Stata Brush Up Course

Session 5

Sergi Quintana

IDEA

September 5, 2024

Plan for the session

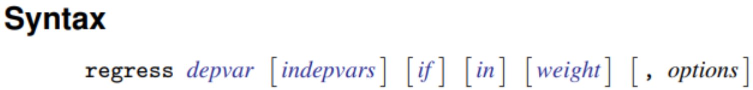
We will see the more advanced data options:

- Regression and options
- Post-estimation commands.
- Panel data
- Other estimation methods.

Regression

The main command for a linear regression is regress.

- regress performs ordinary least-squares linear regression.
- regress can also perform weighted estimation.
- It can compute robust and cluster–robust standard errors, and adjust results for complex survey designs.

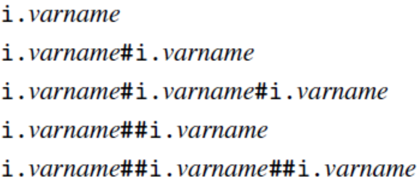


Syntax

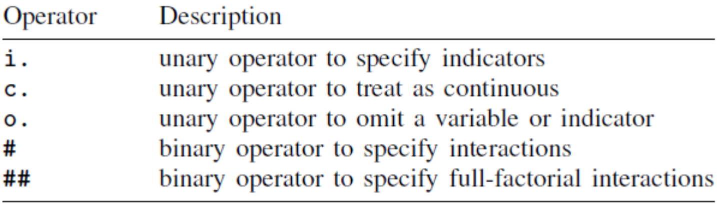
```
regress depvar [indepvars] [if] [in] [weight] [, options]
```

Regress - Factor Variables

Remember factor variables:



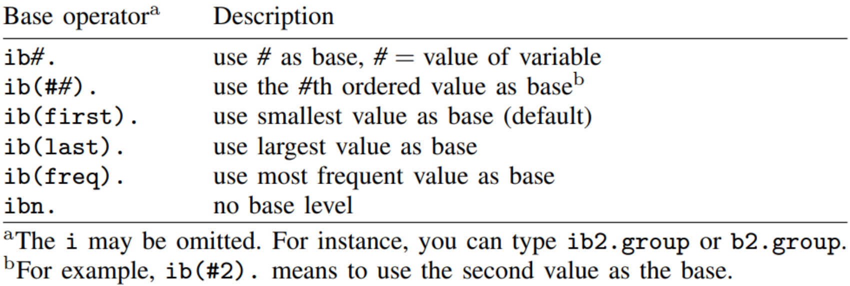
```
i.varname
i.varname#i.varname
i.varname#i.varname#i.varname
i.varname##i.varname
i.varname##i.varname##i.varname
```



Operator	Description
`i.`	unary operator to specify indicators
`c.`	unary operator to treat as continuous
`o.`	unary operator to omit a variable or indicator
`#`	binary operator to specify interactions
`##`	binary operator to specify full-factorial interactions

Regress - Base Levels

The base levels operators are:



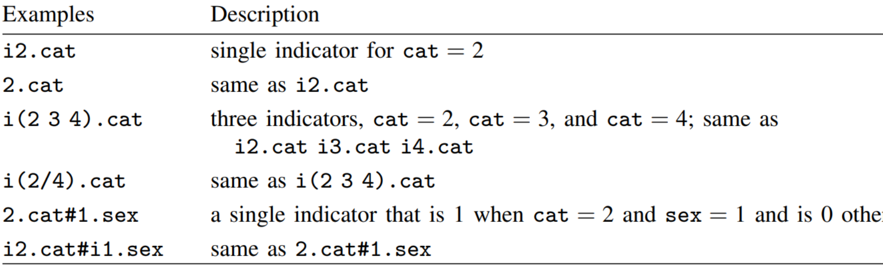
Base operator[a]	Description
`ib#.`	use # as base, # = value of variable
`ib(##).`	use the #th ordered value as base[b]
`ib(first).`	use smallest value as base (default)
`ib(last).`	use largest value as base
`ib(freq).`	use most frequent value as base
`ibn.`	no base level

[a] The `i` may be omitted. For instance, you can type `ib2.group` or `b2.group`.

[b] For example, `ib(#2).` means to use the second value as the base.

