Introduction to Stata

Lecture VIII & IX

Tomas R. Martinez

UC3M

September, 2019

"All models are wrong, but some are useful." George Box

Regressions

- "All models are wrong, but some are useful." George Box
- 99 out of 100 papers in econometrics have some type of regression
- The type of model you want to fit is highly dependent on the data / problem you have
  - Cross-section
  - Time series
  - Panel Data
  - Discrete choice
  - Duration / survival analysis
  - Many others...

Linear Regression Model

- One of the main goals in economics studies is to understand the relationship between some variables
  - Elasticities:
    - What is the effect of increasing my income in 1% on my food expenses?
    - What is the percentage change of my wages if I increase education in one year, everything else constant?
    - What is the effect of Currency Union on the trade between countries?
- We have seen so far in the course how to get correlations between variables
- Problem: We cannot distinguish the direction of the effects

Regression command syntax

- regress varlist [weight] [if exp] [in range] [, options]
- 1st variable in the varlist is the dependent variable, and the remaining are the independent variables.
- You can use Stata's syntax to specify the estimation sample; you do not have to make a special dataset.
- You can, at any time, review the last estimates by typing the estimation command without arguments.
- The level() option to indicate the width of the confidence interval. The default is level(95).
- An important option is **robust**

Example - elemapi.dta

- 400 elementary schools from the California Department of Education's API 2000 dataset. This data file contains a measure of school academic performance as well as other attributes of the elementary schools, such as, class size, enrollment, poverty
- We want to analyze the effect of class size, poverty and teaching quality on academic school performance

corr api00 acs_k3 meals full

Example - elemapi.dta

- 400 elementary schools from the California Department of Education's API 2000 dataset. This data file contains a measure of school academic performance as well as other attributes of the elementary schools, such as, class size, enrollment, poverty
- We want to analyze the effect of class size, poverty and teaching quality on academic school performance

```
corr api00 acs_k3 meals full
```

- Now, let's do a regression

```
regress api00 acs_k3 meals full
```

Example - elemapi.dta

- 400 elementary schools from the California Department of Education's API 2000 dataset. This data file contains a measure of school academic performance as well as other attributes of the elementary schools, such as, class size, enrollment, poverty
- We want to analyze the effect of class size, poverty and teaching quality on academic school performance

corr api00 acs_k3 meals full

- Now, let's do a regression

regress api00 acs_k3 meals full

- let's now use the robust standard errors

reg api00 acs_k3 meals full , robust

Example - elemapi.dta

- How can you access the results of your regression?
- **return list** $\rightarrow$ Nothing here
- **ereturn list** $\rightarrow$ all the information of your regression!
- A easy way to access your betas: _*b*[*varname*]
- **Example:** gen betameals=_ b[meals] $\rightarrow$ generate variable equal the coefficient of *meals*
- Predicting the depend variable and getting residuals:
  - **predict yhat**
  - **predict resid, residuals**

Regression interpretations

- Depending on the functional form, coefficients of the regression has different interpretation
- I will not cover this in class, because you will see this in the class of Introduction to Econometrics
- From the results of the regression before, are we done?
- Can you just handle our exercise/paper ?

Examining the dataset

- All we have seen until now is very important, so we are sure our data is "clear"
- summarize api00 acs_k3 meals full yr_rnd
- tabulate acs_k3
- list snum dnum acs_k3 if acs_k3 < 0
  - Indeed, looks like all observations from district 140 has an "artificial minus"
- Let's correct this!

Examining the dataset

- Let's make some further *graphical* analysis

histogram acs_k3

histogram api00, bin(20) xlabel(300(50)1000)

histogram meals

histogram full

- Looks like we have some problems in the variable full also

tab full

tabulate dnum if full $<= 1$

Some other graphs

- Another useful graphical technique for screening your data is a scatterplot matrix.
- Useful for searching for nonlinearities and outliers in your data

graph matrix api00 acs_k3 meals full, half

twoway (scatter api00 meals) (lfit api00 meals)

- It points to the same problems we already saw
- Now, let's fix these 2 problems!

