A study memo on linear regression

Contents

1 Introduction

This memo is a study memo on estimation and testing in linear regression.

2 Assumptions

In this note, we assume we assume various definitions and facts about random variables, probability measures, definitions and facts about the chi-squared distribution and the t-distribution (See from chapter1 to chapter 3 in [1] and see from chapter1 to chapter 4 in [2]).

Proposition 2.1. *Le A be a N-by-N symteric idempotent matrix and assume $rank(A) = m$ and $\{\epsilon_i\}_{i=1}^{N} \sim N(0, E_N)$. Then*

$$\epsilon^T A \epsilon \sim \chi^2(m) \tag{1}$$

3 General Topics

3.1 Multivariate normal distribution

Remark 3.1. *Let*

- *(S1)* $(\Omega, \mathscr{F}, P)$ *is a probability space.*
- *(S2)* $X := (X_1, ..., X_n)$ *is a vector of random variables.*
- *(S3)* A *is a* (m, n) *matrix.*
- *(A1)* $(X_1, ..., X_n) \sim N(0, E_n)$.

then $cov(AX) = AA^T$.

The following Proposition3.5 is used to prove the Proposition3.3 discussed later.

Proposition 3.1. *Let*

- *(A1)* $X := (X_1, X_2, ..., X_p)^T \sim N(\gamma, BB^T)$, *where* B *is a* (p, q) *matrix.*
- *(S1) Let* $s \in [1, p-1] \cap \mathbb{N}$ *and* $X^{(1)} := (X_1, ..., X_s)$ *and* $X^{(2)} := (X_{s+1}, ..., X_p)$.
- *(A2)* $cov(X^{(1)}, X^{(2)}) = 0$.

then $X^{(1)}$ *and* $X^{(2)}$ *are independent.*

Proof. The following proof consists of two steps.
STEP1. General case

In this step, we will show that it is enough to show the Proposition when $r := rank(B) = p \leq q$. For each $i \in \mathbb{N} \cap [1, p]$, let b_i be the i-th row vector of B. Let V_1 be the vector space generated from $b_1, b_2, ..., b_s$ and let V_2 be the vector space generated from $b_{s+1}, b_{s+2}, ..., b_p$. We can take $\{b_{\sigma(i)}\}_{i=1}^{r_1}$ is a basis of V_1 and $\{b_{\tau(i)}\}_{i=1}^{r_2}$ is a basis of V_2. Since $V_1 \perp V_2$, $\{b_{\sigma(i)}\}_{i=1}^{r_1} \cap \{b_{\tau(i)}\}_{i=1}^{r_2} = \phi$ and $\{b_{\sigma(i)}\}_{i=1}^{r_1} \cup \{b_{\tau(i)}\}_{i=1}^{r_2}$ are linear independent. So it is enough to show $\{b_{\sigma(i)}\}_{i=1}^{r_1}$ and $\{b_{\tau(i)}\}_{i=1}^{r_2}$ are independent when $rank(B)$ is the number of rows of B.

STEP2. Case when $rank(B) = p \leq q$

Let W be the orthogonal complement of the vector space generated from $b_1, b_2, ..., b_p$. We can take $c_1, ..., c_{(q-p)}$ which is an orthonormal basis of W and let

$C := \begin{bmatrix} c_1 \\ c_2 \\ ... \\ c_{(q-p)} \end{bmatrix}$, and let $D := \begin{bmatrix} B \\ C \end{bmatrix}$. By (A1), there are random variables $\{\epsilon\}_{i=1}^{p}$ on $(\Omega, \mathscr{F})$ and random variables $\{\ Y\}_{i=1}^{q-p}$ on $(\Omega, \mathscr{F})$ such that $\boldsymbol{\epsilon} := \{\epsilon\}_{i=1}^{q}$ are $i.i.d$ and $\epsilon_i \sim N(0,1)$ $(\forall i)$

and $Z := \begin{bmatrix} X \\ Y \end{bmatrix} = D\boldsymbol{\epsilon} + \gamma$ and $cov(Z) = DD^T$.

