

EC 339

Problem Set 2

Prof. Santetti

Spring 2023

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

You can find templates for your answer 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 March 8 (W), before class.

Points Possible: 30

- You have 2 weeks to complete this assignment. See our `course syllabus` for late submissions policies.
- 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

The output table on the next page brings results from a regression model whose dependent variable, *math4*, denotes the share (%) of 4th-grade students whose math scores were satisfactory. The control variables are *pctsgle*, the share (%) of 4th-graders living in homes with no married-couple families; *free*, the share (%) of 4th-graders eligible for the school's free-lunch program; and *lexppp*, the natural logarithm of expenditures per pupil.

After checking out the table carefully (standard errors are in parentheses), answer the following questions:

- (a) Are all slope coefficients *individually* statistically significant at $\alpha = 5\%$? Explain. *Hint:* You need not calculate anything to answer this question.
- (b) Are all slope coefficients *individually* statistically significant at $\alpha = 1\%$? Explain. *Hint:* You need not calculate anything to answer this question.
- (c) Are all slope coefficients *jointly* significant at $\alpha = 5\%$? Explain. *Hint:* You need not calculate anything to answer this question.
- (d) How many *degrees-of-freedom* remain after this regression model is estimated?
- (d) Compute a 95% confidence interval for the coefficient on *free*. (You may use 3 decimal points only.)

	Dependent variable:
	math4
pctsgle	-0.259** (0.117)
free	-0.420*** (0.070)
lexppp	8.802** (3.756)
Constant	17.520 (32.246)
Observations	229
R ²	0.472
Adjusted R ²	0.464
Residual Std. Error	11.568 (df = 225)
F Statistic	66.923*** (df = 3; 225)
Note:	* p<0.1; ** p<0.05; *** p<0.01

Problem 2

Anglin and Gençay (1996) estimate several residential housing price models. You can use their data by importing the `house_prices.csv` (available on [theSpring](#)) 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 working environment, replicate the paper's Table III (i.e., look at this regression's output):

$$\begin{aligned} \log(\text{price}_i) = & \beta_0 + \beta_1 \text{driveway}_i + \beta_2 \text{recreation}_i + \beta_3 \text{fullbase}_i + \beta_4 \text{gasheat}_i + \\ & + \beta_5 \text{aircon}_i + \beta_6 \text{garage}_i + \beta_7 \text{prefer}_i + \beta_8 \log(\text{lotsize}_i) + \beta_9 \text{bedrooms}_i + \\ & + \beta_{10} \text{bathrooms}_i + \beta_{11} \text{stories}_i + u_i \end{aligned}$$

(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) Are the two models you've estimated 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 3

For this problem, you will use the `wage.csv` data set, available for download on [theSpring](#). It contains data from the 1976 Current Population Survey. We will analyze the following wage model:

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

[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 $\alpha = 5\%$.
- (e) What is the effect of *gender* on wages?