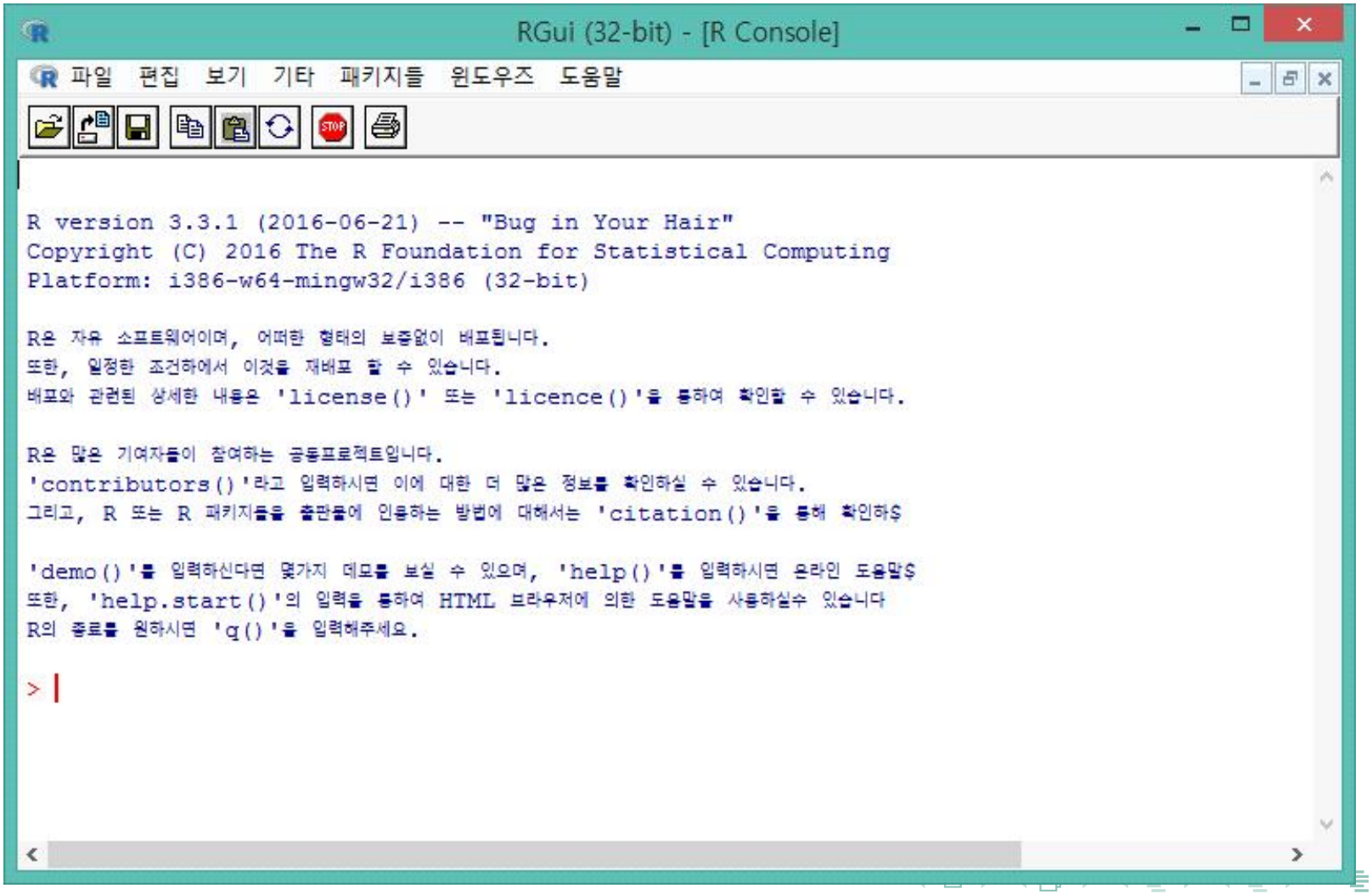
R 기초/선형회귀/로지스틱 회귀모형

이근백

성균관대학교 통계학과

2017년 8월 21일

- R 설치: http://cran.r-project.org/
  download R for windows → base →
  Download R 3.3.1 for Windows



- 기본연산

```
> 3+4
[1] 7
> exp(1)
[1] 2.718282
> log(exp(3))
[1] 3
> sqrt(4)
[1] 2
> 2^10
[1] 1024
```

- 변수지정과 연산

```
> pi
[1] 3.141593
> x=4
> y=6
> x+y
[1] 10
> sin(x*pi)
[1] -4.898425e-16
> exp(x/y)
[1] 1.947734
```

참고: "[1]"은 1행을 의미

- 논리연산자: >, >=, <, <=, ==, !=, &, |, !

```
> 3<4
[1] TRUE
> 3==4
[1] FALSE
> 3!=4
[1] TRUE
```

```
> (3+4==7) & (2+4==6)
[1] TRUE
> (3+4==6) & (2+4==6)
[1] FALSE
> (3+4==6) | (2+4==6)
[1] TRUE
```

참고: TRUE는 1로 간주되고 FALSE는 0으로 간주된다. 예를 들어

```
> (3+4==7)+(2+4==6)
[1] 2
> (3+4==6)-(2+4==6)
[1] -1
```

- 벡터와 벡터연산

```
> x=c(1,2,3)
> y=c(3,4,5)
> y[2]
[1] 4
> y[2]*x[3]
[1] 12
> x+y
[1] 4 6 8
> x/y
[1] 0.33333 0.50000 0.60000
> x>2
[1] FALSE FALSE  TRUE
> z=c(x,4)
> z
[1] 1 2 3 4
> sum(z)
[1] 10
```

```
> prod(z)
[1] 24
> mean(x)
[1] 2
> sd(x)
[1] 1
> max(z)
[1] 4
> min(z)
[1] 1
> seq(1,8)
[1]  1  2  3  4  5  6  7  8
> seq(1,10,2)
[1] 1 3 5 7 9
> rep(c(1,2,3),2)
[1] 1 2 3 1 2 3
```

- 행렬과 행렬연산

```
> matrix(c(1,2,3,4),ncol=2)
     [,1] [,2]
[1,]    1    3
[2,]    2    4
> x=matrix(c(1,2,3,4),ncol=2)
> y=matrix(c(2,4,3,1),ncol=2)
> x
     [,1] [,2]
[1,]    1    3
[2,]    2    4
> y
     [,1] [,2]
[1,]    2    3
[2,]    4    1
> x[,1]
[1] 1 2
> x[1,]
[1] 1 3
```

```
> x[1,2]
[1] 3
> x+y
     [,1] [,2]
[1,]    3    6
[2,]    6    5
> x*y
     [,1] [,2]
[1,]    2    9
[2,]    8    4
> x%*%y
     [,1] [,2]
[1,]   14    6
[2,]   20   10
> solve(x)
     [,1] [,2]
[1,]   -2  1.5
[2,]    1 -0.5
```

- 파일 불러오기
  - txt 파일 불러오기

```
data=read.table(file("d:/MPG.txt"),header=T)
```

  - csv 파일 불러오기

```
data=read.csv(file("d:/MPG.csv"),header=T)
```

- 파일 생성하기
  - "data"라는 변수를 new_data.txt 파일로 저장하기

```
write.table(data,file="d:/new_data.txt")
```

  - "data"라는 변수를 new_data.csv 파일로 저장하기

```
write.csv(data,file="d:/new_data.csv")
```

- 리스트

```
> grade=list(midterm=35, final=86,
             homework=c(18,20,17,19,12))
> grade$midterm
[1] 35
> grade$final
[1] 86
> grade$homework
[1] 18 20 17 19 12
> total=0.3*grade$mid+0.4*grade$final+0.3*sum(grade$ho)
> total
[1] 70.7
```

- apply(): 배열이나 행렬의 주변값(margins)에 함수를 적용시켜서 함수의 연산결과를 반환하는 함수이다.
  apply(x,MARGIN,FUN,...)

```
> mat.l <-matrix(c(1:12),ncol=4)
> mat.l
     [,1] [,2] [,3] [,4]
[1,]    1    4    7   10
[2,]    2    5    8   11
[3,]    3    6    9   12
> apply(mat.l,1,sum)
[1] 22 26 30
> apply(mat.l,2,sum)
[1]  6 15 24 33
> apply(mat.l,1,min)
[1] 1 2 3
```

```
> apply(mat.l,2,max)
[1]  3  6  9 12
> apply(mat.l,2,prod)
[1]    6  120  504 1320
> apply(mat.l,2,mean)
[1]  2  5  8 11
> apply(mat.l,1,quantile)
      [,1]  [,2]  [,3]
0%    1.00  2.00  3.00
25%   3.25  4.25  5.25
50%   5.50  6.50  7.50
75%   7.75  8.75  9.75
100% 10.00 11.00 12.00
```

- lapply(): 리스트 데이터 객체에 함수를 적용시켜서 함수의 연산결과를 반환하는 함수이다.
  lapply(x,FUN,...)

```
> lst <-list(x=c(1:10),y=c(T,F,T,T,F),
+ z=c(3,5,7,11,13,17))
> lst
$x
 [1]  1  2  3  4  5  6  7  8  9 10
$y
[1]  TRUE FALSE  TRUE  TRUE FALSE
$z
[1]  3  5  7 11 13 17
> lapply(lst,mean)
$x
[1] 5.5
$y
[1] 0.6
$z
[1] 9.333333
```

```
> lapply(lst,quantile,
+ probs=1:3/4)
$x
 25%  50%  75%
3.25 5.50 7.75
$y
25% 50% 75%
  0   1   1
$z
 25%  50%  75%
 5.5  9.0 12.5
```

- sapply(): lapply함수처럼 리스트 데이터 객체에 함수를 적용시켜서 연산결과를 벡터(행렬)로 변환하는 함수이다.
  lapply(x,FUN,...,simplify=TRUE, USE.NAMES=TRUE)

```
> sapply(lst,mean)
       x        y        z
5.500000 0.600000 9.333333
> sapply(lst,mean,simplify=F)
$x
[1] 5.5
$y
[1] 0.6
$z
[1] 9.333333
> sapply(lst,quantile,prob=1:3/4)
       x y    z
25% 3.25 0  5.5
50% 5.50 1  9.0
75% 7.75 1 12.5
```

```
> sapply(lst,quantile,
+ prob=1:3/4, simplify=F)
$x
 25%  50%  75%
3.25 5.50 7.75
$y
25% 50% 75%
  0   1   1
$z
 25%  50%  75%
 5.5  9.0 12.5
```

