Lecture 20: Linear model, the LSE, and UMVUE

Linear Models

One of the most useful statistical models is

$$X_i = \beta^\tau Z_i + \varepsilon_i, \qquad i = 1, ..., n,$$

where X_i is the ith observation and is often called the ith response;
β is a p-vector of unknown parameters (main parameters of interest), $p < n$;
Z_i is the ith value of a p-vector of explanatory variables (or covariates);
$\varepsilon_1, ..., \varepsilon_n$ are random errors (not observed).

Data: $(X_1, Z_1), ..., (X_n, Z_n)$.
Z_i's are nonrandom or given values of a random p-vector, in which case our analysis is conditioned on $Z_1, ..., Z_n$.
A matrix form of the model is

$$X = Z\beta + \varepsilon, \tag{1}$$

where $X = (X_1, ..., X_n)$, $\varepsilon = (\varepsilon_1, ..., \varepsilon_n)$, and $Z =$ the $n \times p$ matrix whose ith row is the vector Z_i, $i = 1, ..., n$.

Definition 3.4 (LSE).

Suppose that the range of β in model (6) is $B \subset \mathscr{R}^p$.
A *least squares estimator* (LSE) of β is defined to be any $\widehat{\beta} \in B$ such that

$$\|X - Z\widehat{\beta}\|^2 = \min_{b \in B} \|X - Zb\|^2.$$

For any $l \in \mathscr{R}^p$, $l^\tau \widehat{\beta}$ is called an LSE of $l^\tau \beta$.
Throughout this book, we consider $B = \mathscr{R}^p$ unless otherwise stated.
Differentiating $\|X - Zb\|^2$ w.r.t. b, we obtain that any solution of

$$Z^\tau Z b = Z^\tau X$$

is an LSE of β.

Full rank Z

If the rank of the matrix Z is p, in which case $(Z^\tau Z)^{-1}$ exists and Z is said to be of full rank, then there is a unique LSE, which is

$$\widehat{\beta} = (Z^\tau Z)^{-1} Z^\tau X.$$

Not full rank Z

If Z is not of full rank, then there are infinitely many LSE's of β.
Any LSE of β is of the form

$$\widehat{\beta} = (Z^\tau Z)^- Z^\tau X,$$

where $(Z^\tau Z)^-$ is called a *generalized inverse* of $Z^\tau Z$ and satisfies

$$Z^\tau Z (Z^\tau Z)^- Z^\tau Z = Z^\tau Z.$$

Generalized inverse matrices are not unique unless Z is of full rank, in which case $(Z^\tau Z)^- = (Z^\tau Z)^{-1}$

Assumptions

To study properties of LSE's of β, we need some assumptions on the distribution of X or ε (conditional on Z if Z is random and ε and Z are independent).

A1: ε is distributed as $N_n(0, \sigma^2 I_n)$ with an unknown $\sigma^2 > 0$.

A2: $E(\varepsilon) = 0$ and $\mathrm{Var}(\varepsilon) = \sigma^2 I_n$ with an unknown $\sigma^2 > 0$.

A3: $E(\varepsilon) = 0$ and $\mathrm{Var}(\varepsilon)$ is an unknown matrix.

Remarks

- Assumption A1 is the strongest and implies a parametric model.
- We may assume a slightly more general assumption that ε has the $N_n(0,\sigma^2 D)$ distribution with unknown σ^2 but a known positive definite matrix D.
  Let $D^{-1/2}$ be the inverse of the square root matrix of D.
  Then model (6) with assumption A1 holds if we replace X, Z, and ε by the transformed variables $\tilde{X} = D^{-1/2}X$, $\tilde{Z} = D^{-1/2}Z$, and $\tilde{\varepsilon} = D^{-1/2}\varepsilon$, respectively.
- A similar conclusion can be made for assumption A2.
- Under assumption A1, the distribution of X is $N_n(Z\beta,\sigma^2 I_n)$, which is in an exponential family $\mathscr{P}$ with parameter $\theta = (\beta,\sigma^2) \in \mathscr{R}^p \times (0,\infty)$.
- However, if the matrix Z is not of full rank, then $\mathscr{P}$ is not identifiable (see §2.1.2), since $Z\beta_1 = Z\beta_2$ does not imply $\beta_1 = \beta_2$.

