Problem Set: Natural Experiments

Problem 1. *In 1985, neither Florida nor Georgia had laws banning open alcohol containers in vehicle passenger compartments. By 1990, Florida had passed such a law, but Georgia had not.*

(i) Suppose you can collect random samples of the driving-age population in both states for 1985 and 1990. Let `arrest` *be a binary variable equal to 1 if a person was arrested for drunk driving during the year. Without controlling for any other factors, write down a linear model that allows you to test whether the open container law reduced the probability of being arrested for drunk driving. Which coefficient in your model measures the effect of the law?*

(ii) Why might you want to control for other factors in the model? What might some of these factors be?

(iii) Now suppose that you can only collect data for 1985 and 1990 at the county level *for the two states. The dependent variable is the fraction of licensed drivers arrested for drunk driving during the year. How does this data structure differ from the individual-level data in part (i)? What econometric method would you use?*

Problem 2. *Consider a $T = 2$ difference-in-differences setting with panel data. It is sometimes argued that one may include* baseline (pre-treatment) control variables *as controls in the first-differenced regression. This can be justified by extending the basic DID model.*

Let $w_{i1} = 0$ (no treatment in period 1) and w_{i2} be the binary treatment indicator in period 2. Let $x_{i1}, x_{i2}, \ldots, x_{ik}$ be control variables that are time-invariant for each unit i.

(i) Consider the model

$$y_{it} = a + d\,\mathbf{1}\{t = 2\} + c\,w_{it} + b_1 x_{i1} + \cdots + b_k x_{ik} + u_{it},$$

for $t = 1, 2$. Show that after subtracting period 1 from period 2, we obtain

$$\Delta y_i = d + c\,\Delta w_{i2} + \Delta u_i.$$

(ii) Now expand the model to include interactions between the year dummy and each x_{ij}:

$$y_{it} = a + d\,\mathbf{1}\{t = 2\} + c\,w_{it} + b_1 x_{i1} + \cdots + b_k x_{ik} + e_1(\mathbf{1}\{t = 2\} \cdot x_{i1}) + \cdots + e_k(\mathbf{1}\{t = 2\} \cdot x_{ik}) + u_{it}.$$

This extension allows trends to vary with observed characteristics (i.e., violates strict parallel trends, but only as a function of x_{ij}). Show that first-differencing gives:

$$\Delta y_i = d + c\,\Delta w_{i2} + e_1 x_{i1} + \cdots + e_k x_{ik} + \Delta u_i.$$

In other words, in the regression using Δy_i as the dependent variable, the time-invariant covariates x_{ij} now do *appear along with the treatment indicator.*

(iii) *Now allow the* treatment effect *to vary with the x_{ij} as well:*

$$\begin{aligned} y_{it} =& a + d\,\mathbf{1}\{t=2\} + c\,w_{it} + b_1 x_{i1} + \cdots + b_k x_{ik} + e_1(\mathbf{1}\{t=2\}\cdot x_{i1}) + \cdots + e_k(\mathbf{1}\{t=2\}\cdot x_{ik}) \\ &+ q_1(w_{it}\cdot x_{i1}) + \cdots + q_k(w_{it}\cdot x_{ik}) + u_{it}. \end{aligned}$$

Show that first-differencing yields:

$$\Delta y_i = d + c\,\Delta w_i + e_1 x_{i1} + \cdots + e_k x_{ik} + q_1(w_{i2}\cdot x_{i1}) + \cdots + q_k(w_{i2}\cdot x_{ik}) + \Delta u_i.$$

Note: Why do some DID papers include baseline controls in first differences if differencing normally removes them? If you allow heterogeneous trends (year × covariate interactions), then those covariates will appear in the differenced equation. If you allow heterogeneous treatment effects (treatment × covariate interactions), then those interactions also appear in the differenced equation. In short, when the model allows for heterogeneous trends or heterogeneous treatment effects based on those covariates, including baseline covariates in the DID model is justified.

We would want to center *the x_{ij} before interacting them with w_{i2} (e.g., transform x_{ij} into $x_{ij} - \bar{x}_j$).*

Problem 3. *Use the data in KIELMC (Description) for this exercise. Kiel and McClain (1995) studied the effect of a new garbage incinerator on housing values in North Andover, Massachusetts. Rumors about a possible incinerator began after 1978. Construction started in 1981, and the facility began operation in 1985. We will use samples of houses sold in 1978 and in 1981. The hypothesis is that the prices of homes located closer to the incinerator site should fall relative to the prices of more distant homes.*

(i) *Let* ***dist*** *denote the distance (in feet) from each home to the incinerator site. Consider the model:*

$$\log(\textit{\textbf{price}}) = \beta_0 + \delta_0\, y81 + \beta_1 \log(\textit{\textbf{dist}}) + \delta_1(y81\cdot\log(\textit{\textbf{dist}})) + u.$$

If the incinerator reduces the value of homes closer to the site, what is the expected sign of δ_1? What does it mean if $\beta_1 > 0$?

(ii) *Estimate the model in part (i) and report the results. Interpret the coefficient on $y81\cdot\log(\textit{\textbf{dist}})$. What do you conclude?*

(iii) *Add the following controls to the regression:* ***age***, ***age***^{2}, ***rooms***, ***baths***, $\log(\textit{\textbf{intst}})$, $\log(\textit{\textbf{land}})$, *and* $\log(\textit{\textbf{area}})$. *After adding these variables, what do you conclude about the effect of the incinerator on housing values?*

(iv) *The coefficient on $\log(\textit{\textbf{dist}})$ is positive and statistically significant in part (ii), but not in part (iii). Why does this occur? What does this imply about the importance of the controls used in part (iii)?*