The distribution of Z has the density function $f_q : \mathbb{R}^q \ni x \mapsto c{\cdot}exp(x^T DD^T x) \in \mathbb{R}$, where c is a constant. By (A2) and the definition of C,

$DD^T = \begin{bmatrix} \Sigma_1 & 0 & 0 \\ 0 & \Sigma_2 & 0 \\ 0 & 0 & E_{(q-p)} \end{bmatrix}$, where Σ_1 and Σ_2 are symmetric positive definite matrixies. So the distribution of X has the density function $f_p : \mathbb{R}^p \ni x \mapsto d \cdot exp({x^{(1)}}^T \Sigma_1 x^{(1)}) \cdot exp({x^{(2)}}^T \Sigma_1 x^{(2)}) \in \mathbb{R}$, where d is a constant and $x^{(1)} = (x_1, ..., x_s)$ and $x^{(2)} = (x_{s+1}, ..., x_p)$. By the format of f_p, $X^{(1)}$ and $X^{(2)}$ are independent. □

3.2 Preliminaries for linear regression

Throughout this section, we assume the following settings.

Setting 3.1 (Linear regression)**.** *Let*

- *(S1)* $(\Omega, \mathscr{F}, P)$ *is a probability space.*
- *(S2) Let* $X := \{X_{i,j}\}_{\{1 \leq i \leq N, 1 \leq j \leq K\}}$ *be a* (N, K) *matrix.*
- *(A1)* $X^T X$ *is a regular matrix of order* $(K+1)$*.*
- *(S3) Let* $\epsilon := \{\epsilon_i\}_{\{1 \leq i \leq N\}}$ *be* N *random variables.*
- *(A2)* $\{\epsilon_i\}_{\{1 \leq i \leq N\}} \overset{iid}{\sim} N(\mathbf{0}, \Sigma_{i=1}^{N} \sigma^2 E_N)$*, where* $\sigma > 0$*.*
- *(S4) Let* $\{\beta_i\}_{\{1 \leq i \leq K\}}$ *be a real* K*-dimension vector.*
- *(S5) Let* $y := \{y_i\}_{\{1 \leq i \leq N\}}$ *be* N *random variables which are defined by the following equation.*

$$y = X\beta + \epsilon \tag{2}$$

Remark 3.2. *By (A1),*

$$rank(X) = K \tag{3}$$

Definition 3.1 (Least squares estimate)**.** *Let*

$$\hat{\beta} := (X^T X)^{-1}(X^T y) \tag{4}$$

We call $\hat{\beta}$ the least squares estimate of (2).
And let

$$\hat{y} := X\hat{\beta} \tag{5}$$

We call $\hat{y}$ the predicted values of (2).
Lastly let

$$\hat{e} := y - \hat{y} \tag{6}$$

We call $\hat{e}$ the residual of (2).

Remark 3.3. *$\hat{\beta}$ is the point which minimize $\mathbb{R}^K \ni z \mapsto |y - Xz|^2 \in [0, \infty)$. And*

$$\hat{\beta} := \beta + (X^T X)^{-1} X^T \epsilon \tag{7}$$

and for each i $\hat{\beta}_i \sim N(\beta_i, \sigma^2 \xi_i)$ and $\xi_i > 0$, where ξ_i is (i, i) component of $(X^T X)^{-1}$.

Definition 3.2 (Multivariate normal distribution)**.** *Let X_i be a random variable on $(\Omega, \mathscr{F})$ $(i = 1, 2, ..., N)$. $\{X_i\}_{i=1}^N \sim N(\gamma, \Sigma)$ if there is a natural number l and (N, l) matrix A and there are random variables $\{\epsilon\}_{i=1}^l$ on $(\Omega, \mathscr{F})$ such that $\boldsymbol{\epsilon} := \{\epsilon\}_{i=1}^l$ are i.i.d and $\epsilon_i \sim N(0, 1)$ $(\forall i)$ and $X = A\boldsymbol{\epsilon} + \gamma$ and $\Sigma = AA^T$.*

3.3 Interval estimation of regression coefficients

Proposition 3.2.

$$\frac{|\hat{e}|^2}{\sigma^2} \sim \chi^2(N - K) \tag{8}$$

Proof. The following holds.

