# EC 339

## Problem Set 1

**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 February 22 (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

Consider the following regression model:

$$\widehat{log(price_i)} = 4.39 + 0.036\ sqft_i$$

where *price*$_i$ is the selling price of a property (in $1,000), and *sqft*$_i$ is the total interior square footage of a house (in 100 sqft). This data set has 500 observations of single-family home sales in Baton Rouge, LA in 2013.

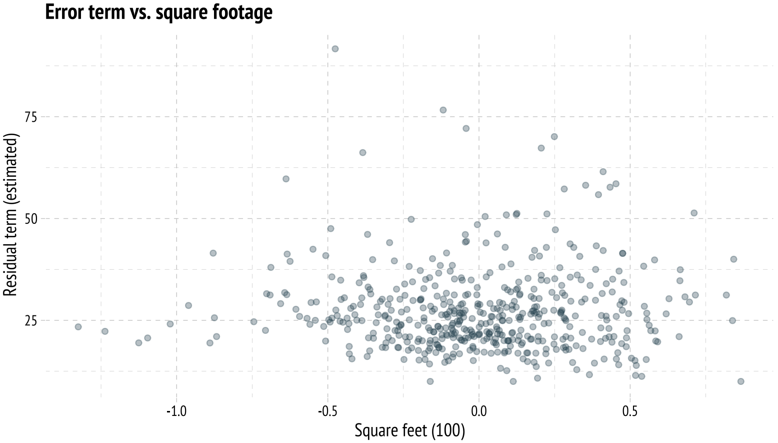
(a) Interpret this regression's *slope* coefficient. (*2 points*)

(b) How many *degrees-of-freedom* remain after this estimation? Explain. (*2 points*)

(c) What other variable(s) would you consider to be included in this regression's *error term*? Explain. (*2 points*)

(d) This regression's *coefficient of determination* ($R^2$) is 0.5417. Interpret its value. (*2 points*)

(e) The scatter plot below charts the above regression model's estimated error term ($\hat{u}$) and independent variable. Based on what you see, is there a clear relationship between the residual and the control variable? Should we expect it? (*2 points*)



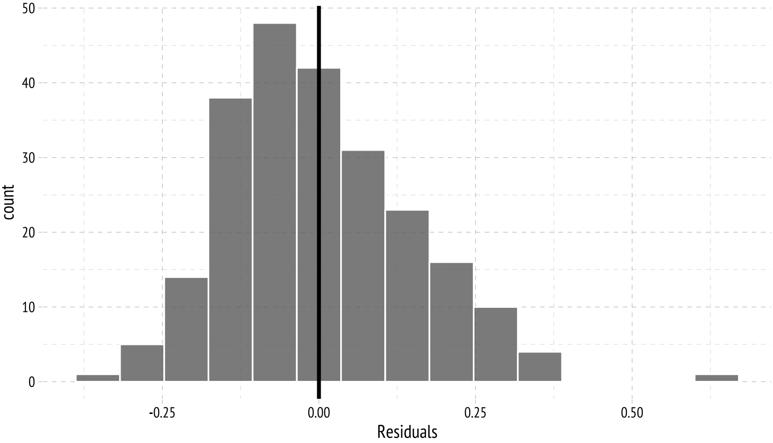
# Problem 2

In 1857, `Ernst Engel` empirically tested the proposition that, as individuals' income increases, their expenditures on food increase *less than proportionally* to the increase in income. Based on these premises, answer the following questions:

(a) Having only access to two variables: individuals' *income* and their *food expenditures* (*foodexp*), how would you **model** the above research question to capture Engel's claim? Present your theoretical regression model.

(b) From your part (a)'s model, the coefficient on the independent variable should represent the proportional change (%) in food expenditures following a change (%) in income. To confirm Engel's hypothesis, the estimated coefficient should lie between what range of values? Explain your reasoning.

(c) Suppose you were lucky enough to get some data and run the regression model you've come up with. Below, a histogram of the model's estimated residual term, with the vertical bar indicating its mean (expected) value.



Does this correspond to your expectations regarding how the residual's histogram should look? Explain your reasoning.

(d) The covariance between the residual term and the independent variable is 0.0000000000000000163. Is this expected? Explain your reasoning.

(e) Based on your model, how do you feel about CLRM Assumption VI, on possible linear associations among the independent variables? Explain your reasoning. *Hint*: this is very easy.

# Problem 3

The `koop_tobias` data set brings a subset of the data used by `Koop and Tobias (2004)`. The sample is restricted to white males who are at least 16 years old and who worked at least 30 weeks and 800 hours during the year of 1987.[1] You can also find a `.txt` file describing the variables.

(a) After importing the data and getting some acquaintance with it, estimate the following regression model: (*2 points*)

$$log(wage_i) = \beta_0 + \beta_1 educ_i + \beta_2 fatheduc_i + \beta_3 exper_i + u_i$$

(b) Interpret the above regression's slope coefficients. (*2 points*)

(c) Now, add a *proxy* variable for worker's *ability* to the previous model, in this data set denoted as *score*. This variable is constructed from the 10 component tests of the Armed Services Vocational Aptitude Battery, administered in 1980, and standardized for age. (*2 points*)

(d) From part (c)'s model, interpret the effect of *score* on the dependent variable. (*2 points*)

(e) Critically compare the *goodness-of-fit* measures between your models from parts (a) and (c). (*2 points*)

[1] This data set is part of Carter-Hill, Griffiths, and Lim, *Principles of Econometrics*, 2018, 5th edition, Wiley.