Spring 2015 Statistics 151 (Linear Models) : Lecture Thirteen

Aditya Guntuboyina

05 March 2015

1 Regression Diagnostics

For regression diagnostics, we need to know about the following quantities:

1. Leverage
2. Standardized Residuals
3. Predicted Residuals
4. Standardized Predicted Residuals
5. Cook's Distance

We looked at Leverages in the last class.

2 Standardized Residuals

The residuals $\hat{e}$ satisfy $var(\hat{e}) = \sigma^2(I - H)$. In particular, it is important to know that the residuals are correlated and have different variances.

For diagnostics, it is useful to look at standardized residuals; defined as

$$r_i = \frac{\hat{e}_i}{\hat{\sigma}\sqrt{1 - h_{ii}}}.$$

Under the assumption of normality on $e_1, \dots, e_n$, we know that the residuals $\hat{e} \sim N(0, \sigma^2(I - H))$. Does the standardized residual r_i have a t-distribution? NO! because $\hat{e}_i$ and $\hat{\sigma}$ are not independent.

3 Predicted Residuals

How does one find outliers in the regression data? A first answer might be to look for subjects having large residuals. But the problem with this approach is that when the outlier also has a large leverage, then the residual will not be that large. Therefore, one needs to look at a combination of leverage and the value of the residual. It turns out that predicted residuals are a natural way of combining the residuals and the leverages.

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.

Let $X_{[i]}$ denote the X-matrix with the ith row deleted. Also, let $Y_{[i]}$ denote the Y-vector with the ith entry deleted and 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 the Sherman-Morrison formula from matrix algebra. I will do this in the next class.