Lecture 21

Identifying Influential Observations as they affect OLS estimates

Reference Wooldridge, pp. 312-317. Section labeled "Outliers and Influential Observations."

Observations that influence OLS estimates by "a large amount" are called "outliers" or "influential" observations. OLS is especially susceptible to outliers because OLS estimates are derived by minimizing sum of squared residuals and large residuals carry a large weight in determining the OLS estimates. As Wooldridge notes (p. 312), outlying observations occur for two reasons:

(1) There is a mistake in entering the data - an added zero; a misplaced decimal point, etc.
It is always good to check the minimum and maximum values of each variable in

your regression to see if such an entry error has occurred. You can use PROC Tabulate or PROC Means in SAS to give you the maximum and minimum values and even histograms of your variables.

(2) Outliers can often arise when sampling from a small population. On the one hand, outlying observations can be very helpful in reducing the standard errors of OLS estimates through adding, often badly needed, variation in the explanatory variable. On the other hand, if the outlying observation really represents an observation from another population, then mixing the outlying observation with the other observations will give a distorted picture of the true population that you are trying to describe with your multiple regression.

Thus, determining whether an observation that is outlying is truly from another population requires further study of the observation especially as it relates to trying to understand why the given observation is so different.

For example consider the SAS program Influence.sas that is posted on the website for this course. The data analyzed are house prices as it relates to various house and lot characteristics. The "influence" option in PROC REG is used to generate various statistics that can be used to detect outlying observations that we might want to single out for further inspection. You should read the comments in the Influence.sas program.

Four influence measures are produced:
(i) leverage, (ii) student, (iii) rstudent, and
(iv) dffits.

Leverage of the i-th observation is defined by the matrix product

$$\text{leverage}_i = \underline{X}_i'(X'X)^{-1}\underline{X}_i \quad \text{where}$$

$\underline{X}_i'$ = the i-th row of the explanatory variables matrix.

The larger the leverage measure, the more influential the observation is. Leverage is related to the degree by which the OLS estimates will change if the i-th observation is dropped from the regression. Belsley, Kuh, and Welsch (1980) suggest that an influential observation is one where $\text{leverage}_i > 2K/N$ where K denotes the number of explanatory variables in the model, including the intercept and N is the number of observations.

A studentized residual, say $\hat{u}_i^{(s)}$, is defined to be the OLS residual divided by the standard error in predicting $E(y_i | X = x_i)$, the conditional mean of y_i:

$$\hat{u}_i^{(s)} = \frac{\hat{u}_i}{\hat{\sigma}\sqrt{1 - X_i'(X'X)^{-1}X_i}} = \frac{\hat{u}_i}{\hat{\sigma}\sqrt{1 - \text{leverage}_i}}$$

where $\hat{\sigma} = \sqrt{SSR/(N-K)}$ is the standard error of the regression. SAS calls $\hat{u}_i^{(s)}$, "student," for studentized residuals. Belsley, Kuh, and Welsch (1980) define outlying observations as those whose studentized residuals are greater than 2 in absolute value, i.e. $|\hat{u}_i^{(s)}| > 2$.

To develop another type of studentized residual, Rstudent, consider the following notation:

Let $h_i = x_i'(X'X)^{-1}x_i$ be the leverage of the i-th observation, and $\hat{\sigma}_{(i)}$ be the standard error of the regression based on all of the observations except the i-th. Let $\hat{u}_i$ denote the OLS residual for the i-th observation based on all of the available observations. Then define

$$RSTUDENT_i = \hat{u}_i/(\hat{\sigma}_{(i)}\sqrt{1-h_i})$$

as the "restricted" studentized residual. Belsley, Kuh, and Welsch define an outlying observation as one where $|RSTUDENT_i| > 2$. Another outlier detection statistic is the so-called DFFITS statistic. This statistic is a scaled measure of the change in the predicted value for the i-th observation and is calculated by deleting the i-th observation. Let $\hat{y}_{(i)}$ be the i-th predicted value of y without using the i-th observation. Then DFFITS is defined by

$$DFFITS = (\hat{y}_i - \hat{y}_{(i)})/(\hat{\sigma}_{(i)}\sqrt{h_i})$$

where $\hat{y}_i$ is the fitted value of y_i using all of the available observations. Large absolute values of DFFITS (i.e. "difference in fits") indicate influential observations. Belsley, Kuh, and Welsch indicate that observations where

$$|DFFITS| > 2\sqrt{K/N}$$

be classified as outlying observations.

The SAS program Influence.sas demonstrates the use of the above concepts in identifying outliers. As it turns out house #77 is an outlying observation where the price of the house is surprising low given the very large lotsize it is located on. Either the housing price or lotsize was misentered in the initial coding of the data or there is something unseen that is affecting the price of the house irrespective of the lotsize like a waste dump on the backside of the lot or something. At any rate some further inquiry needs to be made into the observation to determine if it should remain as part of the housing equation we are trying to build.