**Final exam**

**Due date: June 20, 2020 at 5:00 pm**

**Directions:** Please provide your answers in double spaced and typed document. Your answers should be simple and brief; if you say anything incorrect, you will get partial credit, so it is better to say less and be correct, than to say more but include incorrect information. Each part of the questions is worth 1 point. If you score less than 75%, then you will receive a 75%. In other words, no one can get less than a 75% on this exam.

Regression discontinuity

1. Suppose there is a health-based intervention in the United States where if a person’s income fell below $20,000 US dollars, the person and their family are enrolled in a program that provides free food. Assume you are interested in the effect that the program has on child health outcomes.
   1. If you used a RD design for this project, what is the running variable and what is the cutoff?
   2. What does sorting on the running variable mean and how would you evaluate it in this project?
   3. What does Barrecca, Guldi, Lindo and Waddell suggest if you find large heaps of observations at regular intervals, including the cutoff?
   4. Explain what we mean by the smoothness assumption. How would you express that assumption in this application?
   5. Describe two auxiliary tests you would perform to evaluate the credibility of the smoothness assumption in this project.

Instrumental variables

1. Suppose there is a health-based intervention in a small US city which distributes free food to poor families. You are interested in the effect of the food subsidies on child health outcomes. The program is expensive and underfunded. The city runs a lottery and individuals who win the lottery have the option to enroll. No one can participate unless they win the lottery, but conditional on winning, the person still has to choose to participate.
   1. Assume that treatment effects are homogenous. Under what conditions will the bias of 2SLS be centered on the bias of OLS?
   2. How can we provide evidence that the bias of 2SLS is not centered on the bias of OLS?
   3. What are the identifying assumptions for obtaining a consistent estimate of the program’s causal effect on child health outcomes under homogenous treatment effects?
   4. Now assume that treatment effects are heterogenous. What is meant by the terms complier, defier, always taker and never taker sub-populations? Interpret each in the context of this example.
   5. Are there any always-takers in this program? Why/why not?
   6. Which of the four sub-populations in part (d) contribute to the parameter estimate when using an IV design?
   7. Which parameter is estimated using an IV design with heterogenous treatment effects?
   8. Why should you use IV to estimate the causal effect of the health intervention on child health outcomes as opposed to merely comparing people on the program to those not on the program?

Difference-in-differences

1. Suppose there was a health-based intervention in twenty USA cities in 1985 (i.e., no differential timing) that provided free food to poor families with children (based on 1984 incomes) and that you are interested in estimating the effects of this intervention on earnings. The data runs from 1980 to 1990.
   1. If you estimated the causal effect of the program on earnings using twoway fixed effects, what are the identifying assumptions needed to obtain an unbiased estimate of the program’s average treatment effect on the treated?
   2. If there are dynamic treatment effects, will twoway fixed effects yield an unbiased estimate of the ATT? Why/why not?
2. Now suppose there is a health-based intervention in twenty US cities but the programs are introduced to cities at different points in time. The first group of five cities gets treated in 1985, the second group of five in 1990, the third group of five in 1995 and the fourth group of five in 2000. Assume your dataset runs from 1980 to 2005.
   1. Write out Goodman-Bacon’s 2018 decomposition of twoway fixed effects theorem and explain what each element of the decomposition means.
   2. How many individual 2x2 estimates are there if you estimate this program’s effect using twoway fixed effects?
   3. Calculate the variance of treatment for each group. Which group’s variance of treatment is largest?
   4. Which individual group’s 2x2 weight will be largest and why?
   5. Goodman-Bacon (2018; 2019) warns against using early groups as controls for later groups. What is the risk of doing so and is it possible to avoid this using twoway fixed effects?
   6. Under what assumptions must twoway fixed effects yield the ATT?
   7. What are the implications of treatment dynamics when estimating this difference-in-differences model with twoway fixed effects?
   8. What is the main parameter of interest in the Callaway and Sant’anna framework?
   9. Write down the identifying assumptions of the Callaway and Sant’anna estimator.
   10. How many individual group-time ATT parameters are there in this dataset?
   11. Write down the Callaway and Sant’anna estimator. Explain what each individual element of the estimator means.
   12. Since there is no “never treated” in this dataset, what do Callaway and Sant’anna suggest you use as controls?
   13. Can you estimate, using Callaway and Santa’anna, the last group’s group-time ATT? Why/why not?

Synthetic control

1. Suppose there was a health-based intervention in Houston Texas in 1985 that provided nutritious food to poor families with children (based on 1984 incomes) and that we are interested in estimating the effects of this intervention on earnings. The data runs from 1970 to 1995.
   1. What are the advantages of using synthetic control to estimate the effect of this program versus a traditional quantitative case study approach?
   2. In your own words, explain the synthetic control estimator.
   3. What is the minimization problem, what are the constraints, what values can the vector of weights take on?
   4. Explain *precisely* how Abadie, Diamond and Hainmueller (2011) recommend calculating p-values with this estimator.