Remarks

- Suppose that the rank of Z is $r \le p$.
  Then there is an $n \times r$ submatrix Z_* of Z such that

$$Z = Z_* Q \tag{2}$$

  and Z_* is of rank r, where Q is a fixed $r \times p$ matrix, and

$$Z\beta = Z_* Q\beta.$$

- $\mathscr{P}$ is identifiable if we consider the reparameterization $\tilde{\beta} = Q\beta$.
- The new parameter $\tilde{\beta}$ is in a subspace of $\mathscr{R}^p$ with dimension r.
- In many applications, we are interested in estimating some linear functions of β, i.e., $\vartheta = l^\tau \beta$ for some $l \in \mathscr{R}^p$.
- From the previous discussion, however, estimation of $l^\tau \beta$ is meaningless unless $l = Q^\tau c$ for some $c \in \mathscr{R}^r$ so that

$$l^\tau \beta = c^\tau Q\beta = c^\tau \tilde{\beta}.$$

The following result shows that $l^\tau\beta$ is estimable if $l = Q^\tau c$, which is also necessary for $l^\tau\beta$ to be estimable under assumption A1.

Theorem 3.6

Assume model (6) with assumption A3.

(i) A necessary and sufficient condition for $l \in \mathscr{R}^p$ being $Q^\tau c$ for some $c \in \mathscr{R}^r$ is $l \in \mathscr{R}(Z) = \mathscr{R}(Z^\tau Z)$, where Q is given by (2) and $\mathscr{R}(A)$ is the smallest linear subspace containing all rows of A.

(ii) If $l \in \mathscr{R}(Z)$, then the LSE $l^\tau\widehat{\beta}$ is unique and unbiased for $l^\tau\beta$.

(iii) If $l \notin \mathscr{R}(Z)$ and assumption A1 holds, then $l^\tau\beta$ is not estimable.

Proof

(i) Note that $a \in \mathscr{R}(A)$ iff $a = A^\tau b$ for some vector b.
If $l = Q^\tau c$, then

$$l = Q^\tau c = Q^\tau Z_*^\tau Z_*(Z_*^\tau Z_*)^{-1}c = Z^\tau[Z_*(Z_*^\tau Z_*)^{-1}c].$$

Hence $l \in \mathscr{R}(Z)$.

Proof (continued)

If $l \in \mathscr{R}(Z)$, then $l = Z^\tau \zeta$ for some ζ and

$$l = (Z_* Q)^\tau \zeta = Q^\tau c, \qquad c = Z_*^\tau \zeta.$$

(ii) If $l \in \mathscr{R}(Z) = \mathscr{R}(Z^\tau Z)$, then $l = Z^\tau Z \zeta$ for some ζ and by $\widehat{\beta} = (Z^\tau Z)^- Z^\tau X$,

$$E(l^\tau \widehat{\beta}) = E[l^\tau (Z^\tau Z)^- Z^\tau X] = \zeta^\tau Z^\tau Z (Z^\tau Z)^- Z^\tau Z \beta = \zeta^\tau Z^\tau Z \beta = l^\tau \beta.$$

If $\bar{\beta}$ is any other LSE of β, then, by $Z^\tau Z \bar{\beta} = Z^\tau X$,

$$l^\tau \widehat{\beta} - l^\tau \bar{\beta} = \zeta^\tau (Z^\tau Z)(\widehat{\beta} - \bar{\beta}) = \zeta^\tau (Z^\tau X - Z^\tau X) = 0.$$

(iii) Under A1, if there is an estimator $h(X,Z)$ unbiased for $l^\tau \beta$, then

$$l^\tau \beta = \int_{\mathscr{R}^n} h(x,Z)(2\pi)^{-n/2} \sigma^{-n} \exp\left\{-\tfrac{1}{2\sigma^2}\|x - Z\beta\|^2\right\} dx.$$

Differentiating w.r.t. β and applying Theorem 2.1 lead to

$$l^\tau = Z^\tau \int_{\mathscr{R}^n} h(x,Z)(2\pi)^{-n/2} \sigma^{-n-2} (x - Z\beta) \exp\left\{-\tfrac{1}{2\sigma^2}\|x - Z\beta\|^2\right\} dx,$$

which implies $l \in \mathscr{R}(Z)$.

