EC 339

Problem Set 2

Prof. Santetti

Fall 2022

INSTRUCTIONS: Carefully read all problems. You must submit a single R script (Section 001) and `Stata` do-file (Section 002) with your *first name* (mine would be `marcio.R` and `marcio.do`). In case you submit your files with different names, you will lose 1 point.

You can find templates for your answer scripts/do-files on `theSpring`, under the "Templates" module. Please consider using it.

I should be able to fully replicate your code to answer the questions, as well as fully understand your written interpretations to the proposed problems.

Avoid using unnecessary code in your submission files. It is totally fine to do other things by yourself that may help you better understand the data and the problems. However, for grading purposes, I am only interested in the commands and interpretations that actually answer the questions. You may keep a separate file for yourself with your additional explorations.

`Assignment due 10/26, before class`.
`Points Possible: 30`

- You have 2 weeks to complete this assignment. In accordance with our `course syllabi`, no late submissions will be accepted.
- Be honest. Don't cheat.
- As a Skidmore student, always recall your votes of academic integrity, and the **`Honor Code`** you have abided by:

> *"I hereby accept membership in the Skidmore College community and, with full realization of the responsibilities inherent in membership, do agree to adhere to honesty and integrity in all relationships, to be considerate of the rights of others, and to abide by the college regulations."*

Have fun!

Problem 1

`Anglin and Gençay (1996)` estimate several residential housing price models. You can use their data by importing the `house_prices.csv` data set into your working environment. Make sure to check out the data description `here`.

(a) After the data set is properly loaded into your work space, replicate the paper's Table III (i.e., look at this regression's output):

$$\begin{aligned} log(price_i) = \beta_0 + \beta_1 driveway_i + \beta_2 recreation_i + \beta_3 fullbase_i + \beta_4 gasheat_i + \\ + \beta_5 aircon_i + \beta_6 garage_i + \beta_7 prefer_i + \beta_8 log(lotsize_i) + \beta_9 bedrooms_i + \\ + \beta_{10} bathrooms_i + \beta_{11} stories_i + u_i \end{aligned}$$

(b) Interpret the coefficients on the following variables: *driveway*, *garage*, *lotsize*, and *bathrooms*. *Hint*: pay attention, some variables are binary, some are not.

(c) Now re-estimate (b)'s model removing the *two least significant* variables. What happens to the *goodness-of-fit* measures (i.e., R^2 and adjusted R^2) when comparing the two models?

(d) Is this model linear in parameters? Explain.

(e) Verify CLRM Assumption II, regarding the mean of the error term for part (a)'s regression model. Is this assumption satisfied?

Problem 2

For this problem, you will use the `wage.csv` data set. It contains data from the 1976 Current Population Survey. We will analyze the following wage model:

$$lwage_i = \beta_0 + \beta_1 female_i + \beta_2 educ_i + \beta_3 female_i * educ_i + \beta_4 exper_i + \beta_5 exper_i^2 + \beta_6 tenure_i + \beta_7 tenure^2 + u_i$$

`This page` describes all variables. Estimate it before answering the next questions.

(a) Does this model allow for *experience* and *tenure* to have a *diminishing* effect on wages? Explain your reasoning.

(b) Interpret the effect of *educational attainment* on wages.

(c) Run the appropriate test for the *joint* significance of the squared terms of the *experience* and *tenure* variables. Report the (*i*) test statistic, (*ii*) the p-value, and (*iii*) your inference from it.

(d) Verify CLRM Assumption VII—on the normality of the regression's error term—by running a Shapiro-Wilk test. Assume α = 5%.

(e) What is the effect of *gender* on wages?

Problem 3

For this problem, you will work with *artificial* data. To keep your results consistent, make sure to set a **seed number** whenever you include random components into a variable.

(a) Create 1,000 observations of the following *two* variables:

$$
\begin{aligned}
x_i &= 0.8 + \gamma_i && \text{where } \gamma_i \sim \text{Uniform}[0, 100] \\
y_i &= 120 + 5.5x_i + \epsilon_i && \text{where } \epsilon_i \sim \mathcal{N}(0, 0.5x_i^2)
\end{aligned}
$$

(b) Estimate a regression model for *y*, controlling for *x*.

(c) Generate a scatter diagram, with part (b)'s independent variable on the horizontal, and the regression's residuals on the vertical axis.

(d) Given your plot from part (c), what do you conclude about CLRM Assumption V, regarding the variance of the error term? Explain your answer.

(e) Based on your answer to part (d), what may have caused the behavior observed in the scatter plot? Explain.