- tapply(): tapply (table apply)는 함수를 적용하여 분할표 등 집계 테이블을 만드는 함수이다.
  tapply(x,,INDEX,FUN=NULL,...,simplify=TRUE)

```
> head(warpbreaks)
  breaks wool tension
1     26    A       L
2     30    A       L
3     54    A       L
4     25    A       L
5     70    A       L
6     52    A       L
> dim(warpbreaks)
[1] 54  3
> levels(warpbreaks$wool)
[1] "A" "B"
```

```
> levels(warpbreaks$tension)
[1] "L" "M" "H"
> tapply(warpbreaks$breaks,
+ warpbreaks[,-1],sum)
    tension
wool   L   M   H
   A 401 216 221
   B 254 259 169
> tapply(warpbreaks$breaks,
+ warpbreaks[,3,drop=F],max)
tension
 L  M  H
70 42 43
```

- if...else 문

```
if (test expression) {
   statement 1
   } else {
   statement 2}
```

→ "test expression"이 맞으면 "statement 1"을 실행하고 그렇지 않으면 "statement 2"를 실행

- 예제

```
> x=3
> if (x>0) {
+ print("Positive number")
+ } else {
+ print("Negative number")
+ }
[1] "Positive number"
```

```
> x=-5
> if (x>0) {
+ print("Positive number")
+ } else {
+ print("Negative number")
+ }
[1] "Negative number"
```

- `while(condition) {statement}` 문: "condition"이 만족하는 동안 "statement" 반복적으로 실행

```
> x=1
> while (x<5) {x=x+1}
> x
[1] 5
```

- `for (i in 1:10) {statement}` 문: i를 1부터 10까지 증가시키면서 "statement"를 반복적으로 실행

```
> x=0
> for (i in 1:10) {x=x+i}
> x
[1] 55
```

- 예제: 1부터 100까지의 자연수중에 7의 배수들의 합

```
> x=0
> for (i in 1:100)
+ {
+ if (i%%7==0) x=x+i
+ }
> x
[1] 735
```

혹은

```
> x=seq(1:100)
> sum(x[which(x%%7==0)])
[1] 735
```

- R 함수 만들기

```
function_name=function (input1, input2,...)
{
expressions...
return(object)
}
```

- 예제: 1부터 y까지의 자연수중 7의 배수의 합을 구하는 함수

```
test=function(y)
{
  x=0
  for (i in 1:y)
  {
      if (i%%7==0) x=x+i
  }
  return(x)
}
```

- package: 여러함수들을 모아놓은 꾸러미
- package 설치하기

```
> install.packages("package_name")
```

- package 불러오기

```
> library("package_name")
```

상관분석

두 변수 간에 어떤 관계가 존재할 때 이러한 관계가 선형(linear)관계이면 상관관계가 있다고 하며, 그 크기 정도를 상관계수라고 한다.

- 표본공분산(sample covariance)

$$S_{xy} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$$

  - $S_{xy} > 0$이면 X와 Y 사이에 양의관계이고, $S_{xy} < 0$이면 X와 Y 사이에 음의관계가 있다.
  - 공분산은 관계의 강도(strength)가 얼마나 되는지를 쉽게 알 수 없다.
- 표본상관계수(sample correlation coefficient): 공분산의 표준화 형태

$$r = \frac{S_{xy}}{\sqrt{S_x S_y}} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

여기서 S_x와 S_y는 X와 Y의 표본 분산이다.

상관분석

- 상관계수의 특성
  - -1과 1사이의 값을 가진다.
  - -1 또는 1의 값이면 두 변수 사이에는 완벽한 선형관계가 존재한다.
  - 상관계수는 두 변수간의 상관관계만을 나타낸다.
- 상관계수의 해석: 두 변수 사이의 선형관계가 얼마나 강한가를 측정하는 지수이다.
  - 산포도를 그려봄으로써 두 변수 사이의 대략적인 관계를 파악한다.
  - 표본상관계수를 구하고 해석을 내린다.
    - $0.1 \sim 0.7$ $(-0.1 \sim -0.7)$인 경우: 매우 강한 관련성
    - $0.7 \sim 0.4$ $(-0.7 \sim -0.4)$인 경우: 상당한 관련성
    - $0.4 \sim 0.2$ $(-0.4 \sim -0.2)$인 경우: 약간의 관련성
    - $0.2 \sim 0.0$ $(-0.2 \sim -0.0)$인 경우: 관련성이 없음

상관분석의 가설

두 변수 사이의 선형관계가 통계적으로 유의한지 여부를 검정한다. 이 가설을 검정하기 위하여 두 변수 모두 정규분포를 따른다는 기본가정이 있어야 한다.

$$H_0 : \rho_{xy} = 0, \text{ 두 변수 간에 상관관계는 없다.}$$
$$H_1 : \rho_{xy} \neq 0, \text{ 두 변수 간에 상관관계가 있다.}$$

```
> cov(x,y) #상관계수
> cor.test(x,y) #상관계수의 가설검정
```

회귀분석의 개념

- 두 변수간의 인과관계를 조사하는 방법
  - 독립변수 (설명변수 or 예측변수): 영향을 미칠 것으로 생각되는 변수
  - 종속변수 (결과변수 or 응답변수): 영향을 받을 것으로 생각되는 변수
  - 두 변수간의 관계에서 독립변수와 종속변수의 설정
- 두 변수의 관계에서 종속변수의 설정
  - 논리적 타당성을 토대로 하여 독립변수의 변화에 따라 종속변수의 변화가 있을 것으로 나타나면 비로소 두 변수 간의 인과관계가 있은 것으로 추정할 수 있다.
  - 논리적 근거없이 어떤 임의의 두 변수를 독립변수 및 종속변수로 설정하여 회귀분석 했을 때에 독립변수의 계수가 통계적으로 유의미하더라도 두 변수 간에 인과관계가 있다고 추정할 수 없다.

회귀분석이란?

- 예제: 차량 연비와 차량무게의 상관관계?
  - Y: 차량 연비, X: 차량 무게
  - 모형: $Y = m(X)$
  - Y: 반응변수 혹은 종속변수
  - X: 설명변수 혹은 독립변수
  - m: X로부터 Y를 설명하는 함수
- 회귀분석은 변수들간의 상관관계를 분석하기 위한 통계적 방법으로 위의 예에서 $m(X)$를 찾는 문제
- 선형회귀모형: Y와 X의 관계가 선형 즉 $m(X) = \beta_0 + \beta_1 X$로 가정한 모형
- 현실적으로는 X로부터 Y를 정확히 설명할 수 있는 모형은 존재할수 없지만 가장 잘 설명할수 있는 모형은 존재
- 선형 회귀분석 모형: $Y = \beta_0 + \beta_1 X + \epsilon$, ϵ: 랜덤 오차(random error)

회귀분석의 기본가정

- 독립변수와 종속변수 간의 선형적 관계를 가정한다.
- 오차의 등분산성과 정규성: 오차항(random error)이란 종속변수의 관측치와 모평균 간의 차이를 나타낸다. 이 오차항은 일정한 분산을 갖는 정규분포를 이룬다고 가정한다.
- 오차항의 독립성: 오차값들은 서로 독립이라고 가정한다.

단순회귀모형으로 이를 모형화 하면 아래와 같다.

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

여기서 $\epsilon_i \sim^{iid} N(0, \sigma^2)$ for $i = 1, \cdots, n$

- 단순선형회귀: 독립변수가 하나뿐인 모형
  즉, $Y = \beta_0 + \beta_1 X + \epsilon$
- 다중선형회귀: 독립변수가 두개이상인 모형
  즉, $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$
- 선형회귀분석에서 고려해야 할 사항들
  - 어떻게 β_0와 β_1을 찾을까?
  - 찾은 직선에 대하여 어떻게 통계적 해석을 해야 할까?
  - 찾은 직선을 얼마나 신뢰할수 있을까?
  - 가정된 회귀모형은 적절한가?
  - 설명변수가 많을 경우 어떻게 유용한 설명변수를 고를수 있을까?
  - 설명변수가 연속이 아닌 경우?
  - 반응변수가 연속이 아닌 경우?

