R을 이용한 기초통계학

8강: 회귀분석

수원대학교
데이터과학부
김 진 흠

서울대학교
보건환경연구소
이 보 라

The 19th ASIAN INSTITUTE IN STATISTICAL GENETICS AND GENOMICS

INTRODUCTORY STATISTICS USING R

회귀분석

- 특정 현상과 이에 영향을 미치는 변수 간의 함수 관계를 이론적 근거나 경험적 판단에 따라 **모형화**하고, 관측 자료를 이용하여 이를 **추정·예측하는** 통계 분석 방법

회귀분석의 변수

- 종속변수(dependent variable)
  - 관심의 대상이 되는 특정한 현상을 나타내는 변수. 반응변수(response variable) 또는 결과변수(outcome variable)라 하며 일반적으로 Y로 표시
- 독립변수(independent variable)
  - 종속변수에 영향을 미칠 수 있는 변수. 설명변수(explanatory variable) 또는 예측변수(predictor)라 하며 $x_1, x_2, \ldots$ 등으로 표시

사례

- 젖소를 대상으로 **출산 경험**에 따라 수유 기간 동안 생산되는 **우유량**과 **수유 기간**의 관계를 예측하는 연구를 실시
- 자료 출처: Wood(Nature, 1967)

우유량(Y)	수유기간(x_1)	출산경험(x_2)
27.4	23	1
30.5	54	1
⋮	⋮	⋮

Wood(1967) 자료 플롯

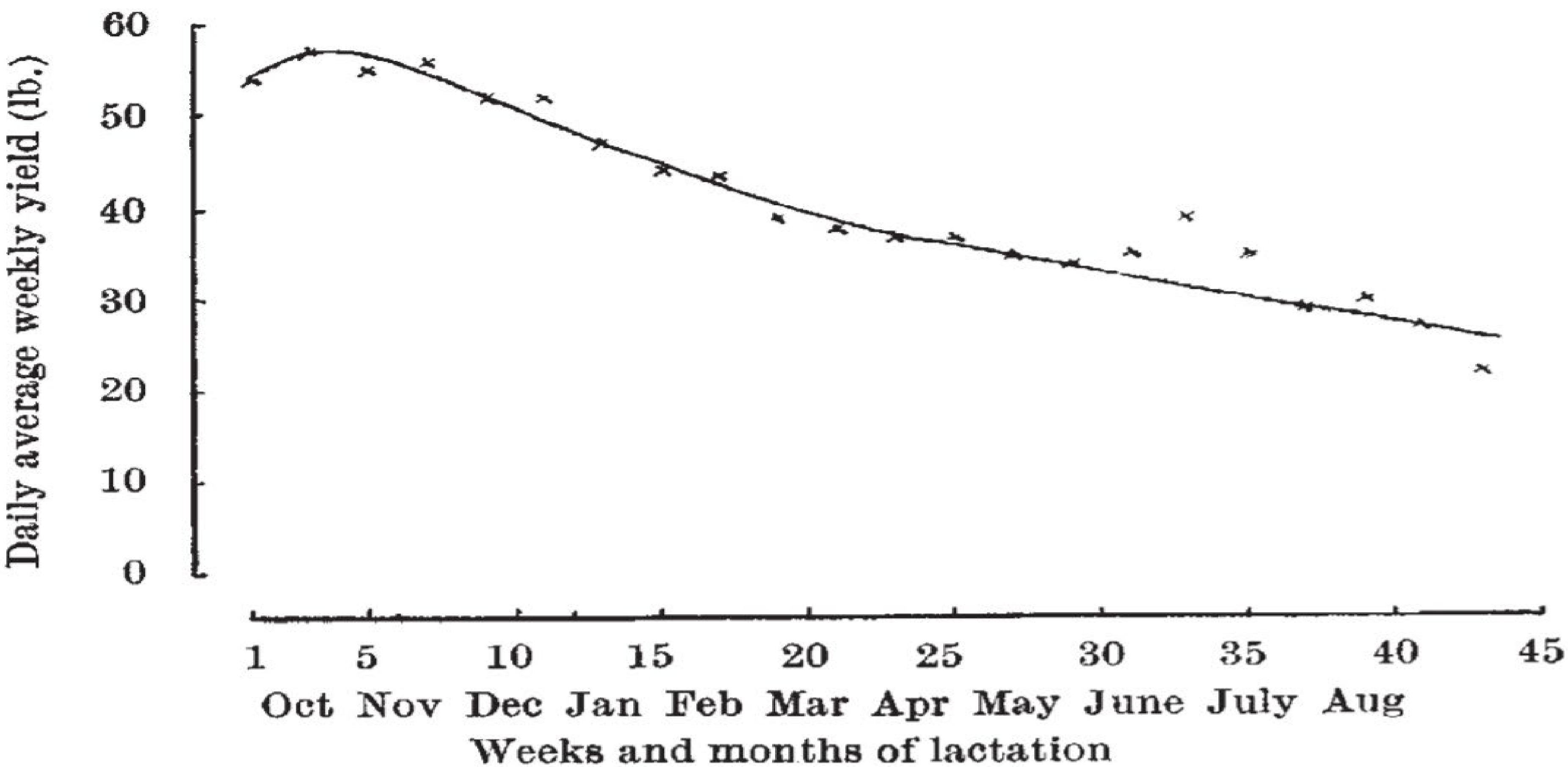


Fig. 1. Regression curve $y = 56{\cdot}62\, n^{0{\cdot}03996} \exp(-0{\cdot}00942\, n)$ fitted to a single Friesian lactation.

Wood, P.D.P (1967) 자료 플롯

- Wood function: $Y = ax^b e^{-cx}$
  - a: the constant level of initial yield of the buffalo milk
  - b: the rate of increase to peak
  - c: the rate of decline after peak
- Transformation: $\log Y = \log a + b \log x - cx$

단순선형회귀모형

- 단순선형회귀모형: 독립변수가 **한** 개인 경우
- 모형의 정의
  - $Y_i = \beta_0 + \beta_1 \boldsymbol{x_i} + \epsilon_i,\ \epsilon_i \sim iid\ N(0, \sigma^2)$ 혹은
  - $E(Y_i) = \beta_0 + \beta_1 x_i$
- **가정**
  - 정규성(normality)
  - 독립성(independence)
  - 등분산성(equal variance)

모형의 적합: 최소제곱법

- 오차의 제곱합을 최소화 하는 β_0와 β_1을 구하는 추정법
