Chapter 3

General Linear Models

3.1 The Linear Model

Let $\mathbf{Y}$ denote an $n \times 1$ vector valued random variable and $\mathbf{X}$ denote an $n \times p$ matrix of known constants.

- The general linear model specifies the mean vector of $\mathbf{Y}$ as a linear combination of the column vectors of $\mathbf{X}$. The coefficients in this linear combination are unknown quantities (parameters) denoted $\beta_1, \ldots, \beta_p$.
- Thus, if $\boldsymbol{\beta}$ is the $p \times 1$ vector with elements $\beta_1, \ldots, \beta_p$, we can write

$$E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$$

This system of n equations may be rewritten in two forms:

- In terms of the columns of $\mathbf{X}$ as

  $$E(\mathbf{Y}) = \mathbf{X}_1^c\beta_1 + \cdots + \mathbf{X}_p^c\beta_p$$

  where $\mathbf{X}_j^c$ is the jth column of $\mathbf{X}$

- In terms of the rows of $\mathbf{X}$ as

  $$\begin{aligned} E(Y_i) &= x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots + x_{ip}\beta_p \\ &= \mathbf{x}_i^T\boldsymbol{\beta}; \text{ for } \ i = 1, 2, \ldots, n \end{aligned}$$

  where $\mathbf{x}_i^T$ is the ith row of $\mathbf{X}$

Each of these representations of the linear model is useful in discussing different aspects of the linear model. The first is useful for discussion of estimation and in determining the importance of covariates. The second is useful in model diagnostics.

Let $\mathbf{V}$ be an $n \times n$, positive semi-definite matrix of known constants. The general linear model specifies the covariance matrix of $\mathbf{Y}$, denoted $\mathbf{\Sigma_Y}$, as

$$\text{var}\,(\mathbf{Y}) = \mathbf{\Sigma_Y} = \sigma^2\mathbf{V}$$

where σ^2 is an unknown, positive constant (parameter).

The general linear model, stated under (so-called) wide sense assumptions, thus specifies forms for only the first two moments of the observation vector

$$\begin{aligned} E(\mathbf{Y}) &= \mathbf{X}\boldsymbol{\beta} \\ \boldsymbol{\Sigma}_{\mathbf{Y}} &= \sigma^2\mathbf{V} \end{aligned}$$

Every other aspect of the distribution of $\mathbf{Y}$, is arbitrary and unspecified under these wide-sense assumptions.

There are two special forms for $\mathbf{V}$ that are particularly useful.

- In the first special case, $\mathbf{V} = \mathbf{I}$.
  - This implies that the variance of each component of $\mathbf{Y}$ is the same (i.e. var $(Y_i) = \sigma^2$), which is called **homoscedasticity**.
  - It also implies that different components of $\mathbf{Y}$ are uncorrelated.
- The second special form for $\mathbf{V}$ is $\mathbf{V} = \text{diag}\,(V_1, \ldots, V_n)$.
  - In this case, the components of $\mathbf{Y}$ are still uncorrelated but their variances are no longer necessarily equal which is called **heteroscedasticity**.

We first concentrate on the case $\mathbf{V} = \mathbf{I}$. We then show that results for general $\mathbf{V}$ can be obtained by transforming the problem to obtain another general linear model but with a simpler covariance structure.

After properties of least squares estimation are developed under wide sense assumptions, the additional assumption of multivariate normality of $\mathbf{Y}$ will be imposed.

- With the addition of the normality assumption we have a complete specification of the distribution of $\mathbf{Y}$.
- Under these narrower assumptions, stronger properties of least squares estimation procedures are available. and more complete inferences can be made.

3.2 Least Squares Estimates

Definition: Let $\mathbf{Y}$ be an $n \times 1$ vector valued random variable with

$$E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$$

where $\mathbf{X}$ is an $n \times p$ known matrix and $\boldsymbol{\beta}$ is a $p \times 1$ vector of (unknown) parameters.

