AAEC 4804/5804G, STAT 4804: Fundamentals of Econometrics

Your Name Here

Spring 2025 – Homework #6

Instructions

This homework is intended to help you review the material covered in Lecture 8. This week we explore some real world issues presented in journals and policy papers.

You are strongly encouraged to work with your classmates, but you must submit your own Solutions.

Question 1: Policy Evaluations using Difference-in-Differences

The U.S. government has a long history of subsidizing housing for low-income families. These policies are designed to help low-income families afford housing, reduce the number of people living in substandard housing, and, by extension, reduce the number of people living in poverty. In recent times, such support programs have become increasingly important as housing and rent costs have risen significantly.

While the policy has its advantages, it is not without criticism and drawbacks. One significant concern is that these housing programs may lead to declining property values in the surrounding neighborhood. This occurs because the subsidies can attract low-income families to the area, which may lower property values and create challenges for homeowners trying to sell their houses. Furthermore, the influx of new residents may increase traffic congestion, crime, and other social issues, further depressing property values.

One influential paper is "Does Federally Subsidized Rental Housing Depress Neighborhood Property Values?" by Ellen, Schwartz, Voicu, and Schill (2007), published in the Journal of Policy Analysis and Management (see `Literature` folder in GitHub). They found that "federally subsidized developments have not typically led to reductions in property values and have, in fact, led to increases in some cases. Impacts are highly sensitive to scale, though patterns vary across programs."

Unfortunately, I could not find their data online. Instead, I will use some simulated data. The file of interest is `Affordable-housing.csv` and is stored in the Data folder on GitHub.

The data contains information on the following variables: - `Price`: The price of the house. - `Close`: House close to a federally subsidized housing site (`=1`, our treated group) and not (`=0`, our control group). - `Post_Treat`: House prices before construction of new federally subsidized housing (`=0`) and after construction (`=1`).

(-) Go ahead and load the data into `R` from the `Data` folder on GitHub.

(a) We are interested in whether affordable housing projects reduce the prices of houses in the surrounding neighborhood. To do this, we will use a difference-in-differences (DiD) approach.

We will first manually compute the average price of houses in the treated group before and after the new federally subsidized housing construction. Next, we will use these to compute the DiD estimator.

Tasks:

- Run the following regression model over the pre- and post-treatment periods:

$$Price = \beta_0 + \beta_1 Close + \varepsilon$$

- Use the `stargazer` package to report the regression results. Ensure that both models are appropriately labeled. **Report the coefficients, along with their standard errors and t-stats, in that order. Omit the F-statistic and standard error from the output.**

- The DiD estimator is the difference in the slope coefficients of the `Close` variable in the pre- and post-treatment periods. **Calculate** and **interpret** the value on the DiD estimator.

Nic sees your results and asks, "But is this statistically significant?" Admittedly, you are not sure, but you have been in the class long enough to know that you could estimate a model with an interaction to get the same results as above and perform a direct t-test.

(b) Using the `lm` function, estimate the following model and show that the DiD estimator is δ_1:

$$Price = \beta_0 + \beta_1 Close + \beta_2 Post\_Treat + \delta_1 Close \cdot Post\_Treat + \varepsilon$$

(ii) Discuss whether the price differential is statistically significant at the 5% level.

(iii) By now, you ought to recognize that each coefficient in the model in (b) speaks to a different hypothesis/question. Answer the following:

(iv) Interpret the intercept.

(v) Do houses in the treatment group sell for more than the control group before new federally subsidized housings are constructed? Is this difference statistically significant? What is the average price of houses in the treated group before the new builds?

(vi) Interpret the coefficient on `Post_Treat`. Also, what is the average price of the houses in the control group after the construction of the new federally subsidized housing?

