Fall 2013 Statistics 151 (Linear Models) : Lecture Six

Aditya Guntuboyina

17 September 2013

We again consider $Y = X\beta + e$ with $\mathbb{E}e = 0$ and $Cov(e) = \sigma^2 I_n$. β is estimated by solving the normal equations $X^T X\beta = X^T Y$.

1 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.

2 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 $H = X(X^TX)^{-1}X^T$. 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

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

3 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

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

As a result $\hat{e}^TXu = 0$ 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.

4 The Residual Sum of Squares

The sum of squares of the residuals is called RSS:

$$RSS = \sum_{i=1}^{n} \hat{e}_i^2 = \hat{e}^T \hat{e} = Y^T (I - H) Y = e^T (I - H) e.$$

What is the residual sum of squares where there are no explanatory variables in the model (the model in this case only contains the intercept term)? Ans: $\sum_{i=1}^{n} (y_i - \bar{y})^2$ where $\bar{y} = (y_1 + \cdots + y_n)/n$. This quantity is called the TSS (Total Sum of Squares). The vector $(y_1 - \bar{y}, \ldots, y_n - \bar{y})$ has $n - 1$ degrees of freedom (because this is a vector of size n and it satisfies the linear constraint that sum is zero).

What is the residual sum of squares in simple linear regression (when there is exactly one explanatory variable)? Check that in simple linear regression:

$$RSS = \left(1 - r^2\right) \sum_{i=1}^{n} (y_i - \bar{y})^2$$

where r is the sample correlation between $(y_1, \ldots, y_n)$ and $x = (x_1, \ldots, x_n)$. Because $1 - r^2 \leq 1$, the RSS in simple linear regression is smaller than the RSS in the linear model with no explanatory variables.

In general, RSS decreases (or remains the same) as we add more explanatory variables to the model.

5 The Coefficient of Determination

This is more commonly referred to as R-squared. It is defined as

$$R^2 := 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}.$$

What is the point of this definition? One of the goals of regression is to predict the value of the response variable for future subjects. For this purpose, we are given data $y_1, \ldots, y_n$ and x_{ij} for $i = 1, \ldots, n$ and $j = 1, \ldots, p$.

Suppose we are told to predict the response of a future subject **without** using any of the data on the explanatory variables i.e., we are only supposed to use $y_1, \ldots, y_n$. In this case, it is obvious that our prediction for the next subject would be $\bar{y}$. The error of this method of prediction on the ith subject is $y_i - \bar{y}$ and the total error is therefore:

$$TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2.$$

If, on the other hand, we are allowed to use data on the explanatory variables, then the prediction will be given by

$$\hat{\beta}_0 + x_1 \hat{\beta}_1 + \cdots + x_p \hat{\beta}_p.$$

The error of this prediction on the ith subject is the residual $\hat{e}_i$ and the total error is the Residual Sum of Squares:

$$RSS = \sum_{i=1}^{n} \hat{e}_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.$$

Because using the explanatory variables is always better than not using them, RSS is always smaller than or equal to TSS (**this fact is crucially reliant on the fact that there is an intercept in our model**).

If RSS is very small compared to TSS, it means that the explanatory variables are really useful in predicting the response. On the other hand, if RSS is only a little bit smaller than TSS, it means that we are not really gaining much by using the explanatory variables. The quantity R^2 tries to quantify how useful the explanatory variables are in predicting the response. It always lies between 0 and 1

1. **If R^2 is high, it means that RSS is much smaller compared to TSS and hence the explanatory variables are really useful in predicting the response.**
2. **If R^2 is low, it means that RSS is only a little bit smaller than TSS and hence the explanatory variables are not useful in predicting the response.**

It must be noted that R^2 is an *in-sample* measure of prediction accuracy. In other words, the predictions are checked on the subjects already present in the sample (as opposed to checking them on new subjects). In particular, these are the same subjects on whom the model is fitted (or trained), so R^2 can be made to look very good by fitting models with lots of parameters.

Because RSS decreases when more parameters are added to the model, R^2 increases when more parameters are added to the model.