```
* Exercise # 10 (below is #12, on Exercise #11 we downloaded panel data and
learned how to upload it in STATA)

*** Steps in Empirical Study

**1. sum and tabs
**2. covariance matrix and graphs
**3. choice of methodology and model, arranging variables
**4. estimation
**5. output tables for paper

**************************
*1. sums and tabs

sum totwage educat exper marst sex age
*** observing the std.dev. we want to exclude extremes

tab totwage
*** based on distribution we exclude all above 5,510
drop if totwage>5510

*** we noticed extreme min in variable age after sum command
tab age
*** in this case we need more information about legal age limits
*** those starting to work and limits for retirement (Ministry of Labor)
*** here we assume that all below 14 and above 65 can not work.
drop if age<14
drop if age>65

*2. covariance matrix and graphs

*** covariance matrix needed for highly correlated independent variables.
corr totwage educat exper marst sex age
*** we want to see significant correlations
pwcorr totwage educat exper marst sex age, sig

*** graphs
*** graphs are usually done between:
*** 1. dependent var. and independent var. of interest (ref. to research
question)
twoway scatter totwage educat
*** 2. highly correlated variables (these are not reported in final papers)
twoway scatter exper age

*** due to high correlation we exclude age and keep exper

*3. choice of methodology and model, arranging variables

*** choice of methodology: look at the dataset, what is observation?
**** do we have time dimension? (is it cross-section or panel data? time
series?)
**** in our case we have cross-section data therefore we use OLS (ordinary
least square)
**** (side-note: if panel data we use fixed effects! (also random and first
differencing))

*** choice of model:
```

```
**** dependent variable: wage - due to non-linearity problem we need to use
log of wage.
gen logtotwage = log(totwage)
**** independent variables: 1. experience - due to diminishing marginal
returns we use
*************************** quadratic variable.
gen exper2 = exper^2
*************************** 2. marital status(0,1,2)- due to interpretation
of the effect of
***************************marital status on wage, we need dummy (0,1)
gen married =0
replace married =1 if marst==1
*************************** 3.  sex (male=1, female=0)- if we choose men (1
male, 0 female) then
rename sex men
**************************** if we choose females then (1 female, 0 male)
gen female=0
replace female = 1 if sex==0

*** additional step: labeling variables

label var logtotwage "Log(Wage)"
label var educat "Education Level"
label var exper "Experience"
label var married "Married"


*4. estimation

reg logtotwage educat exper exper2 married men

*** we need to cluster by state because of standard errors that affect t-
values
reg logtotwage educat exper exper2 married men, cluster(state)


* 5. tables
*** refer to notes from exercise with tables.
*** classifying by gender or by state, etc.

quietly reg logtotwage educat exper exper2 married if men==1 , cluster(state)
estimates store MEN

quietly reg logtotwage educat exper exper2 married if men==0 , cluster(state)
estimates store WOMEN

*** tables usually contain beta parameters and p-values indicating
significance.
estimates table MEN WOMEN, b(%5.3f) p(%5.3f)

*** 5.1. Interpretation
***** Education. Increasing education by one level wage increases 54% for men
and 57% for women. (% due to logtotwage).
***** Experience. Increasing experience by one year increases wage by app. 5%
for men and 6% for women.
***************** With diminishing marginal returns the highest return on
experience is with 24 years of experience for men, and 30 for women.
```

```
* First order condition with parameter values of estimated model
** men
display 0.048/0.002
** women
display 0.061/0.002

****** Marital status. Being married decreases wage for men for 10%. For
women we have no significant result. Being married does not affect wage for
women.
****** parameter beta zero is higher for men relative to women, indicating
higher base wage for men.

****** Further research steps. Since return of education and experience is
higher for women which is in the clash with standard literature.
******** Next step is to find the cause of such difference: type of industry
where men work more often than women (construction sights).
******** This is incorrect assumption because the regression is divided for
men and women. Women are compared with women based on education, experience,
etc.
gen bluec=0
replace bluec=1 if wknic==0|wknic==1|wknic==2|wknic==4

reg logtotwage educat exper exper2 married bluec men , cluster(state)
**** in this case men are compared with women. We see this in the parameter
value of dummy variable men (ref.graph with two different bases, beta and
sigma base)



** Exercise #12

set mem 100m

* estimating panel data models: fixed effect, random and first differencing

*****************************************************************

*** Econometric model

*** country's' FDI inflow due to development and school enrollment?

**** choice of variables  fdiflow_in corru goveff polstab regquality
ruleoflaw voicacc diss enrollmentrate under14

**** next steps: sums and tabs, graphs, estimations, tables

**** Fixed effects

iis unit_id
xtreg  fdiflow_in corru goveff polstab regquality ruleoflaw voicacc diss
enrollmentrate under14, fe

* which problem exists in our model?
*** high correlation between independent variables

pwcorr  fdiflow_in corru goveff polstab regquality ruleoflaw voicacc diss
enrollmentrate under14, sig
```

```
*** from theory and from literature (and dataset) corru goveff polstab
regquality ruleoflaw voicacc are variables from the same group.
**** therefore we need to choose one or two which are the least correlated.
pwcorr   corru goveff polstab regquality ruleoflaw voicacc, sig

* Fixed effect model with political stability
iis unit_id
xtreg  fdiflow_in  polstab  diss enrollmentrate under14, fe

tab polstab

*** problem with data is a lot of unobserved per country

sum fdiflow_in
sum polstab

* exclude all below 1996
drop if year<1996

tab fdiflow_in
drop if fdiflow_in>30630
drop if fdiflow_in<-3000

** when we re-estimate the model without outliers our significance levels
drop, indicating that extremes caused previous results.
*** our analysis is not robust, therefore our model is not correct. We need
to add more variables or change hypothesis.

** Random Effects
*** the model in STATA is the same as FE, just option istead of fe we use re

xtreg  fdiflow_in  polstab  diss enrollmentrate under14, re
```