Chapter 3

Multiple Linear Regression

Chapter 3b

Properties of least-squares estimators (pages 3-6)

Gauss Markov theorem (page 7)

Residual sum of squares and estimation of σ^2 (pages 8-9)

- Example – Cholesterol age data (pages 10-11)
  - An example to show that MS_{Res} is model dependent
- Example – Delivery time data (page 12-15)

Inadequacy of scatter diagrams in multiple regression (page 16)

Maximum-likelihood estimation (pages 17-18)

Distribution theory – An important theorem (pages 19-20)

Properties of least-squares estimators

nxn identity matrix

1. Multiple linear regression model

   $\boldsymbol{y}_{n\times 1} = \boldsymbol{X}_{n\times(k+1)}\boldsymbol{\beta}_{(k+1)\times 1} + \boldsymbol{\epsilon}_{n\times 1}$ where $\boldsymbol{E}(\boldsymbol{\epsilon}) = \boldsymbol{0}$ and $\boldsymbol{Var}(\boldsymbol{\epsilon}) = \sigma^2\boldsymbol{I}_n$

   Least-squares normal equations

   $\boldsymbol{X'X}\hat{\boldsymbol{\beta}} = \boldsymbol{X'y}$

   Least-squares estimators

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

2. Note: Let $\boldsymbol{A}$ be a $k \times k$ matrix of constants and $\boldsymbol{y}$ be a $k \times 1$ random vector with mean $\boldsymbol{E}(\boldsymbol{y}) = \boldsymbol{\mu}$ and non-singular variance-covariance matrix $\boldsymbol{Var}(\boldsymbol{y}) = \boldsymbol{V}$. Then

   (a) $\boldsymbol{E}(\boldsymbol{Ay}) = \boldsymbol{A\mu}$

   $Var(aY) = a^2 Var(Y)$

   (b) $\boldsymbol{Var}(\boldsymbol{Ay}) = \boldsymbol{AVA'}$

3. $\hat{\boldsymbol{\beta}}$ is an unbiased estimator of $\boldsymbol{\beta}$

$$
\begin{aligned}
\boldsymbol{E}(\hat{\boldsymbol{\beta}}) &= \boldsymbol{E}[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}] \\
&= \boldsymbol{E}[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\epsilon})] \\
&= \boldsymbol{E}[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}+(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\epsilon}] \\
&= \boldsymbol{\beta}+(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{E}[\boldsymbol{\epsilon}] \\
&= \boldsymbol{\beta}
\end{aligned}
$$

$y = X\beta + \varepsilon$

assumption that $E[\varepsilon] = 0$

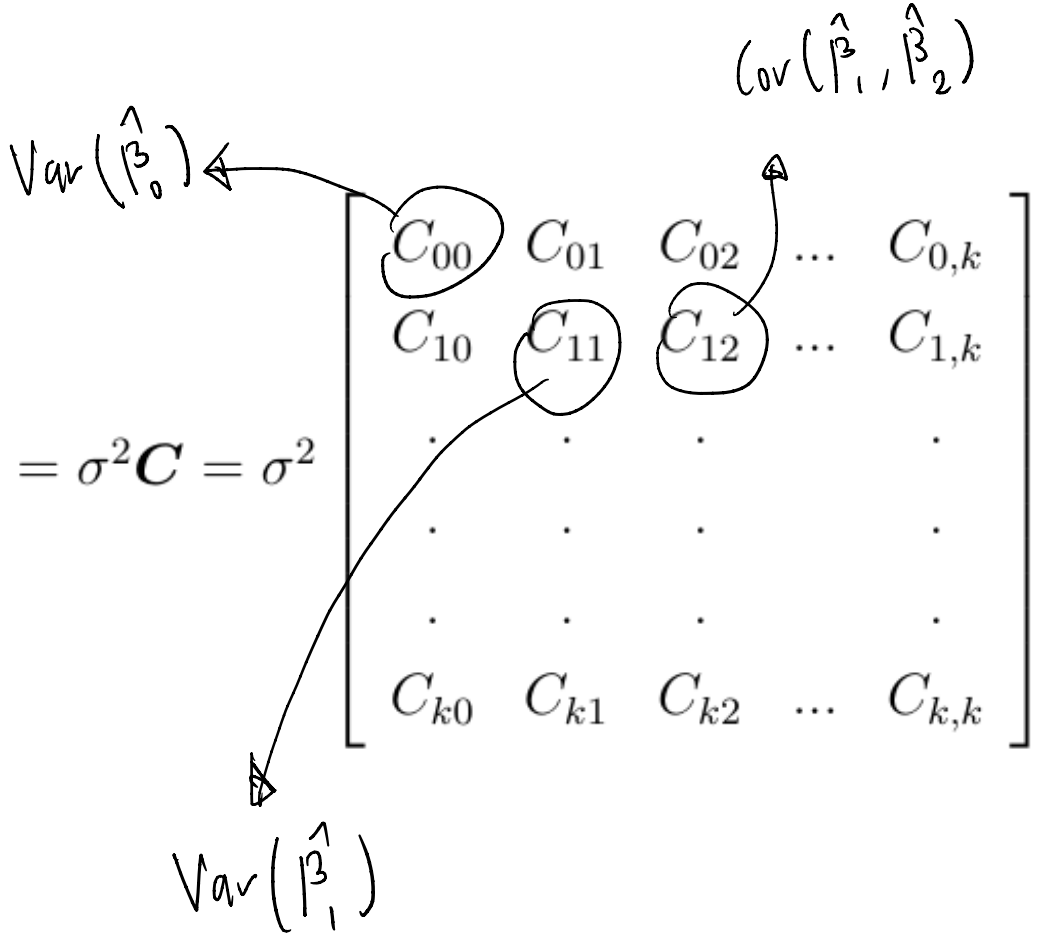
4. Variance-covariance matrix of $\hat{\boldsymbol{\beta}}$