- When we observe $\mathbf{y}$, the least squares estimate of $\boldsymbol{\beta}$ is defined to be the set of $p \times 1$ vectors $\mathbf{b}$, for which $\|\mathbf{y} - \mathbf{X}\mathbf{b}\|^2$ is minimized.
- Note that

$$\|\mathbf{y} - \mathbf{X}\mathbf{b}\|^2 = \sum_{i=1}^{n} (y_i - b_1 x_{i1} - \cdots - b_p x_{ip})^2$$

- Let $\hat{\boldsymbol{\eta}} = \mathbf{X}\mathbf{b}$ where $\mathbf{b}$ varies.
    - ○ From the results on projections in vector spaces we know there is a unique

    $$\hat{\boldsymbol{\eta}} \in \mathbf{Sp}(\mathbf{X}_1^c, \ldots, \mathbf{X}_p^c)$$

    at which $\|\mathbf{y} - \hat{\boldsymbol{\eta}}\|^2$ is minimized.
    - ○ Therefore, the set of least squares estimates (LSE) of $\boldsymbol{\beta}$ correspond to the solutions of the equations $\hat{\boldsymbol{\eta}} = \mathbf{X}\mathbf{b}$
        - ∗ i.e. $\mathbf{b}$ is a LSE of $\boldsymbol{\beta}$ if and only if $\hat{\boldsymbol{\eta}} = \mathbf{X}\mathbf{b}$ where $\hat{\boldsymbol{\eta}}$ is the projection of $\mathbf{y}$ onto $\mathbf{Sp}(\mathbf{X})$.

- Because of the orthogonality relationships for projections we have the following properties of any LSE of $\boldsymbol{\beta}$.
  - ○ For any $\mathbf{b}$ such that $\hat{\boldsymbol{\eta}} = \mathbf{X}\mathbf{b}$ we have
    $$(\mathbf{y} - \mathbf{X}\mathbf{b}) \perp \mathbf{Sp}(\mathbf{X}_1^c, \ldots, \mathbf{X}_p^c)$$
    thus
    $$\mathbf{X}^T(\mathbf{y} - \mathbf{X}\mathbf{b}) = \mathbf{0}$$
  - ○ Rewriting the last system of equations, we have
    $$\mathbf{X}^T\mathbf{X}\mathbf{b} = \mathbf{X}^T\mathbf{y}$$
    for any LSE, $\mathbf{b}$, of $\boldsymbol{\beta}$.
- Thus any LSE, $\mathbf{b}$, of $\boldsymbol{\beta}$ satisfies the system of equations
  $$\mathbf{X}^T\hat{\mathbf{y}} = \mathbf{X}^T\mathbf{y} \ \text{ where } \hat{\mathbf{y}} = \text{fit}$$
- These equations are called the "normal equations" or the "least squares equations".
  - ○ Provided this system of equations is consistent (i.e. $\mathbf{X}^T\mathbf{y}$ must be in the subspace spanned by the columns of $\mathbf{X}^T\mathbf{X}$), then we know that the general form for solutions for this system is (from results on g-inverses):
    $$\mathbf{b} = (\mathbf{X}^T\mathbf{X})^-\mathbf{X}^T\mathbf{y} + [\mathbf{I} - (\mathbf{X}^T\mathbf{X})^-(\mathbf{X}^T\mathbf{X})]\mathbf{z} \ \text{ for any } \ \mathbf{z} \in \mathbf{R}^p$$
  - ○ To show that the normal equations are always consistent, we note that the subspace spanned by the columns of $\mathbf{X}^T\mathbf{X}$ is the same as the subspace spanned by the columns of $\mathbf{X}^T$.

Consider first the case when $\mathbf{X}$ is of full rank.

- Then

$$\begin{aligned} \mathbf{b} &= (\mathbf{X}^T\mathbf{X})^-\mathbf{X}^T\mathbf{y} + [\mathbf{I} - (\mathbf{X}^T\mathbf{X})^-(\mathbf{X}^T\mathbf{X})]\mathbf{Z} \\ &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} + [\mathbf{I} - (\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{X})]\mathbf{Z} \\ &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \end{aligned}$$

  so that there is a unique solution to the normal equations.

- Since the unique LSE is a linear function of the vector $\mathbf{Y}$, we can calculate the first two moments of $\mathbf{b}$:

$$\begin{aligned} E(\mathbf{b}) &= E[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}] \\ &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T E(\mathbf{Y}) \\ &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} \\ &= \boldsymbol{\beta} \end{aligned}$$

$$\begin{aligned} \text{var}\,(\mathbf{b}) &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\Sigma}_{\mathbf{Y}}\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \end{aligned}$$

- Therefore, $\mathbf{b}$ is unbiased and in the uncorrelated, homoscedastic case, where $\mathbf{V} = \mathbf{I}\sigma^2$, var $(\mathbf{b})$ reduces to

$$\text{var}\,(\mathbf{b}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$$

In the general case where $\mathbf{X}$ is not necessarily of full rank, the non-uniqueness of $\mathbf{b}$ poses problems.