- $S = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i)^2$ 을 최소화 하는 β_0와 β_1을 구함
- 최소제곱해
  - $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$
  - $\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$

모형의 적합: 추정 회귀직선

- 적합된 회귀직선은 $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$
- 사례분석
  - 출산경험=1 인 경우: $\hat{y} = 31.083 - 0.021x$
  - 출산경험=2 인 경우: $\hat{y} = 42.354 - 0.061x$
  - 출산경험=3 인 경우: $\hat{y} = 45.556 - 0.080x$

회귀직선의 유의성 검정: F−검정

- 분산분석표

Source	Sum of Square	df	Mean Square	F−값
Regression	SSR = $\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$	1	MSR	MSR/MSE
Residual	SSE = $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$	$n-2$	MSE	
Total	SST = $\sum_{i=1}^{n}(y_i - \bar{y})^2$	$n-1$		

- F−검정법
  - $F = \dfrac{\mathrm{MSR}}{\mathrm{MSE}}$
  - P−값: $P(F > f_0 | H_0), F \sim F(1, n-2)$

결정계수

- $R^2 = \frac{\text{SSR}}{\text{SST}} = \frac{\text{회귀모형에 의한 변동}}{\text{반응변수의 총변동}}$
- 모형이 자료의 변동성을 얼마나 잘 설명하는지를 나타내는 척도.
- $0 < R^2 < 1$
  - $R^2 \approx 1 \Rightarrow$ 적합한 회귀모형 (설명력 높음)
  - $R^2 \approx 0 \Rightarrow$ 부적합한 회귀모형 (설명력이 낮음)

결정계수: 사례

- 출산경험 = 1 인 경우: $R^2 = 0.601$
- 출산경험 = 2 인 경우: $R^2 = 0.851$
- 출산경험 = 3 인 경우: $R^2 = 0.898$

회귀직선의 유의성 검정: t–검정

- Hypothesis: $H_0 : \beta_1 = 0$ 대 $H_0 : \beta_1 \neq 0$
- Test statistic: $T = \dfrac{\hat{\beta}_1 - \mathbf{0}}{\sqrt{\dfrac{\mathrm{MSE}}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}}$
- P–값: $P(|T| > |t_0||H_0), T \sim t(n-2)$