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

$$
\begin{aligned}
&\left[(X'X)^{-1}X'\right]' \\
&= (X')'\left[(X'X)^{-1}\right]' \\
&= X\cdot(X'X)^{-1}
\end{aligned}
$$

$X'X$ is symmetric $\Rightarrow (X'X)^{-1}$ is also symmetric

$$\boldsymbol{Var}\begin{bmatrix}\hat{\beta}_0\\ \hat{\beta}_1\\ \hat{\beta}_2\\ \cdot\\ \\ \cdot\\ \\ \hat{\beta}_k\end{bmatrix} = \sigma^2(\boldsymbol{X'X})^{-1} = \sigma^2\boldsymbol{C} = \sigma^2\begin{bmatrix} C_{00} & C_{01} & C_{02} & \cdots & C_{0,k}\\ C_{10} & C_{11} & C_{12} & \cdots & C_{1,k}\\ \cdot & \cdot & \cdot & & \cdot\\ \cdot & \cdot & \cdot & & \cdot\\ \cdot & \cdot & \cdot & & \cdot\\ C_{k0} & C_{k1} & C_{k2} & \cdots & C_{k,k}\end{bmatrix}$$



$$Var(\hat{\beta}_j) = \sigma^2 C_{jj}$$

$$Cov(\hat{\beta}_i, \hat{\beta}_j) = \sigma^2 C_{ij}$$

Gauss-Markov theorem

$\hat{\boldsymbol{\beta}}$ is the best linear unbiased estimator of $\boldsymbol{\beta}$

$\hat{\boldsymbol{\beta}}$ is also the maximum-likelihood estimator of $\boldsymbol{\beta}$.

Residual sum of squares and estimation of σ^2

Let SS_{Res} be the residual sum of squares.

$$
\begin{aligned}
SS_{Res} &\equiv \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \\
&= \sum_{i=1}^{n} e_i^2 \\
&= \boldsymbol{e'e} \\
&= (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) \\
&= (\boldsymbol{y}' - \hat{\boldsymbol{\beta}}'\boldsymbol{X}')(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) \\
&= \boldsymbol{y'y} - \hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{y} - \boldsymbol{y}'\boldsymbol{X}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}} \\
&= \boldsymbol{y'y} - 2\hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{y} + \hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}} \\
&= \boldsymbol{y'y} - \hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{y} \qquad \because \boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}'\boldsymbol{y}
\end{aligned}
$$

$$SS_{Res} = \boldsymbol{y'y} - \hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{y}$$