Regress - Base Levels

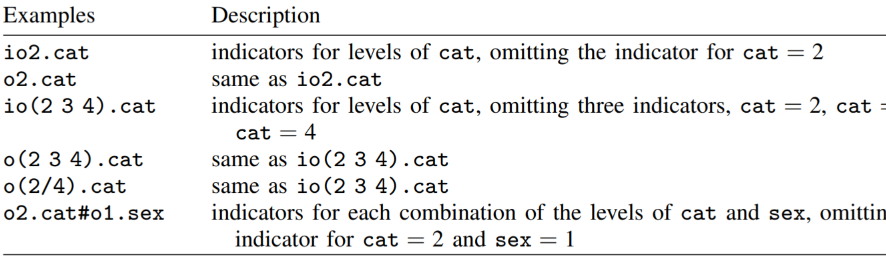
Set the base category:



Examples	Description
`i2.cat`	single indicator for `cat` = 2
`2.cat`	same as `i2.cat`
`i(2 3 4).cat`	three indicators, `cat` = 2, `cat` = 3, and `cat` = 4; same as `i2.cat i3.cat i4.cat`
`i(2/4).cat`	same as `i(2 3 4).cat`
`2.cat#1.sex`	a single indicator that is 1 when `cat` = 2 and `sex` = 1 and is 0 othe
`i2.cat#i1.sex`	same as `2.cat#1.sex`

Regress - Base Levels

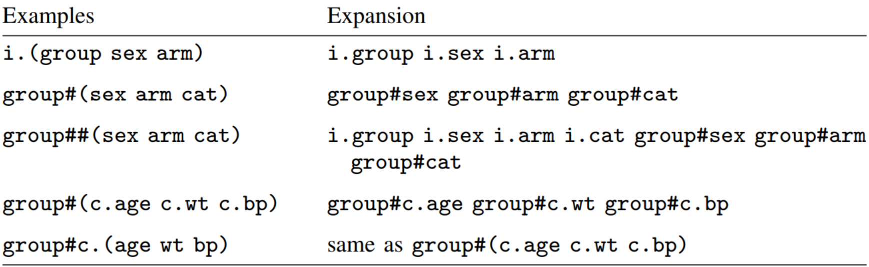
We can choose to omit some categories:



Examples	Description
`io2.cat`	indicators for levels of `cat`, omitting the indicator for `cat` = 2
`o2.cat`	same as `io2.cat`
`io(2 3 4).cat`	indicators for levels of `cat`, omitting three indicators, `cat` = 2, `cat` = `cat` = 4
`o(2 3 4).cat`	same as `io(2 3 4).cat`
`o(2/4).cat`	same as `io(2 3 4).cat`
`o2.cat#o1.sex`	indicators for each combination of the levels of `cat` and `sex`, omittin indicator for `cat` = 2 and `sex` = 1

Regress - Base Levels

We can apply the factor operation to many variables:



Examples	Expansion
i.(group sex arm)	i.group i.sex i.arm
group#(sex arm cat)	group#sex group#arm group#cat
group##(sex arm cat)	i.group i.sex i.arm i.cat group#sex group#arm group#cat
group#(c.age c.wt c.bp)	group#c.age group#c.wt group#c.bp
group#c.(age wt bp)	same as group#(c.age c.wt c.bp)

Regress - Options

Replaying prior results

We can do it by typing reg without any arguments.

Cataloging Estimation Results

After an estimation command, we can run esimtates store name. This will store the results in memory.

Saving Estimation Results

The command estimates save name will save the estimation results to our working directory.

Regress - Options

Width confidence intervals

The option level() allows to change the width of the interval.

Obtaining the variance-covariance matrix

We just need to type estat vce after the estimation has been done.

Predicted values

The predict command will allow to calculate predictions, residuals, and influence statistics.

Regress- Options

Accessing coefficients and standard errors

Just type _*b*[*varname*] or _*se*[*varname*] to access the coefficient or the standard error after estimation. Visit here for more.

Obtaining combinations of coefficients

To obtain linear combinations we use the lincom command. To obtain non linear ones we use the nlcom command.

