The data set on the website, classsizedata.raw, contains information on a sample of 2,049 different fourth-grade classes in Israel. The data come from a paper published in the Quarterly Journal of Economics, and was written by Angrist and Lavy (1999). I encourage you to read through the paper for additional context (a link is provided on the website). Some of the material in the paper you may not yet understand, but it will hopefully become clear as we discuss additional topics in the course. The authors are primarily interested in determining the causal effect of class sizes on student achievement, as measured by reading and mathematics scores.

The following variables are available in the data set, with description of each below. These follow the ordering of the columns in classsizedata:

- **ReadingScore** = Class-level score on a standardized reading test.
- **MathScore** = Class-level score on a standardized mathematics test.
- **ClassSize** = Number of students in the class
- **PrctDisadv** = Percentage of students in the school who are disadvantaged
- **Enroll** = Total enrollment in the school.
- **MaimonidesIV** = Instrumental Variable, as discussed in the paper ($f_{sc}$). We won’t use this or worry about it until later in the course when discussing IV estimation.

(1a) We will begin by replicating some of the results found in Table II of the Angrist / Lavy paper. To begin very simply, use the data to calculate the mean Reading and Math scores, and the standard deviation of those scores, as reported in Table II under columns 8 and 11. (Note that there appears to be a typo on the Math score, and the number appears correct later in Table III).

(1b) Write a MATLAB program to output some basic regression results. Your program should report: (a) OLS coefficient estimates, (b) standard errors associated with the coefficient estimates, (c) $R^2$ and (d) $\hat{\sigma}^2$. Print out and submit the m-file you used to perform these calculations.

Once you have this program written, use it to replicate the results of Table II, columns (7)
and (10). You will do so by running the regressions:

\[ \text{ReadingScore}_i = \beta_0 + \beta_1 \text{ClassSize}_i + \epsilon_i \]
\[ \text{MathScore}_i = \theta_0 + \theta_1 \text{ClassSize}_i + u_i \]

Note that your point estimates of the parameters and \( R^2 \) value should exactly reproduce those reported in the table. Standard errors reported in Table II will be a little different than those you calculate, however, as the authors control for correlation in outcomes across classes within the same school.

(1c) Interpret your point estimates. Are the results consistent with what you would expect? Discuss your beliefs about the appropriateness of the mean-independence assumption in this context.

(1d) Now, replicate the results in Columns (9) and (12) of Table II. Note that this will involve adding \( \text{PrctDisadv} \) and \( \text{Enroll} \) to the regressions.

(1e) How has your assessment of the impact of class size changed? Does this change make sense to you? What is your opinion of the appropriateness of the mean-independence assumption for this regression?

(1f) As discussed in the lectures, let’s get the coefficient on \( \text{ClassSize} \) in Column (9) in a different way. Specifically, perform the following steps:

- Regress \( \text{ClassSize} \) on a constant (intercept), \( \text{PrctDisadv} \) and \( \text{Enroll} \) and get the residuals.
- Regress \( \text{ReadingScore} \) on those residuals

and verify that the same coefficient is obtained.

(1g) Based on estimates reported in Column 9, predict the Reading Score outcome of a class with 25 students, located in a school with 100 students in total where 30 percent of those students are disadvantaged. Also report a standard error for your point estimate.

(1h) Suppose that you are the principal of a school, and you have learned that a company will establish a new firm in your district. This new firm will create new jobs in the local labor market, which you expect will lower the percentage of disadvantaged families by 20
percent. In addition, the firm has offered your school a donation to hire an additional teacher, and as a result, you expect that class sizes will reduce by 10 students.

What is your point estimate of the impact of this new firm on the change in reading scores?

(2) Suppose you estimate the model:

\[ y = X\beta + u \]

and obtain the residuals \( \hat{u} \).

If \( y \) is then regressed on the residuals \( \hat{u} \) (without an intercept term), what will you obtain as the estimated slope coefficient from this regression? Why?