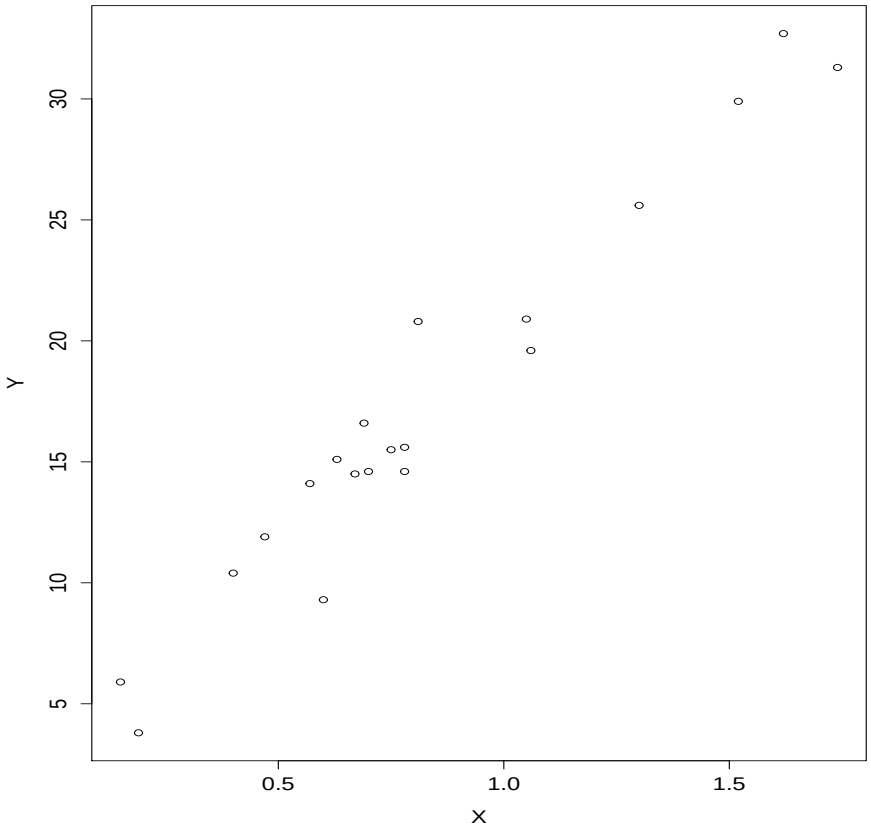
단순선형회귀모형

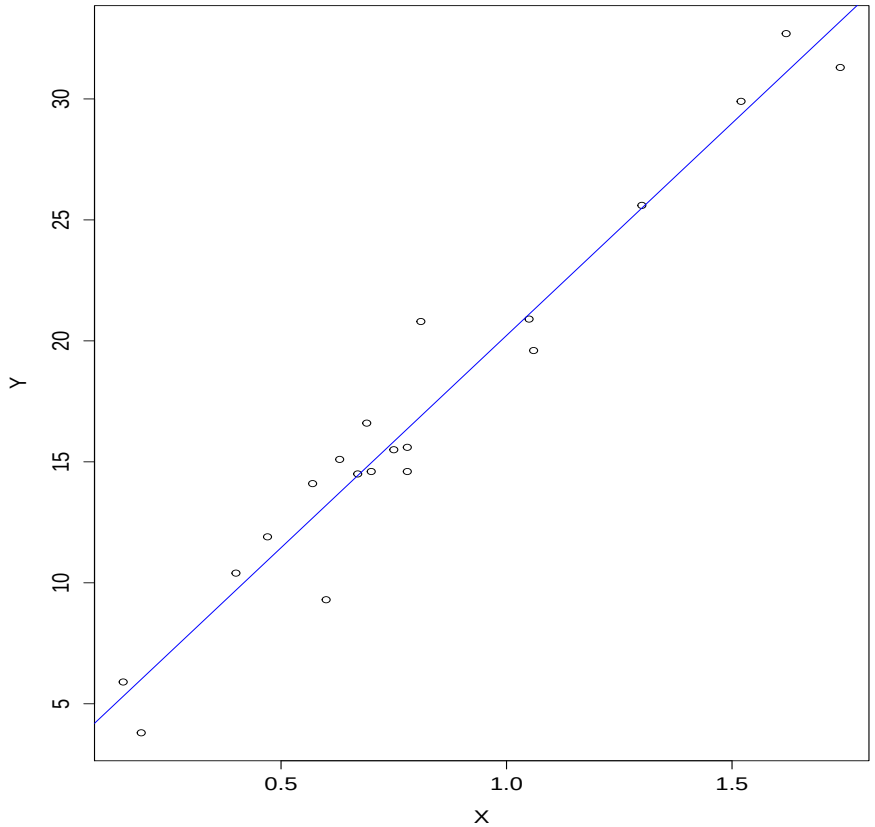
- X: 강유역에서 도로면적이 차지하는 비중
  Y: 강물의 염분농도

y	3.8	5.9	14.1	10.4	14.6	14.5	15.1	11.9	15.5	9.3
x	0.19	0.15	0.57	0.40	0.70	0.67	0.63	0.47	0.75	0.60
y	15.6	20.8	14.6	16.6	25.6	20.9	29.9	19.6	31.3	32.7
x	0.78	0.81	0.78	0.69	1.30	1.05	1.52	1.06	1.74	1.62

```
> x<-c(0.19, 0.15, 0.57, 0.4, 0.7, 0.67, 0.63, 0.47, 0.75, 0.6,
     0.78, 0.81, 0.78, 0.69, 1.3, 1.05, 1.52, 1.06, 1.74, 1.62)
> y<-c(3.8, 5.9, 14.1, 10.4, 14.6, 14.5, 15.1, 11.9, 15.5, 9.3,
     15.6, 20.8, 14.6, 16.6, 25.6, 20.9, 29.9, 19.6, 31.3, 32.7)
> plot(x,y) # 산점도 그리기
> lm(y~x)   # 선형모형 적합
> abline(lm(y~x),col="blue") # 산점도 위에 회귀직선 그리기
```

산점도와 회귀직선





최소제곱추정법

- 자료: $(X_1, Y_1), \ldots, (X_n, Y_n)$
- 최소제곱추정법: 직선과 Y_i의 수직 거리의 제곱합을 최소로 하는 직선을 찾는 방법
  다시 말해서,

$$L = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

  을 최소로 하는 β_0 와 β_1을 찾는다.
- 이렇게 얻어진 추정량을 최소제곱 추정량이라 부르고 이들을 $\hat{\beta}_0$와 $\hat{\beta}_1$으로 표기한다.

통계적 추론

- 예측값: $X = x$ 일 경우 Y의 예측값은 $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x$.
- 잔차: $e_i = Y_i - \hat{Y}_i \rightarrow$ 예측값과 실제값의 차
- 구한 직선이 얼마나 자료를 잘 설명하는가?
  $\Rightarrow$ 반응변수의 변동량중 회귀직선으로 설명되는 변동량이 얼마나 되는가?
- 반응변수의 변동량: $SST = \sum_i (Y_i - \bar{Y})^2$
- 회귀직선으로 설명되는 변동량: $SSR = \sum_i (\hat{Y}_i - \bar{Y})^2$
- 회귀직선으로 설명되지 않는 변동량: $SSE = \sum_i (Y_i - \hat{Y}_i)^2 = \sum_i e_i^2$
- R^2: 반응변수의 변동량중 회귀직선으로 설명되는 변동량의 비중

$$R^2 = \frac{SSR}{SST}$$

참고: SST=SSR+SSE

- X와 Y사이에 서로 통계적으로 유의한 선형관계가 있는가?
  $\Rightarrow H_0 : \beta_1 = 0$ versus $H_0 : \beta_1 \neq 0$
- 위 검정은 t-test를 이용하며 만약 t-test 결과 p-value가 유의수준 α(통상 0.05사용) 보다 작으면 귀무가설 H_0을 기각, 즉 $\beta_1 \neq 0$라는 것을 의미하여 통계적으로 유의한 선형관계가 있다고 판단
- 염분농도 예제에서

```
> result=lm(y~x)
> summary(result)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.6765     0.8680   3.084  0.00641 **
x            17.5467     0.9346  18.774 2.86e-13 ***
---
Residual standard error: 1.791 on 18 degrees of freedom
Multiple R-squared:  0.9514,  Adjusted R-squared:  0.9487
F-statistic: 352.5 on 1 and 18 DF,  p-value: 2.863e-13
```

다중선형회귀모형

- Data: $(x_{11}, \cdots, x_{p1}, y_1), \cdots, (x_{1n}, \cdots, x_{pn}, y_n)$
- Model: $y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} + \epsilon_i,\ i = 1, \cdots, n$
- 최소자승 추정량: 다음의 식을 최소화 하는 $\beta_0, \cdots, \beta_p$

$$L = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{1i} - \cdots - \beta_p x_{pi})^2$$

- β_j의 의미: X_j 이외의 다른 설명변수가 고정되어 있을경우에 X_j가 반응변수 Y에 주는 효과

예제: 다음은 연비(MPG), 차량무게(Weight), 주행거리(Odometer)에 관한 자료이다. 이때 MPG를 반응변수로 하는 선형 회귀모형을 고려해보자.

MPG	Weight	Odometer
7.28	10.5	15
5.63	23	71
5.26	27.5	36
6.58	14.5	113
5.01	30.5	39
6.73	14	97
5.37	21	195
7.28	8.5	8
4.85	26	84
5.08	26.5	25
5.51	15	124
4.75	30	25
6.03	15	75
5.26	22.5	192
5.6	16	139

```
> mpg=read.csv(file("e:/MPG.csv"),header=T) #자료읽어오기
> mpg_fit=lm(MPG~Weight+Odometer,mpg) #선형모형 적합하기
> summary(mpg_fit)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.240678   0.287141  28.699 1.99e-12 ***
Weight      -0.108257   0.011944  -9.064 1.02e-06 ***
Odometer    -0.003925   0.001406  -2.792   0.0163 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1

Residual standard error: 0.3182 on 12 degrees of freedom
Multiple R-squared:  0.8777,  Adjusted R-squared:  0.8573
F-statistic: 43.05 on 2 and 12 DF,  p-value: 3.35e-06
```

- 추정된 회귀식: MPG$= 8.24 - 0.11$Weight-0.003925Odometer
- $R^2 = 0.8777$
- 설명변수들의 유의성:
  - Weight와 MPG는 매우 유의한 선형관계 존재(p-value$<$0.0001)
  - Odometer와 MPG도 유의한 선형관계 존재 (p-value$=$0.0163)
- Weight$=$10이고 Odometer$=$30인 차량의 예측 연비(MPG)는?
  $\widehat{MPG} = 8.24 - 0.11 \times 10 - 0.003925 \times 30 \approx 7.0$

```
> newx=data.frame(Weight=10,Odometer=30)
> predict.lm(mpg_fit,newx,interval="confidence")
       fit      lwr      upr
1 7.040363 6.674768 7.405958
> predict.lm(mpg_fit,newx,interval="predict")
       fit      lwr      upr
1 7.040363 6.256505 7.824221
```

설명변수가 범주형일경우

- 설명변수중 하나가 "성별" 등과 같이 숫자 값을 가지는 것이 아니라 범주를 나타내는 경우에는 이를 가변수로 변환해주어야 한다.
- 예를 들어 설명변수 X가 "남자" 와 "여자" 의 값을 가지는 변수라면 이 변수를 다음과 같이 변환한다.

$$X^* = \begin{cases} 1 & \text{“남자”} \\ 0 & \text{“여자”} \end{cases}$$

- 설명변수가 3개 이상의 범주를 가질때(예: 중학생, 고등학생, 대학생)
  $\Rightarrow$ 전체범주수-1 의 가변수 생성 필요

$$X_1^* = \begin{cases} 1 & \text{“고등학생”} \\ 0 & \text{“고등학생이 아닌 경우”} \end{cases}, \quad X_2^* = \begin{cases} 1 & \text{“대학생”} \\ 0 & \text{“대학생이 아닌 경우”} \end{cases}$$

