Labor Economics with STATA

Liyousew G. Borga

CERGE EI

November 4, 2015

Introduction to Regression Diagnostics

Outline

Major assumptions of the Linear Regression Model

$$Y_i = \beta_0 + X_{i1} + \cdots + X_{iT} + \varepsilon_i$$

We should always check our fitted models to make sure that the following assumptions have not been violated

- The relationship between the outcomes and the predictors is (approximately) linear
- The error term ε has zero mean
- The error term ε has constant variance
- The errors are uncorrelated
- The errors are normally distributed or we have an adequate sample size to rely on large sample theory

Methods for detecting violations of assumptions are often referred to as "diagnostics" in that they are used to diagnose or reveal problems in the data.

- Departures from the underlying assumptions cannot be detected using any of the summary statistics (e.g. t or F statistics, R^2)
- In fact, tests based on these statistics may lead to incorrect inference since they are based on many of the assumptions above
- Common diagnostics tools include:
  - Identifying outliers and influential observations
  - Residual analysis

Outliers:

- Outliers are atypical data points that do not fit with the rest of the data.
- An outlier may arise due to some sort of contamination or error, or may be a valid but very extreme observation
- Outliers may have a dramatic impact on results of regression analyses, potentially having major impact on effects sizes and regression coefficients
  - may cause a weak (or zero) linear relationship to appear to be a strong linear relationship,
  - may have the opposite effect by masking a strong linear relationship
- Outliers tend to have a stronger effect when n is small than when n is large

Three types of detection measures are commonly used:

- Leverage: Extremity of each case on the independent variables
  - Leverage is a measure of how far an observation deviates from the mean of that variable.
  - These leverage points can have an effect on the estimate of regression coefficients
- Discrepancy: Extremity of each case on the dependent variable
- Influence: Influence of each case on regression results
  - An observation is said to be influential if removing the observation substantially changes the estimate of coefficients
  - Influence can be thought of as the product of leverage and outlierness

How to verify regression assumptions and detect potential problems using Stata

- Install some user written convenience commands from the net

```
ssc install commandname
```

- Some of these commands are: indexplot, rvfplot2, rdplot, qfrplot, ovfplot
- Several of these commands are readily available from:

```
net from http://www.ats.ucla.edu/stat/stata/ado/analysis
net install commandname
```

Scatter plots

- Use "crime.dta" to to build a linear regression model between the response variable `crime` and the independent variables `pctmetro`, `poverty` and `single`

```
use http://www.ats.ucla.edu/stat/stata/webbooks/reg/crime.dta, clear
describe
summarize crime murder pctmetro pctwhite pcths poverty single
graph matrix crime pctmetro poverty single
scatter crime pctmetro, mlabel(state)
scatter crime poverty, mlabel(state)
scatter crime single, mlabel(state)
```

- The scatter plots of crime against each of the predictor variables will give us some ideas about potential problems
- A scatterplot matrix of these variables is also a quick summary

Residual Analysis

- The diagnostic methods we'll be exploring are based primarily on the residuals
- If the model is appropriate, it is reasonable to expect the residuals to exhibit properties that agree with the stated assumptions
- Methods for standardizing residuals:
  - Standardized residuals
  - Studentized residuals
  - Jackknife residuals

Residual Analysis

Let's try the regression command predicting `crime` from `pctmetro poverty` and `single`

```
regress crime pctmetro poverty single
predict r, rstudent
stem r
sort r
list sid state r in 1/10
list sid state r in -10/l
hilo r state
list state r crime pctmetro poverty single if abs(r) > 2
```

- Studentized residuals are a type of standardized residual that can be used to identify outliers
- We should pay attention to studentized residuals that exceed +2 or -2, and get even more concerned about residuals that exceed +2.5 or -2.5 and seriously concerned with residuals that exceed +3 or -3.

Leverage:

- identify observations that will have potential great influence on regression coefficient estimates.

```
predict lev, leverage
stem lev
hilo lev state, show(5) high
```

- Rule of thumb: carefully examine a point with leverage greater than $(2k+2)/n$, where k is the number of predictors and n is the number of observations

```
display (2*3+2)/51
list crime pctmetro poverty single state lev if lev >.156
```

- Such points are potentially the most influential

Leverage:

- We can make a plot that shows the leverage by the residual squared and look for observations that are jointly high on both of these measures
- We can do this using the lvr2plot command (leverage versus residual squared plot)
- It is a quick way of checking potential influential observations and outliers at the same time

```
lvr2plot, mlabel(state)
list state crime pctmetro poverty single if state=="dc" | state=="ms"
```

Measures of Influence:

