 -0.66602292 -0.07397708 0.0136804
4-3 -0.11 -0.40602292  0.18602292 0.6943908
```

```
plot(test3)
```

N.1.6.4. 세 방법에서의 p-값 비교

위에서 살펴본 수정을 하지 않은 LSD 방법, Tukey의 HSD 방법과 본페로니 방법에서 계산된 p-값을 아래 표에서 비교하였다.

표 N.2.: LSD, Bonferoni, HSD 방법의 p-값 비교

평균의 비교 조합	LSD	HSD	Bonf
1-2	0.0004	0.0017	0.0021
1-3	0.0397	0.1509	0.2383
1-4	0.2520	0.6363	1.0000
2-3	0.0229	0.0924	0.1374
2-4	0.0030	0.0137	0.0179
3-4	0.2916	0.6944	1.0000

N.2. 예제 7.1 - 일원배치

이제 교재 예제 7.1 (271 페이지)에서 관고비 자료 adsale에서 판매액을 예측하는 회귀식을 고려해 보자. 먼저 adsale 데이터프레임에서 두 변수 광고비 ad 와 매체 media 변수의 차이점을 알아보자. 먼저 광고비 ad는 수치 변수(numeric variable)로서 함수 class()를 이용하면 정수(integer) 형태인 것을 알 수 있다. 수치변수는 정수, 실수, 복소수 등을 의미한다.
반면 매체 media 는 범주형 변수로서 함수 class()를 이용하면 범주형(factor) 형태인 것을 알 수 있다. levels는 범주형 변수의 항목을 나타내는 것으로서 범주형 변수 media는 두 개의 항목 방송과 신문으로 이루어 졌음을 알 수 있다.

```
head(adsale)
```

```
  sale ad media
1   39  4  방송
2   42  6  신문
3   45  6  방송
4   47  8  신문
5   50  8  방송
6   50  9  신문
```

```
adsale$ad
```

```
 [1]  4  6  6  8  8  9  9 10 12 12
```

```
adsale$media
```

```
 [1] 방송 신문 방송 신문 방송 신문 방송 방송 신문 방송
Levels: 방송 신문
```

```
class(adsale$ad)
```

```
[1] "integer"
```

```
class(adsale$media)
```

```
[1] "factor"
```

그룹별 기초통계량을 계산해보자.

```
adsalesum <- adsale %>% group_by(media)  %>%  summarise(mean=mean(ad), median= median(ad), s
adsalesum
```

```
# A tibble: 2 x 6
  media  mean median    sd   min   max
  <fct> <dbl>  <dbl> <dbl> <int> <int>
1 방송  8.17    8.5  2.86     4    12
2 신문  8.75    8.5  2.5      6    12
```

함수 lm()으로 선형모형을 적합하는 경우 set-to-zero 조건을 적용하며 자료에 나타난 처리의 수준들 중 순위가 가장 낮은 수준의 효과를 0으로 지정한다 set-to-zero 조건을 강제로 지정하려면 다음과 같은 명령문을 먼저 실행한다.

```
options(contrasts=c("contr.treatment", "contr.poly"))
```

이제 광고비 ad 와 매체 media 를 포함한 회귀식을 적합시켜 보자. R의 lm 함수는 범주형변수를 자동적으로 가변수로 바꾸어 준다. 회귀식에 사용된 디자인행렬을 보면 media에 해당하는 열이 0 과 1로 이루어진 벡터로 바뀌었음을 알 수 있다.

```
fit1 <- lm(sale~ ad + media, data=adsale)
summary(fit1)
```

```

Call:
lm(formula = sale ~ ad + media, data = adsale)

Residuals:
     Min       1Q   Median       3Q      Max
-0.47902 -0.24720  0.02727  0.22832  0.39091

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 29.21888    0.38588   75.72 1.84e-11 ***
ad           2.56503    0.04408   58.19 1.16e-10 ***
media신문  -2.66294    0.22114  -12.04 6.21e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3403 on 7 degrees of freedom
Multiple R-squared:  0.998, Adjusted R-squared:  0.9974
F-statistic:  1707 on 2 and 7 DF,  p-value: 3.875e-10
```

