Problem Set 7
Due Monday, June 1 at the beginning of lecture
ECON 381-2, Northwestern University, Spring 2015

You may work in groups of 2 or 3 as long as each group member turns in their own handwritten copy and all group members are clearly indicated at the top of every copy. No credit is given if no work is shown. For Stata/R problems include both your code and the output.

1. Suppose that we have an i.i.d. sample $(Y_i, X_i, G_i, T_i), i = 1, \ldots, N$ where Y_i has bounded fourth moments and G_i, T_i, X_i are group, time and treatment binary (0 or 1) random variables with strictly positive variances. We know that only group 1 receives the treatment, X_i, and only in time period 1, so that $X_i = G_i T_i$. Assume that

$$Y_i = \alpha + \beta_i X_i + U_i,$$

where the causal effect, β_i, is treated as an unobservable random variable. Suppose that we maintain the standard difference-in-differences assumption that $\mathbb{E}(U_i \mid G_i, T_i) = \pi_0 + \pi_G G_i + \pi_T T_i$. Is it possible to consistently estimate the average treatment on the treated, i.e. $\overline{\beta}_{att} \equiv \mathbb{E}(\beta_i \mid X_i = 1)$ without making additional assumptions? If so, find a consistent estimator of $\overline{\beta}_{att}$ and justify its consistency. If not, explain why it is not possible under the given assumptions.

2. Consider the linear model

$$Y_i = \alpha + \beta X_i + U_i,$$

where $X_i \in \{0, 1\}$ is binary. We observe an i.i.d. sample $(Y_i, X_i, R_i), i = 1, \ldots, N$ with bounded fourth moments, where R_i is a variable such that

$$\mathbb{P}[X_i = 1 | R_i \geq r_0] > \mathbb{P}[X_i = 0 | R_i < r_0]$$

at some known point r_0. Assume that $\mathbb{E}[U_i | R_i] = 0$. Find a consistent estimator of β and justify its consistency.

3. Suppose that we collect an i.i.d. cross section of data on houses for sale in a particular city in years 0, 1. Between year 0 and year 1 the city decides to build a garbage incinerator at a known spot in the city. The incinerator becomes operational in year 1. In both years our cross-section consists of the sale price P_i, the distance in miles from the incinerator site, D_i, and other characteristics, W_i, about the house (e.g. number of bedrooms) and the neighborhood (e.g. crime). We use a variable $T_i \in \{0, 1\}$ to denote what time period (year) the observation is from. All variables have bounded fourth moments. The goal of our study is to determine the causal effect of being close to the incinerator on house prices. For this purpose we use a causal model

$$\log(P_i) = \alpha + \beta D_i T_i + U_i.$$

(a) How should we interpret β?

(b) Suppose we only use data from year 1. Would we expect $\widehat{\beta}_{lr}$ from a regression of $\log(P_i)$ on $D_i T_i$ and a constant to be consistent for β? Why or why not?

(c) Suppose that we assume

$$\mathbb{E}\left(U_i \mid D_i, T_i\right) = \pi_0 + \pi_T T_i + \pi_D D_i.$$

Show how we can use this assumption to estimate β using data from both years.

(d) Discuss reasons why the assumption in (c) might fail.

(e) Discuss how we could use the additional covariates $\boldsymbol{W}_i$ to modify the assumption in (c) to make it more credible.

Stata/R

4. This question is based on the article "Credit Elasticities in Less-Developed Economies: Implications for Microfinance" by Karlan and Zinman, which was published in *The American Economic Review* in 2008. The dataset the authors used is available in `karlan.dta` on Canvas. In Table 3, the authors compute what they vaguely refer to as "marginal effects" for several different specifications of a probit model. It turns out that these are not the average partial effects that we talked about, but rather a related and generally less useful quantity. When I computed average partial effects for their specifications I obtained the following (broadly similar) results for each column:

   (1) : -.0029662 with a standard error of .0004802.

   (2) : -.0366428 with a standard error of .0059016.

   (3) : -.018188 with a standard error of .0015575.

   (4) : .0010605 with a standard error of .000826.

   (5) : .0053492 with a standard error of .0050479.