- We clearly are not able to estimate everything that we wish. However, there are some quantities that we can estimate.
- Recall that $\mathbf{Xb}$ $(= \hat{\boldsymbol{\eta}})$ is unique for any LSE $\mathbf{b}$, regardless of the rank of $\mathbf{X}$ (i.e. projections are unique).
- Therefore, we can at least get unique estimates of quantities that can be expressed as linear combinations of the components of $\mathbf{X}\boldsymbol{\beta}$.
- This leads us to the concept of linearly estimable functions due to R.C.Bose.

Definition: Let $\boldsymbol{\ell}$ be a $p \times 1$ vector. A linear parametric function $\boldsymbol{\ell}^T\boldsymbol{\beta}$ is said to be (linearly) estimable if there exists a linear unbiased estimator for $\boldsymbol{\ell}^T\boldsymbol{\beta}$. i.e. $\boldsymbol{\ell}^T\boldsymbol{\beta}$ is linearly estimable if there exists an $n \times 1$ vector $\mathbf{c}$, such that

$$E[\mathbf{c}^T\mathbf{Y}] = \boldsymbol{\ell}^T\boldsymbol{\beta} \text{ for all } \boldsymbol{\beta}$$

More generally, a set of q linearly independent parametric functions $\mathbf{L}^T\boldsymbol{\beta}$ (where $\mathbf{L}$ is a $p \times q$ matrix with rank $\mathbf{L} = q$ is said to be linearly estimable if a $n \times q$ matrix, $\mathbf{C}$, exists such that

$$E[\mathbf{C}^T\mathbf{Y}] = \mathbf{L}^T\boldsymbol{\beta} \text{ for all } \boldsymbol{\beta}$$

The following result provides several alternative characterizations of linear estimability.

Theorem 3.1: Let $\mathbf{L}$ be a $p \times q$ matrix of rank $q \leq r$ where r is the rank of $\mathbf{X}$. The following three statements are equivalent:

(1) $\mathbf{L}^T\boldsymbol{\beta}$ is linearly estimable

(2) There exists an $n \times q$ matrix $\mathbf{C}$ such that $\mathbf{C}^T\mathbf{X} = \mathbf{L}^T$

(3) $\mathbf{L}^T[(\mathbf{I} - (\mathbf{X}^T\mathbf{X})^-(\mathbf{X}^T\mathbf{X})] = \mathbf{0}$.

- From the above theorem, we see that if rank $(\mathbf{X}) = r$, we can have at most r linearly independent estimable functions of $\boldsymbol{\beta}$.

Corollary 3.2: All linear functions of the form $\boldsymbol{\ell}^T\boldsymbol{\beta}$ are estimable if and only if rank $(\mathbf{X}) = p$.

From the above corollary, if $\mathbf{X}$ has full rank, then we see that $\boldsymbol{\beta}$ itself is estimable by setting $\mathbf{L}^T = \mathbf{I}$.

Lemma 3.3: Let $\mathbf{L}^T\boldsymbol{\beta}$ be q linearly independent estimable functions.

- Then there exists a unique linear unbiased estimate of $\mathbf{L}^T\boldsymbol{\beta}$, say $\mathbf{A}^T\mathbf{Y}$ with the columns of $\mathbf{A} \in \mathbf{Sp}(\mathbf{X})$.
- If $\mathbf{A}^T\mathbf{Y}$ is any linear unbiased estimator of $\mathbf{L}^T\boldsymbol{\beta}$, then the columns of $\mathbf{A}$ are the projections of the columns of $\mathbf{A}$ onto $S(\mathbf{X})$.

Note that the estimator of $\mathbf{L}^T\boldsymbol{\beta}$, $\mathbf{L}^T\mathbf{b} = \mathbf{L}^T(\mathbf{X}^T\mathbf{X})^{-}\mathbf{X}^T\mathbf{Y}$ is of the form $\mathbf{A}^T\mathbf{Y}$ with the columns of $\mathbf{A}$ in $\mathbf{Sp}(\mathbf{X})$.