$E(\varepsilon) = 0$

$Var(\varepsilon) = \sigma^2 I_n$

$$SS_{Res} = \boldsymbol{y'y} - \hat{\boldsymbol{\beta}}' \boldsymbol{X'y}$$

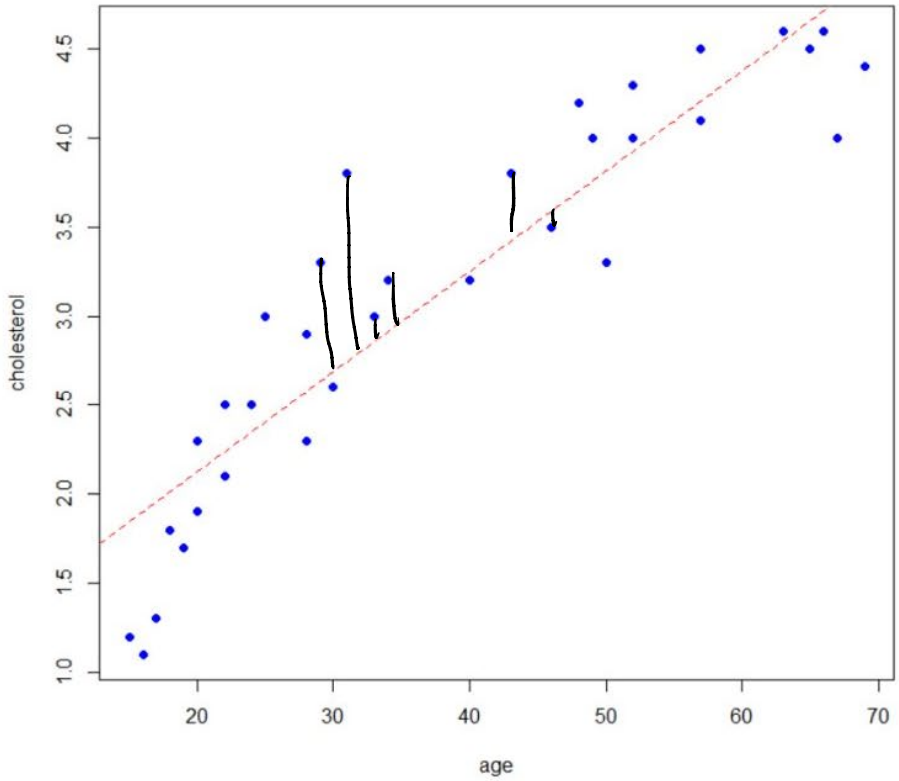
$p = k+1$

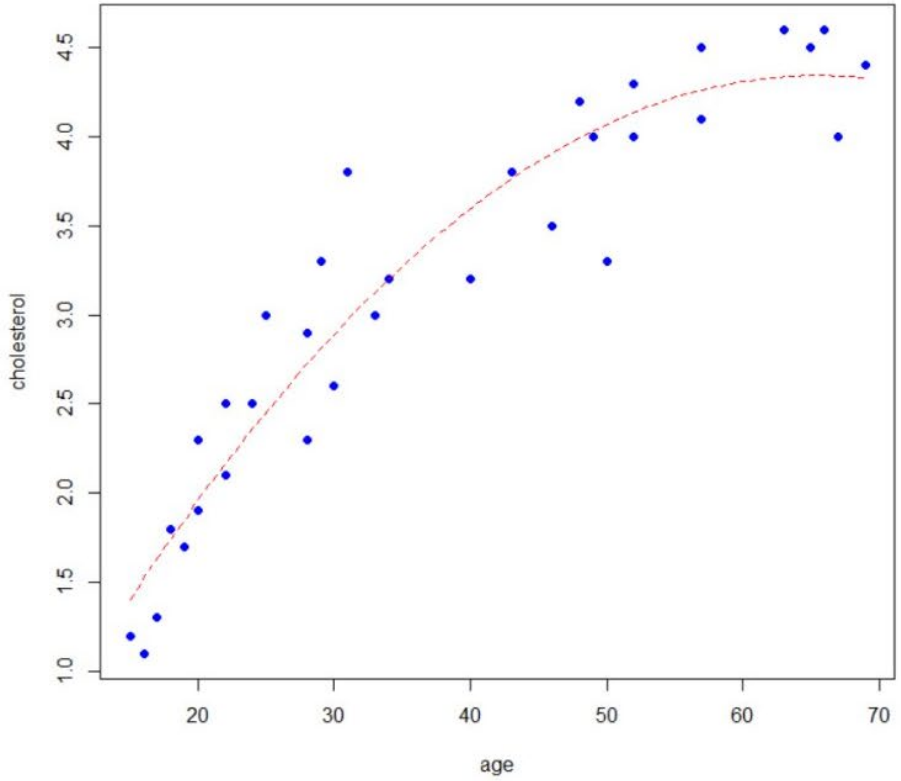
$$\hat{\sigma}^2 = MS_{Res} \equiv \frac{SS_{Res}}{n-p}$$

Linear model theory: $E(MS_{Res}) = \sigma^2$

MS_{Res} is model-dependent because MS_{Res} changes with the model.