Example 3.12 (Simple linear regression)

Let $\beta = (\beta_0, \beta_1) \in \mathscr{R}^2$ and $Z_i = (1, t_i)$, $t_i \in \mathscr{R}$, $i = 1, ..., n$.
Then model (6) is called a *simple linear regression* model.
It turns out that

$$\begin{pmatrix} n & \sum_{i=1}^n t_i \\ \sum_{i=1}^n t_i & \sum_{i=1}^n t_i^2 \end{pmatrix}.$$

This matrix is invertible iff some t_i's are different.
Thus, if some t_i's are different, then the unique unbiased LSE of $l^\tau \beta$ for any $l \in \mathscr{R}^2$ is $l^\tau (Z^\tau Z)^{-1} Z^\tau X$, which has the normal distribution if assumption A1 holds.

The result can be easily extended to the case of *polynomial regression* of order p in which $\beta = (\beta_0, \beta_1, ..., \beta_{p-1})$ and $Z_i = (1, t_i, ..., t_i^{p-1})$.

Example 3.13 (One-way ANOVA)

Suppose that $n = \sum_{j=1}^m n_j$ with m positive integers $n_1, ..., n_m$ and that

$$X_i = \mu_j + \varepsilon_i, \qquad i = k_{j-1} + 1, ..., k_j,\ j = 1, ..., m,$$

where $k_0 = 0$, $k_j = \sum_{l=1}^j n_l$, $j = 1, ..., m$, and $(\mu_1, ..., \mu_m) = \beta$.

Let J_m be the m-vector of ones.
Then the matrix Z in this case is a block diagonal matrix with J_{n_j} as the jth diagonal column.
Consequently, $Z^\tau Z$ is an $m \times m$ diagonal matrix whose jth diagonal element is n_j.
Thus, $Z^\tau Z$ is invertible and the unique LSE of β is the m-vector whose jth component is

$$\frac{1}{n_j} \sum_{i=k_{j-1}+1}^{k_j} X_i, \qquad j = 1, ..., m.$$

Sometimes it is more convenient to use the following notation:

$$X_{ij} = X_{k_{i-1}+j},\ \varepsilon_{ij} = \varepsilon_{k_{i-1}+j}, \qquad j = 1, ..., n_i, i = 1, ..., m,$$

and

$$\mu_i = \mu + \alpha_i, \qquad i = 1, ..., m.$$

Then our model becomes

$$X_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \qquad j = 1, ..., n_i, i = 1, ..., m, \tag{3}$$

which is called a *one-way analysis of variance* (ANOVA) model.

Under model (3), $\beta = (\mu, \alpha_1, ..., \alpha_m) \in \mathscr{R}^{m+1}$.
The matrix Z under model (3) is not of full rank.

An LSE of β under model (3) is

$$\widehat{\beta} = \left(\bar{X}, \bar{X}_{1\cdot} - \bar{X}, ..., \bar{X}_{m\cdot} - \bar{X}\right),$$

where $\bar{X}$ is still the sample mean of X_{ij}'s and $\bar{X}_{i\cdot}$ is the sample mean of the ith group $\{X_{ij}, j = 1, ..., n_i\}$.

The notation used in model (3) allows us to generalize the one-way ANOVA model to any s-way ANOVA model with a positive integer s under the so-called factorial experiments.

Example 3.14 (Two-way balanced ANOVA)

Suppose that

$$X_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}, \quad i = 1, ..., a, j = 1, ..., b, k = 1, ..., c, \qquad (4)$$

where a, b, and c are some positive integers.
Model (4) is called a two-way balanced ANOVA model.

If we view model (4) as a special case of model (6), then the parameter vector β is

$$\beta = (\mu, \alpha_1, ..., \alpha_a, \beta_1, ..., \beta_b, \gamma_{11}, ..., \gamma_{1b}, ..., \gamma_{a1}, ..., \gamma_{ab}). \qquad (5)$$

One can obtain the matrix Z and show that it is $n \times p$, where $n = abc$ and $p = 1 + a + b + ab$, and is of rank $ab < p$.

