Mngt 917

Regression Diagnostics in Stata

Stata offers a number of very useful tools for diagnosing potential problems with your regression. Simply type one or more of these commands after you estimate a regression model. You can refer to the Stata reference manual, under regression diagnostics, to learn more about these tools.

vif This calculates the *variance inflation factor*, a measure of potential multicollinearity. The **vif** command computes a vif for each variable and for the overall regression. (There is no hard and fast rule about acceptable vif's). However, if a predictor variable has a vif in excess of about 20, then it may be collinear with another predictor. If the average vif across all predictors is "substantially higher" than 1 (perhaps 5 or higher), then there is multicollinearity across all the predictors. The vif score may be grossly inflated if you use categorical variables, interactions, or exponents. The vif command is rarely used in practice, perhaps because experienced statisticians can recognize multicollinearity in other ways.

hettest This performs the Cook-Weisberg test for heteroscedasticity. If the test statistic is significant, then there is unspecified heteroscedasticity, which you can correct by estimating with the **robust** option to the **regress** command.

hettest varname Stata determines if the error variance is correlated with the specified variable. If so, then you may use should weighted least squares instead of OLS. You may use both **WLS** and **,robust** in the same model.

Note: Like many diagnostic tests, **hettest** tends to reject the null hypothesis a bit too much for some tastes. Use a rigorous threshold of signficance, say, $p < .01$. Use an even tougher threshold if you have a lot of observations.

lvr2plot This produces a graph known as a *L-R plot*. The x-axis contains the squared values of the residuals (normalized so that the standard deviation of the residuals is 1). The y-axis contains a statistic called the *leverage statistic*, which is a measure of a particular observation's influence on the regression results. You should carefully examine observations with small residuals and large leverage. The computer worked very hard to fit these observations, and the estimated coefficients would change by a lot if you were to remove them. Be sure that these observations belong in the model – often you will find that they contain noisy data, or do not belong in the same data set at all!

You can recover the leverage statistic directly with the following command:

predict varname, leverage

The observations with the highest leverage had the greatest influence on the regression coefficients.