Econometric Softwares: Stata

Jérémy Do Nascimento Miguel

Class 5 Spring 2024

Session 5:Outline

- Hypothesis testing in Stata.
- Basic regression methods.
- Regressions with interactions, high-dimensional FE.
- Post-estimation commands.
- Exporting results.

Hypothesis testing

It is an educated guess about what is happening around us: for instance in-person stata class is more effective than via zoom

- It is a statistical method to identify whether our data or its parameters support a specific hypothesis
- Econometric/Statistics: we deal with two hypothesis
  - → **H0**: Null hypothesis, the one we want to test. For instance, men and women earn the same wage
  - → **H1**: Alternative hypothesis, the case if **H0** is rejected content...

Key Parameters in Hypothesis testing

- **Standard error:** the estimator of standard deviation (i.e. gives information about the variance in parameter of interest)
- **t-statistic:** standardized distance of an estimated parameter from its hypothesized value (i.e. value specified in H0)
- **p-value:** probability of randomly drawing data and observing a test statistic that would support what is stated in H0

Basic and common statistical tests

- **t-test** is based on a test statistic called the t-value, which is calculated by dividing a sample parameter (e.g. sample mean) by an estimate of the standard error of this parameter.
- **One-sample t-test:** to test whether a sample mean significantly differs from a hypothesized value. E.g. test whether the average monthly expenditure differs significantly from 500 EUR.
- **Two-sample t-test:** to compare the means of variable for two independent groups. E.g. balance tests in RCT analysis

t-tests in Stata

- ttest varname == #: one-sample t-test, i.e. test if mean of varname is equal to value specified in #
- ttest varname, by(groupvar): two-sample t-test, i.e. test if means of two groups are equal
- ttest varname1 == varname2 : two-sample t-test using variables, i.e. test if means of two variables are equal

Balance tests:

Why?

- Context of causal inference: check if there is any selection bias in terms of observed characteristics
- For instance in RCT: verify whether there is any statistical differences between treatment and control group

How?

- Estimate means and standard errors of observed characteristics for treatment and control group
- t-test on the difference in means of observed characteristics (two-sample t-test)
- Alternatively: regress variable of interest on treatment dummy (more flexible way if fixed effects or clustered standard errors are needed)

All round command for balacne tests

iebaltab command from **ietoolkit**. Install ietoolkit: ssc install ietoolkit

Synthax: iebaltab varlist, grpvar(treatment variable) save("outputfile.xls")

Exercise 5.1

1. Open the data 'hh comm.dta'.
2. Test whether overall poverty rate (mean of variable **poor**) is larger than 40
3. Test whether overall extreme poverty rate (mean of variable **expoor**) is larger than 20
4. Test whether poverty rate is equal in urban/rural locations (use dummy variable **urban**).
5. Generate a table, using command iebaltab, which shows differences in means of the following variables in poor vs. non-poor households (based on dummy variable poor): **hhsize, nchild, nmigrant, land, empratio, hhh female, hhh educ**.

Regression analysis

[bysort groupvar :] command depvar indepvars [if] [in] [weight] [, options]

- command: regression command: regress for OLS, logit, probit
- depvar: dependent variable
- indepvars: independent variables

Fancier regressions:

- Robust standard errors or clustered SE: add: rob or cluster(group) as option
- Interaction terms: **i.** for categorical variables – i.varname1 #i.varname2; **c.** for continuous c.varname1 #c.varname2

Regression with high dimensional FE

General approach: regress depvar indepvars i.groupvar

- i.groupvar automatically generates dummy variables for each value of groupvar But takes long time to run if there are many groups specified in groupvar

Faster alternative: areg or reghdfe. Do not forget to install them

- areg allows only one group variable: areg depvar indepvars, absorb(groupvar)
- reghdfe allows many group variables (and also their interactions): reghdfe depvar indepvars, absorb(groupvarlist)

Post-estimation Analysis

- Predicting fitted values or residuals
- Testing on estimated coefficients
- Estimating linear/non-linear combination of coefficients
- Calculating marginal effects (e.g. in logit/probit estimations) or elasticity
- Other tests (Hausman's specification test, heteroskedasticity tests etc.)

Steps:

1. Run estimation command (e.g. regress, logit etc.)
2. Run post-estimation command (you can check which post-estimation commands are allowed for each estimation command by typing: estcommand postestimation E.g. regress postestimation

Predicting fitted values/residuals

predict newvar [if], statistic where statistic can be:

- xb: linear prediction of dependent variable (fitted value)
- residuals: residuals
- stdp: standard error of linear prediction
- stdr: standard error of residuals
- ... (check full list in help menu for predict)

Testing on estimated coefficients

Syntax: test *coeflist*; test *exp* = *exp*

- Equality of two coefficients (i.e. equality of effects)
- Joint significance (i.e. F-test)
- Equality of coefficients with specific values

Linear/non-linear combination of parameters

- Linear combination: lincom exp
- Non-linear combination: nlcom exp

Why do we need it?