- 많은 소프트웨어들은 입력된 자료가 숫자가 아닌 경우 이를 자동으로 변환해주는데 결과 해석에 유의

- 예제
  - Y=칠면조 무게
  - X=나이(개월)
  - 생산지(Origin)=Georgia,Virginia,Wisconsin
  - 자료

X	Y	Origin	X	Y	Origin
28	13.3	G	26	11.8	V
20	8.9	G	21	11.5	W
32	15.1	G	27	14.2	W
22	10.4	G	29	15.4	W
29	13.1	V	23	13.1	W
27	12.4	V	25	13.8	W
28	13.2	V			

- R에서 범주형 설명변수 자동인식

```
> turkey=read.csv(file("e:/Turkey.csv"),header=T)
> fit1=lm(Y~X+Origin,turkey)
> summary(fit1)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.48750    0.67340  -0.724    0.487
X            0.48676    0.02574  18.908 1.49e-08 ***
OriginV     -0.27353    0.21844  -1.252    0.242
OriginW      1.91838    0.20180   9.506 5.45e-06 ***
---

Residual standard error: 0.3002 on 9 degrees of freedom
Multiple R-squared:  0.9794,    Adjusted R-squared:  0.9726
F-statistic: 142.8 on 3 and 9 DF,  p-value: 6.6e-08
```

- R에서 범주형 설명변수 수동변환

```
> turkey=read.csv(file("e:/Turkey.csv"),header=T)
> turkey$Vir=ifelse(turkey$Origin=="V",1,0)
> turkey$Wis=ifelse(turkey$Origin=="W",1,0)
> fit2=lm(Y~X+Vir+Wis,turkey)
> summary(fit2)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.48750    0.67340  -0.724    0.487
X            0.48676    0.02574  18.908 1.49e-08 ***
Vir         -0.27353    0.21844  -1.252    0.242
Wis          1.91838    0.20180   9.506 5.45e-06 ***
---
Residual standard error: 0.3002 on 9 degrees of freedom
Multiple R-squared:  0.9794,    Adjusted R-squared:  0.9726
F-statistic: 142.8 on 3 and 9 DF,  p-value: 6.6e-08
```

- 회귀계수의 해석:
  - $\hat{\beta}_1 = 0.49$: 나이가 1개월 증가하면 무게가 0.48파운드 증가
  - $\hat{\beta}_2 = -0.27$: Georgia산에 비해 Virginia산 칠면조의 무게가 0.27만큼 덜 나간다.
    하지만, $\hat{\beta}_2$은 통계적으로 유의하지 않기 때문에 Georgia산에 비해 Virginia산 칠면조의 무게는 통계적으로 차이가 있다고 보기 어렵다.
  - $\hat{\beta}_3 = 1.91$: Georgia산에 비해 Wisconsin산 칠면조의 무게가 1.91만큼 더 나간다.
- 주의: 대부분의 소프트웨어는 입력된 설명변수가 문자일 경우 자동으로 이를 가변수로 변환하여 적합하지만 만약 입력된 설명변수가 숫자일경우 프로그램은 연속형 변수로 간주
- 입력된 설명변수를 범주형으로 인식시키기 위해서는
  `fit1=lm(Y~X+Origin,turkey)` 대신
  `fit1=lm(Y~X+factor(Origin),turkey)` 사용

잔차분석 I

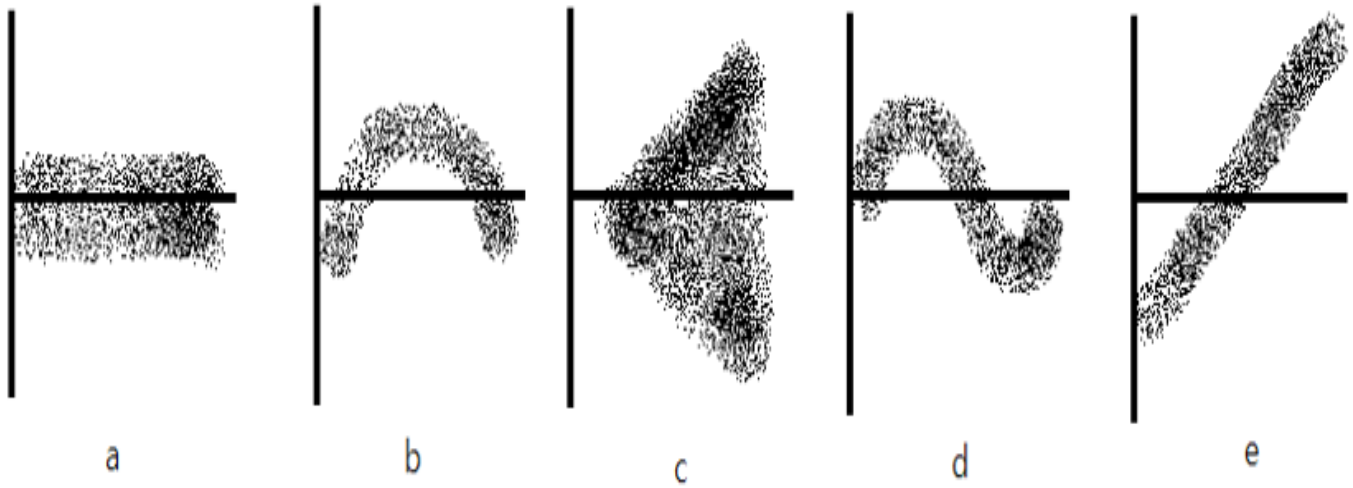
- 회귀분석의 기본가정

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \epsilon_i$$

  - 선형성: $E(Y_i|x_i) = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip}$
  - 등분산성: $var(\epsilon_1) = \cdots = var(\epsilon_n) = \sigma^2 > 0$
  - 독립성: $\epsilon_1, \cdots, \epsilon_n$은 서로 독립
  - 정규성:$\epsilon_i \sim^{iid} N(0, \sigma^2)$ for $i = 1, \cdots, n$

잔차분석 II

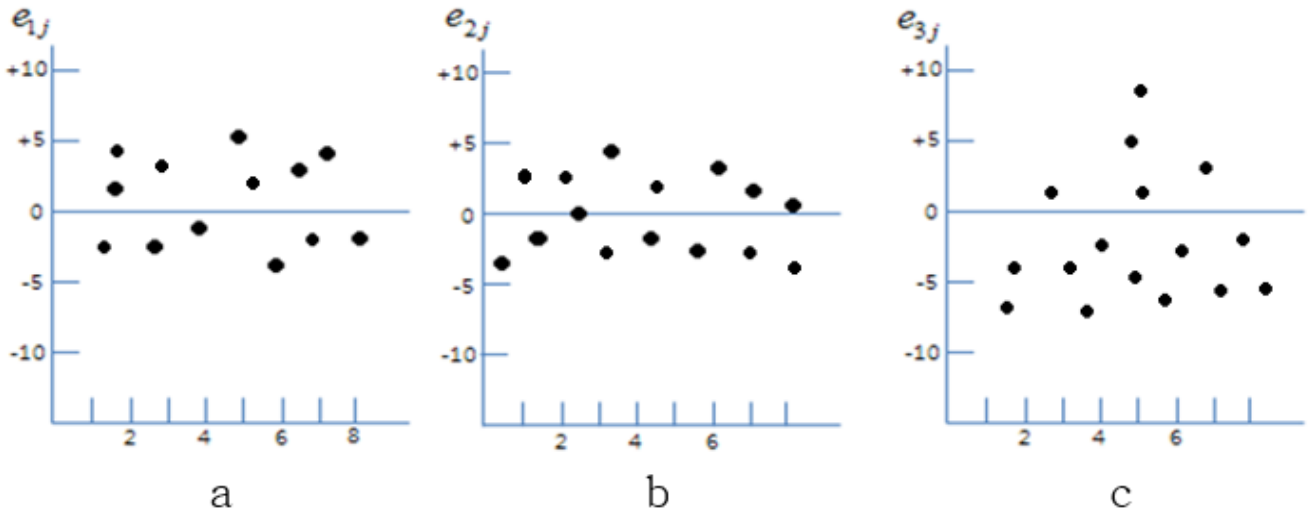
- 회귀함수의 선형성



(a) 잔차들이 0을 중심으로 랜덤하게 분포하고 있으므로 회귀모형이 타당하고 오차의 등분산성이 성립함을 알 수 있음; (b) 이차고선식이 타당해 보임; (c) 분산이 점점 증가하는 형태로 등분산성을 위배한 경우임; (d) 3차곡선식이 적절하다고 판단됨; (e) 잔차들이 계속 증가하고 있음

잔차분석 III

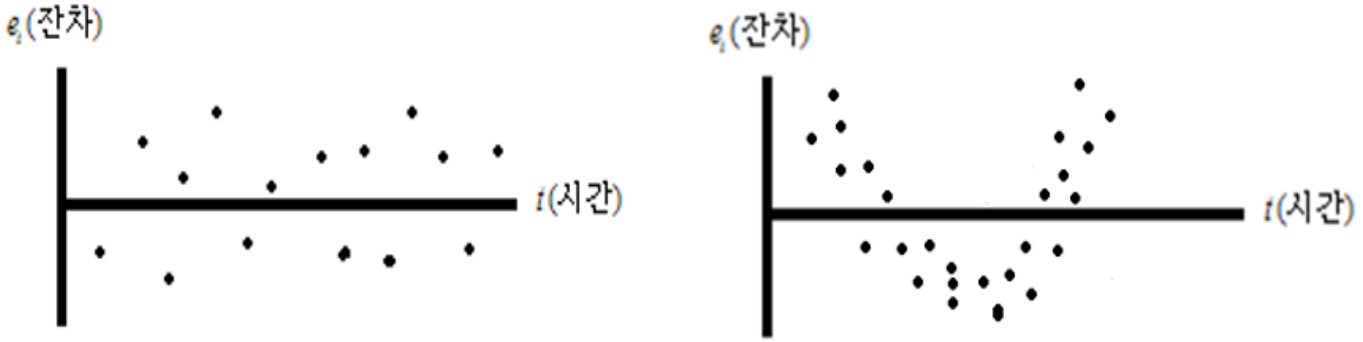
- 등분산성



(a)와 (b)의 경우 잔차값들이 0을 중심으로 흩어지는 경향이 비슷하므로 분산은 동일하다고 할 수 있음. 반면 (c)의 경우 오차의 분산이 등분산이 아닌 경우임.