   (6) : -.0095549 with a standard error of .0065639.

   (7) : .0004208 with a standard error of .0006357.

   (8) : -.0399369 with a standard error of .0106693.

   (9) : -.0121816 with a standard error of .0060631.

   Also, for columns (3), (6), and (9), I found that two of the variables the authors said they included were dropped due to perfect multicollinearity. Replicate my results.
   Hint #1: The notes below Table 3 say that "Robust standard errors reported in parentheses and are clustered within branch." This type of clustering is similar to the clustering we discussed when studying panel data. To implement it use the option `cluster(branch)` *for the* `probit` *command.*
   Hint #2: Completing this problem does not require reading the entire paper closely. Read enough of the paper to understand Table 3, and then attempt to reproduce my results.

5. Suppose that

$$Y_i = \alpha + \beta X_i + U_i,$$

where $\alpha = 0$, $\beta = 2$, $R_i \sim N(0, .5^2)$, $X_i = \mathbb{1}[R_i \geq 0]$ and $U_i = R_i^3 + V_i$ with $V_i \sim N(0, .5^2)$ independently of R_i. We can see from this setup that $\mathbb{E}(U_i \mid R_i) = R_i^3$ is not a linear function of R_i. But suppose a researcher does not know this and estimates a regression of Y_i on X_i, R_i and a constant for the subsample of observations with $R_i \in [-\epsilon, \epsilon]$. Run Monte Carlos with $M = 5000$ replications under each of the following conditions.[1]
Hints: 1) The second argument in Stata's `rnormal` *function is the standard deviation of the desired normal random variable, not the variance. 2) There are several ways to run a regression on a subsample of observations. One is to include an* `if` *statement after the variable list in the regression command (but before the options). Another way is to simply keep the observations you want or drop the observations you do not want, by using the* `keep` *or* `drop` *commands.*

   (a) Let $N = 200$ and $\epsilon = 10^6$. Record the average fraction of observations you use in the regression. What is the bias of $\widehat{\beta}_{mlr}$? What is its standard deviation?

   (b) Let $N = 800$ and $\epsilon = 10^6$. Record the average fraction of observations you use in the regression. What is the bias of $\widehat{\beta}_{mlr}$? What is its standard deviation? Explain your results relative to (a).

   (c) Repeat (a) and (b) with $\epsilon = .5$. Explain your results.

   (d) Repeat (a) and (b) with $\epsilon = .05$. Explain your results.

   (e) Based on your results in parts (a)-(d), discuss what factors a researcher should take into consideration when choosing ϵ.

   (f) Repeat the previous parts but assume that the researcher now regresses Y_i on $X_i, R_i, R_i^2, R_i^3, R_i^4$ and a constant. How do the results change? Explain what is happening.

   (g) In practice we do not know what the functional form of $\mathbb{E}(U_i \mid R_i)$ actually is. In light of the results of these Monte Carlo simulations, discuss what trade-offs a researcher faces when choosing how to model $\mathbb{E}(U_i \mid R_i)$ in conjunction with choosing ϵ.

6. This problem uses the dataset `injury2.dta` on Canvas. The data contains a repeated cross-section of injured workers receiving worker's compensation benefits in Kentucky. In 1980, Kentucky raised the cap on weekly earnings covered by worker's compensation. This affected high-income workers but not low-income workers. We want to determine what the causal effect of the policy change was on the duration that a worker stays away from work and collects worker's compensation benefits.

   (a) Familiarize yourself with the data.

[1] If $M = 5000$ takes too long then you can try $M = 1000$ or $M = 500$.

(b) Regress log duration on a dummy variable for whether the worker is high- or low-income using only data from after 1980. Is this a convincing estimate of the causal effect of the policy? Why or why not?

(c) Regress log duration on a dummy variable for the time period (before or after the change) for the subsample of high-income workers. Is this a convincing estimate of the causal effect of the policy? Why or why not?

(d) Construct a difference-in-differences estimator of the causal effect of the policy. Interpret your results.

(e) Run the same difference-in-differences estimator but now control for gender, marital status, and whether the worker is in manufacturing or construction. Compare your findings to c).