회귀직선의 유의성 검정: 사례

출산경험	t−검정	F−검정	P−값	결정
1	−3.238	10.485	0.014	H_0을 기각
2	−6.770	45.838	<0.001	H_0을 기각
3	−8.372	70.088	<0.001	H_0을 기각

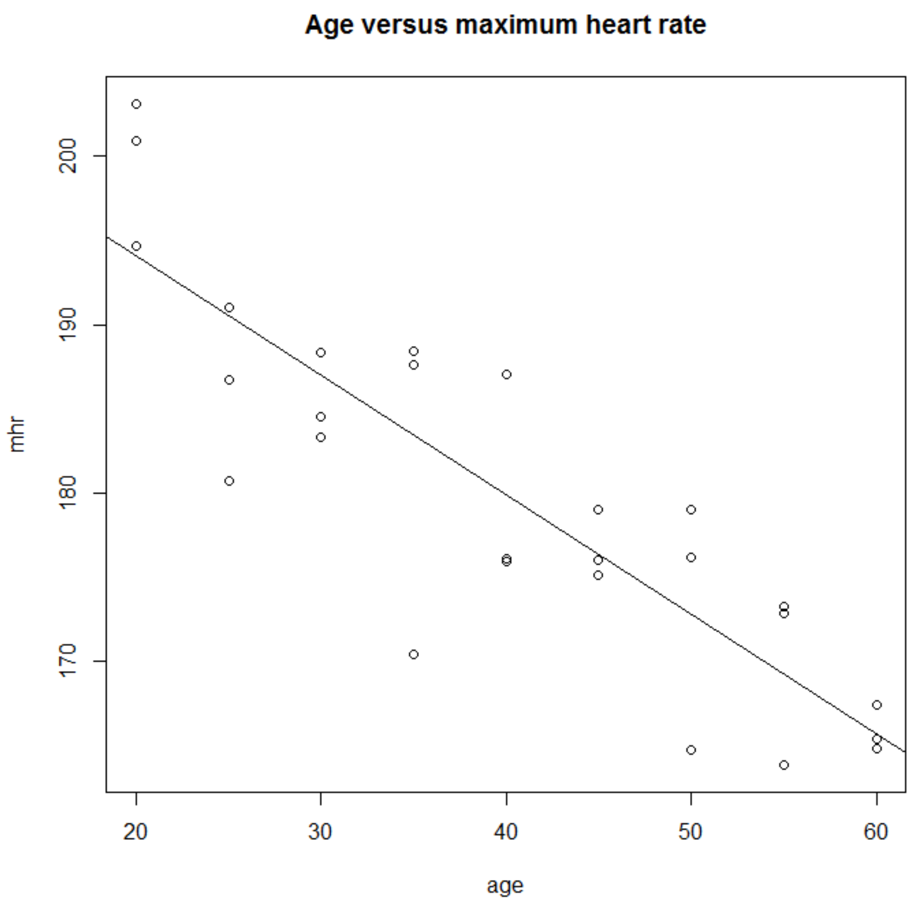
잔차분석

- 모형을 적합한 후, 회귀모형의 가정인 ① 정규성 ② 독립성 ③ 등분산성이 크게 위배되지 않는지를 잔차를 이용해 검토
- 잔차: $e_i = y_i - \hat{y}_i$
- 추정값이나 독립변수에 대해 잔차 산포도를 관찰할 때, 잔차가 0을 중심으로 랜덤하게 분포되어 있는지를 살펴봄

lm() 다루기 ➔ R code로 이동

```
> age <- rep(seq(20,60,by = 5),3)
> mhr <- 209 - 0.7 * age + rnorm(length(age),sd = 4)
> plot(mhr ~ age,main = "Age versus maximum heart rate")
> res.mhr <- lm(mhr ~ age)
> res.mhr

Call:
lm(formula = mhr ~ age)

Coefficients:
(Intercept)          age
    208.245       -0.709

> abline(res.mhr)
```