$$\hat{e} = (E_N - X(X^T X)^{-1} X^T)\epsilon \tag{9}$$

Let $A := (E_N - X(X^T X)^{-1} X^T)$ then A is symmetric and idempotent. So each eigenvalue of A is 0 or 1. And $tr(A) = N - tr(X(X^T X)^{-1} X^T) = N - tr((X^T X)^{-1} X^T X) = N - K$ so $rank(A) = N - K$. So by Proposition2.1, $\frac{|\hat{e}|^2}{\sigma^2} \sim \chi^2(N - K)$. □

Proposition 3.3. *$\hat{\beta}$ and $\hat{e}$ are independent.*

Proof. By (7) and (9), $cov(\hat{e}, \hat{\beta}) = 0$. So by Proposition3.3 $\hat{\beta}$ and $\hat{e}$ are independent. □

By Remark and Proposition3.2 and Proposition3.2 and Proposition3.3, the folloing Proposition holds.

Proposition 3.4. *For each* $i \in \mathbb{N} \cap [1, K]$,

$$\frac{(\hat{\beta}_i - \beta_i)\sqrt{(N-K)}}{|\hat{e}|\sqrt{\xi_i}} \sim t(N-K) \tag{10}$$

In the above equation, t_{N-K} *is the t-distribution whose degrees of freedom is* $N-K$ *and* ξ_i *is* (i,i) *component of* $(X^T X)^{-1}$.

The following is a remark.

Proposition 3.5.

$$E(\frac{|\hat{e}|^2 \xi_i}{N-K}) = V(\hat{\beta}_i) \ (\forall i) \tag{11}$$

Proof. By Proposition3.2, $E(\frac{|\hat{e}|^2 \xi_i}{N-K}) = \sigma^2 \xi_i$. By Remark3.3, $V(\hat{\beta}_i) = \sigma^2 \xi_i$ □

By the above remak, $\frac{|\hat{e}|\sqrt{\xi_i}}{\sqrt{N-K}}$ is denoted by $se(\hat{\beta}_i)$.

3.4 Decomposition of TSS

Proposition 3.6.

$$(\hat{y}, \hat{e}) = 0 \tag{12}$$

Proof. By (7),

$$X^T \hat{y} = X^T X \hat{\beta} = X^T (X\beta + \boldsymbol{\epsilon}) = X^T y \tag{13}$$

So

$$\begin{aligned} (\hat{y}, \hat{e}) &= \beta^T X^T \hat{e} \\ &= \beta^T X^T (y - \hat{y}) \\ &= 0 \end{aligned}$$

□

Proposition 3.7. *Let*

(A1) There is a K*-by-*K *matrix* B *such that the first column of* XB *is* 1_N

then

$$\overline{\hat{y}} = \overline{y} \tag{14}$$

Proof. By (7),

$$X^T \hat{y} = X^T X \hat{\beta} = X^T (X\beta + \boldsymbol{\epsilon}) = X^T y \tag{15}$$

So the following holds.

$$B^T X^T \hat{e} = 0 \tag{16}$$

The fitst component of the $B^T X^T \hat{e}$ is $\overline{\hat{y}} - \overline{y}$. So $\overline{\hat{y}} = \overline{y}$. □