```
model.matrix(fit1)
```

```
   (Intercept) ad media신문
1            1  4         0
```

```
2            1  6         1
3            1  6         0
4            1  8         1
5            1  8         0
6            1  9         1
7            1  9         0
8            1 10         0
9            1 12         1
10           1 12         0
attr(,"assign")
[1] 0 1 2
attr(,"contrasts")
attr(,"contrasts")$media
[1] "contr.treatment"
```

```
data.frame(media=adsale$media, z=model.matrix(fit1)[,3])
```

```
   media z
1   방송 0
2   신문 1
3   방송 0
4   신문 1
5   방송 0
6   신문 1
7   방송 0
8   방송 0
9   신문 1
10  방송 0
```

N.3. 예제 7.2 - 교호작용

```
fit2 <- lm(sale ~ ad + media + ad:media, data=adsale)
summary(fit2)
```

```

Call:
lm(formula = sale ~ ad + media + ad:media, data = adsale)

Residuals:
    Min      1Q  Median      3Q     Max
-0.3674 -0.1400 -0.1043  0.2194  0.4490
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  29.00000    0.46388  62.516 1.13e-09 ***
ad            2.59184    0.05411  47.901 5.55e-09 ***
media신문  -1.93333    0.85628  -2.258   0.0647 .
ad:media신문 -0.08517    0.09645  -0.883   0.4112
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3458 on 6 degrees of freedom
Multiple R-squared:  0.9982,    Adjusted R-squared:  0.9973
F-statistic:  1102 on 3 and 6 DF,  p-value: 1.298e-08
```

```
anova(fit2)
```

```
Analysis of Variance Table

Response: sale
          Df Sum Sq Mean Sq   F value    Pr(>F)
ad         1 378.50  378.50 3166.1379 2.116e-09 ***
media      1  16.79   16.79  140.4378 2.183e-05 ***
ad:media   1   0.09    0.09    0.7797    0.4112
Residuals  6   0.72    0.12
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
model.matrix(fit2)
```

```
   (Intercept) ad media신문 ad:media신문
1            1  4         0            0
2            1  6         1            6
3            1  6         0            0
4            1  8         1            8
5            1  8         0            0
6            1  9         1            9
7            1  9         0            0
8            1 10         0            0
9            1 12         1           12
10           1 12         0            0
attr(,"assign")
[1] 0 1 2 3
attr(,"contrasts")
```

```
attr(,"contrasts")$media
[1] "contr.treatment"
```

N.4. 예제 7.3 - 일원배치 2

```
english1
```

```
   score grade
1     81     1
2     75     1
3     69     1
4     90     1
5     72     1
6     83     1
7     65     2
8     80     2
9     73     2
10    79     2
11    81     2
12    69     2
13    72     3
14    67     3
15    62     3
16    76     3
17    80     3
18    89     4
19    94     4
20    79     4
21    88     4
```

```
class(english1$grade)
```

```
[1] "integer"
```

위의 결과로 보면 변수 grade는 수치형 변수이다. 따라서 이를 범주형 변수로 바꾸어 주어야 한다. 위의 명령문은 함수 factor()를 사용하여 수치 변수인 grade를 항목의 순서가 4,1,2,3 (levels=c(4,1:3)) 인 범주형 변수로 바꾸어 주는 것이다.

```
english1$grade <- factor(english1$grade, levels=c(4, 1:3), labels = c("4학년", "1학년", "2학년"
english1$grade
```

```
 [1] 1학년 1학년 1학년 1학년 1학년 1학년 2학년 2학년 2학년 2학년 2학년 2학년
[13] 3학년 3학년 3학년 3학년 3학년 4학년 4학년 4학년 4학년
Levels: 4학년 1학년 2학년 3학년
```

그룹별 기초통계량을 계산해보자.

```
english1sum <-english1 %>% group_by(grade)  %>%  summarise(mean=mean(score), median= median(
english1sum
```

```
# A tibble: 4 x 6
  grade  mean median    sd   min   max
  <fct> <dbl>  <dbl> <dbl> <int> <int>
1 4학년 87.5   88.5  6.24    79    94
2 1학년 78.3   78    7.79    69    90
3 2학년 74.5   76    6.57    65    81
4 3학년 71.4   72    7.13    62    80
```

이제 일원배치모형을 적합시키고 ANOVA F-검정을 수행해 보자.

