# EC 339

## Problem Set 4

**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 12/09, midnight`**.
**`Points Possible: 30`**

- You have 10 days 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

Suppose a shopper is deciding between purchasing Coke or Pepsi. The `coke.csv/.dta` file (available on `theSpring`) contains data on 1,140 individuals who have made this choice.

(a) After importing this data set into your working environment, estimate the following linear probability model (LPM):

$$\mathbb{E}(coke_i = 1) = \beta_0 + \beta_1 pratio_i + \beta_2 disp\_coke_i + \beta_3 disp\_pepsi_i + u_i$$

where `coke` equals 1 if the individual chooses to buy Coke, and 0 if she buys Pepsi.

(b) Based on your results from part (a), *ceteris paribus*, if the price of Coke relative to Pepsi increases, is the individual more or less likely to buy Coke? How can you tell?

(c) Interpret the estimated coefficient for `disp_coke`.

(d) Now, estimate a `probit` model, with the same variables as in (a).

(e) Based on (d)'s model, compute and interpret the *average marginal effect* of `pratio` on the dependent variable.

# Problem 2

The `loanapp.csv/.dta` data set (available on `theSpring`) brings data from `Hunter and Walker (1996)`, who evaluate loan officers' actions on creditworthiness of minority applicants relative to white applicants. `This page` has a description of the variables.

(a) `approve` is a binary indicator which equals 1 if an individual's mortgage loan was approved, and 0 if not. First, regress this variable on `white` using OLS. Interpret the result.

(b) From your answer to part (a), is there evidence of discrimination against nonwhites in mortgage applications? Explain.

(c) Now, regress `approve` on `white`, `obrat`, `hrat`, and `emp` using OLS. Interpret all estimated slope coefficients.

(d) Now, estimate a `logit` model using the same variables as in (c). Just looking at the model coefficients' signs, what do they indicate for mortgage approval?

(e) Finally, compute *average marginal effects* for all independent variables from (d). Interpret all of them. Re-answer part (b), based on your new results.

# Problem 3

Let us explore some of the factors predicting costs at American universities using the `college_data.csv/.dta` data set (available on `theSpring`). It contains 1,628 panel observations on 203 institutions at different years. Also make sure to check out the `.txt` file describing its variables.

(a) For this problem, we will only use a *cross-section* of these data. Filter only observations for 2005, keeping all column variables.

(b) From this cross-section for 2005, estimate the following model:

$$log(tc_i) = \beta_0 + \beta_1 ftug_i + \beta_2 ftgrad_i + \beta_3 ftef_i + \beta_4 cf_i + \beta_5 ftenap_i + \beta_6 private_i + u_i$$

(c) Does this model suffer from heteroskedasticity? Estimate *Breusch-Pagan* and *White* tests. Use *α = 0.05*.

(d) Given your results for part (c), is our inference reliable the way the model is right now? Explain your reasoning.

(e) Given your answer for the previous part, take the necessary actions (if any) to have a reliable inference from this model. Then, interpret *all* slope coefficients and tell me which variables are statistically significant and which are not.

# Problem 4

**Note**: This is an *extra-credit* opportunity. You need not answer this problem if you do not want to.

This problem is especially designed for those who want to have a *5*-point boost in their final exams. It also works as a good way to *study* for the Final Exam.

To get some extra points, you are asked to *redo* all the problems or parts where you have an incorrect answer in the *Midterm* exam.

To get *full* extra credit, you have to indicate what you did wrong, then telling me the right way to write the code (in case of coding mistake), or the correct answer for the questions (in case of an incorrect interpretation).