Proposition 3.8. *Let*

(S1) $TSS := |y - \overline{y}1_n|^2$

(S2) $RSS := |\hat{y} - \overline{y}1_n|^2$

(S3) $ESS := |y - \hat{y}|^2$

(A1) (A1) in Proposition3.7

then

$$TSS = RSS + ESS \tag{17}$$

Proof. Because

$$TSS = y^T(E - \frac{1}{N}1_{N,N})y \tag{18}$$

and

$$RSS = y^T(X^T(X^TX)^{-1}X - \frac{1}{N}1_{N,N})y \tag{19}$$

and

$$ESS = y^T(E - X^T(X^TX)^{-1}X)y \tag{20}$$

$TSS = RSS + ESS$. □

3.5 Cochran's theorem

Proposition 3.9. *Let*

(S1) $m \in \mathbb{N}$ *and* A_i*:N-by-N symmetric matrix* $(i = 1, 2, ..., m)$

(A1) $E_N = \Sigma_{i=1}^{m} A_i$

(A2) $N = \Sigma_{i=1}^{m} rank(A_i)$

then

$$A_iA_j = \delta_{i,j}A_i \ (\forall i, \ \forall j) \tag{21}$$

where $\delta_{i,j}$ *is a Kronecker delta.*

Proof. Let $V_i := A_i\mathbb{R}^N$ and $n_i := rank(A_i)$ and $\{v_{i,j}\}_{1 \le j \le n_i}$ be a basis of V_i $(i = 1, 2, ..., m)$. By (A1) and (A2), $\{v_{i,j}\}_{1 \le i \le m, 1 \le j \le n_i}$ is a basis of $\mathbb{R}^N$. and

$$\mathbb{R}^N = \bigoplus_{i=1}^{m} V_i \tag{22}$$

Let fix arbitary $i \in \{1, 2, ..., N\}$ and fix arbitary $x \in \mathbb{R}^N$. $A_ix = (\Sigma_{i=1}^{m} A_i)A_ix = (A_i)^2x + (\Sigma_{j \ne i}A_jA_ix$. By (22), $A_ix = {A_i}^2x$ and $A_jA_ix = 0$. □

By Proposition3.9 and Proposition2.1 and Proposition, the following theorem holds.

Proposition 3.10 (Cochran's theorem)**.**
We take over (S1) and (A1) in Proposition3.9. And let

(S2) $(\Omega, \mathscr{F}, P)$ *is a probability space.*

(A1) $\boldsymbol{\epsilon} \sim N(0, E_N)$

(S3) $Q_i := \boldsymbol{\epsilon}^T A_i \boldsymbol{\epsilon}$ $(i = 1, 2, ..., m)$

then $Q_i \sim \chi^2(rankA_i)$ $(\forall i)$ *and* Q_i *and* Q_j *are independent for all* $(i,j) \in \{(i,j)|i \neq j\}$

3.6 Testing

Throughout this subsection, we assume

$$\beta = (\beta_0, 0, 0, ..., 0)^T \tag{23}$$

and

$$X = \begin{pmatrix} 1 & x_{1,1} & x_{1,2} & ... & x_{1,L} \\ 1 & x_{2,1} & x_{2,2} & ... & x_{2,L} \\ ... & ... & ... & ... & \\ 1 & x_{N,1} & x_{N,2} & ... & x_{N,L} \end{pmatrix} \tag{24}$$

Then

$$X\beta = \beta_0 1_{N,1} \tag{25}$$

So

$$\begin{aligned} \hat{y} &= X(X^TX)^{-1}X^Ty \\ &= X(X^TX)^{-1}X^T(X\beta + \epsilon) \\ &= \beta_0 1_{N,1} + X(X^TX)^{-1}X^T\epsilon \end{aligned} \tag{26}$$

And

$$\bar{y}1_{N,1} = \beta_0 \frac{1}{N} 1_{N,1} + 1_{N,N}\epsilon \tag{27}$$

Consequently,

$$RSS = \epsilon^T(X(X^TX)^{-1}X^T - \frac{1}{N}1_{N,1})\epsilon \tag{28}$$

Because $X(X^TX)^{-1}X^T$ is symmetric, $X(X^TX)^{-1}X^T$ and $\frac{1}{N}1_{N,1}$ are commutative.

And because $X(X^TX)^{-1}X^T$ is idempotent and symmetric, $(X(X^TX)^{-1}X^T - \frac{1}{N}1_{N,1})$ is idempotent and symmetric.

$rank(X(X^TX)^{-1}X^T - \frac{1}{N}1_{N,1}) = tr(X(X^TX)^{-1}X^T - \frac{1}{N}1_{N,1}) = L$

So by Proposition3.10, RSS and ESS are independent and $RSS \sim \chi^2(L)$ and $ESS \sim \chi^2(N - L - 1)$.