- Cook's distance or Cook's D is a commonly used estimate of the influence of a data point - it measures the effect of deleting a given observation
- DFFITS is a diagnostic meant to show how influential a point is in a statistical regression
- Conventional cut-off points
  - Cook's D: cut-off point is $4/n$
  - DFFITS: cut-off point for DFITS is $2 * sqrt(k/n)$

```
predict d, cooksd
list crime pctmetro poverty single state d if d>4/51
predict dfit, dfits
list crime pctmetro poverty single state dfit if abs(dfit)>2*sqrt(3/51)
dfbeta
scatter _dfbeta_1 _dfbeta_2 _dfbeta_3 sid, ylabel(-1(.5)3) yline(.28
    -.28) mlabel(state state state)
```

- Partial-regression plot (added-variable plot)

```
avplot single, mlabel(state)
```

Normality of Residuals

OLS regression requires that the residuals (errors) be identically and independently distributed.

- After we run a regression analysis, we can use the predict command to create residuals and then use commands such as kdensity, qnorm and pnorm to check the normality of the residuals
- The "pnorm" command graphs a standardized normal probability (P-P) plot
- The "qnorm" plots the quantiles of a variable against the quantiles of a normal distribution
- Use the `elemapi2.dta` data file for these analyses

```
use elemapi2.dta, clear
regress api00 meals ell emer
predict r, resid
kdensity r, normal
pnorm r
qnorm r
swilk r
```

- "pnorm" is sensitive to non-normality in the middle range of data and
- "qnorm" is sensitive to non-normality near the tails.

Homoscedasticity of Residuals

- One of the main assumptions for the ordinary least squares regression is the homogeneity of variance of the residuals
- If the model is well-fitted, there should be no pattern to the residuals plotted against the fitted values
- If the variance of the residuals is non-constant then the residual variance is said to be "heteroscedastic"
- There are graphical and non-graphical methods for detecting heteroscedasticity
- A commonly used graphical method is to plot the residuals versus fitted (predicted) values, issuing the rvfplot command
- Two popular commands that test for heteroscedasticity are: imest and hettest

```
use elemapi2.dta, clear
regress api00 acs_k3 grad_sch col_grad some_col

rvfplot, yline(0)
```

- There are two ways to deal with the problem of heteroskedasiticty,
  - using heteroskedasticity-robust standard errors,
  - using appropriate transformations (variance stabilizing techniques)
- In practice it is recommended to use heteroskedasticity-robust standard errors by using the option robust in the regress command

```
regress api00 acs_k3 grad_sch col_grad some_col, r
```

Checking for Multicollinearity

- When there is a perfect linear relationship among the predictors, the estimates for a regression model cannot be uniquely computed.
- We can use the vif [variance inflation factor] command after the regression to check for multicollinearity.

```
regress api00 meals ell emer
vif
regress api00 acs_k3 avg_ed grad_sch col_grad some_col
vif
```

- The collin command displays several different measures of collinearity

```
collin acs_k3  avg_ed grad_sch col_grad some_col
collin acs_k3 grad_sch col_grad some_col
```

- The condition number is a commonly used index of the global instability of the regression coefficients - a large condition number, 10 or more, is an indication of instability.

Checking Linearity

- When we do linear regression, we assume that the relationship between the response variable and the predictors is linear
- If this assumption is violated, the linear regression will try to fit a straight line to data that does not follow a straight line
- Checking the linear assumption in the case of simple regression is straightforward, but a bit subtle in the case of multiple regression

```
use elemapi2.dta, clear
//simple regression//
regress api00 enroll
twoway (scatter api00 enroll) (lfit api00 enroll) (lowess api00 enroll)
//multiple regression//
regress api00 meals some_col
predict r, resid
scatter r meals
scatter r some_col
acprplot meals, lowess lsopts(bwidth(1))
acprplot some_col, lowess lsopts(bwidth(1))
```

Correcting for nonlinearity

```
use nations.dta, clear
describe
regress birth gnpcap urban
acprplot gnpcap, lowess
acprplot urban, lowess

graph matrix birth gnpcap urban, half
kdensity gnpcap, normal
generate lggnp=log(gnpcap)
label variable lggnp "log-10 of gnpcap"
kdensity lggnp, normal

regress birth lggnp urban
acprplot lggnp, lowess
```

How do we know we have included all variables we need to explain Y?

- A model specification error can occur when one or more relevant variables are omitted from the model, or one or more irrelevant variables are included in the model
- Testing for omitted variable bias is important for our model since it is related to the assumption that the error term and the independent variables in the model are not correlated
- If we are missing variables in our model and, it is correlated with the included regressor, and the omitted variable is a determinant of the dependent variable, then our regression coefficients are inconsistent

Model Specification

```
use elemapi2.dta, clear

regress api00 acs_k3
linktest
ovtest

regress api00 acs_k3 full
linktest
ovtest

regress api00 acs_k3 full meals
linktest
ovtest
```