- Calculate combined effect of multiple independent variables
- Calculate total effects for groups when there are interaction terms which usually show difference in the effect between groups
- Calculate difference in the effects of independent variables
- Calculate marginal effects manually if needed (i.e. using derivatives)

Marginal effects

Marginal effects = derivatives: amount of change in dependent variable in response to one unit change in independent variable Basic syntax: margins, margeffecttype;

margeffecttype can be:

- dydx(*varlist*): marginal effects of variables in varlist
- eyex(*varlist*): elasticity of variables in varlist
- dyex(*varlist*): semi-elasticity of variables in varlist (= d(y)/d(ln x))
- eydx(*varlist*): semi-elasticity of variables in varlist (= d(ln y)/d(x))

margins command can do more than estimating marginal efffects or elasticities (e.g. estimating marginal means for subsamples, margins at values of covariates etc.). Note that OLS estimates are already marginal effects.

Accessing regression results

- Stata temporarily stores estimation results as **e-class**
- After ereturn list (or check the end of help menu of estimation command)
- And you can display/use stored results with e(*)
- Usually stored parameters:
  - → Estimated coefficients (in matrix or scalar form)
  - → Variance matrix
  - → Sometimes standard errors/CI as scalars
  - → Number of observations used in estimation
  - → R-squared, F-stat, degrees of freedom, sum of squares (SS)

Type mat list e(b) or mat list e(V) to display or use the stored results

Accessing regression results: few tricjs

- Coefficients and its standard errors can be directly accessed with _b[varname] and _se[varname]
- Main part of regression output with coefficients, SE, p-value, t-stat and CI is also stored as matrix r(table): use mat list r(table) to access them

Exporting regression results

Key (and popular) commands: esttab or outreg2

- esttab is a powerful command with a lot of options to design complex tables, yet still complicated to use. Might be more useful if you want to export tables to Latex which is directly put in the paper
- outreg2 is easier to use and offers capabilities sufficient to export results in publication-style tables, despite limitations compared to esttab in terms of flexibility

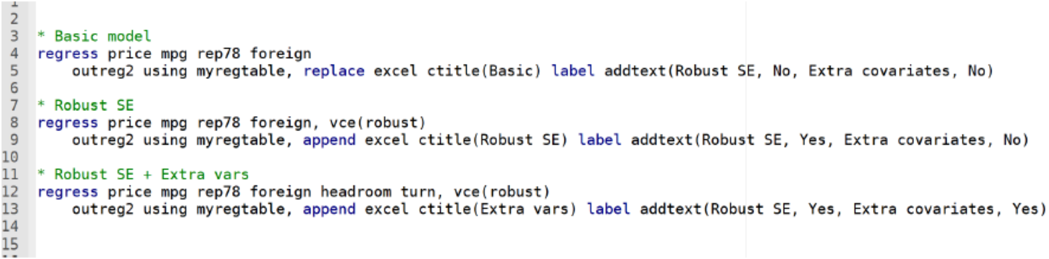
Exporting regression results: outreg2

Basic syntax: outreg2 using filename, [excel | tex | dta | word] [replace | append] [label] [ctitle(columntitle)] [addnote(notes)] [keep(varlist)] [drop(varlist)] [addtext(text)] [nocons] [noobs] [sortvar(varlist)]

- file formats supported: Excel, Word, TEX, DTA (Stata data file)
- replace replaces the existing file
- append adds results in a new column
- ctitle() defines column title
- addnote() adds notes at the bottom of table
- keep() or drop() keeps/drops variables for which coefficients are reported
- addtext() adds additional rows of text in table (e.g. to differentiate columns)
- nocons/noobs drops constant or number of observations in table
- sortvar() provides order of variables from table

Exporting regression results: outreg2

Combining results from multiple regressions



```
* Basic model
regress price mpg rep78 foreign
    outreg2 using myregtable, replace excel ctitle(Basic) label addtext(Robust SE, No, Extra covariates, No)

* Robust SE
regress price mpg rep78 foreign, vce(robust)
    outreg2 using myregtable, append excel ctitle(Robust SE) label addtext(Robust SE, Yes, Extra covariates, No)

* Robust SE + Extra vars
regress price mpg rep78 foreign headroom turn, vce(robust)
    outreg2 using myregtable, append excel ctitle(Extra vars) label addtext(Robust SE, Yes, Extra covariates, Yes)
```

Exercise 5.2 I

1. Open the data 'hh comm.dta'.
2. Run a regression of poverty dummy (poor) on variables the following variables: hhsize, nchild, nmigrant, empratio, land, urban, hhh female. Add option for robust standard errors. This will be your baseline model
3. Run following robustness checks:
   - → Use clustered standard error, with cluster being primary sampling unit (psuid)
   - → Clustered SE + additionally HH head education dummies (hint: use i.varname with HH head education variable)
   - → Clustered SE + HH head education dummies + community-level distance variables: distcapital, distregadm, distlocadm.
   - → Use the last specification, but change your dependent variable to extreme poverty dummy expoor.

Exercise 5.2 II

- → Use the last specification, but change your dependent variable to log per capita consumption logpccd.
- → Clustered SE + HH head education dummies + community FE (use commands for regressions with high-dimensional FE).

4. Export all your results to an Excel sheet using outreg2. At the bottom of the table, add rows of text to differentiate specifications in different columns. Make sure that variables are labelled in the exported table.
5. Now estimate the specification 'Clustered SE + HH head education dummies + community-level distance variables' using logit and probitand calculate marginal effects. If possible, export your results to an Excel file.