Cholesterol age data





```
48  # Compare straight line and quadratic models
49  # in terms of MS_Res
50  anova(lm(cholesterol~age))
51  anova(lm(cholesterol~age + I(age^2)))
```

48:1 (Top Level)

Console | Background Jobs

R 3.4.1 ·

```
> # Compare straight line and quadratic models
> # in terms of MS_Res
> anova(lm(cholesterol~age))
Analysis of Variance Table

Response: cholesterol
          Df  Sum Sq Mean Sq F value    Pr(>F)
age        1 30.7370 30.7370  161.52 7.856e-14 ***
Residuals 31  5.8993  0.1903
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> anova(lm(cholesterol~age + I(age^2)))
Analysis of Variance Table

Response: cholesterol
          Df  Sum Sq Mean Sq F value    Pr(>F)
age        1 30.7370 30.7370  262.39 2.234e-16 ***
I(age^2)   1  2.3851  2.3851   20.36 9.198e-05 ***
Residuals 30  3.5143  0.1171
```

Residual mean square is smaller for a better fitted model

$MS_{RES} = \hat{\sigma}^2$

Delivery Time Data

Column of response

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \qquad y' = \text{transpose of } y = (y_1\ y_2\ \cdots\ y_n)$$

$$\mathbf{y}'\mathbf{y} = \sum_{i=1}^{25} y_i^2 = 18{,}310.6290$$

$$\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} = \begin{bmatrix} 2.34123115 & 1.61590721 & 0.01438483 \end{bmatrix} \begin{bmatrix} 559.60 \\ 7{,}375.44 \\ 337{,}072.00 \end{bmatrix}$$

$$= 18{,}076.90304$$

$$SS_{\text{Res}} = \mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y}$$

$$= 18{,}310.6290 - 18{,}076.9030 = 233.7260$$

MS_{Res}^2 $\hat{\sigma}^2 = \dfrac{SS_{\text{Res}}}{n-p} = \dfrac{233.7260}{25-3} = 10.6239$

p: $k+1$

```r
# matrices and finding betas
one <- rep(1,length(time))
X <- array(c(one,case,dist), dim=c(length(time),3))
y <- array(time, dim=c(length(time),1))
XPX <- t(X) %*% X
XPX
XPy <- t(X) %*% y
XPy
betahat <- solve(XPX, XPy)
betahat

# estimator of sigma^2
bxy <- t(betahat) %*% t(X) %*% y
bxy
SSRes <- t(y) %*% y - bxy
SSRes
sigma2hat <- SSRes/(length(time)-3)
sigma2hat

# alternative method
anova(fitted.model)
```

$X'X\hat{\beta} = X'y$ use solve function

```
> # matrices and finding betas
> one <- rep(1,length(time))
> X <- array(c(one,case,dist), dim=c(length(time),3))
> y <- array(time, dim=c(length(time),1))
> XPX <- t(X) %*% X
> XPX
      [,1]   [,2]    [,3]
[1,]    25    219   10232
[2,]   219   3055  133899
[3,] 10232 133899 6725688
> XPy <- t(X) %*% y
> XPy
           [,1]
[1,]    559.60
[2,]   7375.44
[3,] 337071.69
> betahat <- solve(XPX, XPy)
> betahat
             [,1]
[1,] 2.34123115
[2,] 1.61590721
[3,] 0.01438483
>
```

```
> # estimator of sigma^2
> bxy <- t(betahat) %*% t(X) %*% y
> bxy
         [,1]
[1,] 18076.9
> SSRes <- t(y) %*% y - bxy
> SSRes
          [,1]
[1,] 233.7317
> sigma2hat <- SSRes/(length(time)-3)
> sigma2hat
          [,1]
[1,] 10.62417
> # alternative method
> anova(fitted.model)
Analysis of Variance Table

Response: time
           Df Sum Sq Mean Sq F value    Pr(>F)
case        1 5382.4  5382.4 506.619 < 2.2e-16 ***
dist        1  168.4   168.4  15.851 0.0006312 ***
Residuals 22  233.7    10.6
```

