POLS0012 Causal Analysis: Tutorial Exercises Week 9

Question 1

This question uses data from Malesky, Nguyen and Tran's (2014) study of the effect of centralising control of public services in Vietnam.[1] Scholars have long speculated that de-centralised control of public services in developing countries can lead to corruption and 'capture' by local politicians who fail to use tax revenue to invest in public services. Centralisation of control may therefore lead to increased public investment. This is difficult to test because control of government spending rarely changes hands. But in 2009, Vietnam decided to experiment with re-centralising control of its public services in some areas but not others, placing them under central government control instead of control by local district authorities in around one-fifth of districts nationwide. The authors use a difference-in-differences framework to ask what impact this had on infrastructure spending, comparing changes in treated districts (where spending was re-centralised) to untreated districts (where it remained under local control). They have data on infrastructure spending before the change (in 2006 and 2008) and after the change (in 2010), for small areas called 'communes.'

You'll need to install the `lmtest` and `multiwayvcov` packages for this tutorial. The dataset is called "malesky.Rda" and contains the following variables by commune:

- *year*: 2006,2008 or 2010
- *treatment*: =1 if the commune is in a treated district, 0 otherwise
- *infra*: index measuring infrastructure investment ranging from 0 to 5: higher values indicate greater investment
- *lnpopden*: population of commune, logged
- *district*: district ID number

a) Create a new dataset for the years 2008 and 2010 only. In this dataset, create:
(i) a dummy variable called "time" equalling 1 if the year is 2010
(ii) a variable for the interaction between time and treatment

```
m3 <- m[m$year>2006,]
m3$time <- ifelse(m3$year==2010,1,0)

m3$time_treat <- m3$time*m3$treatment
```

[1] Edward Malesky, Cuong Viet Nguyen and Anh Tran (2014). "The Impact of Recentralization on Public Services: A Difference-in-Differences Analysis of the Abolition of Elected Councils in Vietnam." *American Politican Science Review* 108 (1), pp. 144-168.

b) Use the dataset and variables you created in c) to calculate a difference-in-difference estimate of the causal effect of centralisation on infrastructure investment. Interpret the resulting coefficient and its statistical significance.

Code Hint: Remember the formula: $Y_i = \mu + \gamma G_i + \delta T_i + \tau G_i T_i$. Where: $G_i = 1$ for the treatment group and 0 otherwise; $T_i = 1$ for the time period when treatment occurs and 0 otherwise

Code:

```
summary(lm(infra ~ time + treatment + time_treat,data=m3))
```

Answer:

The estimated treatment effect is 0.27 with a p-value of 0.005, meaning that it is statistically significant at the 1% level. It implies that the centralisation of control over spending led to an increase of 0.27 in infrastructure investment, measured on a scale from 0 to 6.

c) Repeat b), this time adding in a control for commune population. Why might it be sensible to include this control variable?

Code:

```
model1 <- lm(infra ~ time + treatment + time_treat + lnpopden,data=m3)
summary(model1)
```

Answer:

This gives a treatment effect of 0.25, slightly smaller than in (b). It is sensible to control for population because it is an obvious potential time-variable confounder. Infrastructure investment may increase in places experiencing rapid population growth. Controlling for it therefore helps make the parallel trends assumption more credible.

d) Because the program was rolled out at the district level, there may be serial correlation in the standard errors across districts. To account for this, estimate cluster-robust standard errors for the model you estimated in c). How different is the estimated p-value for the difference-in-differences causal effect?

Code Hints: Use two stages:

i. Use `cluster.vcov()` in the `multiwayvcov` package to estimate the cluster-robust variance-covariance matrix. Supply the function with the regression model name and cluster name

ii. Plug this matrix into the `coeftest` function in the `lmtest` package. Supply the function with the regression model name and the matrix name

Code:

```
install.packages("lmtest")
install.packages("multiwayvcov")
```

```
library(lmtest)
library(multiwayvcov)

coeftest(model1, cluster.vcov(model1,m3$district))
```

Answer:
The p-value increases more than five-fold from approximately 0.01 in part (c) to 0.055. This is very common when adjusting for clustering

e) Difference-in-differences estimation relies on the 'parallel trends' assumption. Explain what that assumption means in this study.

It means that in the absence of treatment, the treated and untreated groups would have experienced the same changes in the outcome variable. It requires that there be no time-varying confounders; confounders that are time-invariant do not matter in this framework.

