University of Bamberg

Migration and Labor Market

STATA II:

An Introduction into Panel Regression Models

Bamberg, June, 15, 2020

Dr. Ehsan Vallizadeh

Department of Economics – Integration of European Labor Markets

Last Meeting

1 Generating dummy variables

- Gen command, replace command
- Advanced generation (forvalues 1/3 {….})

2 Organise your work with globals

- glo name varlist
- $name

3 Summary statistics

- sum varlist

Last Meeting

University of Bamberg

Making Graphs

- One Y-axis:
- graph twoway line var1 var2 if ed == 1 & ex ==1

- Two Y-axes:
- graph twoway line (var1 var2) (var3 var2, yaxis(2)) if…

- Scatter plots:
- graph scatter var1 var2 if …

- Scatter plots with regression line:
- graph scatter var1 var2 || lfit var1 var2 if …

Today’s Meeting

University of Bamberg

1 Discussing your ‘homework’
- Generating the data sets and dummy variable
- Descriptive statistics
- Graphs
- Problems?

2 Introduction into regressions commands in STATA

3 Discussing the next steps

What is a linear regression model?

University of Bamberg

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki} + \varepsilon_i$$

The general econometric model:

y_i	indicates the dependent (or: endogenous) variable
$x_{1i,ki}$	exogenous variable, explaining the independent variable
β_0	constant or the y-axis intercept (if x = 0)
$\beta_{1,2,k}$	regression coefficient or parameter of regression
ε_i	residual, disturbance term.

The OLS estimator

- The **ordinary least square (OLS)** estimator minimizes the squared deviations from the linear regression line
- Under the assumption that (i) the **error term** is **normally distributed** and has an **expected mean value of zero**, and (ii) the **variance** of the **error term** is **constant** and limited, the OLS estimator delivers **unbiased** results.
- Depending on the sample size, you can draw inference on the **total population**
- For the example, you can use the standard errors or t-statistics of the coefficients of the explanatory variables to test the **null-hypothesis** whether the **estimated coefficient** is zero

The OLS estimator: problems

- **Endogeneity (simultaneous equation bias):** if your explanatory variables are not truly exogenous but correlated with the explanatory variable, the mean value of the error term isn't zero and you obtain biased estimates.
  - Solutions: Instrumental Variable (IV) estimation, natural experiments
- **Omitted variable bias:** If you have omitted relevant explanatory variables (i) the estimates of the remaining coefficients might be biased and (ii) the error term might be not zero.

The OLS estimator: problems (cont.)

- **Multicollinearity:** If your explanatory variables are correlated, you may have problems to identify the true correlation coefficient for the individual regressors, in particular when your sample is small. Is less serious.

- **Heteroscedasticity:** If the variance is not constant across groups in your sample (e.g. education groups) the standard errors are not properly estimated and the coefficients might be biased as well.
  Solutions: Robust Standard Error estimates, Generalized Least Square (GLS) estimation, and others.

The OLS estimator: problems (cont.)

- **Contemporary correlation of the error term:** If your error term across the panels is contemporaneously correlated, e.g. due to simultaneous time shocks, you obtain biased estimates
- Solutions: GLS-, Generalized Methods of Moments estimators

- **And many, many others ….**

Regression Analysis with STATA

University of Bamberg

General Syntax

- The standard OLS regression syntax in STATA is:
  - **regress depvar [list of indepvar] [if], [options]**

Example:

- Regress log wage on migration share, controlling for education, working experience and time
  - **regress ln_wqjt m_qjt $D_i $D_j $D_t**

Borjas (2003)-Modell in STATA Syntax

$$y_{ijt} = \theta p_{ijt} + s_i + x_j + \pi_t + (s_i \times x_j) + (s_i \times \pi_t) + (x_j \times \pi_t) + \varphi_{ijt},$$

This model in STATA Syntax:

```
regress ln_wqjt  m_qjt $D_i $D_j $D_t $D_ed_ex $D_ed_t $D_ex_t
```

where

- ln_wqjt: dependent variable (log wage)
- mqjt: migration share in educatipn-experience cell
- $D_i: global for education dummies
- $D_j: global for experience dummies
- $D_t: global for time dummies
- $D_ed_ex: global for interaction education-experience dummies
- $D_ed_t : global for education-time interaction dummies
- $D_ex_t: global for experience-time interaction dummies

STATA: Running a regression model

University of Bamberg

```
***********************************************
***********  Simple Regression  *************
***********************************************

regress ln_wqjt mqjt $D_i $D_j $D_t
```

