Spring 2015 Statistics 151 (Linear Models) : Lecture Three

Aditya Guntuboyina

27 January 2015

1 Estimation of β in the Linear Model

The linear model is

$$Y = X\beta + e \qquad \text{with } \mathbb{E}e = 0 \text{ and } Cov(e) = \sigma^2 I_n$$

where Y is $n \times 1$ vector containing all the values of the response, X is $n \times (p+1)$ matrix containing all the values of the explanatory variables (the first column of X is all ones) and $\beta = (\beta_0, \beta_1, \dots, \beta_p)^T$ (β_0 is the intercept).

As we have seen last time, β is estimated by minimizing $S(\beta) = ||Y - X\beta||^2$. Taking derivatives with respect to β and equating to zero, one obtains the normal equations

$$X^T X\beta = X^T Y.$$

The normal equations always have solutions. But there is no uniqueness unless $X^T X$ is invertible. If $X^T X$ is invertible (i.e., if X has full column rank), then the normal equations have a unique solution, $\hat{\beta}_{ls}$, which we call THE least squares estimate of β:

$$\hat{\beta}_{ls} := (X^T X)^{-1} X^T Y.$$

If $X^T X$ is not invertible, then any solution of the normal equations is called a least squares estimate of β.

When $X^T X$ is not invertible, I argued last class that some linear combinations of β cannot be estimated. This leads to the following definition: A linear combination $\lambda^T \beta = \lambda_0 \beta_0 + \lambda_1 \beta_1 + \cdots + \lambda_p \beta_p$ is said to be estimable if the vector λ lies in the column space of X^T.

Because the column spaces of X^T and $X^T X$ are always equal, we can equivalently define estimability by requiring that $\lambda \in \mathcal{C}(X^T X)$.

Result 1.1. *If $\lambda^T \beta$ is estimable, then $\lambda^T \hat{\beta}_{ls}$ is the same for every solution $\hat{\beta}_{ls}$ of the normal equations. In other words, the least squares estimate of $\lambda^T \beta$ is unique.*

Proof. Since $\lambda^T \beta$ is estimable, the vector λ lies in the column space of $X^T X$ and hence $\lambda = X^T X u$ for some vector u. Therefore,

$$\lambda^T \hat{\beta}_{ls} = u^T X^T X \hat{\beta}_{ls} = u^T X^T Y$$

where the last equality follows from the fact that $\hat{\beta}_{ls}$ satisfies the normal equations. Since u only depends on λ, this proves that $\lambda^T \hat{\beta}_{ls}$ does not depend on the particular choice of the solution $\hat{\beta}_{ls}$ of the normal equations. $\square$

Thus when $\lambda^T\beta$ is estimable, it is estimated by $\lambda^T\hat{\beta}_{ls}$ for any least squares estimate of β (it does not matter which least squares estimate is used). When $\lambda^T\beta$ is not estimable, it of course does not make sense to try to estimate it.

2 Special Case: Simple Linear Regression

Suppose there is only one explanatory variable x. The matrix X would then of size $n \times 2$ where the first column of X consists of all ones and the second column of X equals the values of the explanatory variable $x_1, \dots, x_n$. Therefore

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \qquad X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \qquad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}.$$

Check that

$$X^TX = \begin{pmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{pmatrix} = \begin{pmatrix} n & n\bar{x} \\ n\bar{x} & \sum_{i=1}^n x_i^2 \end{pmatrix}$$

where $\bar{x} = \sum_i x_i/n$. Also let $\bar{y} = \sum_i y_i/n$. Because

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \frac{1}{ad-bc}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix},$$

we get

$$(X^TX)^{-1} = \frac{1}{n\sum_{i=1}^n (x_i-\bar{x})^2}\begin{pmatrix} \sum_{i=1}^n x_i^2 & -n\bar{x} \\ -n\bar{x} & n \end{pmatrix}.$$

Also

$$X^TY = \begin{pmatrix} n\bar{y} \\ \sum_{i=1}^n x_iy_i \end{pmatrix}$$

Therefore

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = \frac{1}{n\sum_{i=1}^n (x_i-\bar{x})^2}\begin{pmatrix} \sum_i x_i^2 & -n\bar{x} \\ -n\bar{x} & n \end{pmatrix}\begin{pmatrix} n\bar{y} \\ \sum_{i=1}^n x_iy_i \end{pmatrix}.$$