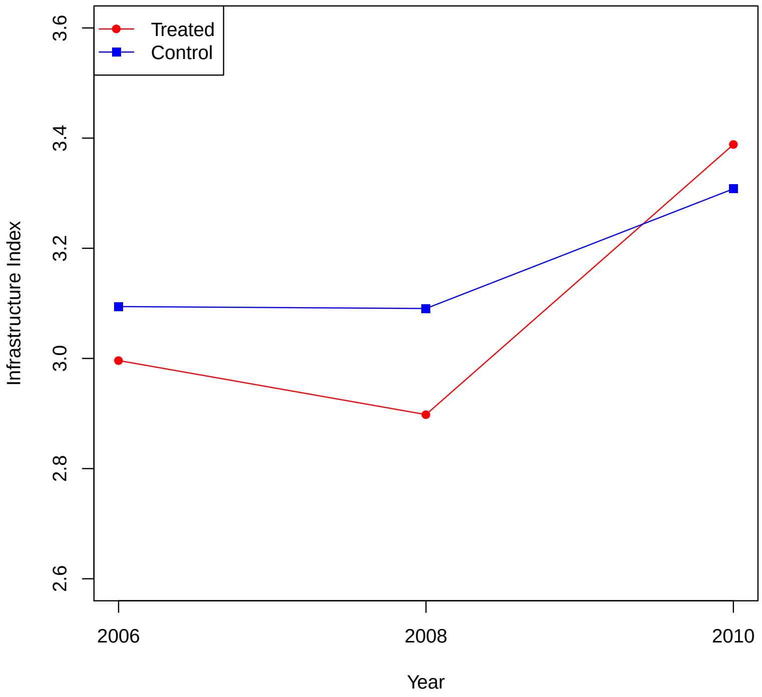
f) Now, we'll assess the parallel trends assumption graphically. Return to the full dataset and estimate means of the *infra* variable for 2006, 2008 and 2010 for both the treated and control groups (six means in total). Plot these means separately over time for the treated and control groups. Do you think that the parallel trends assumption is satisfied here?

```
  # means, treated
means.t <- c(mean(m$infra[m$year==2006&m$treatment==1]),
mean(m$infra[m$year==2008&m$treatment==1]),
mean(m$infra[m$year==2010&m$treatment==1]))

  #means,control
means.c <- c(mean(m$infra[m$year==2006&m$treatment==0]),
mean(m$infra[m$year==2008&m$treatment==0]),
mean(m$infra[m$year==2010&m$treatment==0]))

  # plot
plot(means.t,
     ylim=c(2.6,3.6),
     type="o",
     pch=16,
     col="red",
     xaxt="n",
     xlab="Year",
     ylab="Infrastructure Index")
lines(means.c,type="o",pch=15,col="blue")
```

```
axis(1,at=c(1,2,3),lab=c(2006,2008,2010))
legend("topleft",
       c("Treated","Control"),
       col=c("red","blue"),
       pch=c(16,15),
       lty=c(1,1))
```

Answer:
See Figure 1. Although not completely parallel, there is not much evidence of divergence between the treated and untreated groups prior to treatment

Figure 1: **Trends in the Outcome Variable for Treated and Untreated Units**



g) Create a new dataset for the years 2006 and 2008 only, and use it to estimate a placebo difference-in-differences effect before the treatment occurred. What do you conclude about the parallel trends assumption?
Code Hints: You'll need to create variables again. Don't worry about cluster-robust standard errors (it is not possible because the authors don't supply district IDs before 2008)

Code:

```
m4 <- m[m$year<2010,]
```

```
m4$time <- ifelse(m4$year==2008,1,0)
m4$time_treat <- m4$time*m4$treatment

summary(lm(infra ~ time + treatment + time_treat + lnpopden ,data=m4))
```

Answer:
The placebo treatment effect is close to zero and no longer statistically significant (with clustering it would be even less significant). This provides a more formal statistical justification for the parallel trends assumption shown in (f)

Question 2

In this question, we'll revisit Leah Stokes' study of Canadian wind farms that we introduced in Tutorial 7, on instrumental variables. Recall that she examines the impact of wind farms on support for the incumbent party. Wind farms were built in some districts but not others between 2007 and 2011. In her paper, she actually uses fixed effects estimation as well as instrumental variables. For the fixed effects part of her study, she has data on electoral districts for the years 2003, 2007 and 2011.

