Fall 2013 Statistics 151 (Linear Models) : Lecture Five

Aditya Guntuboyina

12 September 2013

1 Least Squares Estimate 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 $\beta = (\beta_0, \beta_1, \ldots, \beta_p)^T$ (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

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

If $X^T X$ is invertible (this is equivalent to the rank of X being equal to $p+1$), then the solution to the normal equations is unique and is given by

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

This is the least squares estimate of β.

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 $x_1, \ldots, x_n$. Therefore

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

Check that

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

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

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

Therefore

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

Simplify to obtain

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

Thus

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

and

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

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 $\hat{\beta} = (X^T X)^{-1} X^T Y$ as an estimator of β in the linear model. For this we need a few facts about means and covariances.

Let $Z = (Z_1, \ldots, Z_k)^T$ 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., $\mathbb{E}Z = (\mathbb{E}Z_1, \mathbb{E}Z_2, \ldots, \mathbb{E}Z_k)^T$.

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 $W = (W_1, \ldots, W_m)^T$ 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. $Cov(AZ + c, BW + d) = ACov(Z, W)B$ 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),

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

4 Properties of the Least Squares Estimator

Assume that $X^T X$ is invertible (equivalently, that X has rank $p + 1$) and consider the least squares estimator

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

What properties does $\hat{\beta}$ have as an estimator of β?

4.1 Linearity

An estimator of β is said to be linear if it can be written as AY for some matrix A. Clearly $\hat{\beta} = (X^T X)^{-1} X^T Y$ is of this form and hence it is a linear estimator of β.

4.2 Unbiasedness

An estimator for a parameter is said to be unbiased if its expectation equals the parameter (for all values of the parameter).

The expectation of the least squares estimator is (using the formula for expectation: $\mathbb{E}AZ = A\mathbb{E}Z$)

$$\mathbb{E}\hat{\beta} = \mathbb{E}((X^T X)^{-1} X^T Y) = (X^T X)^{-1} X^T \mathbb{E}Y = (X^T X)^{-1} X^T X\beta = \beta$$

In particular, this means that $\mathbb{E}\hat{\beta}_i = \beta_i$ for each i which implies that each $\hat{\beta}_i$ is an unbiased estimator of β_i. More generally, for every vector λ, the quantity $\lambda^T\hat{\beta}$ is an unbiased estimator of $\lambda^T\beta$.

4.3 Covariance Matrix

The Covariance matrix of the estimator $\hat{\beta}$ can be easily calculated using the formula: $Cov(AZ) = ACov(Z)A^T$:

$$Cov(\hat{\beta}) = Cov((X^TX)^{-1}X^TY) = (X^TX)^{-1}X^TCov(Y)X(X^TX)^{-1} = \sigma^2(X^TX)^{-1}$$

In particular, the variance of $\hat{\beta}_i$ equals σ^2 multiplied by the ith diagonal element of $(X^TX)^{-1}$. Once we learn how to estimate σ, we can use this to obtain standard errors for $\hat{\beta}_i$.

4.4 Optimality - The Gauss-Markov Theorem

The Gauss-Markov Theorem states that $\hat{\beta}$ is BLUE (Best Linear Unbiased Estimator). This means that $\hat{\beta}$ is the "best" estimator among all **linear and unbiased** estimators of β. Here, "best" is in terms of variance. This implies that $\hat{\beta}_i$ has the **smallest variance** among all linear and unbiased estimators of β_i.