lm() 다루기


Age versus maximum heart rate
200
190
180
170
mhr
20
30
40
50
60
age

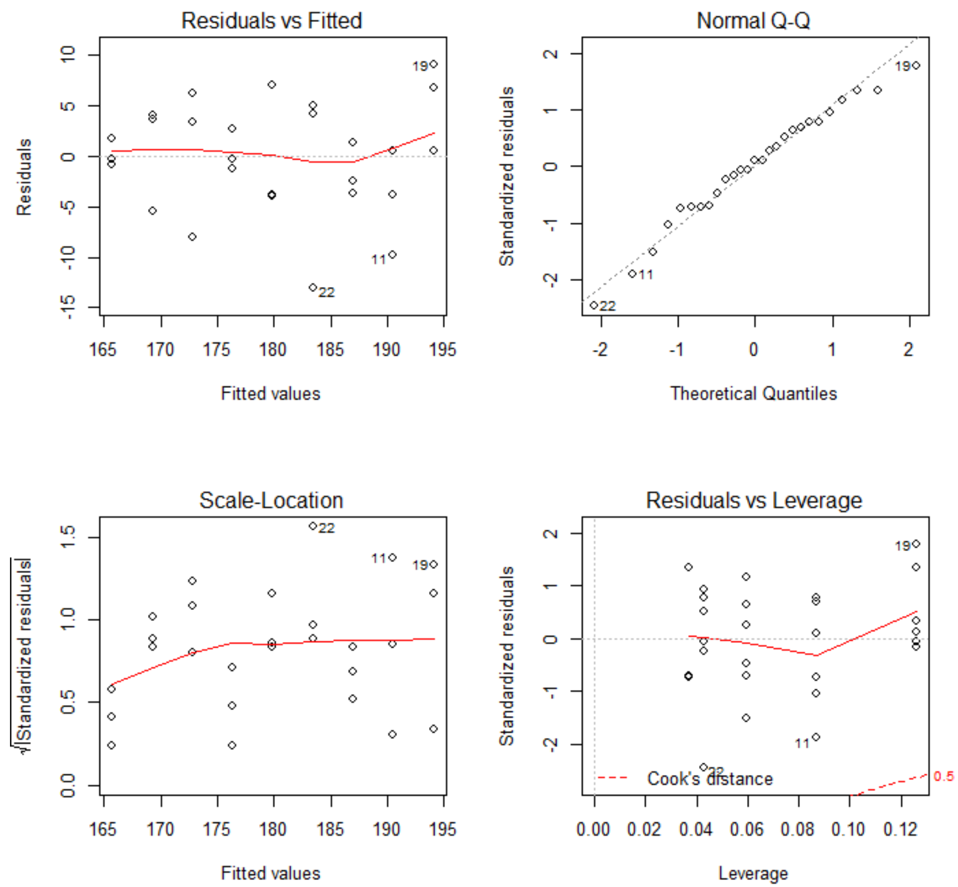
lm() 다루기

```
> summary(res.mhr)

Call:
lm(formula = mhr ~ age)

Residuals:
     Min       1Q   Median       3Q      Max 
-12.9992  -3.7145   0.4931   3.8411   9.0657 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 208.24496    3.40777  61.109  < 2e-16 ***
age          -0.70896    0.08108  -8.744 4.47e-09 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.439 on 25 degrees of freedom
Multiple R-squared:  0.7536,    Adjusted R-squared:  0.7437 

F-statistic: 76.46 on 1 and 25 DF,  p-value: 4.47e-09
```

lm() 다루기

```
> anova(res.mhr)
Analysis of Variance Table

Response: mhr
          Df Sum Sq Mean Sq F value    Pr(>F)
age        1 2261.8 2261.78  76.463 4.47e-09 ***
Residuals 25  739.5   29.58
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> par(mfrow = c(2,2))
> plot(res.mhr)
```

모형 검토 ➔ R code로 이동



변수 변환

- 회귀모형에서 회귀계수에 대한 추정과 가설검정은 종속변수Y의 **정규성**을 전제로 하기 때문에, 이 조건이 충족되는지를 확인하는 과정이 필요
- 만약 Y의 정규성이 **만족되지** 않는다면,
  - **변수 변환**
  - 비모수 방법

변수 변환: 사례

- 우유량과 수유기간이 **승법적**인 관계가 (즉, $y = ae^{-cx} \Leftrightarrow \log y = \log a - cx$) 있는 경우에는 **우유량**을 **로그변환**하여 회귀모형을 적합하여 선형적인 모형으로 변환할 수 있음
- 출산경험=1 인 경우: $\log \hat{y} = 3.444 - 0.001x_1$
- 출산경험=2 인 경우: $\log \hat{y} = 3.780 - 0.002x_1$
- 출산경험=3 인 경우: $\log \hat{y} = 3.872 - 0.003x_1$