You'll need to install the `plm` package for this question. The dataset is called "Stokes_fe.Rda" and contains the following variables by district-year:

- *op*: =1 if a wind farm was operational in the district-year observation, 0 otherwise
- *perc_lib*: Percentage vote for the incumbent (liberal) party
- *year*: 2003, 2007 and 2011
- *master_id*: ID variable indicating district
- *treat_o*: =1 if a wind farm was ever operational in the district regardless of year, 0 otherwise
- *p_uni_degree*: proportion of district with a university degree
- *log_pop_denc*: district population density, logged
- *unemploy_rate*: district unemployment rate
- *log_median_inc*: district median income. logged
- *p_immigrant*: proportion of immigrants in district

a) For this study, what is:

   i) The treatment?

   ii) The outcome?

   iii) The group variable for fixed effects?

i) *op*

ii) *perc_lib*

iii) *master_id*

b) Using the framework of fixed effects estimation and the code provided in the slides/lecture notes, estimate a suitable model for the causal effect of wind farms on support for the incumbent party. Explain how you chose your model. Carefully interpret the resulting causal effect and its statistical significance

```
install.packages("plm")
library(plm)
mod_fe2 <- plm(perc_lib ~ op + p_uni_degree + log_pop_denc + unemploy_rate +
                      log_median_inc +  p_immigrant,
              data = s, index = c("master_id", "year"), effect = "twoways")
coeftest(mod_fe2, vcov=vcovHC(mod_fe2, cluster="group", type="HC1"))
```

Answer:

There is no good reason not to include *all* the time-varying controls, here, since they make the parallel trends assumption more plausible. The results suggest that a wind farm becoming operational in the district led to a 9 percentage-point drop in support for the incumbent party

c) What is the key assumption needed for valid estimation of this causal effect? In theory (without doing any estimation), how reasonable do you think this assumption is?

Answer:

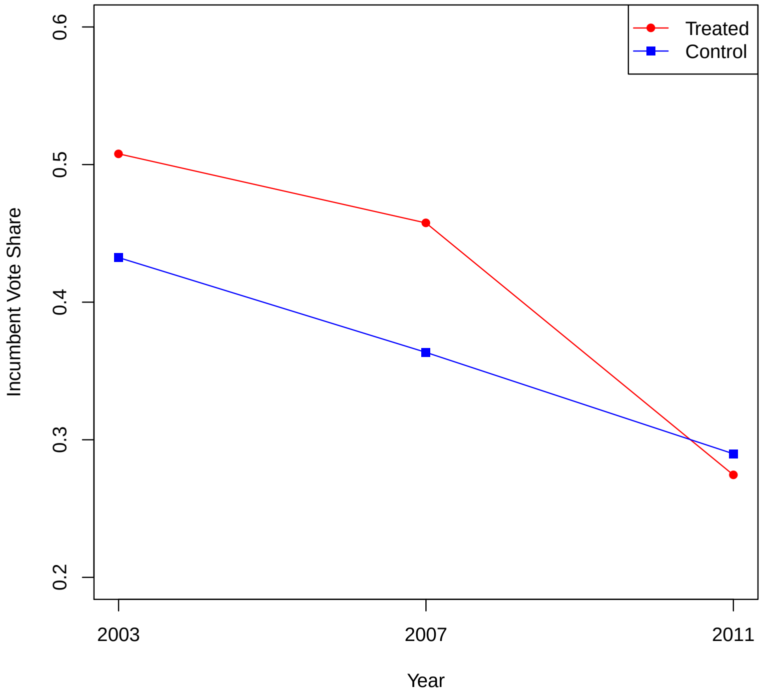
The key assumption needed is parallel trends between areas that did and did not receive wind farms. This ensures that the untreated areas provide a valid counterfactual for the treated areas. It is difficult to say how plausible this assumption is without more formal tests, but it is certainly helped by the inclusion of a large number of time-varying confounders that probably have a strong impact on voting, like median income.

d) Using a graphical approach, assess whether or not you think the assumption in c) is satisfied
Hint: Think carefully about what to use as your treatment variable

```
  # means, treated
means.t <- c(mean(s$perc_lib[s$year==2003&s$treat_o==1]),
             mean(s$perc_lib[s$year==2007&s$treat_o==1]),
             mean(s$perc_lib[s$year==2011&s$treat_o==1]))

#means,control
means.c <- c(mean(s$perc_lib[s$year==2003&s$treat_o==0]),
             mean(s$perc_lib[s$year==2007&s$treat_o==0]),
             mean(s$perc_lib[s$year==2011&s$treat_o==0]))

# plot
plot(means.t,
     ylim=c(0.2,0.6),
     type="o",
     pch=16,
     col="red",
     xaxt="n",
     xlab="Year",
     ylab="Incumbent Vote Share")
lines(means.c,type="o",pch=15,col="blue")
axis(1,at=c(1,2,3),lab=c(2003,2007,2011))
legend("topright",
       c("Treated","Control"),
       col=c("red","blue"),
       pch=c(16,15),
       lty=c(1,1))
```

Answer:
The treated and untreated districts have near-perfect parallel trends before treatment. Note that the treatment and control groups are defined for all periods of the data, so you need to use the *treat_o* variable.

Figure 2: **Trends in the Outcome Variable for Treated and Untreated Units**


Treated
Control
Incumbent Vote Share
Year