So,

$$\frac{\frac{RSS}{L}}{\frac{ESS}{N-L-1}} \sim F(L, N - L - 1) \tag{29}$$

4 Simple linear regression

Throughout this subsection, we set

$$T_x = \sum_{i=1}^{n} x_i,\ T_y = \sum_{i=1}^{n} y_i,\ T_{x,x} = \sum_{i=1}^{n} x_i^2,\ T_{x,y} = \sum_{i=1}^{n} x_i y_i \tag{30}$$

4.1 Case1: there is intercept

Throughout this subsection, we assume

$$X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ ... & ... \\ 1 & x_n \end{pmatrix} \tag{31}$$

Then

$$\begin{aligned} \hat{\beta} &= \begin{pmatrix} \hat{\alpha} \\ \hat{\gamma} \end{pmatrix} \\ &= (X^T X)^{-1} X^T y \\ &= (\begin{pmatrix} 1 & 1 & ... & 1 \\ x_1 & x_2 & ... & x_n \end{pmatrix} \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ ... & ... \\ 1 & x_n \end{pmatrix})^{-1} X^T y \\ &= \begin{pmatrix} n & T_x \\ T_x & T_{x,x} \end{pmatrix}^{-1} X^T y \\ &= \frac{1}{nT_{x,x} - T_x^2} \begin{pmatrix} T_{x,x} & -T_x \\ -T_x & n \end{pmatrix} \begin{pmatrix} T_y \\ T_{x,y} \end{pmatrix} \end{aligned} \tag{32}$$

So

$$\begin{aligned} \hat{\gamma} &= \frac{nT_{x,y} - T_x T_y}{nT_{x,x} - T_x^2} \\ &= \frac{T_{x,y} - \frac{1}{n} T_x T_y}{T_{x,x} - \frac{1}{n} T_x^2} \\ &= \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \end{aligned} \tag{33}$$

Consequently,

$$\hat{\gamma} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{34}$$

4.2 Case2: there is no intercept

Throughout this subsection, we assume

$$X = (x_1, x_2, ..., x_n)^T \tag{35}$$

Then

$$\hat{\beta} = \frac{T_{x,y}}{T_{x,x}} \tag{36}$$

5 Estimation about population mean

Throughout this section, we assume $X = 1_N$ is one and we define μ by $\beta = \mu 1_1$. The followings hold.

$$X^T X = N \tag{37}$$

$$Y := X(X^T X)^{-1} X^T = \frac{1}{N} 1_{N,N} \tag{38}$$

$$\hat{e} := y - \overline{y} 1_N \tag{39}$$

$$\frac{|\hat{e}|^2}{\sigma^2} \sim \chi^2(N-1) \tag{40}$$

$$\frac{(\mu - \overline{y})\sqrt{N(N-1)}}{|y - \overline{y}|} \sim t(N-1) \tag{41}$$

6 Estimation about difference between two population means

Throughout this section, we assume

$$X = \begin{pmatrix} 1_M & 0 \\ 0 & 1_N \end{pmatrix} \tag{42}$$

and

$$\beta = \begin{pmatrix} \mu_1 1_M \\ \mu_2 1_N \end{pmatrix} \tag{43}$$

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} := y \tag{44}$$

$$\begin{pmatrix} \epsilon_1 \\ \epsilon_2 \end{pmatrix} := \epsilon \tag{45}$$

Then the followings hold.

$$X^T X = \begin{pmatrix} M & 0 \\ 0 & N \end{pmatrix} \tag{46}$$

$$Y := \begin{pmatrix} \frac{1}{M} 1_{M,M} & 0 \\ 0 & \frac{1}{N} 1_{N,N} \end{pmatrix} \tag{47}$$

$$\mu_1 = (\hat{y_1})_1 = \overline{y_1} + \overline{\epsilon_1} \tag{48}$$

$$\mu_2 = (\hat{y_2})_1 = \overline{y_2} + \overline{\epsilon_2} \tag{49}$$

So, by reproductive property of normal distribution,

$$\mu_1 - \mu_2 - (\overline{y_1} - \overline{y_2}) \sim N(0, (\frac{1}{M} + \frac{1}{N})\sigma^2) \tag{50}$$

And the following holds.

$$|\hat{e}|^2 = |y_1 - \mu_1 1_M|^2 + |y_2 - \mu_2 1_N|^2 \tag{51}$$

By Proposition3.3, $(\mu_1 - \mu_2 - (\overline{y_1} - \overline{y_2}))$ and $|y_1 - \mu_1 1_M|^2 + |y_2 - \mu_2 1_N|^2$ are independent.