It can also be shown that an LSE of β is given by the right-hand side of (5) with μ, α_i, β_j, and γ_{ij} replaced by $\widehat{\mu}$, $\widehat{\alpha}_i$, $\widehat{\beta}_j$, and $\widehat{\gamma}_{ij}$, respectively, where

$$\widehat{\mu} = \bar{X}_{\cdots},$$

$$\widehat{\alpha}_i = \bar{X}_{i\cdot\cdot} - \bar{X}_{\cdots},$$

$$\widehat{\beta}_j = \bar{X}_{\cdot j\cdot} - \bar{X}_{\cdots},$$

$$\widehat{\gamma}_{ij} = \bar{X}_{ij\cdot} - \bar{X}_{i\cdot\cdot} - \bar{X}_{\cdot j\cdot} + \bar{X}_{\cdots},$$

and a dot is used to denote averaging over the indicated subscript, e.g., with a fixed j,

$$\bar{X}_{\cdot j\cdot} = \frac{1}{ac} \sum_{i=1}^{a} \sum_{k=1}^{c} X_{ijk}$$

Theorem 3.7 (UMVUE).

Consider model

$$X = Z\beta + \varepsilon \tag{6}$$

with assumption A1 (ε is distributed as $N_n(0, \sigma^2 I_n)$ with an unknown $\sigma^2 > 0$).
Then

(i) The LSE $l^\tau \widehat{\beta}$ is the UMVUE of $l^\tau \beta$ for any estimable $l^\tau \beta$.

(ii) The UMVUE of σ^2 is $\widehat{\sigma}^2 = (n-r)^{-1} \|X - Z\widehat{\beta}\|^2$, where r is the rank of Z.

Proof of (i)

Let $\widehat{\beta}$ be an LSE of β.
By $Z^\tau Zb = Z^\tau X$,

$$(X - Z\widehat{\beta})^\tau Z(\widehat{\beta} - \beta) = (X^\tau Z - X^\tau Z)(\widehat{\beta} - \beta) = 0$$

and, hence,

$$\begin{aligned}\|X - Z\beta\|^2 &= \|X - Z\widehat{\beta} + Z\widehat{\beta} - Z\beta\|^2 \\ &= \|X - Z\widehat{\beta}\|^2 + \|Z\widehat{\beta} - Z\beta\|^2 \\ &= \|X - Z\widehat{\beta}\|^2 - 2\beta^\tau Z^\tau X + \|Z\beta\|^2 + \|Z\widehat{\beta}\|^2.\end{aligned}$$

Using this result and assumption A1, we obtain the following joint Lebesgue p.d.f. of X:

$$(2\pi\sigma^2)^{-n/2}\exp\left\{\frac{\beta^\tau Z^\tau x}{\sigma^2} - \frac{\|x - Z\widehat{\beta}\|^2 + \|Z\widehat{\beta}\|^2}{2\sigma^2} - \frac{\|Z\beta\|^2}{2\sigma^2}\right\}.$$

By Proposition 2.1 and the fact that $Z\widehat{\beta} = Z(Z^\tau Z)^- Z^\tau X$ is a function of $Z^\tau X$, the statistic $(Z^\tau X, \|X - Z\widehat{\beta}\|^2)$ is complete and sufficient for $\theta = (\beta, \sigma^2)$.

Note that $\widehat{\beta}$ is a function of $Z^\tau X$ and, hence, a function of the complete sufficient statistic.

If $l^\tau\beta$ is estimable, then $l^\tau\widehat{\beta}$ is unbiased for $l^\tau\beta$ (Theorem 3.6) and, hence, $l^\tau\widehat{\beta}$ is the UMVUE of $l^\tau\beta$.

Proof of (ii)

From $\|X - Z\beta\|^2 = \|X - Z\widehat{\beta}\|^2 + \|Z\widehat{\beta} - Z\beta\|^2$ and $E(Z\widehat{\beta}) = Z\beta$ (Theorem 3.6),

$$
\begin{aligned}
E\|X - Z\widehat{\beta}\|^2 &= E(X - Z\beta)^\tau (X - Z\beta) - E(\beta - \widehat{\beta})^\tau Z^\tau Z(\beta - \widehat{\beta}) \\
&= \mathrm{tr}\left(\mathrm{Var}(X) - \mathrm{Var}(Z\widehat{\beta})\right) \\
&= \sigma^2 [n - \mathrm{tr}\left(Z(Z^\tau Z)^- Z^\tau Z (Z^\tau Z)^- Z^\tau\right)] \\
&= \sigma^2 [n - \mathrm{tr}\left((Z^\tau Z)^- Z^\tau Z\right)].
\end{aligned}
$$

Since each row of $Z \in \mathscr{R}(Z)$, $Z\widehat{\beta}$ does not depend on the choice of $(Z^\tau Z)^-$ in $\widehat{\beta} = (Z^\tau Z)^- Z^\tau X$ (Theorem 3.6).