Simplify to obtain

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = \frac{1}{\sum_{i=1}^n (x_i-\bar{x})^2}\begin{pmatrix} \bar{y}\sum_{i=1}^n x_i^2 - \bar{x}\sum_{i=1}^n x_iy_i \\ \sum_{i=1}^n x_iy_i - n\bar{x}\bar{y} \end{pmatrix}.$$

Thus

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n x_iy_i - n\bar{x}\bar{y}}{\sum_{i=1}^n (x_i-\bar{x})^2} = \frac{\sum_{i=1}^n (x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^n (x_i-\bar{x})^2}$$

and

$$\hat{\beta}_0 = \frac{\bar{y}\sum_{i=1}^n x_i^2 - \bar{x}\sum_{i=1}^n x_iy_i}{\sum_{i=1}^n (x_i-\bar{x})^2} = \bar{y} - \hat{\beta}_1\bar{x}$$

If we get a new subject whose explanatory variable value is x, our prediction for its response is

$$y = \hat{\beta}_0 + \hat{\beta}_1 x. \tag{1}$$

If the predictions given by the above are plotted on a graph (with x plotted on the x-axis), then one gets a line called the **Regression Line**.

The Regression Line has a much nicer expression than (1). To see this, note that

$$y = \hat{\beta}_0 + \hat{\beta}_1 x = \bar{y} - \bar{x}\hat{\beta}_1 + \hat{\beta}_1 x = \bar{y} + \hat{\beta}_1(x - \bar{x})$$

This can be written as

$$y - \bar{y} = \hat{\beta}_1(x - \bar{x}) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}(x - \bar{x}) \tag{2}$$

Using the notation

$$r := \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}, \quad s_x := \sqrt{\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2}, \quad s_y := \sqrt{\frac{1}{n}\sum_{i=1}^n (y_i - \bar{y})^2},$$

we can rewrite the prediction equation (2) as

$$\frac{y - \bar{y}}{s_y} = r\frac{x - \bar{x}}{s_x}. \tag{3}$$

r is the correlation between x and y which is always between -1 and 1.

As an implication, note that if $(x - \bar{x})/s_x = 1$ i.e., if the explanatory variable value of the subject is one standard deviation above the sample mean, then its response variable is predicted to be only r standard deviations above its mean. Francis Galton termed this **regression to mediocrity** which is where the name regression comes from.

3 Basic Mean and Covariance Formulae for Random Vectors

We next want to explore properties of $\hat{\beta} = (X^TX)^{-1}X^TY$ as an estimator of β in the linear model. For this we need a few facts about means and covariances.

Let $Z = (Z_1, \dots, Z_k)^T$ be a random vector. Its expectation $\mathbb{E}Z$ is defined as a vector whose ith entry is the expectation of Z_i i.e., $\mathbb{E}Z = (\mathbb{E}Z_1, \mathbb{E}Z_2, \dots, \mathbb{E}Z_k)^T$.

The covariance matrix of Z, denoted by $Cov(Z)$, is a $k \times k$ matrix whose (i, j)th entry is the covariance between Z_i and Z_j.

If $W = (W_1, \dots, W_m)^T$ is another random vector, the covariance matrix between Z and W, denoted by $Cov(Z, W)$, is a $k \times m$ matrix whose (i, j)th entry is the covariance between Z_i and W_j. Note then that, $Cov(Z, Z) = Cov(Z)$.

The following formulae are very important:

1. $\mathbb{E}(AZ + c) = A\mathbb{E}(Z) + c$ for any constant matrix A and any constant vector c.
2. $Cov(AZ + c) = ACov(Z)A^T$ for any constant matrix A and any constant vector c.
3. $Cov(AZ + c, BW + d) = ACov(Z, W)B^T$ for any pair of constant matrices A and B and any pair of constant vectors c and d.

The linear model is

$$Y = X\beta + e \qquad \text{with } \mathbb{E}e = 0 \text{ and } Cov(e) = \sigma^2 I_n.$$

Because of the above formulae (remember that X and β are fixed),

$$\mathbb{E}Y = X\beta \quad \text{and} \quad Cov(Y) = \sigma^2 I_n.$$