Homework 2

Economics 7103

You have access to imaginary data on an energy-efficiency retrofit program in Atlanta *kwh.csv* and you are interested in whether the program reduced energy use. In your dataset is the following information:

Variable	Description
electricity	kWh of electricity used by the household in the month
sqft	Square feet of the home
retrofit	= 1 if the home received a retrofit
temp	The outdoor average temperature (°F) during the month at the home's location

Table 1: Variable descriptions for homework 1.

After recruiting the households for the program, you assigned them to treatment and control groups. Treatment homes received the retrofits on the first of the month and control homes did not have any work done.

1 Python

1. Check for balance between the treatment and control groups using Python. Create a table that displays each variable's sample mean, sample standard deviation, and p-values for the two-way t-test between treatment and control group means. Your table should have four columns: one with variable names, one with sample mean and standard deviation for the control group, one with sample mean and standard deviation for the treatment group, and one with the p-value for the difference-in-means test. Does it appear that the randomization worked? If so, what can we say about the simple difference-in-means estimate?

2. Provide graphical evidence that the retrofits worked. Plot kernel density plots of the electricity use for treated group and control group on the same graph using Python. Make sure to label the histogram appropriately.

3. Suppose you want to estimate the linear equation $Y = \beta X + \varepsilon$ where Y is an $n \times 1$ vector of the dependent variable, X is an $n \times p + 1$ matrix of the predictor variables in table 1 and a column of ones, and ϵ is an $n \times 1$ vector of unobserved random error. Use the following methods to estimate $\hat{\beta}$, presenting coefficients in a single table with a column for each estimation technique (note I am not requiring that you present confidence intervals):

   (a) OLS by hand. Use the Numpy package in Python to create an array X that is the $n \times p+1$ matrix of the predictor variables in table 1 and a column of ones and an array Y that is the $n \times 1$ vector of the dependent variable. Use matrix operations to calculate $\hat{\beta}$. Recall that $\hat{\beta} = (X'X)^{-1}X'Y$ is the closed-form solution to the least-squares minimization problem.

   (b) OLS by simulated least squares. Use the Scipy.optimize.minimize() function in Python to numerically minimize the sum of squares objective function. Recall that the sum of squares is $\sum_{i=1}^{n} (y_i - \beta x_i)^2$ where y_i and x_i are (1×1) and $(1 \times p + 1)$ vectors respectively.

   (c) OLS using a canned routine. Use the StatsModels package in Python using the OLS routine.

2 Stata

1. Check for balance between the treatment and control groups using Stata. Create a table that displays each variable's sample mean, sample standard deviation, and p-values for the two-way t-test between treatment and control group means. Your table should have four columns: one with variable names, one with sample mean and standard deviation for the control group, one with sample mean and standard deviation for the treatment group, and one with the p-value for the difference-in-means test. Hint: https://www.statalist.org/forums/forum/general-stata-discussion/general/1519721-summarized-statistics-table-with-t-test-for-difference-in-means contains useful code.

2. Create a two-way scatterplot with electricity consumption on the y-axis and square feet on the x-axis using Stata's `twoway` command. Make sure to label the axes.

3. Estimate the same regression as in #3 above using Stata's `regress` command, estimating heteroskedasticity-robust standard errors. Report the results in a new LaTeX table (including standard errors) using Stata's `outreg2` command.