Mngt 469

Influential Observations and the Leverage Statistic

Suppose you have a regression with 100 observations. You might think that each observation contributes 1% towards the computation of the regression coefficients. You are wrong! *Some observations are more important than others*. This is because OLS regression tries to minimize SSE and so it pays more attention to those observations for which changes in the coefficients produce the largest reductions in the SSE. We say that these observations have the most *influence* or *leverage*. In practical terms, if an observation has a lot of leverage, then if you remove it, the coefficients will change noticeably.

It is often helpful to identify the most influential observations. If you had some doubts as to the validity of the data for these observations or whether they belonged in the model to begin with, you should be aware of their influence. Do you really want your results to hinge on questionable observations? You can be comforted if, after removing these questionable yet influential observations, your key results remain unchanged.[1]

There are a variety of ways to measure leverage. Three measures available in Stata are the DFITS, Cook's Distance, and Welsch Distance. In my experience, the Cook's Distance is used most often. To compute the Cook's Distance, simply follow your regression with **predict varname if e(sample), cooksd**.[2] This will create a new variable (varname) that measures each observation's leverage.

You can now identify the observations with the highest leverage (e.g., sort by varname and examine the first few rows of your data). If you are in doubt as to whether to include these observations, try rerunning your model without them!

Steps for dealing with influential observations:

1) Determine whether you are concerned about undue influence – do you have doubts about including all the observations in your data?
2) After your regression, type
predict varname if e(sample), cooksd
sort varname
3) Examine the observations with the highest values of varname
4) If you concerned about these observations, rerun your regression without them (e.g., regress Y X if varname < Z, where Z is some cutoff value of the leverage statistic that you think is too high.)

[1] Don't remove observations simply because they have influence! Some observation must always have the most influence. This does not mean it should be removed from the data.
[2] The term **if e(sample)** assures that Stata computes the Cook's d only for observations used in the regression.