Spring 2015 Statistics 151 (Linear Models) : Lecture Fourteen

Aditya Guntuboyina

10 March 2015

1 Regression after deleting the ith subject

Recall the definition of the ith predicted residual from last class. The ith predicted residual is defined as follows. First throw away the ith subject and fit the linear model. Using that linear model, predict the value of y_i based on the explanatory variable values of the ith subject. The difference between y_i and this predicted value is called the ith predicted residual, denoted by $\hat{e}_{[i]}$.

We will show here that the ith predicted residual can be calculated via a simple formula involving $\hat{e}_i$ (the usual residual) and the leverage h_{ii}. For this, consider fitting the linear model to all the subjects except the ith one. Let $X_{[i]}$ and $Y_{[i]}$ denote the X-matrix and the Y-vector respectively in this regression. In other words, $X_{[i]}$ denotes the X-matrix with the ith row deleted and $Y_{[i]}$ denotes the Y-vector with the ith entry deleted. Also, let x_i^T denote the ith row of the original X matrix.

The estimate of β after deleting the ith row is:

$$\hat{\beta}_{[i]} = \left(X_{[i]}^T X_{[i]}\right)^{-1} X_{[i]}^T Y_{[i]}.$$

The ith predicted residual is defined as

$$\hat{e}_{[i]} = y_i - x_i^T \hat{\beta}_{[i]}.$$

It might seem that to calculate $\hat{e}_{[i]}$ for different i, one would need to perform many regressions deleting each subject separately. Fortunately, one can calculate these in a simpler way using $\hat{e}_i$ and h_{ii}. For this, we need the Sherman-Morrison Formula from matrix algebra.

Theorem 1.1 (Sherman-Morrison Formula)**.** *Suppose A is an $n \times n$ matrix and u and v are $n \times 1$ vectors. Then*

$$(A - uv^T)^{-1} = A^{-1} + \frac{A^{-1}uv^T A^{-1}}{1 - v^T A^{-1} u} \tag{1}$$

provided all the inverses above make sense.

Because $X^T X = X_{[i]}^T X_{[i]} + x_i x_i^T$, we get from (1) that

$$(X_{[i]}^T X_{[i]})^{-1} = (X^T X - x_i x_i^T)^{-1} = (X^T X)^{-1} + \frac{(X^T X)^{-1} x_i x_i^T (X^T X)^{-1}}{1 - x_i^T (X^T X)^{-1} x_i} = (X^T X)^{-1} + \frac{(X^T X)^{-1} x_i x_i^T (X^T X)^{-1}}{1 - h_{ii}}.$$

Also check that

$$X_{[i]}^T Y_{[i]} = X^T Y - y_i x_i.$$

Both of these give

$$\hat{\beta}_{[i]} = \hat{\beta} - \frac{\hat{e}_i}{1 - h_{ii}} (X^T X)^{-1} x_i. \tag{2}$$

As a result

$$\hat{e}_{[i]} = y_i - x_i^T \hat{\beta}_{[i]} = y_i - x_i^T \hat{\beta} + \frac{\hat{e}_i}{1-h_{ii}} x_i^T (X^T X)^{-1} x_i = \hat{e}_i + \frac{h_{ii}}{1-h_{ii}} \hat{e}_i = \frac{\hat{e}_i}{1-h_{ii}}.$$

Therefore, the predicted residual $\hat{e}_{[i]}$ is the usual residual divided by 1 minus the leverage. One can thus see, if the leverage of the ith subject, h_{ii}, is very large, then the residual $\hat{e}_i$ will be small, but the predicted residual might be very large.

Under the assumptions of the linear model (i.e., under $Y = X\beta + e$ with $\mathbb{E}e = 0$ and $Cov(e) = \sigma^2 I$), what are $\mathbb{E}\hat{e}_{[i]}$ and $var(\hat{e}_{[i]})$? It is easy to check that $\mathbb{E}\hat{e}_{[i]} = 0$. For the variance,

$$var(\hat{e}_{[i]}) = var\left(\frac{\hat{e}_i}{1-h_{ii}}\right) = (1-h_{ii})^{-2} var(\hat{e}_i) = \frac{\sigma^2}{1-h_{ii}}.$$

The predicted residuals therefore have different variances (if h_{ii} is large, then $\hat{e}_{[i]}$ has high variance). We can thus standardize them and we get what are called Standardized Predicted Residuals. These are described next.

