Introduction to the General Linear Model

Consider the General Linear Model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where ...

$\boldsymbol{y}$ is an $n \times 1$ random vector of responses that can be observed.

The values in $\boldsymbol{y}$ are values of the response variable.

$\boldsymbol{X}$ is an $n \times p$ matrix of known constants.

Each column of $\boldsymbol{X}$ contains the values for an explanatory variable that is also known as a predictor variable or a regressor variable in the context of multiple regression.

The matrix $\boldsymbol{X}$ is sometimes referred to as the design matrix.

$\boldsymbol{\beta}$ is an unknown parameter vector in $\mathbb{R}^p$.

We are often interested in estimating a LC of the elements in $\boldsymbol{\beta}$ $(\boldsymbol{c}'\boldsymbol{\beta})$ or multiple LCs $(\boldsymbol{C}\boldsymbol{\beta})$ for some known $\boldsymbol{c}$ or $\boldsymbol{C}$.

ε is an $n \times 1$ random vector that cannot be observed.

The values in ε are called errors. Initially, we assume only $E(\varepsilon) = \mathbf{0}$.

Note that

$$\begin{aligned} E(\boldsymbol{\varepsilon}) = \mathbf{0} \Rightarrow E(\boldsymbol{y}) &= E(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) \\ &= \boldsymbol{X}\boldsymbol{\beta} + E(\boldsymbol{\varepsilon}) \\ &= \boldsymbol{X}\boldsymbol{\beta} \in \mathcal{C}(\boldsymbol{X}). \end{aligned}$$

Thus, this general linear model simply says that $\boldsymbol{y}$ is a random vector with mean $E(\boldsymbol{y})$ in the column space of $\boldsymbol{X}$.

Example:

Suppose 5 hogs are fed diet 1 for three weeks. Let y_i be the weight gain of the i^{th} hog for $i = 1, \ldots, 5$.

Suppose that we assume $E(y_i) = \mu \ \forall \ i = 1, \ldots, 5$. Then

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \mu + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \end{bmatrix}$$

where $E(\varepsilon_i) = 0 \ \forall \ i = 1, \ldots, 5$.

Example:

Suppose 10 hogs are randomly divided into two groups of 5 hogs each. Group 1 is fed diet 1 for three weeks, group 2 is fed diet 2 for three weeks.

Let y_{ij} denote the weight gain of the j^{th} hog in the i^{th} diet group $(i = 1, 2;\ j = 1, \ldots, 5)$.

If we assume that $E(y_{ij}) = \mu_i$ for $i = 1, 2;\ j = 1, \ldots, 5$, then

$$
\begin{bmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{14} \\ y_{15} \\ y_{21} \\ y_{22} \\ y_{23} \\ y_{24} \\ y_{25} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} + \begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{13} \\ \varepsilon_{14} \\ \varepsilon_{15} \\ \varepsilon_{21} \\ \varepsilon_{22} \\ \varepsilon_{23} \\ \varepsilon_{24} \\ \varepsilon_{25} \end{bmatrix}
$$

where $E(\varepsilon_{ij}) = 0 \; \forall \; i = 1, 2; \; j = 1, \ldots, 5$.

Alternatively, we could assume $E(y_{ij}) = \mu + \tau_i \; \forall \; i = 1, 2; \; j = 1, \ldots, 5$.

Then the design matrix and parameter vector become $\ldots$

$$\begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \tau_1 \\ \tau_2 \end{bmatrix} = \begin{bmatrix} \mu + \tau_1 \\ \mu + \tau_1 \\ \mu + \tau_1 \\ \mu + \tau_1 \\ \mu + \tau_1 \\ \mu + \tau_2 \\ \mu + \tau_2 \\ \mu + \tau_2 \\ \mu + \tau_2 \\ \mu + \tau_2 \end{bmatrix}.$$

These models are equivalent because both design matrices have the same column space:

$$\left\{ \begin{bmatrix} a \\ a \\ a \\ a \\ a \\ b \\ b \\ b \\ b \\ b \end{bmatrix} : a, b \in \mathbb{R} \right\}.$$

Example:

10 steer carcasses were assigned to be measured for pH at one of five times after slaughter. Data are as follows (Schwenke & Milliken(1991), Biometrics, 47, 563-573.)

Steer	Hours after Slaughter	pH
1	1	7.02
2	1	6.93
3	2	6.42
4	2	6.51
5	3	6.07
6	3	5.99
7	4	5.59
8	4	5.80
9	5	5.51
10	5	5.36

$\forall\ i = 1, \ldots, 10$, let

$$x_i = \text{ measurement time (hours after slaughter) for steer } i$$

$$y_i = \text{ pH for steer } i.$$

Suppose $y_i = \beta_0 + \beta_1 \log(x_i) + \varepsilon_i$ where $E(\varepsilon_i) = 0\ \forall\ i = 1, \ldots, 10$.

Determine $\boldsymbol{y}$, $\boldsymbol{X}$, and $\boldsymbol{\beta}$.

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \\ y_8 \\ y_9 \\ y_{10} \end{bmatrix} = \begin{bmatrix} 1 & \log(x_1) \\ 1 & \log(x_2) \\ 1 & \log(x_3) \\ 1 & \log(x_4) \\ 1 & \log(x_5) \\ 1 & \log(x_6) \\ 1 & \log(x_7) \\ 1 & \log(x_8) \\ 1 & \log(x_9) \\ 1 & \log(x_{10}) \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \\ \varepsilon_7 \\ \varepsilon_8 \\ \varepsilon_9 \\ \varepsilon_{10} \end{bmatrix}$$

$$\boldsymbol{y} = \begin{bmatrix} 7.02 \\ 6.93 \\ 6.42 \\ 6.51 \\ 6.07 \\ 5.99 \\ 5.59 \\ 5.80 \\ 5.51 \\ 5.36 \end{bmatrix}, \ \boldsymbol{X} = \begin{bmatrix} 1 & \log(1) \\ 1 & \log(1) \\ 1 & \log(2) \\ 1 & \log(2) \\ 1 & \log(3) \\ 1 & \log(3) \\ 1 & \log(4) \\ 1 & \log(4) \\ 1 & \log(5) \\ 1 & \log(5) \end{bmatrix}, \ \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

Chapter 1 of our text contains many more examples. Please read the entire chapter.