Spring 2015 Statistics 151 (Linear Models) : Lecture Four

Aditya Guntuboyina

29 January 2015

1 Properties of the Least Squares Estimator

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

$$\hat{\beta} = (X^TX)^{-1}X^TY.$$

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

1.1 Linearity

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

1.2 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: $\mathbb{E}AZ = A\mathbb{E}Z$)

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

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 $\lambda^T\beta$.

1.3 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 $\hat{\beta}_i$.

1.4 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.

Next we look at some very commonly used quantities in regression.

2 The Regression Plane

If we get a new subject whose explanatory variable values are $x_1, \ldots, x_p$, then our prediction for its response variable value is

$$y = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots \hat{\beta}_p x_p$$

This equation represents a plane which we call the regression plane.

3 Fitted Values

These are the values predicted by the linear model for the n subjects.

The values of the explanatory variables are $x_{i1}, \ldots, x_{ip}$ for the ith subject. Thus the linear model prediction for the ith subject is

$$\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip}.$$

Because the value of the response variable for the ith subject is y_i, it makes sense to call the above prediction $\hat{y}_i$. Thus

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip} \qquad \text{for } i = 1, \ldots, n.$$

These values $\hat{y}_1, \ldots, \hat{y}_n$ are called fitted values and the vector $\hat{Y} = (\hat{y}_1, \ldots, \hat{y}_n)^T$ is called the vector of fitted values. This vector can be written succinctly as $\hat{Y} = X\hat{\beta}$. Because $\hat{\beta} = (X^T X)^{-1} X^T Y$, we can write

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

The vector of fitted values $\hat{Y}$ is the (orthogonal) projection of Y onto the column space of X.

Let $H = X(X^T X)^{-1} X^T$ so that $\hat{Y} = HY$. Because multiplication by H changes Y into $\hat{Y}$, the matrix H is called the **Hat Matrix**. It is very important in linear regression. It has the following three easily verifiable properties:

1. It is a symmetric $n \times n$ matrix.
2. It is idempotent i.e., $H^2 = H$.
3. $HX = X$.
4. The ranks of H and X are the same.

These can be easily derived from the definition $H = X(X^T X)^{-1} X^T$. Because of these, we get

$$\mathbb{E}\hat{Y} = \mathbb{E}(HY) = H(\mathbb{E}Y) = HX\beta = X\beta = \mathbb{E}Y. \tag{1}$$

Thus $\hat{Y}$ and Y have the same expectation. Also

$$Cov(\hat{Y}) = Cov(HY) = HCov(Y)H^T = H(\sigma^2 I)H = \sigma^2 H.$$

4 Residuals

The difference between y_i and $\hat{y}_i$ is called the residual for the ith subject. $\hat{e}_i := y_i - \hat{y}_i$. The vector $\hat{e} = (\hat{e}_1, \ldots, \hat{e}_n)^T$ is called the vector of residuals. Clearly

$$\hat{e} = Y - \hat{Y} = (I - H)Y.$$

The vector of residuals $\hat{e}$ acts as a proxy for the **unobserved error vector** e.

The most important fact about the residuals in the linear model is that they are orthogonal to the column space of X. This happens because $HX = X$ so that

$$\hat{e}^T X = ((I - H)Y)^T X = Y^T (I - H)X = Y^T (X - HX) = 0.$$

As a result $\hat{e}^T X u = 0$ for every vector u which means that $\hat{e}$ is orthogonal to the column space of X.

The first column of X consists of ones. Because $\hat{e}$ is orthogonal to everything in the column space of X, it must therefore be orthogonal to the vector of ones which means that

$$\sum_{i=1}^{n} \hat{e}_i = 0.$$

$\hat{e}$ is also orthogonal to every column of X:

$$\sum_{i=1}^{n} \hat{e}_i x_{ij} = 0 \qquad \text{for every } j$$

The vector of fitted values belongs to the column space of X because $\hat{Y} = X\hat{\beta}$. Thus, $\hat{e}$ is also orthogonal to $\hat{Y}$.

Because $X^T \hat{e} = 0$, the residuals satisfy $rank(X) = p + 1$ linear equalities. Hence, although there are n of them, they are effectively $n - p - 1$ of them. The number $n - p - 1$ is therefore referred to as the *degrees of freedom* of the residuals $\hat{e}_1, \ldots, \hat{e}_n$.

The expectation of $\hat{e}$ is

$$\mathbb{E}\hat{e} = \mathbb{E}\left((I - H)Y\right) = (I - H)(\mathbb{E}Y) = (I - H)X\beta = (X - HX)\beta = 0.$$

Alternatively $\mathbb{E}\hat{e} = \mathbb{E}(Y - \hat{Y}) = \mathbb{E}Y - \mathbb{E}\hat{Y} = 0$ by (1).

The Covariance matrix of $\hat{e}$ is

$$Cov(\hat{e}) = Cov((I - H)Y) = (I - H)Cov(Y)(I - H) = \sigma^2(I - H). \tag{2}$$

Note that the residuals have different variances (even though $e_1, \ldots, e_n$ are assumed to have the same variance σ^2).