

## Semester Project SYA 6305

See the 'Data Management in Stata' document.

### Part I

#### **Hypotheses**

- Present appropriate hypotheses to guide your project.

#### **Project folder**

- Keep all of the data and files for this, or any, project in its own exclusive folder.
- Save a backup copy of the original ('master') data set.
  - *Never save changes made to the original ('master') data set, but rather make the changes by means of do-files and save the changes as new, working versions of the data sets (see below on nested do-files).*

#### **Log file, command file, and do-file**

- Make a log file, command file, and do-file for every step.
  - Nest the do-files: nested, shorter do-files are more manageable than long do-files (see 'Nested Do-Files').
- *Every step you do must be documented and replicable.*

***If helpful, restrict the working version of the data set to the variables that are relevant to your project (including variables that will help you explore the data).***

- Log, command, and do files are necessary for this step, as well as for all of the other steps.
- help keep; help drop

***The Stata commands 'inspect' and 'codebook' provide excellent overviews of potential problems.***

### Part II

#### ▪ **Key questions to answer about the data**

- Who funded the collection of the data, and who collected the data and for what objectives? Do your answers suggest any possible biases in the data?
- Who do the data represent, how were the data collected, and how many observations are there? In view of these questions, to what extent are the data adequate or not for your study?
- What are the data's variables? In view of your study's purpose, are potentially important variables not included? If so, what variables are absent, and why are they important?
- How are each of the variables defined and measured? To what extent are the definitions and measurements valid or not, and to what extent are the measurements likely to be reliable (if you have some way of judging) (see, e.g., Babbie, *The Practice of Social Research*, on 'validity' and 'reliability')? Insofar as the operationalization could be improved, are there mitigating circumstances, and if there are, what are they? Are there ways that you might be able to improve the operationalization of the variables? What are the implications of the limitations for the study, as well as for the social construction of reality more generally?

- Do the observed values of the variables make sense: minimum and maximum values; variability and distribution of values by categories (e.g., distribution of years of education or income with regard to other evidence); and means/medians?
- Do the relations among the variables make sense? (E.g., do less educated people tend to earn less, rather than more? Are children coded as “adults” or vice versa?)
- Are there missing values, and if so, are they distributed randomly or according to patterns? If there are patterns, how do they potentially bias the data?

### Part III

#### **Data preliminaries & description**

- Label and, if appropriate, date the data set (help label > label data).
- Where necessary or helpful, rename and label the variables (help label > label variables; findit renvars), and for categorical variables label the sub-categories ('value labels', e.g., 1=female 0=male; help label > label define).
- Create an ID-variable, and name and label it appropriately (such as 'id,' 'obs,' 'obsnum', etc.).
- Order the variables in some way that is helpful to you (help order; help aorder; findit placevar).

```
. describe
```

(check variable names & labels, and value labels; check for string variables, & use 'encode x, gen(xnew)' to destring them)

```
. label list
```

(help label > label list)

```
. labelbook, problems
```

(help labelbook)

```
. describe
```

#### **Check for duplicate id's &, after sorting by id #, check that the number of id #'s isn't fewer than the number of observations**

```
. sort id
```

(or whatever you named the id-variable)

```
. duplicates report id
```

(help duplicates)

```
    . duplicates list id
```

```
        .duplicates drop in ...
```

```
. sort id
```

```
. list id
```

(Check that the number of id numbers isn't fewer than the number of observation numbers, which would be a sign of missing observations.)

#### **Scan the variables for an overview of potential problems**

```
. tab cv1-cv5
```

```
. su qv1-qv5, d
```

(or univar qv1-qv5; or tabstat qv1-qv5, stats(n min max mean sd median p25 p75))

- Do any of the summary values (#obs; min, max; mean, sd; median, quartiles) look questionable in terms of what you expect to find (such as impossible

negative or positive values)? Is there any apparent pattern to any such problems?

### **Check the coding of the categorical variables**

```
assert id>=1 & id<=200 (help assert, help codebook)
assert female==0 | female==1 if female<.
codebook female
assert race>=1 & race<=4 if race<.
codebook race
assert ses>=1 & ses<=3 if ses<.
codebook ses
```

### **Check the coding of the quantitative variables**

```
assert read>=20 & read<=80 if read<.
codebook read
```

### **Check (in more depth than before) the distributions for the categorical & quantitative variables**

#### *Categorical*

```
. tab1 cv1 cv2 cv3 cv4 cv5
▪ Do the minimums, maximums, and the distributions of observations by category make sense?
```

#### *Quantitative*

```
. summarize qv1 qv2 qv3 qv4 qv5, detail (or univar qv1...; or tabstat qv1...,
Do the distributions (minimums, maximums, stats(n min max mean sd median
means/medians, sds/quartiles) make sense p25 p75))
in view of other evidence and your knowledge
of the topic? Check that there are no extremely
unusual or impossible values, such as negative
achievement test scores or scores that are higher
than the maximum possible.
```

```
. for varlist qv1-qv5: hist X, discrete
```

- Check that the distributions of observations for the quantitative variables make sense. E.g., for years of education in the U.S., check that the proportions of observations are appropriate for high school, associate's degree, and college.