잔차분석 IV

- 오차항의 독립



독립변수가 시간을 나타내는 변수일 때 또는 시간이 변수로서 직접 고려 된 것은 아니더라도 관측치 y가 시간에 영향을 받을 경우 시간에 따라 잔차를 그려본다. 오차항이 서로 독립이라면 잔차들은 0을 중심으로 램덤하게 변할 것이므로 위의 그림과 같이 만약 랜덤하게 변하지 않으면 오차항들이 서로 독립이라고 할 수 없을 것이다.

잔차분석 V

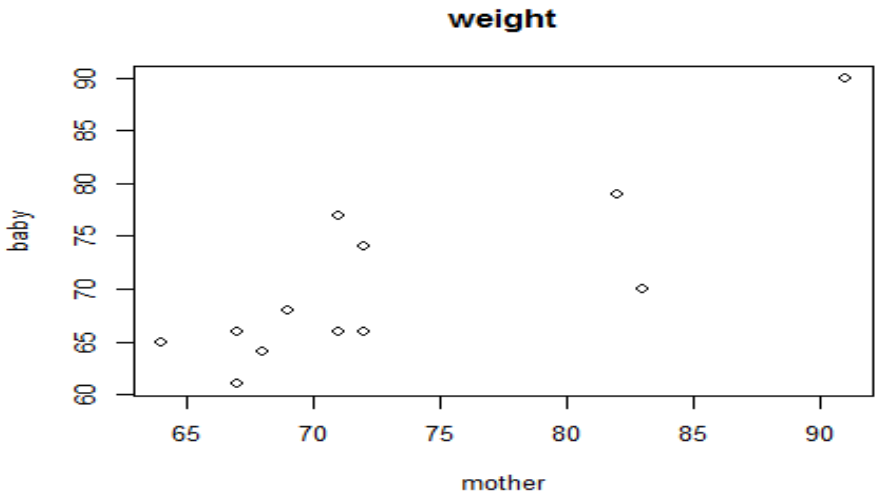
- 잔차를 이용한 가정체크: $e_i = y_i - \hat{y}_i$
  - 선형성 및 등분산성: 표준화잔차를 세로축으로 $\hat{y}_i$ 혹은 독립변수 x_i를 가로축으로 하여 그린 residual plot을 그려서 체크함.
  - 정규성: 잔차의 normal probability plot (Q-Q plot)에서 45도의 기울기를 가진 직선에 가까울수록 정규성을 따름을 알 수 있다.
  - 독립성: 더빈-왓슨(Durbin-Watson)통계량을 통하여 체크함(2에 가까우면 상관관계가 없음을 의미; 4 (0)에 가까우면 음(양)의 상관관계가 있음을 의미함).

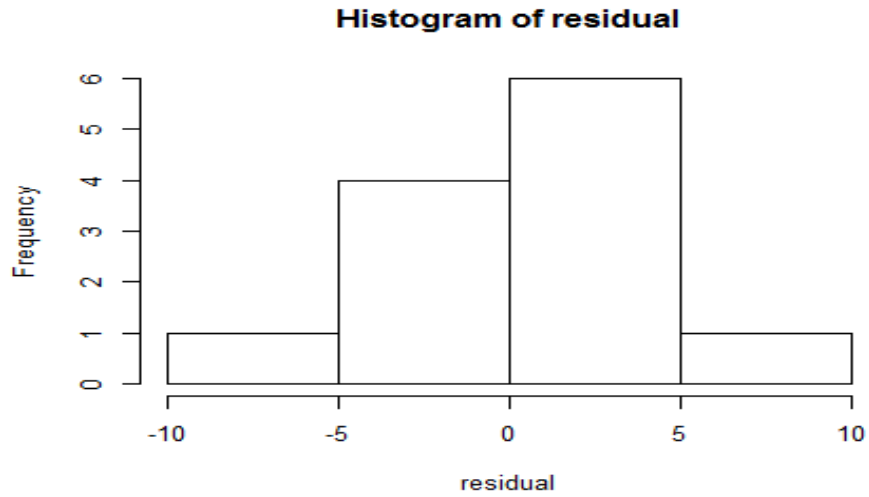
예제: 어미소와 새끼소

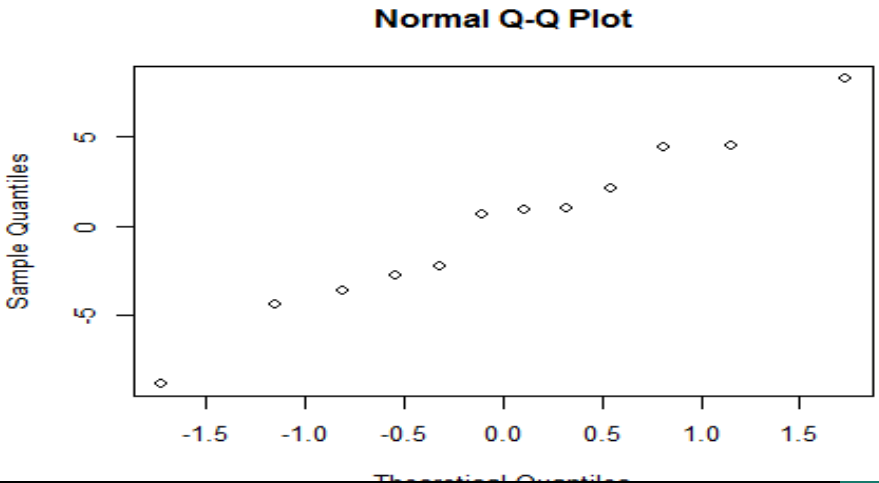
예제) 12마리의 어미소와 새끼소에 대한 몸무게(단위: 10kg)를 축정한 자료이다.

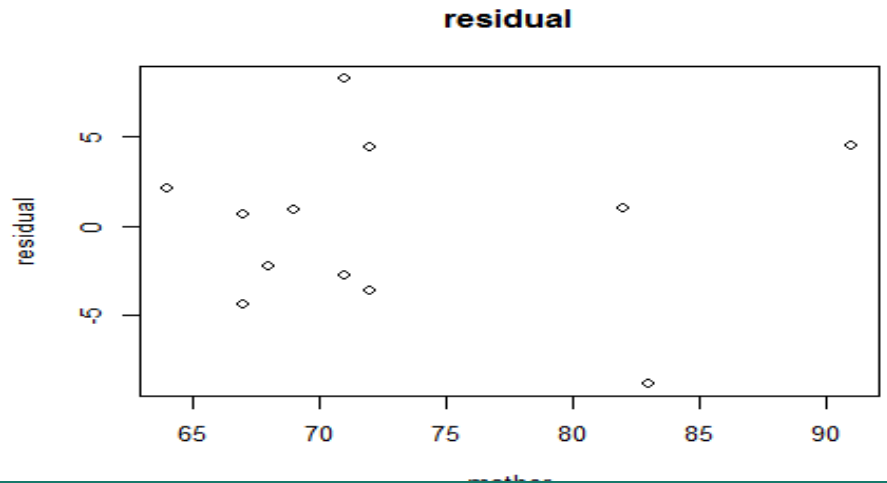
```
mother=c(71,67,83,64,91,68,72,82,71,67,69,72)
baby=c(77,66,70,65,90,64,74,79,66,61,68,66)
cow=lm(baby mother); plot(mother,baby);
abline(cow); summary(cow);
residual=resid(cow); par(mfrow=c(2,2));
plot(mother,baby,main="weight"); hist(residual);
qqnorm(residual); plot(mother,residual,
main="residual");
```

예제: 어미소와 새끼소


weight
baby
mother
60
65
70
75
80
85
90


Histogram of residual
Frequency
residual
0
1
2
3
4
5
6
-10
-5
0
5
10


Normal Q-Q Plot
Sample Quantiles
-5
0
5
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5


residual
residual
-5
0
5
65
70
75
80
85
90

다중공선성

- 다중공선성: 여러개의 설명변수중 하나의 설명변수가 다른 설명변수로 잘 설명이 되는 경우
- 직관적으로 다중공선성이 있다는 말은 설명변수중 불필요한 변수가 중복해서 있는 경우라고 할 수 있으며 이경우 모형적합시 많은 문제를 야기한다.
- 다중공선성이 있는 경우에 일어나는 현상
  - 추정된 회귀계수의 값이나 부호가 상식에서 벗어나는 경우
  - 중요하다고 생각되는 변수가 통계적으로 유의하지 않는 경우
  - 설명변수가 약간만 변해도 회귀계수가 크게 변하는 경우
  - 관측치가 하나 추가되거나 제거되었을때 회귀계수가 크게 변하는 경우

$\Rightarrow$ 이러한 문제는 모두 회귀계수의 추정이 불완전함에서 기인

예제

- 반응변수(ACHV): 학습성취도지수
- 설명변수:
  - FAM: 가정환경지수
  - PEER: 교우관계지수
  - SCHOOL: 학교환경지수
- Model: $ACHV = \beta_0 + \beta_1 FAM + \beta_2 PEER + \beta_3 SCHOOL$