```
fit3 <- lm(score ~ grade, english1)
summary(fit3)
```

```

Call:
lm(formula = score ~ grade, data = english1)

Residuals:
   Min     1Q Median     3Q    Max
-9.500 -5.500  0.600  4.667 11.667

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   87.500      3.513  24.910 8.06e-15 ***
grade1학년  -9.167      4.535  -2.021  0.05927 .
grade2학년 -13.000      4.535  -2.867  0.01069 *
grade3학년 -16.100      4.713  -3.416  0.00329 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.025 on 17 degrees of freedom
Multiple R-squared:  0.4341,    Adjusted R-squared:  0.3342
F-statistic: 4.347 on 3 and 17 DF,  p-value: 0.01905
```

```
anova(fit3)
```

```
Analysis of Variance Table

Response: score
          Df Sum Sq Mean Sq F value  Pr(>F)
grade      3 643.63 214.544   4.347 0.01905 *
Residuals 17 839.03  49.355
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

적합한 모형을 이용하여 최소제곱 평균을 구해보자.

```
emmeans(fit3, "grade")
```

```
 grade emmean   SE df lower.CL upper.CL
 4학년  87.5 3.51 17     80.1     94.9
 1학년  78.3 2.87 17     72.3     84.4
 2학년  74.5 2.87 17     68.4     80.6
 3학년  71.4 3.14 17     64.8     78.0

Confidence level used: 0.95
```

다중비교 방법을 적용하지 않고 각 학년별 평균의 차이를 비교하자.

```
anova.res <- aov(fit3)
test1 <- LSD.test(anova.res, "grade", alpha = 0.05, group = FALSE, console = FALSE, p.adj=c(
test1$comparison
```

```
               difference pvalue signif.        LCL        UCL
1학년 - 2학년  3.833333 0.3579          -4.724208 12.3908748
1학년 - 3학년  6.933333 0.1215          -2.041892 15.9085586
1학년 - 4학년 -9.166667 0.0593         . -18.734289  0.4009556
2학년 - 3학년  3.100000 0.4761          -5.875225 12.0752252
2학년 - 4학년 -13.000000 0.0107         * -22.567622 -3.4323777
3학년 - 4학년 -16.100000 0.0033        ** -26.042965 -6.1570353
```

본페로니 수정(Bonferroni correction)을 적용하여 각 학년별 평균의 차이를 비교하자.

```
test2 <- LSD.test(anova.res, "grade", alpha = 0.05, group = FALSE, console = FALSE, p.adj=c(
test2$comparison
```

```
               difference pvalue signif.        LCL        UCL
```

```
1학년 - 2학년  3.833333 1.0000          -8.270123 15.9367895
1학년 - 3학년  6.933333 0.7292          -5.760879 19.6275452
1학년 - 4학년 -9.166667 0.3556         -22.698742  4.3654087
2학년 - 3학년  3.100000 1.0000          -9.594212 15.7942119
2학년 - 4학년 -13.000000 0.0641        . -26.532075  0.5320754
3학년 - 4학년 -16.100000 0.0197       * -30.162945 -2.0370548
```

다중비교 방법 중 가장 많이 이용되는 Tukey's Honest Significant Difference (HSD) 방법으로 각 학년별 평균의 차이를 비교하자.

```
test3 <- TukeyHSD(anova.res, conf.level = 0.95, ordered=FALSE)
test3
```

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = fit3)

$grade
                   diff       lwr        upr     p adj
1학년-4학년 -9.166667 -22.05714  3.7238087 0.2190188
2학년-4학년 -13.000000 -25.89048 -0.1095246 0.0476922
3학년-4학년 -16.100000 -29.49617 -2.7038250 0.0157259
2학년-1학년 -3.833333 -15.36293  7.6962583 0.7813729
3학년-1학년 -6.933333 -19.02567  5.1590044 0.3891708
3학년-2학년 -3.100000 -15.19234  8.9923378 0.8842429
```