Regress - Exercices

Using the nlswork from webuse:

- Try to find the determinants of the log of wage. Create different models. Try to include a polynomial on age and tenure of degree three. Save the results of the different models and create a table.

- Create a confidence interval at the 90% level. Predict the residuals and create a plot of residuals against fitted values.

- Now perform a regression of the determinant of wage with some regressors and their interaction with grade, but omitting the grade 15.

Regress - Exercices

Using the auto data from webuse:

- Estimate the following model:

  $$price = \beta_0 + \beta_1 mpg + \beta_2 mpg^2 + \beta_3 mpg \times weight + \beta_4 i.rep78 + \beta_5 length$$

- Is there any evidence of *mpg* and mpg^2 having statistically different effects? What is the joint effect of mpg on price? What is the effect of $\beta_1/\beta_2 + \beta_3^2$?

- Generate a new variable which is the predicted price. Now generate another variable which is the predicted price but without using rep78. Show graphically the relationship between both variables.

Regress - Diagnostic Plots

Diagnostic plots are used after estimation to analyze a particular part of the model. Some of them are:

- rvfplot: residual-versus-fitted plot
- avplot: added-variable plot
- avplots : all added-variable plots in one image.
- rvpplot: residual-versus-predictor plot.

To see the full list and documentation visit here. Important! In economics we use less diagnostic plots than in other branches since economic theory should already provide a lot of intuition for potential problems.

Regress - Variance Estimates

The main options are:

- vce(vcetype): specifies the type of standard error reported.

  Some types are:

  - vce(ols), the default, uses the standard variance estimator for ordinary least-squares regression.
  - vce(cluster clustvarlist) specifies that standard errors allow for intragroup correlation within groups defined by one or more variables in clustvarlist, relaxing the usual requirement that the observations be indepenent.
  - vce(robust) uses the robust or sandwich estimator of variance.

Other options can be found here.

Regress - Heteroskedasticity

To deal with heteroskedasticity we have to adjust the variance estimation.

We can also use the hetregress function which estimates a linear regression model in which the variance is an exponential function of the covariates that we specify.

hetregress implements two estimators for the variance:

- a maximum likelihood (ML) estimator and a two-step GLS estimator.
- The ML estimates are more efficient than those obtained by the GLS estimator if the mean and variance function are correctly specified and the errors are normally distributed.
- The two-step GLS estimates are more robust if the variance function is incorrect or the errors are nonnormal.

Exercices - Heteroskedasticty

Here we will use a dataset of 725 faculty members' salaries described in DeMaris (2004) to determine whether there is evidence of a difference in salaries between male faculty and female faculty. In addition to sex (female), other variables that might affect the salaries are prior experience (priorexp), years in rank (yrrank), years at the university (yrbg), and marketability of discipline (salfac).

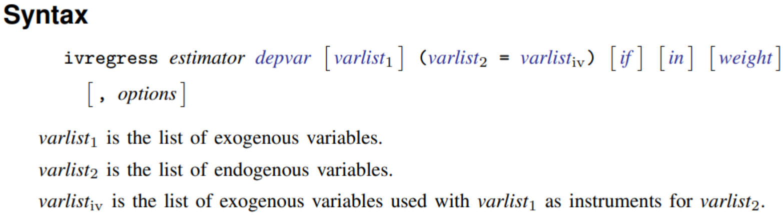
Using data salary from webuse:

- Fit a salary model with regress by including main effects and the interaction terms between female and all other variables.
- Test for the presence of heteroskedasticity across sex. Perform the Breusch-Pagan test but also perform visual investigation with diagnostic graphs.
- Correct for heteroskedasticity using robust standard errors and also using a GLS two-step estimation.

Instrumental Variables

The main command for instrumental variable estimation is ivregress.

ivregress supports estimation via two-stage least squares (2SLS), limited-information maximum likelihood (LIML), and generalized method of moments (GMM).



Syntax

ivregress *estimator* *depvar* [*varlist*$_1$] (*varlist*$_2$ = *varlist*$_{\text{iv}}$) [*if*] [*in*] [*weight*]
[, *options*]

varlist$_1$ is the list of exogenous variables.

varlist$_2$ is the list of endogenous variables.

varlist$_{\text{iv}}$ is the list of exogenous variables used with *varlist*$_1$ as instruments for *varlist*$_2$.