Consequently, the following holds.

$$\frac{(\mu_1 - \mu_2 - (\overline{y_1} - \overline{y_2}))\sqrt{M+N-2}}{\sqrt{(|y_1 - \mu_1 1_M|^2 + |y_2 - \mu_2 1_N|^2)(\frac{1}{M} + \frac{1}{N})}} \sim t(M+N-2) \tag{52}$$

7 One way analysis of variance

Throughout this section we set

$$y := (y_{1,1}, ..., y_{1,n_1}, y_{2,1}, ..., y_{2,n_2}, ..., y_{K,1}, ..., y_{K,n_K})^T \tag{53}$$

$$\beta := (\mu_1, \mu_2, ..., \mu_K)^T \tag{54}$$

$$\bar{y}_{i,\cdot} := \frac{\sum_{j=1}^{n_i} y_{i,j}}{n_i} \ (i = 1, 2, ..., K) \tag{55}$$

$$X := \begin{pmatrix} 1_{n_1} & O & O & O \\ 1_{n_2} & 1_{n_2} & O & O \\ ... & ... & ... & ... \\ 1_{n_K} & O & O & 1_{n_K} \end{pmatrix} \tag{56}$$

Then

$$Y := X(X^TX)^{-1}X^T := \begin{pmatrix} \frac{1}{n_1}1_{n_1,n_1} & O & O & O \\ O & \frac{1}{n_2}1_{n_2,n_2} & O & O \\ ... & ... & ... & ... \\ O & O & O & \frac{1}{n_K}1_{n_K,n_K} \end{pmatrix} \quad (57)$$

In this subsection, hereafter, we assume there is a real number μ such that

$$\beta = \mu 1_K \quad (58)$$

Then the followings holds.

$$TSS = \boldsymbol{\epsilon}^T(E_N - \frac{1}{N}1_{N,N})\boldsymbol{\epsilon} \quad (59)$$

$$ESS = \boldsymbol{\epsilon}^T(Y - \frac{1}{N}1_{N,N})\boldsymbol{\epsilon} \quad (60)$$

$$rank(Y - \frac{1}{N}1_{N,N}) = K - 1 \quad (61)$$

$$RSS = \boldsymbol{\epsilon}^T(E_N - Y)\boldsymbol{\epsilon} \quad (62)$$

$$rank(E_N - Y) = N - K \quad (63)$$

So, by Cohchran's theorem, ESS and RSS are independent, and $ESS \sim \chi^2(K-1)$ and $RSS \sim \chi^2(N-K)$.

Consequently, the following theoem holds.

Theorem 7.1. *Under the setting(56) and the assumption(58)*

$$(ESS/(K-1))/(RSS/(N-K)) \sim F(K-1, N-K) \quad (64)$$

And the followings hold.

$$(X^TX)^{-1} = \begin{pmatrix} \frac{1}{n_1} & 0 & ... & 0 \\ 0 & \frac{1}{n_2} & ... & 0 \\ ... & ... & ... & ... \\ 0 & ... & 0 & \frac{1}{n_K} \end{pmatrix} \quad (65)$$

$$\hat{\beta} = (\bar{y}_{1,\cdot}, \bar{y}_{2,\cdot}, ..., \bar{y}_{K,\cdot})^T \quad (66)$$

So, by Proposition3.4, the following theoem holds.

Theorem 7.2. *Under the setting(56)*

$$(\bar{y}_{i,\cdot} - \mu_i)\sqrt{\frac{(N-K)n_i}{ESS}} \sim t(N-K) \quad (67)$$

References

[1] Tadahisa Funaki, Probability Theory(in Japanese), ISBN-13 978-4254116007.

[2] Akimichi Takemura, Modern Mathematical STATISTICS(in Japanese), ISBN-13 978-4780608601.

[3] Hideki Iwaki, Review of Statistics(in Japanese), lecture notes in kyoto sangyo university. http://www.cc.kyoto-su.ac.jp/ iwaki/Stat0.pdf