$\hat{\sigma}^2$

$\hat{\sigma}^2$

SSRes

Inadequacy of Scatter Diagrams in Multiple Regression

These data were generated from the equation

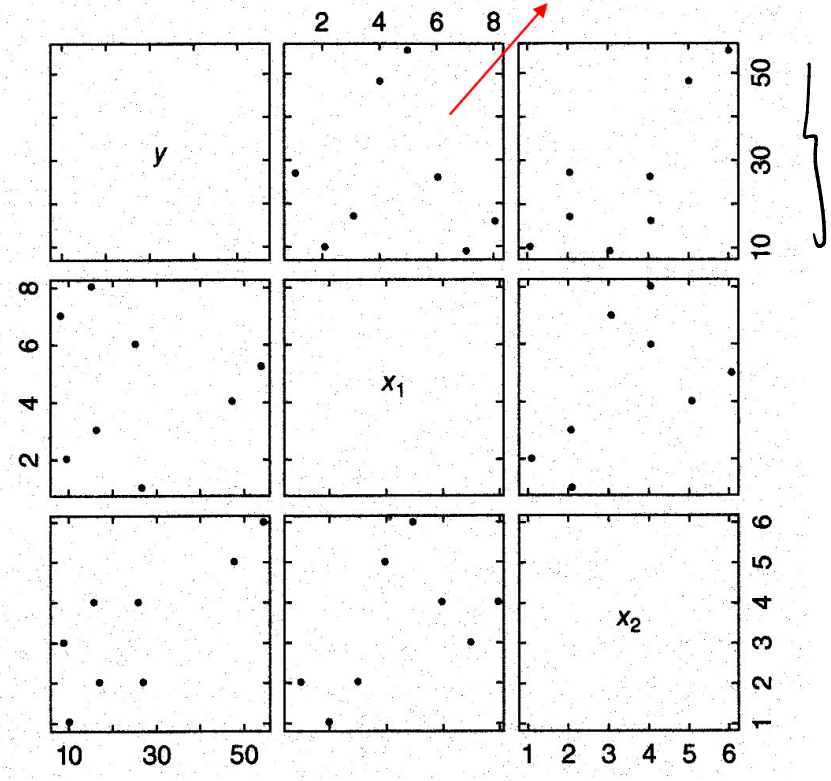
$$y = 8 - 5x_1 + 12x_2$$

The matrix of scatterplots is shown. The y-versus-x_1, plot does not exhibit any apparent relationship between the two variables. The y-versus-x_2 plot indicates that a linear relationship exists, with a slope of approximately 8. Note that both scatter diagrams convey erroneous information. Since in this data set there are two pairs of points that have the same x_2 values ($x_2 = 2$ and $x_2 = 4$), we could measure the x_1 effect at fixed x_2 from both pairs. This gives, $\hat{\beta}_1 = (17-27)/(3-1) = -5$ for $x_2 = 2$ and $\hat{\beta}_1 = (26-16)/(6-8) = -5$ for $x_2 = 4$ the correct results.

No apparent relationship

$y = 8 - 5x_1 + 12x_2$

y	x_1	x_2
10	2	1
17	3	2
48	4	5
27	1	2
55	5	6
26	6	4
9	7	3
16	8	4



simulated data

slope x 8 (real is 12?)

40

individual scatterplot may not tell the picture

Maximum-Likelihood Estimation

$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \Rightarrow y \sim N(X\beta, \sigma^2 I)$

$\boldsymbol{\varepsilon}$ is distributed as $N(\mathbf{0}, \sigma^2\mathbf{I})$

$$f(\varepsilon_i) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}\varepsilon_i^2\right)$$

$Var(\varepsilon) = \sigma^2 \begin{pmatrix} 1 & & 0 \\ & 1 & \\ 0 & & 1 \end{pmatrix}$ identity matrix

$= \begin{pmatrix} \sigma^2 & & \\ & \sigma^2 & \\ & & \sigma^2 \end{pmatrix} \Leftrightarrow Var(\varepsilon_i) = \sigma^2$

$Cov(\varepsilon_i, \varepsilon_j)$, $i \neq j$

$Var(\varepsilon_1)$ $Var(\varepsilon_2)$

The likelihood function is the joint density of $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ or $\prod_{i=1}^{n} f(\varepsilon_i)$

$$L(\boldsymbol{\varepsilon}, \boldsymbol{\beta}, \sigma^2) = \prod_{i=1}^{n} f(\varepsilon_i) = \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left(-\frac{1}{2\sigma^2}\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}\right)$$