Exercice - Instrumental Variables

Using nlswork data from webuse:

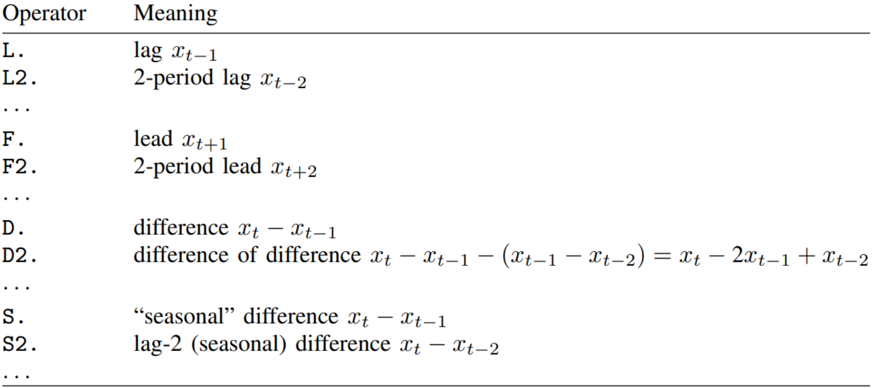
- Fit a model of wages as a function of each woman's age and age squared, job tenure, birth year, and level of education. We believe that random shocks that affect a woman's wage also affect her job tenure, so we treat tenure as endogenous. As additional instruments, we use her union status, number of weeks worked in the past year, and a dummy indicating whether she lives in a metropolitan area.

- Perform an IV estimation strategy to account for the endogeneity. First use an indenfitied model. Which instrument works better? Use now a 2SLS method. Are the instruments weak? Finally estimate it by GMM. Always make sure you account for autocorrelation in the residual, since we have several observations for each woman, so try to cluster at the right level.

Time Series

To unlock all Stata pwoer when dealing with time series we need to declare our data set as being a time series. This is done with the command **tsset**.

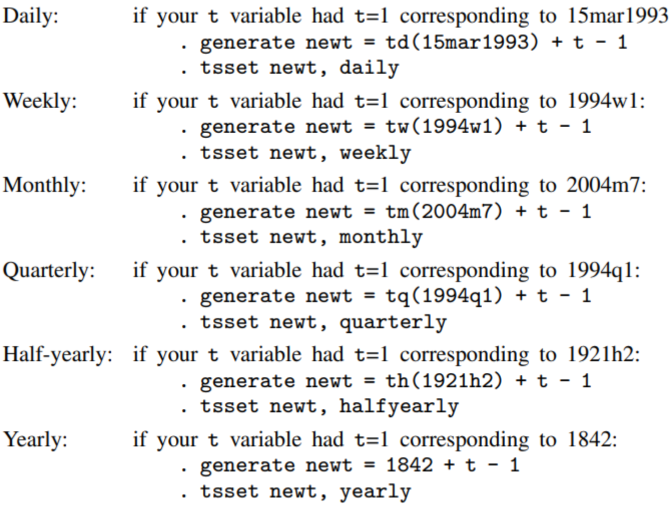
With this, we can use all the time operators:



Operator	Meaning
`L.`	lag x_{t-1}
`L2.`	2-period lag x_{t-2}
...	
`F.`	lead x_{t+1}
`F2.`	2-period lead x_{t+2}
...	
`D.`	difference $x_t - x_{t-1}$
`D2.`	difference of difference $x_t - x_{t-1} - (x_{t-1} - x_{t-2}) = x_t - 2x_{t-1} + x_{t-2}$
...	
`S.`	"seasonal" difference $x_t - x_{t-1}$
`S2.`	lag-2 (seasonal) difference $x_t - x_{t-2}$
...	

