

EC204 Empirical Economics 2 Jimin Oh

HW1 – Part 2 STATA Homework

This Stata problem set consists of 15 questions. Total: 100 points.

Questions 1-10 covers Chp 4 Inference. Use the data for the housing price (hprice.dta) to answer the questions.

Questions 11-15 covers Chp 3 Omitted Variable Bias. Use the data for the wage and intelligence (WAGE2.dta) to answer the questions.

You are required to submit **two** files on the Blackboard course website under the Assignment folder (HW1: Stata Part):

- 1) **A do file (.do)**: this should contain all the commands and comments for the results.
- 2) **A Word file (.docx)**: this should contain all the commands, comments, and the results from running Stata (regression tables, figures, test results, etc.).

Note: when you make comments in the do file, you should put the asterisk mark (*) in front of the line.

Questions 1-10: Answer the following questions by using the dataset: hprice.dta. We would like to discover determinants of the selling price of houses (measured in \$1000s).

1. We first start by clearing the dataset. Drop the observations if the year is 1981. **5 points.** (Please make sure to include the Stata command in the Word file.)
2. Present a table of summary statistics for the following variables: price (measured in \$1000s), rooms, area (measured in sqft), land (measured in sqft), baths, age. **5 points.** (Please make sure to include the Stata command and a table of summary statistics in the Word file.)

3. Draw a histogram of the selling price of houses (price). Comment if the distribution is symmetric or skewed. **5 points.** (Please make sure to include the Stata command, figure, your comments in the Word file.)
4. Now we would like to visualize the relationship between two variables by using a scatter plot. Draw a scatter plot between the selling price (price) and the size of the house (area) where the x-axis variable is area and the y-axis variable is price. Comment on the relationship between the two variables after looking at the scatter plot. **5 points.** (Please make sure to include the command, figure, and comments in the Word file.)
5. Draw/Add a regression line (fitted line) on the scatter plot between price and area. **5 points.** (Please make sure to include both command and figure in the Word file.)
6. Run a regression of price on age. Interpret the estimation results. What happens to the selling price if the size of the house increases by 1 sqft? Is the variable area statistically significant at 1%, 5%, and 10% level? Interpret the R-squared from the regression result. **15 points.** (Please make sure to include the Stata command, comments, and the regression table in the Word file.)
7. Now we add additional independent variables (rooms, land, baths, age) and run a regression of price on area, rooms, land, baths, and age. What happens to the selling price if there is additional one more bathroom in the house? What variables are statistically significant at the 5% level? **10 points.** (Please make sure to include the Stata command, comments, and the regression table in the Word file.)
8. Interpret the R-squared from the regression result in Question 7. Does the R-squared increase, compared to Question 6? If so, why? If not, why not? **5 points.**
9. We would like to see whether the partial effect of the number of rooms on the selling price is the same as the partial effect of the number of bathrooms on the selling price. So we set up the following hypothesis: $H_0: \beta_{rooms} = \beta_{baths}$ and $H_1: \beta_{rooms} \neq \beta_{baths}$. Are we able to reject the null at the 5% level? Show the test results by using Stata. **5 points.** (Please make sure to include the Stata command, comments, and the Stata test results in the Word file.)

10. We set up the following joint hypotheses: $H_0: \beta_{rooms} = 0, \beta_{land} = 0, \beta_{baths} = 0$ and $H_1: H_0$ is not true. Are we able to reject the null at the 5% significance level? Show the test results by using Stata. **5 points.** (Please make sure to include the Stata command, comments, and the Stata test results in the Word file.)

Questions 11-15: Answer the following questions by using the dataset: WAGE2.dta. We would like to discover the omitted variable bias in the wage equation.

11. We would like to decide the direction of the omitted variable bias by using monthly wage data. We'll use IQ as a (imperfect) proxy for the ability. Run a regression of wage on education (educ). And run a regression of wage on education and IQ. Present the regression tables with the commands. **5 points.**
12. According to the regression results in Question 11, there seems to be an omitted variable bias as the estimated coefficient on education changed "a lot" when we included the omitted variable IQ into our regression. Is this bias (seemingly) positive or negative? Why? **5 points.**
13. Calculate the change in the estimated coefficient for education in terms of standard errors. By how many standard errors did the estimated coefficient for education change when we included the variable IQ compared to the case of excluding the variable IQ? **10 points.** (Hint: use the command 'display' to compute the change on Stata.)
14. Present the table of correlation coefficients among the three variables: wage, educ, and IQ. **5 points.**

15. According to the table of correlation coefficients in Question 14, what is the **value** for the correlation coefficient between the two independent variable ($corr(x_1, x_2)$ in our lecture)? Based on the correlation table, what is our educated guess for the **sign** of the relationship between the dependent and the omitted independent variables (β_2 in our lecture)? By combining this two information, what is our educated guess for the direction of the bias? Is it positive or negative? **10 points.**