Hence, we can evaluate $\mathrm{tr}((Z^\tau Z)^- Z^\tau Z)$ using a particular $(Z^\tau Z)^-$.

From the theory of linear algebra, there exists a $p \times p$ matrix C such that $CC^\tau = I_p$ and

$$
C^\tau (Z^\tau Z) C = \begin{pmatrix} \Lambda & 0 \\ 0 & 0 \end{pmatrix},
$$

Then, a particular choice of $(Z^\tau Z)^-$ is

$$(Z^\tau Z)^- = C \begin{pmatrix} \Lambda^{-1} & 0 \\ 0 & 0 \end{pmatrix} C^\tau \tag{7}$$

and

$$(Z^\tau Z)^- Z^\tau Z = C \begin{pmatrix} I_r & 0 \\ 0 & 0 \end{pmatrix} C^\tau$$

whose trace is r.

Hence $\widehat{\sigma}^2$ is the UMVUE of σ^2, since it is a function of the complete sufficient statistic and

$$E\widehat{\sigma}^2 = (n-r)^{-1} E\|X - Z\widehat{\beta}\|^2 = \sigma^2.$$

Residual vector

- The vector $X - Z\widehat{\beta}$ is called the *residual vector* and $\|X - Z\widehat{\beta}\|^2$ is called the *sum of squared residuals* and is denoted by SSR.
- The estimator $\widehat{\sigma}^2$ is then equal to $SSR/(n-r)$.

- The Fisher information matrix is

$$\frac{1}{\sigma^2}\begin{pmatrix} Z^\tau Z & 0 \\ 0 & \frac{n}{2\sigma^2} \end{pmatrix}$$

- The UMVUE $l^\tau \widehat{\beta}$ attains the information lower bound, but not $\widehat{\sigma}^2$.
- Since

$$X - Z\widehat{\beta} = [I_n - Z(Z^\tau Z)^- Z^\tau]X$$

and

$$l^\tau \widehat{\beta} = l^\tau (Z^\tau Z)^- Z^\tau X$$

are linear in X, they are normally distributed under assumption A1.
- Also, using the generalized inverse matrix in (7), we obtain that

$$[I_n - Z(Z^\tau Z)^- Z^\tau]Z(Z^\tau Z)^- = Z(Z^\tau Z)^- - Z(Z^\tau Z)^- Z^\tau Z(Z^\tau Z)^- = 0,$$

which implies that $\widehat{\sigma}^2$ and $l^\tau \widehat{\beta}$ are independent (Exercise 58 in §1.6) for any estimable $l^\tau \beta$.
- $Z(Z^\tau Z)^- Z^\tau$ is a projection matrix, $[Z(Z^\tau Z)^- Z^\tau]^2 = Z(Z^\tau Z)^- Z^\tau$, hence

$$SSR = X^\tau [I_n - Z(Z^\tau Z)^- Z^\tau]X.$$

- The rank of $Z(Z^\tau Z)^- Z^\tau$ is $\text{tr}(Z(Z^\tau Z)^- Z^\tau) = r$.
- Similarly, the rank of the projection matrix $I_n - Z(Z^\tau Z)^- Z^\tau$ is $n - r$.
- From

$$X^\tau X = X^\tau [Z(Z^\tau Z)^- Z^\tau] X + X^\tau [I_n - Z(Z^\tau Z)^- Z^\tau] X$$

and Theorem 1.5 (Cochran's theorem), SSR/σ^2 has the chi-square distribution $\chi^2_{n-r}(\delta)$ with

$$\delta = \sigma^{-2} \beta^\tau Z^\tau [I_n - Z(Z^\tau Z)^- Z^\tau] Z\beta = 0.$$

Thus, we have proved the following result.

Theorem 3.8.

Consider model (6) with assumption A1. For any estimable parameter $l^\tau \beta$, the UMVUE's $l^\tau \widehat{\beta}$ and $\widehat{\sigma}^2$ are independent; the distribution of $l^\tau \widehat{\beta}$ is $N(l^\tau \beta, \sigma^2 l^\tau (Z^\tau Z)^- l)$; and $(n - r)\widehat{\sigma}^2/\sigma^2$ has the chi-square distribution χ^2_{n-r}.