다중회귀모형: 독립변수가 2개 이상인 경우

- 독립변수가 **2개 이상**인 모형
- 모형의 정의
  - $Y_i = \beta_0 + \beta_1 \boldsymbol{x_{1i}} + \beta_2 \boldsymbol{x_{2i}} + \epsilon_i,\ \epsilon_i \sim iid\ N(0, \sigma^2)$ 혹은
  - $E(Y_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}$

모형의 적합

- 단순회귀모형과 동일하게 최소제곱법을 이용하여 회귀계수 $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$를 추정 ➜ 추정회귀식 $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$를 얻음

모형의 적합: 사례

- 우유량(혹은 로그우유량)이 수유기간 외에도 로그수유기간에 의해서도 영향을 받는 것으로 알려져 있으므로 수유기간(x_1)과 로그수유기간(x_2)을 독립변수로 정의하여 회귀모형을 적합

모형의 적합: 사례

- 출산경험=1 인 경우: $\log \hat{y} = 3.003 - 0.004x_1 + 0.229x_2$
- 출산경험=2 인 경우: $\log \hat{y} = 2.991 - 0.004x_1 + 0.257x_2$
- 출산경험=3 인 경우: $\log \hat{y} = 2.744 - 0.002x_1 + 0.198x_2$
- 결정계수의 비교

출산 경험	x_1만 적합한 경우	x_1, x_2를 적합한 경우
1	R^2 = 0.601	R^2 = 0.967
2	R^2 = 0.841	R^2 = 0.986
3	R^2 = 0.888	R^2 = 0.993

회귀분석의 활용: 가변수를 이용한 회귀분석

- 독립변수 중에 범주형 변수가 있는 경우: 가변수를 정의하여 회귀모형을 적합

가변수를 이용한 회귀분석: 사례

- 출산경험을 가변수로 정의하여 출산경험이 우유량에 미치는 영향을 알아볼 수 있음
- 가변수의 정의
  - x_{31} = 1 if 출산경험=1; 0 o.w
  - x_{32} = 1 if 출산경험=2; 0 o.w
- 추정 회귀식:

$$\log \hat{y} = 3.019 - 0.004x_1 + 0.213x_2 - 0.160x_{31} + 0.003x_{32}$$

- 출산경험에 따라 x_1과 x_2의 효과는 같으나 절편은 달라짐

가변수를 이용한 회귀분석: 사례

- 그러나 출산경험이 x_1, x_2와 **교호작용**(interaction)이 있는 경우에는 x_1과 x_2의 **회귀계수도 달라짐**
- 가변수(x_{31}, x_{32})와 x_1, x_2의 교호작용을 고려한 모형을 적합

가변수를 이용한 회귀분석: 사례

- 추정 회귀식:
  $\log \hat{y} = 2.744 - 0.002x_1 + 0.198x_2 + 0.259x_{31} + 0.247x_{32} - 0.002x_1x_{31} - 0.002x_1x_{32} + 0.030x_2x_{31} + 0.059x_2x_{32}$
- 출산경험=1 인 경우 $(x_{31} = 1, x_{32} = 0)$:
  $\log \hat{y} = 3.003 - 0.004x_1 + 0.228x_2$
- 출산경험=2 인 경우 $(x_{31} = 0, x_{32} = 1)$:
  $\log \hat{y} = 2.991 - 0.004x_1 + 0.257x_2$
- 출산경험=3 인 경우 $(x_{31} = 0, x_{32} = 0)$:
  $\log \hat{y} = 2.744 - 0.002x_1 + 0.198x_2$

formula 작성하기

- Syntax: lm(formula, data= ...,)
  - $y = \beta_0 + \epsilon \Leftrightarrow$ lm(y ~ 1)
  - $y = \beta_0 + \beta_1 x + \epsilon \Leftrightarrow$ lm(y ~ x)
  - $y = \beta_1 x + \epsilon \Leftrightarrow$ lm(y ~ x − 1)
  - $y = \beta_0 + \beta_1 x + \beta_2 \boldsymbol{x^2} + \epsilon \Leftrightarrow$ lm(y ~ x + I(x^2))
  - $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 \boldsymbol{x_1 x_2} + \epsilon \Leftrightarrow$ lm(y ~ x1 + x2 + x1:x2) or lm(y ~ x1* x2)
- Wood function? lm(log(y) ~ x + log(x))