$\boldsymbol{\varepsilon}$: matrix; σ^2: scalar

$$L(\mathbf{y}, \mathbf{X}, \boldsymbol{\beta}, \sigma^2) = \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left(-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right)$$

$y = X\beta + \varepsilon$

$\boldsymbol{\varepsilon} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}$

$$\ln L(\mathbf{y}, \mathbf{X}, \boldsymbol{\beta}, \sigma^2) = -\frac{n}{2}\ln(2\pi) - n\ln(\sigma) - \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

Maximize $\ln L(\mathbf{y}, \mathbf{X}, \boldsymbol{\beta}, \sigma^2)$ is the same as

minimize $(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$ for a fixed value of σ.

Least-squares function

Maximum-likelihood estimator of $\boldsymbol{\beta}$ under normal errors is the same as the least-squares estimator $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$.

The maximum-likelihood estimator of σ^2 is

$$\tilde{\sigma}^2 = \frac{\left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right)'\left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right)}{n}$$

$E(\tilde{\sigma}^2) \neq \sigma^2$

$$\hat{\sigma}^2 = \frac{(y - X\hat{\beta})'(y - X\hat{\beta})}{n - p}$$

$E(\hat{\sigma}^2) = \hat{\sigma}_2$

(unbiased ⇒ preferred)

Distribution theory (graduate level course in multiple linear regression)

Theorem DT: Let $\boldsymbol{A}$ be a $k \times k$ matrix of constants, and $\boldsymbol{y}$ be a $k \times 1$ multivariate normal vector with mean $\boldsymbol{\mu}$ and non-singular variance-covariance matrix $\boldsymbol{V}$; thus

$E(y) = \mu$, $Var(y) = V$

$$\boldsymbol{y} \sim \boldsymbol{N}(\boldsymbol{\mu}, \boldsymbol{V}).$$

Let U be the quadratic form defined by $U = \boldsymbol{y'Ay}$. eg. $AA = A$

1. If $\boldsymbol{AV}$ or $\boldsymbol{VA}$ is an idempotent matrix of rank p, then

$$U \sim \chi^2_{p,\lambda}$$

   where $\chi^2_{p,\lambda}$ is the chi-square distribution with p degrees of freedom and non-centrality parameter $\lambda = \boldsymbol{\mu' A \mu}$.

2. Let $\boldsymbol{V} = \sigma^2 \boldsymbol{I}$, which is a typical assumption. If $\boldsymbol{A}$ is an idempotent matrix of rank p, then

$$U/\sigma^2 \sim \chi^2_{p,\lambda} \Rightarrow \frac{(n-p)\,MS_{Res}}{\sigma^2} \sim \chi^2_{n-p}$$

   where $\lambda = \boldsymbol{\mu' A \mu}/\sigma^2$.

Theorem: $X \sim N(0,1)$, $U \sim \chi^2(m)$, X and U independent $\Rightarrow \dfrac{X}{\sqrt{U/m}} \sim t(m)$ (degree of freedom)

3. Let $\boldsymbol{B}$ be a $q \times k$ matrix, and let $\boldsymbol{W}$ be the linear form given by $\boldsymbol{W} = \boldsymbol{By}$. The quadratic form $U = \boldsymbol{y'Ay}$ and $\boldsymbol{W}$ are independent if

MSReg

$$\boldsymbol{BVA} = \boldsymbol{0}.$$

$\hat{\beta}_1 = \sum d_i y_i$ (function of y)

Note: If $\boldsymbol{V} = \sigma^2 \boldsymbol{I}$, then U and $\boldsymbol{W}$ are independent if $\boldsymbol{BA} = \boldsymbol{0}$.

4. Let $\boldsymbol{B}$ be a $k \times k$ matrix. Let $V = \boldsymbol{y'By}$. The two quadratic forms, U and V, are independent if

$$\boldsymbol{AVB} = \boldsymbol{0}.$$

Note: If $\boldsymbol{V} = \sigma^2 \boldsymbol{I}$, then U and V are independent if $\boldsymbol{AB} = \boldsymbol{0}$.

$U \sim \chi^2(m)$

$V \sim \chi^2(n)$

$\dfrac{U/m}{V/n} \sim F_{m,n}$

The End