

Introduction to Stata - Exercise

Matthias Krapf
University of Basel

September 23, 2021

I. The Data

This exercise deals with a manipulated data set from the Stata online source that contains information about individuals' wages, schooling, etc. The data set has an unbalanced panel structure ranging from 1968 until 1988, resulting in 28,534 single observations. During the following exercise, you should not worry about the panel structure. It contains the following variables:

idcode	individual id code
year	year of interview
birthyr	year of birth
race	1=white, 2=black, 3=other
msp	1 if married and spouse present
grade	current grade completed
collgrad	1 if college graduate
south	1 if living in the south
indcode	industry of employment
unionfee	monthly union membership fee
tenure	job tenure in years
hours	usual hours worked
wage	wage

This data set has been artificially manipulated. Hence, do not use it for any research!

II. The Task

a) Prepare the project

- Create project directory.
- Download the file *rawdata.xlsx*
- Open it with Excel, inspect it carefully and save it as *rawdata.csv* in your working directory.
- Create and save a do-file under your desired directory.
- Chose that directory as your working directory.
- Create a log-file

Helpful commands: <i>clear; set mem; clear; set more off; cd; capture log close; log</i>
--

b) Import Data

- Import *rawdata.csv* into Stata, and inspect the data carefully.
- Save the data as *myfirstdata.dta* in your working directory.

Helpful commands: <i>insheet; save; browse; describe; edit</i>
--

c) Data Manipulation

- Generate a new variable called *logwage* which is the natural logarithm of the wage.
- Generate a new variable called *age* that contains the age of the individuals.
- Generate a new variable called *black* that equals 1 for all black and zero else.
- Generate a new variable called *union* that equals 1 for all individuals with a union membership (*payments*>0) and zero else (Take care of the missing values!!!).
- Generate a new variable called *indmwage* that contains the average wage of the industry an individual works in.
- Delete all individuals that have a wage above 50\$ from the data.
- Delete the variable *birthyr* from the data.

Helpful commands: <i>generate; replace; egen; drop; keep; sort; recode</i>
--

d) Descriptive Statistics

- Calculate the mean of the variable *wage*.
- Calculate the mean, the variance, as well as the 5 and 95 percentiles of the variable *wage*.
- Calculate all correlations between the variables *wage*, *age*, *grade*, *tenure*, and *union*.
- So far, Stata does not know what the variable *union* means. Verify this by typing *describe union* and *tabulate union*. Now label first the variable as “Union Membership” and then the values as “No Member” if *union* equals zero and “Union Member” if *union* equals 1. Now repeat *describe union* and *tabulate union*.
- Calculate the mean of *wage* for union members and non-members separately.
- Test whether the difference in the mean of *wage* for union members and non-members is significantly different from zero.

Helpful commands: *summarize*; *correlate*; *describe*; *tabulate*; *label*; *ttest*

e) Graphs

- Plot a histogram of the variable *wage*.
- Plot a scatter plot with *logwage* on the y-axis and *exper* on the x-axis.

Helpful commands: *histogram*; *graph*; *scatter*

f) Regressions

- Perform an OLS-Regression of the following model:

$$\log wage_{i,t} = \beta_0 + \beta_1 age_{i,t} + \beta_2 age_{i,t}^2 + \beta_3 exper_{i,t} + \beta_4 black_{i,t} + \beta_5 tenure_{i,t} + \beta_6 union_{i,t} + u_{i,t}$$

- Repeat the first regression using robust standard errors.
- Test if *age* has a significant effect on *wage* (*age* shows up twice in the equation!).

Helpful commands: *regress*; *predict*; *test*