formula 작성하기 → R code로 이동

```
> x1 <- 1:10; x2 <- rchisq(10, 3); y <- 1 + x1 + x2 + rnorm(10)
> lm(y ~ x1 + x2 + x1:x2)

Call:
lm(formula = y ~ x1 + x2 + x1:x2)

Coefficients:
(Intercept)           x1           x2        x1:x2
    2.61809      0.72936      0.84297      0.04259

> lm(y ~ x1 * x2)

Call:
lm(formula = y ~ x1 * x2)

Coefficients:
(Intercept)           x1           x2        x1:x2
    2.61809      0.72936      0.84297      0.04259
```

가변수를 이용한 회귀분석: 사례

→ R code로 이동

```
...
> library(MASS)
> data("Cars93")
> names(Cars93)
 [1] "Manufacturer"       "Model"
 [3] "Type"               "Min.Price"
 [5] "Price"              "Max.Price"
 [7] "MPG.city"           "MPG.highway"
 [9] "AirBags"            "DriveTrain"
[11] "Cylinders"          "EngineSize"
[13] "Horsepower"         "RPM"
[15] "Rev.per.mile"       "Man.trans.avail"
[17] "Fuel.tank.capacity" "Passengers"
[19] "Length"             "Wheelbase"
[21] "Width"              "Turn.circle"
[23] "Rear.seat.room"     "Luggage.room"
[25] "Weight"             "Origin"
[27] "Make"
```

가변수를 이용한 회귀분석: 사례

```
> str(Cars93$Cylinders)
 Factor w/ 6 levels "3","4","5","6",..: 2 4 4 4 2 2 4 4 4 5 ...
> lm(MPG.highway ~ Cylinders + Horsepower,data = Cars93)

Call:
lm(formula = MPG.highway ~ Cylinders + Horsepower, data = Cars93)

Coefficients:
    (Intercept)       Cylinders4       Cylinders5
       45.10713        -10.60858        -16.88485
     Cylinders6       Cylinders8  Cylindersrotary
      -15.06570        -13.79903        -13.25383
     Horsepower
       -0.02688
```

변수 선택법

- 전진선택법(forward selection)
  - 가장 **간단한** 모형으로부터 시작하여 유의한 변수를 1개씩 **선택**하는 방법
- 후진제거법(backward elimination)
  - 가장 **복잡한** 모형으로부터 시작하여 유의하지 않은 변수를 하나씩 **제거**해 나가는 방법
- 단계선택법(stepwise selection)
  - 가장 간단한 모형으로부터 시작하여 유의한 변수를 1개씩 선택해 나가되 한 번 선택된 변수도 제거될 수 있는 방법

변수 선택법: 사례

➜ R code로 이동

```
> library(UsingR)
> library(MASS)
> data("stud.recs")
> head(stud.recs)
      seq.1 seq.2 seq.3 sat.v sat.m letter.grade num.grade
1528     80    47    76   440   450            B       3.0
14586    74    76    81   410   480           B-       2.7
8295     84    85    90   610   590            D       1.0
8446     85    88    68   300   520            A       4.0
16685    81    75    66   550   560            A       4.0
16398    81    71    67   460   400            C       2.0
```

변수 선택법: 사례

```
> fit1 <- lm(num.grade ~ ., data = d)
> stepAIC(fit1)
Start:  AIC=101.22
num.grade ~ seq.1 + seq.2 + seq.3 + sat.v + sat.m

        Df Sum of Sq    RSS     AIC
- seq.2  1    0.0705 254.70  99.254
- seq.1  1    0.1578 254.78  99.297
- sat.v  1    1.2766 255.90  99.840
<none>               254.63 101.220
- seq.3  1    5.9746 260.60 102.096
- sat.m  1   10.4959 265.12 104.229

Step:  AIC=99.25
num.grade ~ seq.1 + seq.3 + sat.v + sat.m

        Df Sum of Sq    RSS     AIC
- seq.1  1    0.2944 254.99  97.398
- sat.v  1    1.2553 255.95  97.864
<none>               254.70  99.254
- seq.3  1    6.0421 260.74 100.162
- sat.m  1   10.6615 265.36 102.339
```