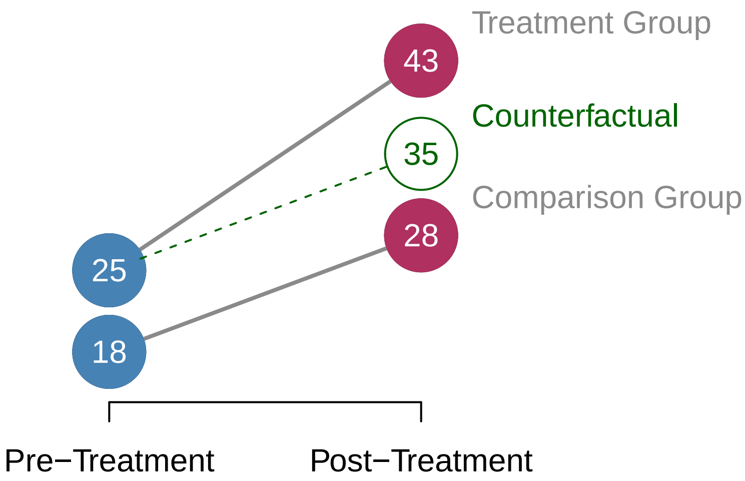
(vii) We did not explicitly explore another relevant concept in class: parallel trends and the counterfactual. In essence, we are asking the question: "What would have happened to the price of houses in the treatment group if the new federally subsidized housing was not built?"

An easy way to get to this is to assume that the price differential (after - before) in the control group is the same as the price differential (after - before) in the treatment group.

For completeness, we will are trying to find $\widetilde{\text{Price}}_{\text{Treated, after}}$ in the equation below:

$$\begin{aligned}\widetilde{\text{Price}}_{\text{Treated, after}} - \text{Price}_{\text{Treated, before}} &= \text{Price}_{\text{Control, after}} - \text{Price}_{\text{Control, before}} \\ \widetilde{\text{Price}} &= \text{Price}_{\text{Treated, before}} + \left[\text{Price}_{\text{Control, after}} - \text{Price}_{\text{Control, before}}\right]\end{aligned}$$

Visually:



Using the results from (c), compute and interpret the counterfactual price of houses in the treatment group.

(e) What is the actual price of the houses in the treatment group after the construction of the new federally subsidized housing? How does this compare to the counterfactual price? **Be sure to give the exact value of the actual price and the difference.**

Question 2: Sex Sells: The Economics of Prostitution

As you might have noticed from our discussions in class, economists examine a wide range of topics, not just the boring ones (yes, Miru, I heard you say, "such as birth weight and number of cigarettes smoked while pregnant"). If a topic has come up in a debate or conversation with a friend, chances are, the topic has either been explored or, with a little reframing, can be explored using econometric techniques.

One important issue is the concept of prostitution. This represents a true market where there are buyers, sellers, and goods or services being offered. Sellers often possess significant bargaining power and operate within a monopolistically competitive structure since the goods and services offered are slightly differentiated. Additionally, there is a risk premium associated with these goods and services that may not be reflected in their prices. This risk premium can be understood as the extra amount that sellers are willing to accept to "throw caution to the wind" and engage in transactions that carry a higher risk of negative outcomes, such as sexually transmitted infections (STIs) or legal consequences.

While it is a controversial topic, it is one that has been studied by economists as it can provide insights into human behavior, market dynamics, and the impact of regulations. One such influential paper is "Risky Business: The Market for Unprotected Commercial Sex" by Gertler, Shah, and Bertozzi (2005) (GSB, henceforth) published in the Journal of Political Economy. The details of their abstract are as follows:

> While condoms are an effective defense against the transmission of HIV, large numbers of sex workers are not using them. We argue that some sex workers are willing to take the risk because clients are willing to pay more to avoid using condoms. Using data from Mexico, we estimate that sex workers received a 23 percent premium for unprotected sex. The premium represents a value of one life year of between $14,760 and $51,832 or one to five times annual earnings. The premium jumped to 46 percent if the sex worker was considered very attractive, a measure of bargaining power.