Theorem 3.4: (Gauss Markov Theorem) In the linear model with

$$E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta} \text{ and } \operatorname{var}(\mathbf{Y}) = \sigma^2\mathbf{I}$$

where $\mathbf{X}$ is an $n \times p$ matrix with rank $(X) = r \leq p$

- The unique best (minimum variance) unbiased linear estimator (called the BLUE or Gauss-Markov Estimator) of the estimable function $\mathbf{L}^T\boldsymbol{\beta}$ ($\mathbf{L}$ is $p \times q$ with rank $\mathbf{L} = q \leq r$) is given by

$$\mathbf{L}^T\mathbf{b} \text{ where } \mathbf{b} \text{ is any LSE of } \boldsymbol{\beta}$$

- and

$$\operatorname{var}(\mathbf{L}^T\widehat{\boldsymbol{\beta}}) = \sigma^2\mathbf{L}^T(\mathbf{X}^T\mathbf{X})^{-}\mathbf{L}$$

  where any g-inverse of $\mathbf{X}^T\mathbf{X}$ may be used to compute $\operatorname{var}(\mathbf{L}^T\widehat{\boldsymbol{\beta}})$

In order to use the LSE of $\mathbf{L}^T\boldsymbol{\beta}$ for inferences we need an estimator of its covariance matrix. From the Gauss Markov theorem, we know its form to be

$$\text{var}\,(\mathbf{L}^T\widehat{\boldsymbol{\beta}}) = \sigma^2\mathbf{L}^T(\mathbf{X}^T\mathbf{X})^-\mathbf{L}$$

Therefore, an estimate of σ^2 is all that is required to obtain an estimate of var $(\mathbf{L}^T\widehat{\boldsymbol{\beta}})$. Since

$$\begin{aligned} \sigma^2 &= \text{var}\,(Y_i) \\ &= \frac{1}{n}\sum_{i=1}^{n}\text{var}\,(Y_i) \\ &= \frac{1}{n}\sum_{i=1}^{n}E[Y_i - \mathbf{x}_i^T\boldsymbol{\beta}]^2 \\ &= \frac{1}{n}E[\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2] \end{aligned}$$

a reasonable estimate of σ^2 might be

$$\widehat{\sigma^2} = \frac{1}{n}\|\mathbf{Y} - \mathbf{X}\mathbf{b}\|^2$$

where $\mathbf{b}$ is any LSE of $\boldsymbol{\beta}$.

We thus consider the expected value of

$$\|\mathbf{Y} - \mathbf{X}\mathbf{b}\|^2$$

to guide us to an estimator of σ^2.

- Let $P_{\mathbf{X}} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-}\mathbf{X}^T$ then

$$\begin{aligned} E[\|\mathbf{Y} - \mathbf{X}\mathbf{b}\|^2] &= E[\|(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\mathbf{Y}\|^2] \\ &= E[\mathbf{Y}^T(\mathbf{I} - \mathbf{P}_{\mathbf{X}})(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\mathbf{Y}] \\ &= E[\mathbf{Y}^T(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\mathbf{Y}] \\ &= \text{tr}\,[(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\text{var}\,(\mathbf{Y})] + E[\mathbf{Y}^T](\mathbf{I} - \mathbf{P}_{\mathbf{X}})E[\mathbf{Y}] \end{aligned}$$

- But

$$\text{tr}\,[(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\text{var}\,(\mathbf{Y})] = \sigma^2[n - \text{rank}\,(\mathbf{P}_{\mathbf{X}})] = \sigma^2[n - \text{rank}\,(\mathbf{X})]$$

and

$$E[\mathbf{Y}]^T(\mathbf{I} - \mathbf{P}_{\mathbf{X}})E[\mathbf{Y}] = \boldsymbol{\beta}^T\mathbf{X}^T(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\mathbf{X}\boldsymbol{\beta} = \mathbf{0}$$

since

$$(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\mathbf{X} = \mathbf{X} - \mathbf{P}_{\mathbf{X}}\mathbf{X} = \mathbf{X} - \mathbf{X} = \mathbf{0}$$

- Therefore, if $r = \text{rank}\,(\mathbf{X})$, we have that

$$\widehat{\sigma^2} = \frac{1}{n-r}[\|\mathbf{Y} - \mathbf{X}\mathbf{b}\|^2]$$

is an unbiased estimate of σ^2.