2 Standardized Predicted Residuals

How do we decide if a subject is an outlier? One idea is to calculate its predicted residual $\hat{e}_{[i]}$ and check it is too high. How do we decide if it is too high? We have just seen that the variance of the predicted residual is given by

$$var(\hat{e}_{[i]}) = \frac{\sigma^2}{1-h_{ii}}.$$

When judging if $\hat{e}_{[i]}$ is high or not, we need to keep this variability in mind. It makes to consider a standardized version of the predicted residual where we divide it by its standard deviation:

$$\frac{\hat{e}_{[i]}\sqrt{1-h_{ii}}}{\sigma}.$$

Because σ is unknown, we cannot compute these. So we need to estimate σ. We have two choices:

1. We can use the Residual Standard Error in the full model: $\sqrt{RSS/(n-p-1)}$. We then get

$$\frac{\hat{e}_{[i]}\sqrt{1-h_{ii}}}{\sqrt{RSS/(n-p-1)}} = \frac{\hat{e}_i}{\sqrt{(1-h_{ii})RSS/(n-p-1)}}.$$

   This is nothing but the standardized residual r_i that we defined previously. Note that these do NOT have the t-distribution (under normality) because the numerator and denominator are not independent.

2. We can use the Residual Standard Error in the model obtained by dropping the ith subject. We then get

$$t_i = \frac{\hat{e}_{[i]}\sqrt{1-h_{ii}}}{\sqrt{RSS_{[i]}/(n-p-2)}}.$$

   Now it can be checked that the numerator and the denominator are independent (why?) which implies that t_i has the t-distribution with $n-p-2$ degrees of freedom. These quantities $\{t_i\}$ are called standardized predicted residuals or externally studentized residuals.

Under the assumption that the linear model is true and normality of the errors, the standardized predicted residual t_i has the $t(n-p-2)$ distribution. This can therefore be used to assess whether the ith subject

is an outlier. Indeed, one can conduct a formal test of whether the ith subject is an outlier or not (what are H_0 and H_1 here?) by looking at t_i and rejecting the null (i.e., declaring ith subject as the outlier) if it is larger in absolute than the $\alpha/2$ critical value for the $t(n-p-2)$ distribution.

One has to be careful with this above test. Suppose that the standardized predicted residuals are $t_1, \ldots, t_n$. Suppose also that the model is true so that $t_1, \ldots, t_n$ have the $t(n-p-2)$ distribution. They are not necessarily independent however. But just for the sake of the discussion, assume that they are roughly independent. Then we would expect about $n\alpha$ number of $t_1, \ldots, t_n$ to be larger in absolute value than the $\alpha/2$ critical value of the $t(n-p-2)$. Therefore, we would be tagging $n\alpha$ of the subjects as outliers even when there are no outliers in the data.

To counter this, one takes a value of α much smaller than 0.05 while testing whether the ith observation is an outlier or not. A particularly conservative value is $\alpha = 0.05/n$. In this case, one can show that the probability that at least one subject is tagged an outlier when in fact there are none is atmost 0.05. This is known as the Bonferroni correction.

3 Cook's Distance

This measures the distance between $\hat{\beta}$ and $\hat{\beta}_{[i]}$. Recall from the last section that $\hat{\beta}_{[i]}$ is the estimate of β with the ith subject removed.

A naive way of measuring the distance between $\hat{\beta}$ and $\hat{\beta}_{[i]}$ is to use the Euclidean distance: $(\hat{\beta} - \hat{\beta}_{[i]})^T(\hat{\beta} - \hat{\beta}_{[i]})$. But this ignores the fact that the different elements of $\hat{\beta}$ have different variances. Because $Cov(\hat{\beta}) = \sigma^2(X^TX)^{-1}$, it makes sense to use the idea underlying Mahalanobis distance to measure the distance between $\hat{\beta}$ and $\hat{\beta}_{[i]}$ by

$$\frac{(\hat{\beta} - \hat{\beta}_{[i]})^T X^T X(\hat{\beta} - \hat{\beta}_{[i]})}{\sigma^2}. \tag{3}$$

Because σ is unknown, we can estimate it by the Residual Standard Error, $\hat{\sigma}$. This gives us the notion of Cook's distance:

$$C_i = \frac{(\hat{\beta} - \hat{\beta}_{[i]})^T X^T X(\hat{\beta} - \hat{\beta}_{[i]})}{(p+1)\hat{\sigma}^2}. \tag{4}$$

Note that there is a division by $p+1$ above which was not there in (3). This does not really matter (as the division is the same for all i). I have kept the divisor because this is the standard way of defining the Cook's distance.

We can use (2) to give the following alternative expression for C_i:

$$C_i = r_i^2 \frac{h_{ii}}{(1-h_{ii})(p+1)}.$$

Check this. Note that C_i depends on r_i and the leverage h_{ii}. If r_i^2 is large and/or h_{ii} is large, then C_i will be large.