**Homework 5: Natural experiments**

**GB 740: Sacks**

Your homework this week asks you to estimate and reflect on difference-in-differences models. In part 1 I give you example code (in the markdown template) and ask you to execute it and interpret the results. In part 2 I ask you to write your own code for a different difference-in-differences model.

Please upload a knitted markdown file (html preferred) which contains your code, findings, and answers.

**Part 1**

Consider a retail company that has launched a store redesign at a set of 50 retail locations in one region of the country.  They are interested to know whether this redesign caused an increase in sales.  They did not randomly select stores for the redesign.  However, they identified a set of 50 comparison stores from another region of the country that they believe should provide a useful "control" for the "treatment" stores.  There were no changes at these 50 comparison stores.

The data set they have available for analysis has 24 months of data, with months 1 through 12 being the year before the change took effect and months 13 through 24 corresponding to the year when the redesign was in place at the stores in the treatment region.  For each of those months they have data on the total sales (in dollars) at each store, for both the 50 stores in the treatment region and the 50 stores in the control region.

The case data is on Canvas as [sales\_DID.csv](https://canvas.wisc.edu/courses/324566/files/26633522/download?wrap=1)

1. The mark down file makes a plot with the time series of outcomes for treatment region and control regions. Describe the plot and note (i) whether it appears consistent with the parallel trends assumption and (ii) whether we should expect to see a positive, negative, or zero DID estimate. (Bonus (iii): use the plot to get a rough guess of the DID estimate. Explain your reasoning.)
2. Suppose your smart and experienced colleague after seeing this graph stated: "Well, I guess it is helpful to have the control region as a comparison to validate that the changes we see in the treatment region were not due to other factors at that time.  But ultimately it looks like the difference-in-difference approach will give us basically the same estimate as just a simple event-study analysis for the treatment region."  **Briefly explain what your colleague means by that statement and what she is looking at in the graph that leads her to that conclusion.**
3. Now run the regression part of the example case code.  Briefly interpret the coefficients in the regression and explain how they relate to what we see in the graph we created.  Focus on the magnitude of the regression coefficients -- what do these numbers mean in practical business terms?  Also report the confidence intervals. Notice that the code reports the coefficient and standard errors ,so you will need to do a bit of math to calculate the confidence intervals. You can do this by using the formula for the confidence interval:

**Part 2.** Ok, now it's your turn to do this yourself.  We want you to analyze some data from a research study that examined the effect of a change to rating algorithms.  In January 2009, Trip Advisor began to allow users to score restaurants on multiple dimensions as well as providing an overall ranking of the restaurant (which was always possible).  The question of interest in the research study was whether this change to allow for multi-dimensional rating had an effect on the overall average rating of restaurants on the platform.  You should adapt the R code from part 1 to this question and data set. Use the data set RatingData.csv.

1. Plot the average rating for each website by month and include 95% confidence intervals on the average rating, similar to the example case.  Make sure to label the axes and legend appropriately so that it is easy to process the information in the graph.

Interpret what you are seeing in the graph.  Does the "parallel trends" assumption between Trip Advisor ratings and Yelp ratings appear valid here?  Is there a discernable effect of the treatment at Trip Advisor that you can see in the graph?

1. Now run a regression (as in the example case) to quantify the difference-in-difference estimate of the effect of the switch to multi-attribute rating on average restaurant ratings. What is the DID estimate and its 95% confidence interval? Is the estimate statistically significant? Business significant?
2. Focusing on the regression estimates and ignoring graphical evidence for the moment, what do you conclude about the effect of the switch to the multivariate rating system on average ratings at Trip Advisor?  Please be specific about the evidence you are using for your conclusion.
3. Briefly discuss how the regression results compare to your visual analysis of the graph you created.  Do you reach a similar conclusion or different?  Which approach would you rely on more in a business situation for this case, and why?