### **Check for missing values & their patterns**

```
. nmissing (findit nmissing)
. mvpat v1 v2 v3 v4 v5, skip (findit mvpat)

. egen mvals=rmiss(_all) (Do appropriate bivariate exploration of
. tab mvals missingness patterns, such as by gender,
class, ethnicity, age category, marital
status)
```

### **Explore the categorical variables**

```
. tab1 cv1 cv2 cv3 cv4
. for varlist cv1 cv2: ci X, binomial wilson (for binary variables only)
```

. tab2 cv1 cv2 cv3 cv4, chi2 (cross-tabulations)  
. bys cv1: tab cv2 cv3, chi2

### **Explore the quantitative variables**

. summarize qv1 qv2 qv3 qv4, detail (or univar or tabstat)  
. for varlist qv1-qv4: ci X  
. ciplot qv1-qv4 (horizontal option)  
. for varlist qv1-qv4: graph box X \ more  
. for varlist qv1-qv4: hist X, norm \ more (or kdensity, qnorm, etc.)  
▪ It's probably less trouble to inspect these graphs on the fly, rather than save each one to inspect later.

. twoway scatter qv1 qv2, ml(id) || qfit qv1 qv2  
. gr save scatter1, replace  
. corr qv1 qv2 (if necessary: spearman qv1 qv2)  
. reg qv1 qv2 (if there's sufficient linearity)

### **Explore the quantitative variables by the categorical variables**

. tabstat qv1-qv4, by(cv1) stats(mean sd median p25 p75 min max) format(%9.2f)  
. gr box qv1-qv4, over(cv1, total) \ more (for each appropriate relationship)  
. gr save box1, replace

. ciplot qv1-qv4, over(cv1) (horizontal option)  
. gr save ciplot1, replace  
. for varlist qv1-qv4: ttest X, by(cv1) \ more (for each appropriate relationship; unequal option)  
. for varlist qv1-qv4: oneway X cv4, tab bonf (if cat var has 3+ categories; help oneway; Bartlett's test must be insignificant)

. twoway scatter qv1 qv2, by(cv1) || qfit qv1 qv2  
. gr save scatter1bycv1, replace  
. bys cv1: corr qv1 qv2  
. bys cv1: reg qv1 qv2

### **Multiple regression analysis**

▪ Use a new do-file, nested within the data exploratory do-file(s).

```
* model 1  
. reg y qv1 qv2  
. eststo  
. linktest (should test insignificant)  
. estat ovtest (should test insignificant)  
. rvfplot, yline(0)  
. gr save rvfplot1, replace
```

```
* model 2  
. reg y qv1 qv2 qv3 cv1  
. eststo  
. linktest  
. estat ovtest  
. rvfplot, yline(0)  
. gr save rvfplot2, replace
```

```

* model 3
. reg y qv1 qv2 qv3 cv1 cv2 cv3
. eststo
. linktest
. estat ovtest
. rvfplot, yline(0)
. gr save rvfplot3, replace
. estat vif
. predict rstu, rstu
. hist rstu, norm
. gr save hist3, replace
. estat hettest, mt(bonf)
. rvpplot, yline(0) (for each quantitative explanatory variable)
. gr save rvpplot3, replace
. lvr2plot, ml(id)
. gr save lvr2plot3, replace

```

\* Regression tables

```

esttab, se b(%9.2f) starlevels(+ .10 * .05 ** .01) r2(%9.2f) ar2(%9.2f) aic bic
nodepvars nomtitles title("OLS Models") addnotes("Note: Put your notes here.")
replace
(Displays in Stata results window)

```

```

esttab using project.rtf, se b(%9.2f) starlevels(+ .10 * .05 ** .01) r2(%9.2f)
ar2(%9.2f) aic bic nodepvars nomtitles title("OLS Models") addnotes("Note: Put your
notes here.") replace
(Click blue-link to display in MS-Word. See 'Editing esttab and outreg2 tables in MS-
Word'; and see 'help esttab,' including for Excel-output option if you prefer.)

```

***Close/exit logs & do-file***