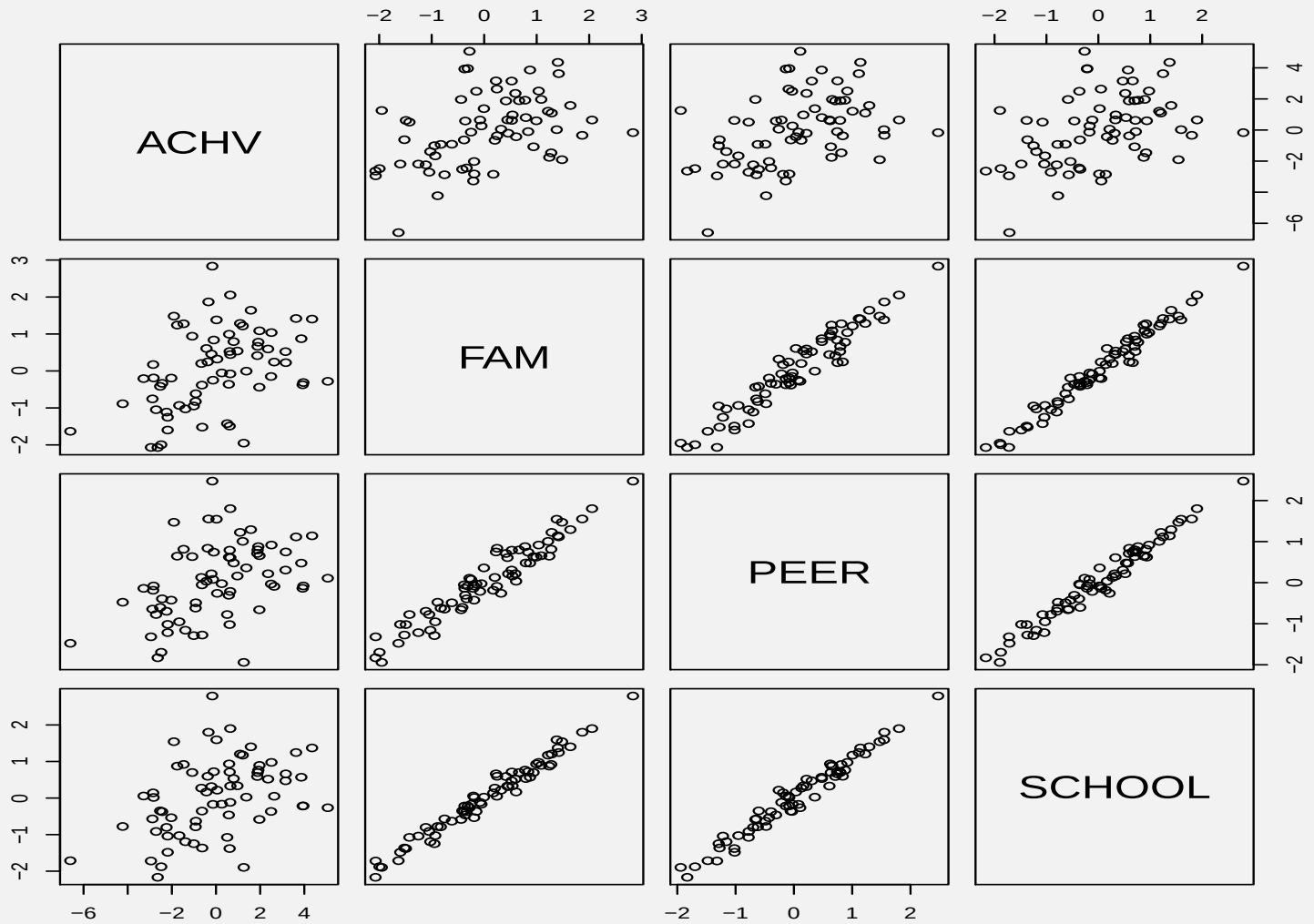
- R에서 자료적합

```
> edu=read.csv(file("e:/achv.csv"),header=T)
> fit=lm(ACHV~FAM+PEER+SCHOOL,edu)
> summary(fit)
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.06996    0.25064  -0.279    0.781
FAM          1.10126    1.41056   0.781    0.438
PEER         2.32206    1.48129   1.568    0.122
SCHOOL      -2.28100    2.22045  -1.027    0.308

Residual standard error: 2.07 on 66 degrees of freedom
Multiple R-squared: 0.2063,     Adjusted R-squared: 0.1702
F-statistic: 5.717 on 3 and 66 DF,  p-value: 0.001535
```

- 산점도 행렬

```
> plot(edu)
```



Variance Inflation Factor (VIF)

- 분산팽창계수(VIF): 하나의 설명변수가 다른 설명변수들에 의해 얼마나 잘 설명되는지를 나타내는 지표로 보통 10이상이면 다중공선성을 의심해볼 수 있다.

$$VIF_j = \frac{1}{1 - R_j^2}$$

여기서 R_j^2은 j번째 독립변수를 종속변수로 지정하고, 나머지를 독립변수로 하여 회귀모형을 적합하였을 때의 결정계수이다.

```
> install.packages("car")
> library("car")
> vif(fit)
     FAM     PEER   SCHOOL
37.58064 30.21166 83.15544
```

- 세변수 모두 분산팽창계수가 매우 크므로 다중공선성이 있다고 판단할 수 있다.

이상점, 지렛값, 영향점 I

- 이상점(Outlier)
  - 표준화잔차의 값이 특정한 범위(± 3.0)을 벗어나면 이상점이 아닌가를 의심할 수 있다.
  - 이상점은 모형이 이상점에 대하여 적절하지 않을 수 있음을 의미한다.
- 지렛값(Leverage)
  - 한 관찰치의 설명변수값이 다른 관찰치로부터 얼마나 떨어져 있는지를 나타내는 지표이다.
  - 큰 지렛값은 회귀식에 영향을 미칠 수 있으므로 자세히 검토해야 한다.
  - 지렛값이 전체지렛값 평균($(p+1)/n$)의 두배이상인 경우 high leverage point 의심한다.

이상점, 지렛값, 영향점 II

- 영향점(Influential observation): 한 관찰값을 모형에 넣었을 때와 그렇지 않았을 때에 추정된 회귀계수의 차이가 큰 관측값
  $\Rightarrow$ Cook's distance가 1이상인 경우 의심

- 표준화잔차의 계산 (연비예제)

```
> rstudent(mpg_fit)
          1           2           3   ......
 0.87535315  0.50105064  0.46100885   ......
```

- 지렛값의 계산 (연비예제)

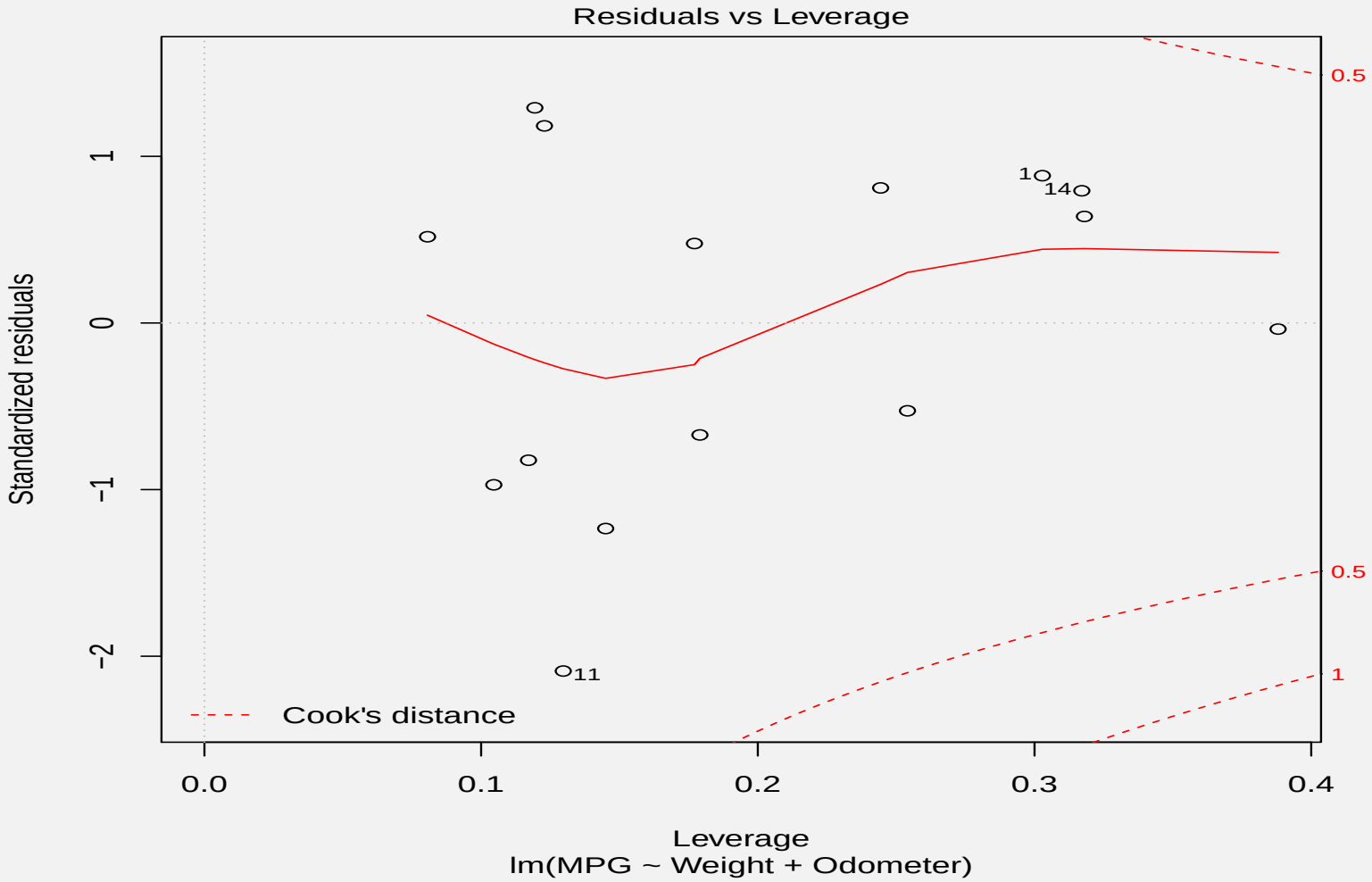
```
> hatvalues(mpg_fit)
         1          2          3   ......
0.30284911 0.08063975 0.17709697   ......
> mean(hatvalues(mpg_fit))
[1] 0.2
```

- Cook's distance 계산 (연비예제)

```
> cooks.distance(mpg_fit)
           1            2            3   ......
0.1131587866 0.0078287708 0.0163168902   ......
```