Globals “$“

Regression command

Dependent variable

Independent variables

STATA: Output a regression model

University of Bamberg

```
. regress ln_wqjt mqjt $D_i $D_j $D_t

      Source |       SS       df       MS              Number of obs =     204
-------------+------------------------------           F( 22,   181) =  798.55
       Model |  21.7438431    22  .988356504           Prob > F      =  0.0000
    Residual |  .224021026   181  .001237685           R-squared     =  0.9898
-------------+------------------------------           Adj R-squared =  0.9886
       Total |  21.9678641   203  .108216079           Root MSE      =  .03518

------------------------------------------------------------------------------
     ln_wqjt |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        mqjt |   .3158438   .0501982     6.29   0.000     .2167948    .4148927
       Ded_2 |   .2226291   .0133158    16.72   0.000     .1963548    .2489033
       Ded_3 |   .7617929   .0141016    54.02   0.000     .7339682    .7896176
       Dex_2 |   .1942821   .0070443    27.58   0.000     .1803825    .2081816
       Dex_3 |   .2962792   .0074409    39.82   0.000     .2815972    .3109612
       Dex_4 |   .3591109   .0072256    49.70   0.000     .3448537    .3733681
       year2 |   .0144838   .0143728     1.01   0.315    -.0138761    .0428437
       year3 |   .0063602   .0143947     0.44   0.659    -.0220429    .0347632
       year4 |   .0179271   .0144253     1.24   0.216    -.0105362    .0463904
       year5 |    .013435   .0144442     0.93   0.354    -.0150657    .0419358
       year6 |   .0059516    .014444     0.41   0.681    -.0225488    .0344519
       year7 |   .0022585   .0144567     0.16   0.876    -.0262668    .0307839
       year8 |    .002628   .0144624     0.18   0.856    -.0259086    .0311646
       year9 |   .0011193   .0144725     0.08   0.938    -.0274373    .0296759
      year10 |   .0078542   .0144864     0.54   0.588    -.0207297     .036438
      year11 |   .0215036   .0144755     1.49   0.139    -.0070589     .050066
      year12 |   .0185784   .0144583     1.28   0.200      -.00995    .0471068
      year13 |   .0131597   .0144415     0.91   0.363    -.0153357    .0416551
      year14 |  -.0159386   .0144263    -1.10   0.271     -.044404    .0125268
      year15 |  -.0394645   .0144193    -2.74   0.007    -.0679161   -.0110129
      year16 |  -.0524586   .0144216    -3.64   0.000    -.0809146   -.0240025
      year17 |  -.0586402   .0144174    -4.07   0.000     -.087088   -.0301924
       _cons |   3.941102   .0172898   227.94   0.000     3.906986    3.975217
------------------------------------------------------------------------------
```

STATA: How to interpret the output of a regression?

variance of model | **degrees of freedom** | **Mean squared** | **ANOVA table**

```
. reg ln_wqkt mqkt

      Source |       SS       df       MS              Number of obs =     800
-------------+------------------------------           F(  1,   798) =  212.53
       Model |  23.4146717     1  23.4146717           Prob > F      =  0.0000
    Residual |  87.9145738   798  .110168639           R-squared     =  0.2103
-------------+------------------------------           Adj R-squared =  0.2093
       Total |  111.329246   799  .139335727           Root MSE      =  .33192

------------------------------------------------------------------------------
     ln_wqkt |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        mqkt |  -1.369118    .093913   -14.58   0.000    -1.553464   -1.184772
       _cons |   4.706176    .017403   270.42   0.000     4.672015    4.740337
------------------------------------------------------------------------------
```

Overall Model Fit

1. Observations
2. fit of the model
3. F-Test
4. R-squared
5. adjusted R-squared
6. Root Mean Standard Error

β_1

β_0

analysis of significance levels

95% confidence interval

STATA: Instrumental Variable (IV) Model

- Instrumental variables (IVs) are variables which are correlated with the (potentially) endogenous explanatory variable, but not the dependent variable and, hence, not the error term
- Syntax of a IV model in stata:

  ivregress depvar indepvar

(endogenous variable = iv)

- Example:

  ivregress ln_wqjt $D_i $D_j … (mqjt = iv1 iv2 …)

STATA: Panel Model

- Very often you use panel models, i.e. models which have a group and time series dimension
- There exist special estimators for this, e.g. fixed or random effects models
  - A **fixed effects** model is a model where you have a fixed (constant) effect for each individual/group. This is equivalent to a dummy variable for each group
  - A **random effects** model is a model where you have a random effect for each individual group, which is based on assumptions on the distribution of individual effects

Recourse: Random Effects vs. Fixed Effects

University of Bamberg

Consider a simple linear model:

unobserved heterogeneity

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \alpha_i + u_{it}$$

- **Assumption:** $\text{Cov}(\alpha_i, X_{it}) = 0$
- If assumption holds:
  - $\hat{\beta}_{RE}, \hat{\beta}_{FE}$ *are consistent*
  - $se(\hat{\beta}_{RE}) < se(\hat{\beta}_{FE})$
- If assumption does not hold
  - $\hat{\beta}_{FE}$ is solely consistent ($\hat{\beta}_{RE}$ no longer consistent)

- Hausman test – RE vs FE

STATA: Panel Model (*cont.*)

University of Bamberg

Preparation for Panel Models:

- For running panel models STATA needs to identify the group(individual) and time series dimension
- Therefore you need an index for each group and an index for each time period
- Then use the **tsset** command to organize you dataset as a panel data set
- Syntax:
  - **tsset index year**
- where **index** is the group/individual index and **year** the time index