Time Series

We can change the time dimension format:



Daily: if your `t` variable had `t=1` corresponding to 15mar1993

```
. generate newt = td(15mar1993) + t - 1
. tsset newt, daily
```

Weekly: if your `t` variable had `t=1` corresponding to 1994w1:

```
. generate newt = tw(1994w1) + t - 1
. tsset newt, weekly
```

Monthly: if your `t` variable had `t=1` corresponding to 2004m7:

```
. generate newt = tm(2004m7) + t - 1
. tsset newt, monthly
```

Quarterly: if your `t` variable had `t=1` corresponding to 1994q1:

```
. generate newt = tq(1994q1) + t - 1
. tsset newt, quarterly
```

Half-yearly: if your `t` variable had `t=1` corresponding to 1921h2:

```
. generate newt = th(1921h2) + t - 1
. tsset newt, halfyearly
```

Yearly: if your `t` variable had `t=1` corresponding to 1842:

```
. generate newt = 1842 + t - 1
. tsset newt, yearly
```

Exercises - Time Series

Using the "ex_1.dta" dataset:

- Set it as a time series.
- Generate the log differences in consumption and investment.
- Generate the exact percentage change in consumption and investment.
- Check how well the log differences approximate the percentage changes.
- Run the following regressions:

$$inv_t = \beta_0 + \beta_1 cons_t + \beta 2 cons_{t-1} + \beta_3 inc_t + \beta_4 inc_{t-1}$$

$$inv_t = \beta_0 + \beta_1 \Delta cons_t + \beta_3 \Delta inc_t$$

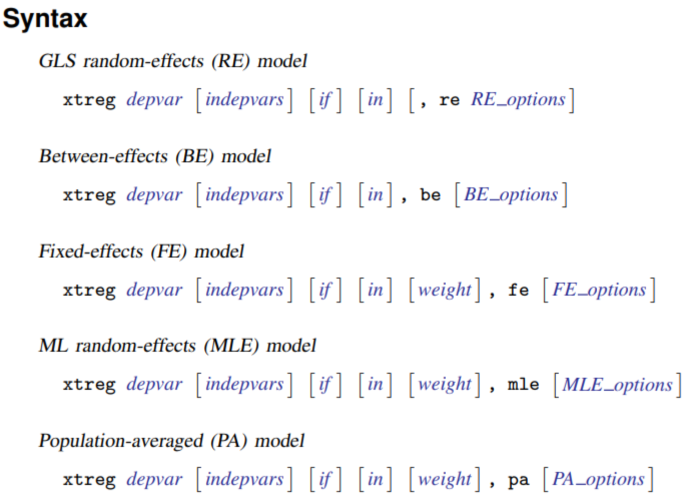
Panel Data Methods

There are many types of panel data and goals of panel-data analysis, leading to different models and estimator for panel data. We will here review the estimation methods for the most common panel data scenarios:

- Fixed-effects model.
- Random-effects model.
- Population-averaged model.

Panel Data

The key for the estimation of panel data methods is the xtreg command.



Syntax

GLS random-effects (RE) model

```
xtreg depvar [indepvars] [if] [in] [, re RE_options]
```

Between-effects (BE) model

```
xtreg depvar [indepvars] [if] [in] , be [BE_options]
```

Fixed-effects (FE) model

```
xtreg depvar [indepvars] [if] [in] [weight] , fe [FE_options]
```

ML random-effects (MLE) model

```
xtreg depvar [indepvars] [if] [in] [weight] , mle [MLE_options]
```

Population-averaged (PA) model

```
xtreg depvar [indepvars] [if] [in] [weight] , pa [PA_options]
```

Panel Data

We can also deal with panel data using other commands. Usually, when data sets are very large we will use:

- areg
- reghdfe. This should be your preferred option.

There are also other estimation techniques:

- IV panel data
- Arellano-Bond and Arellano-Bover for dynamic panels.

Exercices - Panel Data

using firm data:

- Reshape the data from wide to long and drop duplicates and missings. Is the panel balanced?
- Estimate a regression of n on k w y controlling for firms fixed effects. Check that introducing dummies for each firm or using the xreg,fe command is equivalent.
- Estimate now a random effects model and check which one should be used with the corresponding test.
- Estimate now the model absorbing the firm effects with areg. Finally, estimate the model including firm and year effects. Save the firm fixed effects and show the relationship between firm fixed effects and y. You can do it graphically and using regressions.

Exercices - AKM Model

Abowd, Kramarz and Margolis (1999) (AKM) where trying to understand if workers sort into firms according to their skills.

