# EC 339–001

## Problem Set 2

**Prof. Santetti**

Fall 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 October 25 (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 $4^{th}$-grade students whose math scores were satisfactory. The control variables are *pctsgle*, the share (%) of $4^{th}$-graders living in homes with no married-couple families; *free*, the share (%) of $4^{th}$-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?

(e) 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 coefficients presented in the paper's Table III:

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

(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) 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

Use the `cps5_small.dta` data set for this problem. More details on it can be found in the `.txt` file describing its variables.

(a) Estimate a regression model for *wages*, controlling for *education*, *experience*, *squared experience*, *family income*, and whether or not the individual lives in the *South*.

(b) Is the regional factor statistically significant at $\alpha = 5\%$ to explain variations in wages? Explain.

(c) Test for joint significance involving education, experience, its squared term, and the regional factor variable. Use $\alpha = 5\%$. What do you conclude? Explain.

(d) Interpret the effects of (*i*) the regional factor and (*ii*) years of experience on the dependent variable.

(e) Plot the model's estimated residuals (call them *resid_3*, *y-axis*) against years of *education* (*x-axis*). Does something call your attention? Explain. *Hint: CLRM Assumption V*.