

### Assignment 4

**Due Monday Oct. 16 6pm. Hand in online version (QuestromTools→Assignments and a hard copy of this assignment with a revised Project Description, either handed in during class or put under Prof. Kahn's door (518C) or in her mailbox (531)**

(Worth 1 ½ points if completed and handed in on time; late assignments lose 5% per day)

**Part A: Questions on your dependent variable** (if you have > 1, choose the most important one):

1. If you have a numeric dependent variable, create a histogram of your dependent variable in Stata (**histogram** varname). Save it in pdf or graphic form. Insert the histogram here as an object.

If you have a categorical dependent variable, tabulate it with the Stata command:

**tab variablename, missing.**

Then **copy and paste here** the related command line and the Stata results from the Results panel (using ctl-V, ctl-C)

2. What do you learn from this histogram or tabulation? Answer in 1-4 sentences.
3. If you have a numeric dependent variable, get descriptive statistics for your (key) dependent variable in Stata by using **summarize variablename, detail**. **Copy and paste here** the related command line and resulting output from the Stata results panel (formatting as Courier New 9 point).
4. If you have a 2-category dependent variable, make it into a single dummy variable, making sure that any missing values are left as missing. For instance, you could type in Stata: **gen newvar = oldvar== "yes" if oldvar!=**. Then **summarize varname, detail**. **Copy and paste here** the related command lines and resulting output from the Stata results panel (formatting as Courier New 9 point).

Important tip: What if your dependent variable is categorical with more than 2 categories (e.g. happiness, with 1=very happy 2=somewhat happy 3=neither happy nor unhappy, 4=somewhat unhappy, 5=very unhappy)? At this point you have several choices: (1) You can choose to make a single dummy variable. For instance, you might make the variable **happy** (either very or somewhat) v. other. Or, you can choose to make a single dummy variable for **unhappy** (either very or somewhat) v. others. Or something else.... To decide which, think about where the biggest contrast is. However, IF you actually believe that the difference between a 1 and a 2 *is* similar to the difference between a 2 and a 3, etc., you might choose to use it as a numeric variable. However, defend this assumption.

5. What important things do you learn about the distribution of your dependent variable from these descriptive statistics? Answer in 1-4 sentences.
6. Based on this evidence, are there any observations with values that seem like mistakes? Do you think we should drop these observations or correct the mistake? Explain and drop or correct if appropriate.
7. (For numeric variables only) Based on this evidence, is your dependent variable very skewed, and particularly are there any extreme outliers? If so, do you think we should top-code these values (or use logs etc.)? Explain why. Then top-code or change into logs if appropriate, and **copy and paste here** the related command lines and resulting output from Stata (formatting as Courier New 9 point).

**Part B: Questions on your key explanatory variable** (if you have > 1, choose the most important one):

8. If it is a numeric variable, create a histogram of it in Stata (**histogram** varname). Save it in pdf or graphic form. Insert the histogram here as an object.

9. If it is a categorical variable, tabulate it with the Stata command: **tab variablename, missing**. Then copy and paste here the related command line and the Stata results from the results panel (using `ctl-V`, `ctl-C`) (formatting as Courier New 9 point).

10. If it is a numeric variable, get descriptive statistics for it in Stata by using **summarize variablename, detail**. Copy and paste here the related command line and resulting output from the Stata results panel (formatting as Courier New 9 point).

If it is categorical, make it into a single dummy variable. (See important tip above if there are more than 2 categories.). Make sure that any missing values are left as missing. For instance, you could type in Stata:

```
gen newvar = oldvar== "yes" if oldvar!=.  
summarize varname, detail
```

11. What important things do you learn about the distribution of your key explanatory variable from these descriptive statistics? Answer in 1-4 sentences.

12. Based on this evidence, are there any observations with values that seem like mistakes? Do you think we should drop these observations or correct the mistake? Explain, and drop or correct if appropriate.

13. Based on this evidence, is your explanatory variable very skewed, and particularly are there any extreme outliers? If so, do you think we should top-code these values (or use logs etc.)? Explain (and do it). Then top-code or change into logs if appropriate, and **copy and paste** here the related command lines and resulting output from Stata (formatting as Courier New 9 point).

### Part C: Questions on Correlation:

14. Correlate all variables you plan to use. Copy and paste here the related command line and Stata response/output from the Results panel (formatting as Courier New 9 point).

15. What important things do you learn about the relationship between your dependent variable(s) and your key explanatory variable(s) from this correlation table? Answer in 1-4 sentences.

### Part D: Simple Regression:

15. Run a simple regression of your key dependent variable on your key explanatory variable (or one of them, if you have several.) **Copy and paste here** the related command line and the Stata regression output from the results panel (formatting as Courier New 9 point).

16. What important things do you learn about the relationship between your key dependent and explanatory variables from this regression? Answer in 3-8 sentences. In your answer, include a discussion of the explanatory variable's coefficient, its t-statistic and its confidence interval.

### Part E. Project Description

17. Finally, update the Project Description Questions 1-5 if you have changed or added any part of it. Then, fill in the Codebook (question 6) in your Project Description. Hand this in.

**Be sure to save this dataset under a new name for yourself, since you have changed so many variables.**