Testing linear restrictions

- reg api00 ell meals yr_rnd mobility acs_k3 acs_46 full emer enroll , robust
- Test if all coefficients are equal to 0 jointly: **test acs_k3 meals full**
- Test if one coefficient is equal to 0: **test acs_k3**
- Test if coefficient of full is equal to 1: **test full=1**
- Test if full=1 and meals =-3.5:
- **test meals=-3.5, accumulate**

Diagnostic tools

- We know that the mean is sensitive to extreme values
- Problem of outliers
- We can detect them using some techniques
- Leverage-versus-residual squared plot :
  - This plots the leverages of all observations against their squared residuals.
  - Leverage tells you how large the influence of a single observation on the estimated coefficients is. Observations with high values (especially if they also have a large squared residual) could potentially be driving the results obtained.

reg api00 ell meals yr_rnd mobility acs_k3 acs_46 full emer enroll

lvr2plot, mlabel(snum)

Diagnostic tools

- We can also check if the residuals are normally distributed, for example
- We first need to get the residuals

regress api00 ell meals yr_rnd mobility acs_k3 acs_46 full emer enroll, robust

predict api00_hat

predict residuals, residuals

- The Normal Q–Q plot: sensitive to deviations from normality in the tails: **qnorm residuals**
- Also, we can plot the density: **kdensity residuals, normal**

Test for Homoskedasticity

- We can also check for homoskedasticity
- Common diagnostic tool: Plot the residuals against the predict values

rvfplot, yline(0)

- Formal test for heteroskedasticity

estat hettest // Breusch-Pagan test

estat imtest // White's test

Time Series

- A time series is a data set ordered in time
- GDP, consumption, unemployment rate, interest rate, inflation rate...
- **Example:** United States Real GDP (in 2012 prices)
- To Stata understand your data is time-series $\rightarrow$ **tsset**
- *tsset* datevariable [, options]
- Let's try data $\rightarrow$ tsset date
- Stata thinks our data is daily instead of quarterly...

Time Series

- Let's retrieve the quarter and year from our date variable and re-built in a way Stata understands
- **gen yr = year(date)** $\rightarrow$ generate year variable
- **gen qtr = quarterly(date)** $\rightarrow$ generate quarterly variable
- **gen date2 = yq(yr,qtr)** $\rightarrow$ recombine both
- *tsset date2*
- It could be the case your data is in another format or comes in string
  - generate date=quarterly(datestr,QY)
  - look at function date: **help date**

Time Series

- Once you ttset your data you can use time series operators
- **Lag:** l.variable
- **Lead:** f.variable
- **Differences:** d.variable
- **Second Lag:** l2.variable
- **Example: gen lgdp=l.gdpc1**
- Now if you want to estimate an AR(1) process: *reg gdpc1 lgdp1*

Time Series

- Other useful commands in time series:
- **Autocorrelations corrgram**, **ac** and **pac**
- **Example** ac gdp_growth
- Estimating an ARIMA(p,d,q) (or AR(p), MA(q)...)
- **arima depvar, arima(p,d,q)** or **arima depvar, ar(p) ma(q)**
- **Filters:** tsfilter hp cyclicalname = gdp, trend(trendname) $\rightarrow$ HP filter

Panel Data

- Panel data has cross-section dimension, as well as time dimension
- **Example:** Information of multiple countries over time
- Panel data can come in both **long** or **wide** format
- Use the **reshape** command to go from one format to the other
- As in the time series data, you have to set Stata to identify the panel: **xtset panelvar datevar** (your data has to be long)

Panel Data

- **NLSY79:** Follow individuals who were 14-22 years old in 79 over their life (in this sample only men)
- Apply **xtset**, now you can still apply time series operators
- It is usual in panel regressions to have *fixed* and *random* effect models
- Stata has the command **xtreg** for that, I prefer to use the user written command **reghdfe**
- Do some examples

Exporting Results

- Eventually you want to export your results for presentations / papers...
- Two nice user written commands: **outreg2** and **estout**
- You can export your regression output to Excel, txt, LaTeX...
- I will cover **outreg2**, but feel free to explore **estout**
  - ssc install outreg2
  - perform a regression
  - estimates store *regname*
  - outreg2 *regname* using filename.xls
- Let's do some examples