- 그래프로 한번에 보기

```
> plot(mpg_fit, which=5)
```



변수선택

- 변수 선택이 필요한 이유
  - 빅데이터 시대⇒ 너무 많은 수의 변수
  - 불필요한 변수의 추가시 다중공선성의 가능성이 높음
  - 일반적인 통계분석에서는 (불필요한)변수가 많아짐에 따라 추정의 정밀도는 낮아짐
  - 변수가 너무 많은 복잡한 모형은 해석이 어려움
- 전진선택법(Forward selection): 전체 변수중에서 가장 분석에 유용한 변수를 하나씩 차례로 추가하는 방법
- 후진제거법(Backward elimination): 설명변수 전체를 사용한 모형에서 하나씩 가장 불필요한 변수를 제거해 나가는 방법
- 단계적방법(Stepwise): 전진선택법이나 후진 선택법에서는 한번 추가하거나 제거된 변수는 다시 모형에 들어갈수 없는 단점이 있다. 이를 보완하기 위해 전진선택법과 후진제거법을 차례로 적용해서 변수를 선택하는 방법

학업성취도 자료에서의 변수선택

- 가장 큰 모형과 가장 작은 모형적합

```
> null=lm(ACHV~1,edu)
> full=lm(ACHV~.,edu)
```

- 전진선택법

```
> step(null,scope=list(lower=null,upper=full),
      direction="forward")
```

- 후진제거법

```
> step(full,scope=list(lower=null,upper=full),
      direction="backward")
```

- 전진선택법

```
> step(null,scope=list(lower=null,upper=full),
      direction="both")
```

반응변수가 범주형일 경우의 회귀모형

- 방응변수가 범주형인 예
  - Y: 암 발병여부, X_1: 나이, X_2: 성별, X_3: 흡연여부
  - Y: 시험 합격여부, X_1: 나이, X_2: 성별, X_3: TOEIC 점수
- 이러한 경우 Y를 다음과 같이 정의한다.

$$Y = \begin{cases} 1 & \text{for “암발병” 혹은 “합격”} \\ 0 & \text{for “암발병하지 않은 경우” 혹은 “불합격”} \end{cases}$$

- 모형 1: $Y_i = \beta_0 + \beta_1 X_i$?
- 모형 2: $P(Y_i = 1|X_i) = \beta_0 + \beta_1 X_i$?
- 모형 3: $P(Y_i = 1|X_i) = \frac{\exp(\beta_0+\beta_1 X_i)}{1+\exp(\beta_0+\beta_1 X_i)}$?
  $\Rightarrow$ 로지스틱 회귀모형

로지스틱 회귀모형과 오즈

- 로지스틱 회귀모형

$$P(Y_i = 1|X_i) = \frac{\exp(\beta_0+\beta_1 X_i)}{1+\exp(\beta_0+\beta_1 X_i)}$$

$$\Leftrightarrow \log\left(\frac{P(Y_i=1|X_i)}{1-P(Y_i=1|X_i)}\right) = \beta_0 + \beta_1 X_i$$

- 오즈(Odds): $\frac{P(Y_i=1|X_i)}{1-P(Y_i=1|X_i)}$: 실패할 확률에 대한 성공확률의 비
- Odds의 성질: 0부터 ∞ 사이의 값을 가지고 성공확률이 클수록 큰 값을 가진다.
- 로그오즈(log odds): $-\infty$와 ∞사이의 값을 가지며 성공확률이 클수록 큰 값을 가진다.
- 로지스틱 회귀모형: 로그오즈에 대한 선형회귀모형

- 예제: 심장질환과 흡연과의 관계

	심장질환 있음	심장실환 없음	합
흡연자	84	2916	3000
비흡연자	87	4913	5000

- 흡연자의 오즈 : 84/2916=0.0288
  비흡연자의 오즈 : 87/4913=0.0177
- 비흡연자에 대한 흡연자의 오즈비 (Odds ratio):
  $\frac{84/2916}{87/4913} = 1.627$
- 해석: 흡연자의 오즈가 비흡연자의 오즈보다 1.627 배 크다.

예제

- 예제: 대학원 입시 결과
- 반응변수(admit): 합격여부 (합격: 1, 불합격: 0)
- 설명변수: 대학원 입학시험점수(GRE), 학점(GPA), 출신학교등급(rank)
- 자료의 형태는 다음과 같으며 총 400명에 해당되는 자료

	admit	gre	gpa	rank
1	0	380	3.61	3
2	1	660	3.67	3
3	1	800	4.00	1
4	1	640	3.19	4
5	0	520	2.93	4
6	1	760	3.00	2

- R을 통한 로지스틱 회귀모형

```
> mydata=read.csv("./데이터파일/admit.csv")
> fit=glm(admit~gre+gpa+factor(rank),data=mydata,family="binomial")
> summary(fit)

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)   -3.989979   1.139951  -3.500 0.000465 ***
gre            0.002264   0.001094   2.070 0.038465 *
gpa            0.804038   0.331819   2.423 0.015388 *
factor(rank)2 -0.675443   0.316490  -2.134 0.032829 *
factor(rank)3 -1.340204   0.345306  -3.881 0.000104 ***
factor(rank)4 -1.551464   0.417832  -3.713 0.000205 ***
```

주의: rank는 범주형 자료이므로 반드시 가변수로 변환을 하거나 `factor(rank)`를 사용

- 반응변수의 예측값 구하기

$$\hat{Y}_i = \begin{cases} 1, & \text{if } \hat{P}(Y_i = 1|X_i) > a \\ 0, & \text{if } \hat{P}(Y_i = 1|X_i) \leq a \end{cases}$$

a는 보통 0.5를 사용하지만 경우에 따라 다른 값을 설정하기도 한다.

- 예측값과 실제값의 비교

```
> pred_y=ifelse(fit$fit>0.5,1,0)
> table(mydata$admit,pred_y,
          dnn=c("Observed","Predicted"))
          Predicted
Observed    0   1
       0  254  19
       1   97  30
```

- 정분류율: $P(\hat{Y} = Y) = 284/400 = 0.71 \rightarrow$ 예측의 정확도

```
> mean(mydata$admit==pred_y)
[1] 0.71
```

- 오분류율: $P(\hat{Y} \neq Y) = 116/400 = 0.29$
- Sensitivity=$P(\hat{Y} = 1|Y = 1) = 30/127 = 0.11$: 실제 합격인 사람을 합격으로 예측할 확률
- Specficity=$P(\hat{Y} = 0|Y = 0) = 254/273 = 0.93$: 실제 불합격인 사람을 불합격으로 예측할 확률

R을 통한 예측 합격확률 계산: GRE점수 500, GPA=3.25, rank=1인 학생이 합격할 예측확률과 95% 신뢰구간 구하기

- 설명변수 만들기

```
> new=data.frame(gre=500,gpa=3.25,rank=1)
```

- 예측 로그오즈 계산, 즉, $\log\left(\frac{P(Y_i=1|X_i)}{1-P(Y_i=1|X_i)}\right)$

```
> predict(fit,newdata=new,type="link")
-0.2446441
```

- admit=1일 확률 즉, 합격할 확률과 95% 신뢰구간 계산하기

```
> pred=predict(fit,newdata=new,type="response",se.fit=TRUE)
> pred$fit
0.4391422
> pred$se.fit
0.07099108
> c(pred$fit-1.96*pred$se.fit,pred$fit+1.96*pred$se.fit)
0.1807352 0.5247743
```

다항 로지스틱 모형

- 반응변수가 세개이상의 범주를 가지는 경우, 예를 들어 앞의 예제에서 반응변수가 합격, 불합격, 예비합격의 범주를 가지는 경우
- 반응변수가 세개의 범주를 가지는 경우

$$Y = \begin{cases} 0 & \text{“불합격”} \\ 1 & \text{“합격”} \\ 2 & \text{“예비 합격”} \end{cases}$$

- 다항 로지스틱 모형:

$$\log\left(\frac{P(Y_i = 1|X_i)}{P(Y_i = 0|X_i)}\right) = \beta_{01} + \beta_{11}X_i$$
$$\log\left(\frac{P(Y_i = 2|X_i)}{P(Y_i = 0|X_i)}\right) = \beta_{02} + \beta_{12}X_i$$

$\Rightarrow$ 즉, 두개의 모형이 필요하고 이때 기준이 되는 $Y_i = 0$에 해당되는 범주를 baseline-category라고 부른다.

- 앞의 두식을 연립해서 풀면 다음과 같이 각 범주에 속할 확률을 계산할수 있다.

$$P(Y_i = 0|X_i) = \frac{1}{1 + \exp(\beta_{01} + \beta_{11}X_i) + \exp(\beta_{02} + \beta_{12}X_i)}$$

$$P(Y_i = 1|X_i) = \frac{\exp(\beta_{01} + \beta_{11}X_i)}{1 + \exp(\beta_{01} + \beta_{11}X_i) + \exp(\beta_{02} + \beta_{12}X_i)}$$

$$P(Y_i = 2|X_i) = \frac{\exp(\beta_{02} + \beta_{12}X_i)}{1 + \exp(\beta_{01} + \beta_{11}X_i) + \exp(\beta_{02} + \beta_{12}X_i)}$$

당뇨병 자료 예제

- 반응변수(CC): 당뇨병에 대한 임상적 분류 (1: 당뇨병, 2: 준당뇨병, 3: 정상)
- 설병변수: 인슐린반응(IP), 인슐린저항성(SSPG), 상대체중(RW)

```
> install.packages("nnet")
> library("nnet")
> diabetes=read.csv(file("e:/diabetes.csv"),header=T)
> fit=multinom(CC~RW+IR+SSPG,diabetes)
```

```
> summary(fit)
Call:
multinom(formula = CC ~ RW + IR + SSPG, data = diabetes)

Coefficients:
  (Intercept)       RW          IR        SSPG
2   -5.771730 9.341177 0.01694102 -0.02909055
3    1.844214 5.867834 0.01335416 -0.04550422

Std. Errors:
  (Intercept)       RW          IR        SSPG
2    1.551814 1.364353 0.004807741 0.008152940
3    1.401596 1.572040 0.004943845 0.008904468

Residual Deviance: 136.8293
AIC: 152.8293
```

- 각분류에 대한 예측확률

```
> predict(fit,type="probs")
                1             2             3
1   3.074783e-03 3.052557e-02 9.663996e-01
2   3.717075e-03 6.579498e-02 9.304879e-01
3   9.896663e-03 1.066159e-01 8.834874e-01
4   2.764630e-03 1.793762e-01 8.178591e-01
5   8.652487e-03 2.795528e-01 7.117948e-01
```

- 예측 분류

```
> pred_CC=predict(fit,type="class")
> pred_CC
  [1] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 2
 [27] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 1 3
 [53] 3 3 3 3 3 3 3 3 3 2 2 3 2 3 3 1 2 3 2 3 3 3 2 3 3 3
 [79] 3 3 3 2 2 2 3 2 2 3 2 2 2 2 2 2 2 3 2 2 2 2 2 2 2 2
[105] 3 2 3 3 2 3 3 3 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
[131] 2 1 1 3 3 2 2 1 1 1 1 3 1 1 1
```

- 예측값과 관측값의 비교(정분류율=(27+24+69)/145=82.8%)

```
> table(diabetes$CC,pred_CC,dnn=c("Observed","Predicted"))
        Predicted
Observed  1  2  3
       1 27  3  3
       2  0 24 12
       3  2  5 69
```

주의1: 상관관계? 인과관계?

