Lecture 35: The LSE and estimability

One of the most useful statistical models

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

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 $Z_1, ..., Z_n$.
$X = (X_1, ..., X_n)$, $\varepsilon = (\varepsilon_1, ..., \varepsilon_n)$
$Z =$ the $n \times p$ matrix whose ith row is the vector Z_i, $i = 1, ..., n$
A matrix form of model (1) is

$$X = Z\beta + \varepsilon. \tag{2}$$

Definition 3.4. Suppose that the range of β in model (2) is $B \subset \mathcal{R}^p$. A *least squares estimator* (LSE) of β is defined to be any $\hat{\beta} \in B$ such that

$$\|X - Z\hat{\beta}\|^2 = \min_{b \in B} \|X - Zb\|^2. \tag{3}$$

For any $l \in \mathcal{R}^p$, $l^\tau\hat{\beta}$ is called an LSE of $l^\tau\beta$.

Throughout this book, we consider $B = \mathcal{R}^p$ unless otherwise stated.
Differentiating $\|X - Zb\|^2$ w.r.t. b, we obtain that any solution of

$$Z^\tau Z b = Z^\tau X \tag{4}$$

is an LSE of β.
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

$$\hat{\beta} = (Z^\tau Z)^{-1} Z^\tau X. \tag{5}$$

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

$$\hat{\beta} = (Z^\tau Z)^{-} Z^\tau X, \tag{6}$$

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 $(Z^\tau Z)^- = (Z^\tau Z)^{-1}$ and (6) reduces to (5).

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).

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

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

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

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 (2) 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 $\mathcal{P}$ with parameter $\theta = (\beta, \sigma^2) \in \mathcal{R}^p \times (0, \infty)$.
However, if the matrix Z is not of full rank, then $\mathcal{P}$ is not identifiable (see §2.1.2), since $Z\beta_1 = Z\beta_2$ does not imply $\beta_1 = \beta_2$.

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

$$Z = Z_* Q \tag{7}$$

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

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

$\mathcal{P}$ is identifiable if we consider the reparameterization $\tilde{\beta} = Q\beta$.
The new parameter $\tilde{\beta}$ is in a subspace of $\mathcal{R}^p$ with dimension r.

In many applications, we are interested in estimating some linear functions of β, i.e., $\vartheta = l^\tau \beta$ for some $l \in \mathcal{R}^p$.
From the previous discussion, however, estimation of $l^\tau\beta$ is meaningless unless $l = Q^\tau c$ for some $c \in \mathcal{R}^r$ 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 (2) with assumption A3.
(i) A necessary and sufficient condition for $l \in \mathcal{R}^p$ being $Q^\tau c$ for some $c \in \mathcal{R}^r$ is $l \in \mathcal{R}(Z) = \mathcal{R}(Z^\tau Z)$, where Q is given by (7) and $\mathcal{R}(A)$ is the smallest linear subspace containing all rows of A.
(ii) If $l \in \mathcal{R}(Z)$, then the LSE $l^\tau\hat{\beta}$ is unique and unbiased for $l^\tau\beta$.

(iii) If $l \notin \mathcal{R}(Z)$ and assumption A1 holds, then $l^\tau \beta$ is not estimable.
Proof. (i) Note that $a \in \mathcal{R}(A)$ if and only if $a = A^\tau b$ 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 $l \in \mathcal{R}(Z)$. If $l \in \mathcal{R}(Z)$, then $l = Z^\tau \zeta$ for some ζ and

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

with $c = Z_*^\tau \zeta$.
(ii) If $l \in \mathcal{R}(Z) = \mathcal{R}(Z^\tau Z)$, then $l = Z^\tau Z \zeta$ for some ζ and by (6),

$$\begin{aligned} E(l^\tau \hat{\beta}) &= E[l^\tau (Z^\tau Z)^- Z^\tau X] \\ &= \zeta^\tau Z^\tau Z (Z^\tau Z)^- Z^\tau Z \beta \\ &= \zeta^\tau Z^\tau Z \beta \\ &= l^\tau \beta. \end{aligned}$$

If $\bar{\beta}$ is any other LSE of β, then, by (4),

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

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

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

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

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

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

Example 3.12 (Simple linear regression). Let $\beta = (\beta_0, \beta_1) \in \mathcal{R}^2$ and $Z_i = (1, t_i)$, $t_i \in \mathcal{R}$, $i = 1, ..., n$.
Then model (1) or (2) is called a *simple linear regression* model.
It turns out that

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

This matrix is invertible if and only if 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 $l \in \mathcal{R}^2$ 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 $n_j^{-1} \sum_{i=k_{j-1}+1}^{k_j} X_i$, $j = 1, ..., m$.

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

$$X_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \qquad j = 1, ..., n_i, i = 1, ..., m, \tag{8}$$

which is called a *one-way analysis of variance* (ANOVA) model.
Under model (8), $\beta = (\mu, \alpha_1, ..., \alpha_m) \in \mathcal{R}^{m+1}$.
The matrix Z under model (8) is not of full rank.
An LSE of β under model (8) is

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

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 (8) 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

$$X_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}, \quad i = 1, ..., a, j = 1, ..., b, k = 1, ..., c, \tag{9}$$

where a, b, and c are some positive integers.
Model (9) is called a two-way balanced ANOVA model.
If we view model (9) as a special case of model (2), then the parameter vector β is

$$\beta = (\mu, \alpha_1, ..., \alpha_a, \beta_1, ..., \beta_b, \gamma_{11}, ..., \gamma_{1b}, ..., \gamma_{a1}, ..., \gamma_{ab}). \tag{10}$$

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 (10) with μ, α_i, β_j, and γ_{ij} replaced by $\hat{\mu}$, $\hat{\alpha}_i$, $\hat{\beta}_j$, and $\hat{\gamma}_{ij}$, respectively, where $\hat{\mu} = \bar{X}_{\cdots}$, $\hat{\alpha}_i = \bar{X}_{i\cdot\cdot} - \bar{X}_{\cdots}$, $\hat{\beta}_j = \bar{X}_{\cdot j\cdot} - \bar{X}_{\cdots}$, $\hat{\gamma}_{ij} = \bar{X}_{ij\cdot} - \bar{X}_{i\cdot\cdot} - \bar{X}_{\cdot j\cdot} + \bar{X}_{\cdots}$, and a dot is used to denote averaging over the indicated subscript, e.g.,

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

with a fixed j.