To do so, they decomposed individual's income into a firm component and a wage component.

$$log(inc) = firm\_fe + individual\_fe + \epsilon$$

Using the akm_data.dta file, perform the so-called AKM decomposition, and show if there is any sorting of workers into firms.

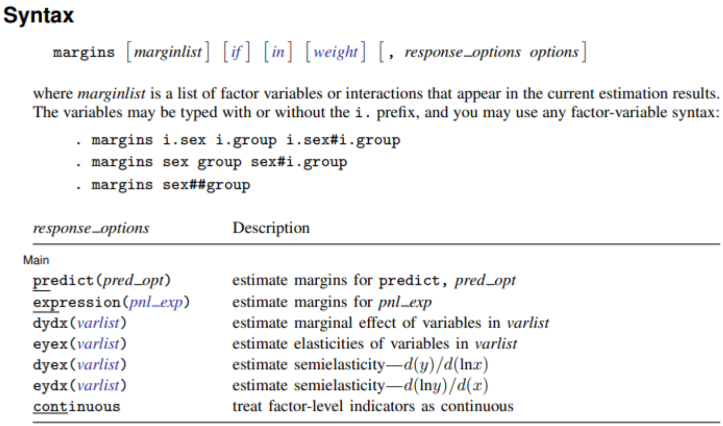
Post Estimation Commands

We will now focus on post-estimation commands:

- margins: to calculate statistics frmo previously fitted models.
- coefplot: plot results from estimated models.
- esttab: allows to store stata outputs to different formats (latex among them).

The Margins Command

Margins are statistics calculated from predictions of a previously fit model at fixed values of some covariates and averaging or otherwise integrating over the remaining covariates.



Syntax

```
margins [marginlist] [if] [in] [weight] [, response_options options]
```

where *marginlist* is a list of factor variables or interactions that appear in the current estimation results. The variables may be typed with or without the `i.` prefix, and you may use any factor-variable syntax:

```
. margins i.sex i.group i.sex#i.group
. margins sex group sex#i.group
. margins sex##group
```

response_options	Description
Main	
`predict(`*pred_opt*`)`	estimate margins for `predict,` *pred_opt*
`expression(`*pnl_exp*`)`	estimate margins for *pnl_exp*
`dydx(`*varlist*`)`	estimate marginal effect of variables in *varlist*
`eyex(`*varlist*`)`	estimate elasticities of variables in *varlist*
`dyex(`*varlist*`)`	estimate semielasticity—$d(y)/d(\ln x)$
`eydx(`*varlist*`)`	estimate semielasticity—$d(\ln y)/d(x)$
`continuous`	treat factor-level indicators as continuous

Post-Estimation Methods: coefplot

- coefplot is a Stata command to plot results from estimation commands or Stata matrices.
- Results from multiple models or matrices can be combined in a single graph.
- The default behavior of coefplot is to draw markers for coefficients and horizontal spikes for confidence intervals. However, coefplot can also produce various other types of graphs.
- Visit here for the full documentation.

Post-Estimation Methods: esttab

The estout package provides tools for making regression tables in Stata. Visit here for the full documentation. The main commands are:

- esttab: a command for publiation-style regression tables that can be either displayed in the Stata window or exported to many formats.
- esttout : a generic program for making tables. It is the engine behid esttab.

Exercices - Other Post-Estimation Methods

Using auto data from webuse:

- Estimate three different models that expain price as a function of other variables. Always include one dummy for each category of rep78.
- Plot the coefficients of the dummies of rep78 for the three different models together.
- Generate a publication style table with the results of the three models. Try to include as many options as possible, so that once in latex you don't need to change anything.

Other Estimation Methods

There are several methods:

- Quantile regression
- Kernel regression.
- Logit and probit.

Exercice - Other Estimation Methods

Using nlswork from webuse:

- Keep only the year 1970 . Estimate a quantile regression model of the effect of tenure and weeks unemployed on wages. Do it also for quantiles 25, 50 and 75. Can you test if the effect of tenure is different at the 25 or 75 quantile?
- Estimate now a kernel regression of the same effect.
- Estimate different logit model on the probability of being married. Keep the one with the higher likelihood value. Compute the margins of race and plot the marginal effects. You might want to take a look at the command marginsplot.