변수 선택법: 사례

```
Step:  AIC=97.4
num.grade ~ seq.3 + sat.v + sat.m

        Df Sum of Sq    RSS     AIC
- sat.v  1    1.2497 256.24  96.004
<none>               254.99  97.398
- seq.3  1    5.8384 260.83  98.205
- sat.m  1   10.8807 265.87 100.579

Step:  AIC=96
num.grade ~ seq.3 + sat.m

        Df Sum of Sq    RSS    AIC
<none>               256.24 96.004
- seq.3  1    5.5494 261.79 96.661
- sat.m  1    9.6323 265.87 98.580

Call:
lm(formula = num.grade ~ seq.3 + sat.m, data = d)

Coefficients:
(Intercept)        seq.3        sat.m
   -1.14078      0.01371      0.00479
```

변수 선택법: 사례

```
> fit2 <- lm(num.grade ~ 1, data = d)
> scope = list(upper = fit1,lower = fit2)
> stepAIC(fit2,direction = "forward",scope = scope)
Start:  AIC=102.43
num.grade ~ 1

        Df Sum of Sq    RSS     AIC
+ sat.m  1   16.9316 261.79  96.661
+ seq.3  1   12.8487 265.87  98.580
+ seq.2  1    5.4308 273.29 101.992
+ seq.1  1    4.8770 273.85 102.243
<none>               278.72 102.432
+ sat.v  1    0.4485 278.27 104.232

Step:  AIC=96.66
num.grade ~ sat.m

        Df Sum of Sq    RSS    AIC
+ seq.3  1    5.5494 256.24 96.004
<none>               261.79 96.661
+ sat.v  1    0.9606 260.83 98.205
+ seq.2  1    0.2324 261.56 98.551
+ seq.1  1    0.0821 261.71 98.622
```

변수 선택법: 사례

```
Step:  AIC=96
num.grade ~ sat.m + seq.3

        Df Sum of Sq    RSS    AIC
<none>                256.24 96.004
+ sat.v  1   1.24967 254.99 97.398
+ seq.1  1   0.28877 255.95 97.864
+ seq.2  1   0.17043 256.07 97.921

Call:
lm(formula = num.grade ~ sat.m + seq.3, data = d)

Coefficients:
(Intercept)        sat.m        seq.3
   -1.14078      0.00479      0.01371
```

종속변수가 범주형인 경우

- $E(Y|x) = \beta_0 + \beta_1 x,$
- $Y = \begin{cases} 1, \text{wp } p \\ 0, \text{wp } (1-p) \end{cases}$

로지스틱회귀모형

- 로지스틱함수 변환을 통하여 로지스틱회귀모형을 적합

$$0 < p < 1 \Rightarrow -\infty < \log\frac{p}{1-p} < \infty$$

- 로지스틱회귀모형은 다음과 같이 정의됨
  - $\log\frac{p(x)}{1-p(x)} = \beta_0 + \beta_1 x$ 또는
  - $\frac{p(x)}{1-p(x)} = e^{\beta_0+\beta_1 x}$ (오즈, odds) 또는
  - $p(x) = P(Y=1|x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$

로지스틱회귀모형: 사례

- 나이가 관상동맥질환에 미치는 영향을 알아보고자 관상동맥질환 여부(Y)와 나이(age)를 측정

변수	회귀계수	표준오차	t−값
Intercept	−5.310	1.134	−4.68
age	0.111	0.024	4.61

로지스틱 회귀모형: 사례

- 관상동맥질환에 걸릴 확률은 다음과 같음
- $\hat{p}(\mathrm{age}) = \dfrac{e^{-5.311+0.111\times\mathrm{age}}}{1+e^{-5.311+0.111\times\mathrm{age}}}$
- 60세인 사람이 관상동맥질환에 걸릴 확률은 79.4%

다중 로지스틱회귀모형

- 독립변수가 **2개 이상인** 경우
- 모형의 정의
  - $\log \frac{p(x)}{1-p(x)} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ 또는
  - $\frac{p(x)}{1-p(x)} = e^{\beta_0+\beta_1 x_1+x_2}$ 또는
  - $p(x) = P(Y = 1|x_1, x_2) = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2}}{1+e^{\beta_0+\beta_1 x_1+\beta_2 x_2}}$

다중 로지스틱회귀모형: 사례

- 저체중 신생아 출산에 영향을 주는 위험인자를 탐색!
- 신생아의 체중이 2500g 미만이면 low=1, 이상이면 low=0 이라고 정의하고,
- 산모의 나이(age), 인종(race), 임신 직전 산모의 체중(LWT), 임신기간 중 의사방문횟수(FTV)를 관측
- Hosmer, D.W., Lemeshow, S., & Sturdivant, R.X. (2013). Applied Logistic Regression. 3rd ed. Wiley.