주의 1: 설명변수 X가 반응변수 Y를 잘 설명한다는 말은 둘 사이의 밀접한 관계를 이야기 하는 것이지 둘 사이의 인과관계를 나타내는 것은 아니다.

- 예제: X=화재현장에 출동한 소방관수 Y=화재 피해액
- 양의 상관관계 혹은 음의 상관관계? $\Rightarrow$ 양의 상관관계
- 출동한 소방관수가 많을수록 화재피해액이 많다?
  $\Rightarrow$ 잘못된 인과관계 추론
- 이경우 제 3의 요인이 존재할 가능성이 있음

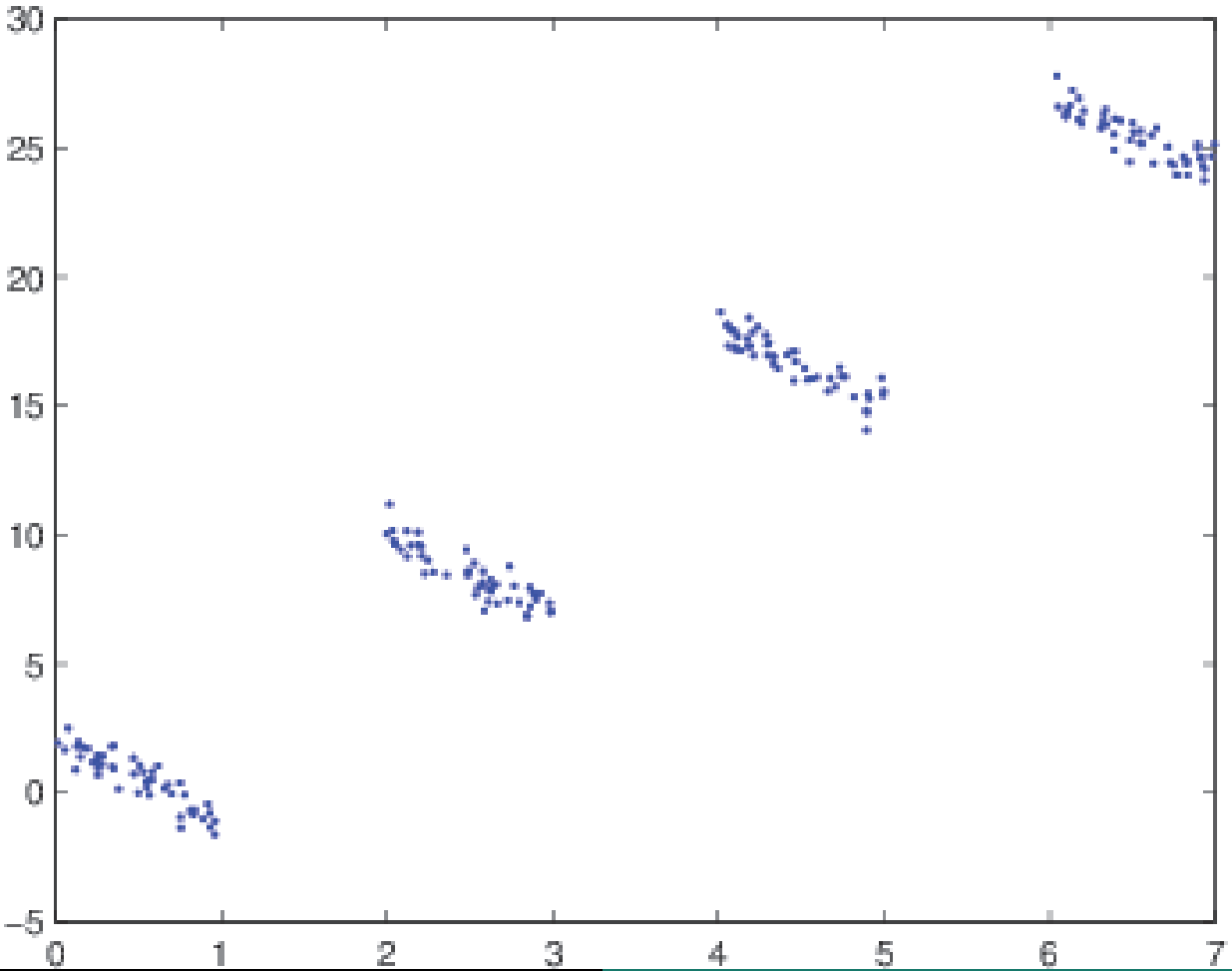
주의 2: 잠재변수

주의 2: 잠재변수의 존재 가능성

- 잠재변수: 관측되지 않은 숨겨진 변수로 관측된 변수들의 연관관계에 중요한 영향을 끼지는 변수
- 예제1: 아이스크림 판매량과 익사사고:
  $\rightarrow$ 잠재변수: _____
- 예제3: 독해력과 발크기:
  $\rightarrow$ 잠재변수: _____

주의 3: Simpson's Paradox

주의 3: 두변수의 상관관계의 방향성이 제3의 변수가 추가한후 달라질수 있다.



주의 4: 과적합

- 앞의 연비 예제에서 $R^2 = 0.8777$로 비교적 높게 나왔다. 하지만 갑은 R^2가 0.9이상이 나오기를 원해서 이를 통계전문가에게 의뢰한 결과 다음과 같은 결과를 주었다.

```
mpg$Weight2=mpg$Weight^2
mpg$Odometer2=mpg$Odometer^2
mpg$Weight3=mpg$Weight^3
mpg$Odometer3=mpg$Odometer^3
mpg_fit=lm(MPG~Weight+Odometer+Weight2+Odometer2+Odometer3
+Weight3+Weight*Odometer+Weight*Odometer2+Weight*Odometer3
+Weight2*Odometer+Weight2*Odometer2+Weight2*Odometer3
+Weight3*Odometer,mpg)
summary(mpg_fit)
```

주의 4: 과적합

```
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)       -6.001e+00  1.753e+01  -0.342    0.790
Weight             2.163e+00  4.036e+00   0.536    0.687
Odometer           1.240e+00  1.576e+00   0.787    0.576
Weight2           -7.367e-02  2.183e-01  -0.337    0.793
Odometer2         -1.779e-03  2.101e-02  -0.085    0.946
Odometer3         -7.146e-05  1.116e-04  -0.640    0.637
Weight3            3.193e-04  3.635e-03   0.088    0.944
Weight:Odometer   -2.157e-01  1.726e-01  -1.250    0.430
Weight:Odometer2   1.078e-03  1.962e-03   0.549    0.680
Weight:Odometer3   3.517e-06  8.836e-06   0.398    0.759
Odometer:Weight2   8.454e-03  5.500e-03   1.537    0.367
Weight2:Odometer2 -3.691e-05  4.701e-05  -0.785    0.576
Weight2:Odometer3 -4.195e-08  1.909e-07  -0.220    0.862
Odometer:Weight3  -8.031e-05  5.274e-05  -1.523    0.370
```

주의 4: 과적합

```
Residual standard error: 0.3977 on 1 degrees of freedom
Multiple R-squared:  0.9841,    Adjusted R-squared:  0.7771
F-statistic: 4.755 on 13 and 1 DF,  p-value: 0.3459
```

Are you happy, now?

실습예제 1: R 기초

1. 5, 23, 6, 10, 15, 4를 벡터형태로 변수명 test로 저장하시오
2. test의 평균을 구하시오
3. test안의 몇번째 숫자가 10이상인지 찾으시오
4. test안의 숫자를 큰 순서부터 나열하시오

실습예제 2: R 함수 만들기

1. 숫자 두개 x와 y를 입력하면 (x+y)/x를 return해주는 함수를 만드시오.
2. 벡터를 입력하면 그 벡터의 숫자중 10이하인 것들의 합을 구해주는 함수를 만드시오.
3. 벡터를 입력하면 그 벡터안의 숫자중 3의 배수인 숫자들의 최대값을 구해주는 함수를 만드시오.

실습예제 3: 회귀분석

- 자료: 1975년 미국 주별 공교육 지출액자료: “education75.csv”
- 반응변수(Y): 일인당 공교육지출액
- 설명변수:
  - X1: 일인당 소득
  - X2: 1000명당 18세미만 인구수
  - X3: 1000명당 도심거주 인구수
  - Region: 지리적 지역(1: 북동부, 2:북중부, 3: 남부, 4: 서부)

1. 모형을 적합시키고 해석하시오.
2. 영향점이 있는지 판단하시오.
3. 다중공선성이 있는지 판단하시오.
4. 남부지역에 있는 일인당 소득이 4200이고 1000명당 18세미만 인구수가 300명이며 1000명당 도심거주인구수가 645명인 주의 일인당 공교육지출액을 구하고 예측신뢰구간을 구하시오.

실습예제 4: 로지스틱 회귀

- 자료: 기업파산자료 "bankruptcy.csv"
- 반응변수(Y): 파산여부(0: 2년뒤 도산, 1: 정상)
- 설명변수:
  - X1: 총보유수입/총자산
  - X2: 세전총수입/총자산
  - X3: 총매출/총자산

1. 모형을 적합시키고 해석하시오.
2. 정분류율을 계산하시오.
3. X1=40, X2=6.0, X3=1.8인 기업이 2년뒤 도산할 확률과 95% 신뢰구간을 계산하시오

Thank you!