For the following questions, we will use the `mexican.csv` file stored in the Data folder on GitHub. This file presents a subset of the data from the paper. There is information on four transactions (`trans`) per worker (`id`). The paper is in the `Literature` folder on GitHub. While you are not expected to read and understand the entire paper, I do encourage you to browse through and use aspects of the paper to help you understand the context of the data and the questions that follow. Last, the `mexican_meta.txt` provides a brief description of the variables in the dataset.

(-) Go ahead and load the data into `R` from the `Data` folder on Github.

(a) Using OLS, estimate the following model:

$$lnprice = \beta_0 + \mathbf{X}\gamma + \varepsilon$$

where *lnprice* is the natural log of the transaction price, **X** is a vector of variables that includes the `Sex Worker`, `Client`, and `Transaction` characteristics.

Store and report the results of the regression.

(b) Discuss and interpret the results on (i) age, (ii) attractive, (iii) regular, (iv) alcohol, (v) nocondom, and (vi) bar. **Please ensure your discussion is about the *level* of y (`price`) and not the *log* of y (`lnprice`). Also, be sure to use precise estimates as appropriate.**

Note: I would like to see you explicitly answer questions such as "Do older workers charge more or less?", "Do attractive workers command a higher price for their services?", "Do regulars appear to enjoy a discount?". You get the gist.

(c) As noted in their abstract, GSB argued that sex workers are willing to engage in risky transactions since clients are willing to spend more not to use a condom. They called this a `risk premium`. Using the results from your regression in (a), report the 95% CI of this risk premium. **Does this risk premium appear to be statistically significant using the OLS model?**

(d) In class, we always discuss the concept of the unobserved heterogeneity error component. We also acknowledged that the assumption that our unobserved heterogeneity is uncorrelated with the regressors must hold for the OLS estimator to be consistent.

(e) What are some of the factors that might drive the unobserved heterogeneity error in the model in (a)?

(ii) Could this assumption be violated in the OLS case above? **A yes/no will not suffice. You must briefly explain your thoughts.**

(e) Using an OLS model, re-estimate the model in part (a) to derive the FE estimator. For this approach, I would like you to demean the variables manually. **Store and report the results of your regression.**

Why did the worker characteristics fall out of the model? **Connect this back to the mechanics of the fixed effects estimator.**

(f) An alternative to manually demeaning the variables is to use the least squares dummy variable (LSDV) approach. This approach involves creating a dummy variable for each worker and including it in the regression model. Using the `lm` and `as.factor()` functions, re-estimate your model in part (e) using the LSDV approach. **Store and report the results of your regression.**

Hint: Because we will have a large number of coefficients, I would like you to report the coefficients (along with standard errors, t values, and p values) on `regular`, `rich`, `alcohol`, `nocondom`, `bar`, and `street` only. You can use the `summary()` function on the model fit and then extract the coefficients of interest by indexing the results passed to the `coef()` function.

(g) Using the `plm` package, re-estimate the model in part (e) and (f) using a fixed effects estimator. **Store and report the results of your regression.**

(h) From your fixed effect regression in (g), how is the price affected by whether the client is (i) a regular, (ii) rich, or (iii) consumes alcohol before the transaction?

Please ensure your discussion is in relation to the *level* of y (`price`) and not the *log* of y (`lnprice`). Also, be sure to use precise estimates as appropriate.

(i) Interpret the `risk premium`. Does it now appear to be correctly signed? Present the 95% CI for the risk premium. **Does this risk premium appear to be statistically significant using the FE model?**

(j) You were presenting your replication study in class and a classmate asked you: "Are you sure that your FE model is appropriate? What if the unobserved heterogeneity error is correlated with the regressors?"

You feel attacked, and his smirk did not make it better. You have `R` open on your computer and decide to humble him. How would you respond to this question?

You must use an appropriate statistical test to support your response. Be sure to discuss what the test is doing– the null and alternative hypotheses– and correctly conclude.