Linear Models

One of the most useful statistical models is

\[ X_i = \beta^\top Z_i + \varepsilon_i, \quad i = 1, \ldots, n, \]

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

Data: \((X_1, Z_1), \ldots, (X_n, Z_n)\).

\( Z_i \)'s are nonrandom or given values of a random \( p \)-vector, in which case our analysis is conditioned on \( Z_1, \ldots, Z_n \).

\( X = (X_1, \ldots, X_n), \varepsilon = (\varepsilon_1, \ldots, \varepsilon_n) \)  
\( Z = \) the \( n \times p \) matrix whose \( i \)th row is the vector \( Z_i, i = 1, \ldots, n \)

A matrix form of the model is

\[ X = Z\beta + \varepsilon. \]
Definition 3.4
Suppose that the range of $\beta$ in model (1) is $B \subset \mathbb{R}^p$. A least squares estimator (LSE) of $\beta$ is defined to be any $\hat{\beta} \in B$ such that
\[
\| X - Z\hat{\beta} \|^2 = \min_{b \in B} \| X - Zb \|^2.
\]
For any $l \in \mathbb{R}^p$, $l^\tau \hat{\beta}$ is called an LSE of $l^\tau \beta$.
Throughout this book, we consider $B = \mathbb{R}^p$ unless otherwise stated. Differentiating $\| X - Zb \|^2$ w.r.t. $b$, we obtain that any solution of
\[
Z^\tau Zb = Z^\tau X
\]
is an LSE of $\beta$.

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
\[
\hat{\beta} = (Z^\tau Z)^{-1} Z^\tau X.
\]
Non full rank $Z$

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

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

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 $\beta$, we need some assumptions on the distribution of $X$ or $\varepsilon$ (conditional on $Z$ if $Z$ is random).

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

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

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 \( \varepsilon \) has the \( N_n(0, \sigma^2 D) \) distribution with unknown \( \sigma^2 \) but a known positive definite matrix \( D \).

Let \( D^{-1/2} \) be the inverse of the square root matrix of \( D \).

Then model (1) with assumption A1 holds if we replace \( X, Z, \) and \( \varepsilon \) 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 \( \theta = (\beta, \sigma^2) \in \mathbb{R}^p \times (0, \infty) \).

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 \( \beta_1 = \beta_2 \).
Remarks

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

$$Z = Z_* Q \quad (2)$$

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 $\mathbb{R}^p$ with dimension $r$.
- In many applications, we are interested in estimating some linear functions of $\beta$, i.e., $\vartheta = l^\tau \beta$ for some $l \in \mathbb{R}^p$.
- From the previous discussion, however, estimation of $l^\tau \beta$ is meaningless unless $l = Q^\tau c$ for some $c \in \mathbb{R}^r$ 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 (1) with assumption A3.

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

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

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

**Proof**

(i) Note that $a \in \mathbb{R}(A)$ 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 $l \in \mathbb{R}(Z)$. 
Proof (continued)

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

\[
l = (Z_* Q)^\tau \zeta = Q^\tau c, \quad c = Z_*^\tau \zeta.
\]

(ii) If \( l \in \mathcal{R}(Z) = \mathcal{R}(Z^\tau Z) \), then \( l = Z^\tau Z \zeta \) for some \( \zeta \) and by \( \hat{\beta} = (Z^\tau Z)^{-1} Z^\tau X \),

\[
E(l^\tau \hat{\beta}) = E[l^\tau (Z^\tau Z)^{-1} Z^\tau X] = \zeta^\tau Z^\tau Z(Z^\tau Z)^{-1} Z^\tau Z \beta = \zeta^\tau Z^\tau Z \beta = l^\tau \beta.
\]

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

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

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

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

Differentiating w.r.t. \( \beta \) and applying Theorem 2.1 lead to

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

which implies \( l \in \mathcal{R}(Z) \).
Example 3.12 (Simple linear regression)

Let $\beta = (\beta_0, \beta_1) \in \mathbb{R}^2$ and $Z_i = (1, t_i)$, $t_i \in \mathbb{R}$, $i = 1, \ldots, n$. Then model (1) is called a \textit{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 $l \in \mathbb{R}^2$ 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 \textit{polynomial regression} of order $p$ in which

$\beta = (\beta_0, \beta_1, \ldots, \beta_{p-1})$

and

$Z_i = (1, t_i, \ldots, t_i^{p-1})$. 
Example 3.13 (One-way ANOVA)

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

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

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

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 $j$th diagonal column.

Consequently, $Z^\tau Z$ is an $m \times m$ diagonal matrix whose $j$th diagonal element is $n_j$.

Thus, $Z^\tau Z$ is invertible and the unique LSE of $\beta$ is the $m$-vector whose $j$th component is

$$\frac{1}{n_j} \sum_{i=k_{j-1}+1}^{k_j} X_i, \quad j = 1, \ldots, m.$$
Example 3.13 (continued)

Sometimes it is more convenient to use the following notation:

\[ X_{ij} = X_{k_{i-1}+j}, \quad \epsilon_{ij} = \epsilon_{k_{i-1}+j}, \quad j = 1, \ldots, n_i, \quad i = 1, \ldots, m, \]

and

\[ \mu_i = \mu + \alpha_i, \quad i = 1, \ldots, m. \]

Then our model becomes

\[ X_{ij} = \mu + \alpha_i + \epsilon_{ij}, \quad j = 1, \ldots, n_i, \quad i = 1, \ldots, m, \tag{3} \]

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

Under model (3), \( \beta = (\mu, \alpha_1, \ldots, \alpha_m) \in \mathbb{R}^{m+1} \).

The matrix \( Z \) under model (3) is not of full rank.

An LSE of \( \beta \) under model (3) is

\[ \hat{\beta} = (\bar{X}, \bar{X}_1 - \bar{X}, \ldots, \bar{X}_m - \bar{X}), \]

where \( \bar{X} \) is still the sample mean of \( X_{ij} \)'s and \( \bar{X}_i \) is the sample mean of the \( i \)th group \( \{X_{ij}, j = 1, \ldots, n_i\} \).
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

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

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 (1), then the parameter vector $\beta$ is

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

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 $ab < p$. 
Two-way balanced ANOVA

It can also be shown that an LSE of $\beta$ is given by the right-hand side of (5) with $\mu$, $\alpha_i$, $\beta_j$, and $\gamma_{ij}$ replaced by $\hat{\mu}$, $\hat{\alpha}_i$, $\hat{\beta}_j$, and $\hat{\gamma}_{ij}$, respectively, where

$$\hat{\mu} = \bar{X}...,$$
$$\hat{\alpha}_i = \bar{X}_{i.} - \bar{X}...,$$
$$\hat{\beta}_j = \bar{X}_{.j} - \bar{X}...,$$
$$\hat{\gamma}_{ij} = \bar{X}_{ij} - \bar{X}_{i.} - \bar{X}_{.j} + \bar{X}...,$$

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

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

with a fixed $j$. 