Steps for simultaneous (OLS) multiple regression using STATA

Simultaneous OLS regression analysis:

**regress <dv> <iv’s>**  (to obtain unstandardized regression solution and test statistics)

**regress <dv> <iv’s> , beta** (to obtain standardized regression coefficients)

Diagnostics: much of this review comes from <https://stats.idre.ucla.edu/stata/webbooks/reg/chapter2/stata-webbooksregressionwith-statachapter-2-regression-diagnostics/>

visualizing relationships among variables & looking for possible non-linear relationships:

**graph matrix <variable names>**

Closer look using scatterplot

**scatter (x,y)**

Matrix of correlations and significance

**pwcorr <var names>, sig**

Identifying problems with multicollinearity using Variance inflation factor (VIF):

**estat vif**

assumptions looking at residuals plots:

**predict yhat** (generating fitted values on Y)

**predict ur, resid** (generating unstandardized residuals)

**predict sr, rstandard** (generating standardized residuals)

**predict r, rstudent** (generating studentized residuals)

---Plots useful for detecting violation of constant variance assumption, looking for non-linear relationships between IV’s and DV that were unspecified in original regression model

**twoway scatter r yhat, yline(0) (**plotting studentized residuals against fitted values on Y)

**twoway scatter r <iv> yline(0) (**plotting studentized residuals against fitted values on Y)

Plot of squared (unstandardized) residuals against fitted values can also be used to identify potential violation of constant variance assumption:

**generate squr=ur^2**

**twoway scatter squr yhat, (**plotting squared residuals against fitted values on Y)

Detecting extreme cases (residuals) using studentized residuals

**generate absVrstu=abs(r)** - using to generate absolute values of studentized residuals

**sort absVrstu (**sorting the absolute value of the studentized residuals from low to high)

**list id r absVrstu** (to get the entire list of absolute studentized residuals)

**list id r absVrstu in 41/50** (to get the largest 10 absolute studentized residuals)

Criterion: Optimally residuals fall within +2 and -2

Testing normality assumption with residuals

**summarize ur, detail** (to obtain general descriptives, including skewness and kurtosis statistics on residuals)

**sktest ur** (provides statistical tests of skewness and kurtosis, and joint normality)

**swilk ur** (best used with observations < 50)

Testing normality of residuals (part B) using P-P and Q-Q plots

**pnorm ur**

**qnorm ur**

tests of constant variance Breush-Pagan

**estat hettest**

tests of constant variance White’s test

**estat imtest**

Obtaining univariate, bivariate, and multivariate normality test for variables in the model:

**mvtest normality SMint MGoal Anxiety Achieve, univariate bivariate**

Detecting influential cases with Cook’s d

**predict d,cooksd**

**sort d** (sorting cases on cooks d from low to high)

**list id d in 41/50** (to get the 10 cases with the largest d values)

--criterion values > 1 (Lomax & Hahs-Vaugn) or values >4/n)

Detecting influential cases with dffits

**predict dffits, dfits** (generating dffits variable)

**generate absVdffits=abs(dffits)**

**sort absVdffits** (sorting on dffits variable)

**list id absVdffits in 41/50** (identifying top 10 cases with largest dffits)

--criterion: cases with dffits > **2\*sqrt(k/n)**.

Detecting influential cases with dfbetas

**predict dfbeta<iv>, dfbeta(iv)** (generating dfbeta variable)

**generate absVdfbeta<iv>= abs(dfbeta<iv>)**

**sort absVdfbeta<iv>** (sorting on dfbeta variable)

**list id absVdfbeta<iv>**  **in 41/50** (identifying top 10 cases with **largest absVdfbeta<iv>**

criterion: cases with absVdfbeta > **2/sqrt(n)**