Preparation: Running the ***tsset*** command

```
*********************************************************
************ Panel Regression **************
*********************************************************

tsset index year

```

```
. do "C:\Users\

. tsset index year
       panel variable:  index (strongly balanced)
        time variable:  year, 1992 to 2008
                delta:  1 unit

.
end of do-file
```

STATA: Panel Model (*cont.*)

- Then you can use panel estimators, e.g. the **xtreg** estimator
- Syntax
  - **xtreg depvar [list of indepvar] [if], [options]**
  - **xtreg ln_wqjt m_qjt, fe**
- i.e. in the example we run a simple fixed effects panel regression model which is equivalent to include a dummy variable for each group (in this case education-experience group),
  <u>but:</u> no time or interaction dummies

STATA: Panel Model (*cont.*)

University of Bamberg

Running a Panel Regression: command

```
*************************************************
*********** Panel Regression *************
*************************************************

tsset index year

xtreg ln_wqjt mqjt, fe
```

STATA: Panel Regression Output

```
. xtreg ln_wqjt mqjt, fe

Fixed-effects (within) regression               Number of obs      =       204
Group variable: index                           Number of groups   =        12

R-sq:  within  = 0.1569                         Obs per group: min =        17
       between = 0.2948                                        avg =      17.0
       overall = 0.2471                                        max =        17

                                                F(1,191)           =     35.54
corr(u_i, Xb)  = -0.6051                        Prob > F           =    0.0000

------------------------------------------------------------------------------
     ln_wqjt |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        mqjt |   .4030928   .0676135     5.96   0.000     .2697277     .536458
       _cons |   4.466355    .010334   432.20   0.000     4.445971    4.486738
-------------+----------------------------------------------------------------
     sigma_u |  .37037033
     sigma_e |  .03825941
         rho |  .98944169   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0:     F(11, 191) =   1009.83            Prob > F = 0.0000

.
end of do-file
```

STATA: Panel Regression Model

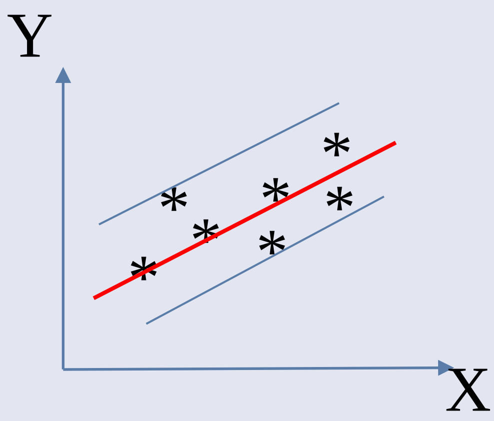
University of Bamberg

- There are other features of panel estimators which are helpful
- **Heteroscedasticity:**
  - Heteroscedasticity: the variance is not constant, but varies across groups
  - **xtpcse , p(h)** corrects for heteroscedastic standard errors
  - **xtgls , p(h)** corrects coefficient and standard errors for panel heteroscedasticity, but may produce biased results depending on the group and time dimension of the panel
  - *Note:* p(h) after the comma is a so-called option in the STATA syntax

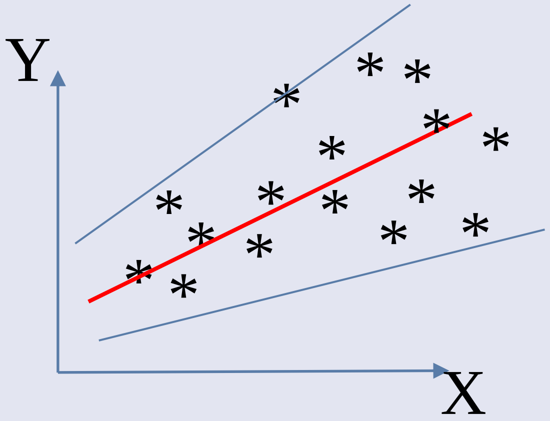
Recourse: Homo- vs Heteroscedasticity

- Variance of errors given independent variable is constant: $Var(u_i|X_i) = \sigma$
- Variance of errors given independent variable varies: $Var(u_i|X_i) = f(X_i)$

Homoscedasticity



Heteroscedasticity



STATA: Panel Regression Model

- **Contemporary correlation across cross-sections**
  - Contemporary correlation: the error terms are contemporarily correlated across cross-sections, e.g. due to macroeconomic disturbances
  - **xtgls , p(c)** corrects for contemporary correlation and panel heteroscedasticity, but may produce biased results depending on the group and time dimension of the panel.

Plan for the presentation

- **Introduction into your country and outline of research question**
- **[Very brief sketch of state of research]**
- **Theoretical hypothesis/hypotheses**
- **Description of data**
  - Data sources
  - Descriptive Statistics
  - Graphical presentation of descriptive evidence
- **Econometric findings**
  - Regression model
  - Regression results
  - Interpretation
- **Conclusions**

Next Meting: June 29

University of Bamberg

- Begin: 12:00 – 14:00
- Topic: Descriptive data analysis

THANKS FOR YOUR ATTENTION!