**Final Exam Practice Questions**

1. Propensity Score

Assume that we are interested in modelling the effect of a medical procedure on patient-level outcomes. The dataset also includes demographic information of age, sex, insurance status, comorbidities, and year.

1. Describe in words or in an equation how you would construct the propensity score for the probability of receiving the medical procedure.
2. Is the propensity score different depending on the outcome being evaluated?
3. Describe in words (you can include equations if you want) a method you would use to construct an estimate of the effect of the medical procedure on a patient-level outcome from the estimated propensity score.
4. There is some concern that the patients that receive this particular medical procedure may be different in unobservable ways than patients that receive a less novel procedure for the same condition. Can we use propensity scores to correct this problem?

2. Fixed Effects

See the regression output for Question 2. The treatment variable is “cigarettes per day” (*cig\_per\_day*) and the outcome is continuous birth weight *(dbwt)*. For the purpose of this question, ignore the fact that heteroscedasticity-robust standard errors are not used (they should be).

1. The second regression output adds state fixed effects.
   1. State the key assumption (in words or in an equation) under which the approach recovers valid causal estimate of the effect of smoking on birth weight.
   2. Provide a specific example of something that state fixed effects may control for in this regression, which could plausibly affect both the birth weight and smoking.
   3. Does controlling for state fixed effects change the association between the treatment and the outcome?
2. The third regression output adds year fixed effects.
   1. State the key assumption (in words or in an equation) under which the approach recovers valid causal estimate of the effect of smoking on birth weight.
   2. Provide a specific example of something that year fixed effects may control for in this regression, which could plausibly affect both the outcome and the treatment.

3. Differences-in-Differences

We are interested in modelling the effect of the Hospital Readmission Reduction Program (HRRP) on reducing hospital-level readmissions. This program started in year 2012 and provided additional payments or levied penalties on hospitals that performed above or below expectations in reducing preventable patient readmissions. The size of the incentives and penalties increased over time in the years after 2012. This program was applied to all acute care hospitals.

A possible control group may be Veteran’s Affairs (VA) hospitals that were not subject to HRRP. We have annual data from 2008 to 2015 for all VA and acute care hospitals. We know the readmission rate of each hospital (**readm**, which has a range of 8-25), as well as its status as an acute care or VA hospital (**hospital type**, which does not vary over time). We also have information on the patient **case mix** of the hospitals, which varies over time. You also know the **year** that the data was collected.

1. Please write a differences-in-differences model that you would use to estimate the association that the HRRP program has on readmission rate. Please make sure all the variables or vectors in your model are clearly labeled or defined. Pay particular attention to the subscripts you use.
2. Describe in words the key assumption needed for the model you wrote down to provide a valid causal estimate of the effect of the HRRP on hospital-level readmissions.
3. Draw a graph that provides an example of when the above assumption holds. Label all relevant components of the graph.
4. Name one technique we learned in class that would allow us to relax the assumption you stated above (you do NOT need to provide an equation). Draw a graph that represents a scenario in which this technique would yield a valid causal estimate of the effect of interest.

4. Instrumental Variables

We are interested in modelling the effect of an immunization program on an individual’s probability of getting a similar disease to that being vaccinated against. We have longitudinal data on 100,000 individuals, 50% of whom received the vaccination. Based on comparing the demographics of those that have and have not received the vaccination, it’s apparent that there was self-selection into receiving the vaccination (rather than pure randomization). We have the following information for each individual in the data:

* Whether they received the vaccination at any point
* Later-life presence of the similar disease (those who developed the disease before the vaccination have been dropped from the data)
* Basic demographic information (age, gender, education, family income)
* State-year level advertising dollars to support the vaccination campaign
* State, year and county of each observation

1. Which variable would you consider as a candidate for the instrumental variable in this case?
2. Please use the variables/vectors you have to write out the four equations. Make sure you properly label or define each variable/vector in your equations. Pay particular attention to the subscripts you use.

Standard OLS:

Reduced Form:

First-Stage of IV Model:

Second-Stage of IV Model:

1. Describe in words or in equations the two assumptions necessary for the instrument to be valid. Explain why you think that the instrument may or may not fail these two assumptions in this situation.
2. Assuming the instrument is valid, describe in words the kind of individuals among whom the IV method would estimate the average treatment effect.