다중 로지스틱회귀모형: 사례

- 백인이면 race1=race2=0,
  흑인이면 race1=1, race2=0,
  그 외 인종이면 race1=0, race2=1로 정의

변수	회귀계수	표준오차	t−값
Intercept	1.295	1.071	1.21
age	−0.024	0.034	−0.71
LWT	−0.014	0.0065	−2.18
race1	1.004	0.498	2.02
race2	0.433	0.362	1.20
FTV	−0.049	0.167	−0.30

다중 로지스틱회귀모형: 사례

- **변수선택**방법을 적용했을 때 유의한 위험인자만으로 이루어진 모형

변수	회귀계수	표준오차	t−값
Intercept	0.806	0.845	0.95
LWT	−0.015	0.0064	−2.36
race1	1.081	0.488	2.22
race2	0.481	0.357	1.35

- 추정 회귀식:

$$\log \frac{\hat{p}(x)}{1-\hat{p}(x)} = 0.806 - 0.015 \times \mathrm{LWT} + 1.081 \times \mathrm{race1} + 0.481 \times \mathrm{race2}$$

로지스틱 회귀모형의 해석: 오즈비

- LWT가 10파운드 감소하면 저체중 신생아를 출산할 오즈가 $\frac{e^{-0.015\times(x-10)}}{e^{-0.015\times x}} = e^{0.15} = 1.16$배 증가하고,
- 인종에 따른 오즈는 백인을 기준으로 흑인 산모는 $\frac{e^{1.081\times1+0.481\times0}}{e^{1.081\times0+0.481\times0}} = e^{1.081} = 2.95$배, 타 인종 산모는 $\frac{e^{1.081\times0+0.481\times1}}{e^{1.081\times0+0.481\times0}} = e^{0.481} = 1.62$배 높음

glm() 다루기

- Syntax: glm(formula, family=…,)
- family= “binomial”

glm() 다루기 ➜ R code로 이동

```
> data("birthwt")
>
> # race를 factor로 변환
> birthwt$race <- factor(birthwt$race,
+                        levels = c(1, 2, 3),
+                        labels = c("White", "Black", "Other"))
>
> summary(birthwt$race)
White Black Other
   96    26    67
>
```

glm() 다루기

```
> # 초기 풀모형
> fit_logit <- glm(low ~ age + lwt + race + ftv,
+                  family = binomial,
+                  data = birthwt)
>
> summary(fit_logit)

Call:
glm(formula = low ~ age + lwt + race + ftv, family = binomial,
    data = birthwt)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.295366   1.071443   1.209   0.2267
age         -0.023823   0.033730  -0.706   0.4800
lwt         -0.014245   0.006541  -2.178   0.0294 *
raceBlack    1.003898   0.497859   2.016   0.0438 *
raceOther    0.433108   0.362240   1.196   0.2318
ftv         -0.049308   0.167239  -0.295   0.7681
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

glm() 다루기

```
> # AIC 기준 stepwise selection
> final_model <- stepAIC(fit_logit, trace = 0)
>
> summary(final_model)

Call:
glm(formula = low ~ lwt + race, family = binomial, data = birthwt)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.805753   0.845167   0.953   0.3404
lwt         -0.015223   0.006439  -2.364   0.0181 *
raceBlack    1.081066   0.488052   2.215   0.0268 *
raceOther    0.480603   0.356674   1.347   0.1778
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

glm() 다루기

```
> print(result_table)
               Estimate        OR     X2.5..     X97.5..
(Intercept)   0.8057535 2.2383824 0.4494580 12.5233475
lwt          -0.0152231 0.9848922 0.9717796  0.9967361
raceBlack     1.0810662 2.9478208 1.1273979  7.7651592
raceOther     0.4806033 1.6170496 0.8031808  3.2667497
>
```

요약

- 단순선형회귀모형
  - 모형적합 검토
  - 잔차분석
- 회귀분석의 활용
  - 가변수를 포함하는 회귀모형
- 다중선형회귀모형
  - 변수선택방법
- 로지스틱 회귀모형