Pol Sci 630: Problem Set 11 - Causal Inference Techniques - Differences-in-differences

Prepared by: Jan Vogler (jan.vogler@duke.edu)

Grading Due Date: Friday, November 18th, 1.40 PM (Beginning of Lab)

Insert your comments on the assignment that you are grading above the solution in bold and red text. For example write: "GRADER COMMENT: everything is correct!" Also briefly point out which, if any, problems were not solved correctly and what the mistake was. See below for more examples.

In order to make your text bold and red, you need to insert the following line at the beginning of the document:

```
\usepackage{color}
```

and the following lines above the solution of the specific task:

```
\textbf{\color{red} GRADER COMMENT: everything is correct!}
```

R Programming

Problem 1

```
setwd("C:/Users/Jan/OneDrive/Documents/GitHub/ps630_lab/ps630_f16/W5/")
library(foreign)
vote1 = read.dta("VOTE1.dta")
summary(vote1)

library(reporttools)

## Loading required package:  xtable

varsVote <- vote1[, c("voteA", "expendA", "expendB", "prtystrA", "lexpendA",
    "lexpendB", "shareA")]
capVote <- "Descriptive Statistics: Vote Dataset"
tableContinuous(vars = varsVote, cap = capVote, lab = "tab: cont1", longtable = F,
    prec = 2)
```

Variable	n	Min	q_1	$\tilde{x}$	$\bar{x}$	q_3	Max	s	IQR	#NA
voteA	173	16.00	36.00	50.00	50.50	65.00	84.00	16.78	29.00	0
expendA	173	0.30	81.63	242.78	310.61	457.41	1470.67	280.99	375.78	0
expendB	173	0.93	60.05	221.53	305.09	450.72	1548.19	306.28	390.66	0
prtystrA	173	22.00	44.00	50.00	49.76	56.00	71.00	9.98	12.00	0
lexpendA	173	-1.20	4.40	5.49	5.03	6.13	7.29	1.60	1.72	0
lexpendB	173	-0.07	4.10	5.40	4.94	6.11	7.34	1.57	2.02	0
shareA	173	0.09	18.87	50.85	51.08	84.26	99.50	33.48	65.39	0

Table 1: Descriptive Statistics: Vote Dataset

Problem 2

```
setwd("C:/Users/Jan/OneDrive/Documents/GitHub/ps630_lab/ps630_f16/W11/")

load("Data2015.Rdata")
load("Data2016.Rdata")
```

```r
# We need to append the second dataset to the first one In order to do this,
# we first need to

YearComplete = c(Data2015$Year, Data2016$Year)
RestOwnComplete = c(Data2015$RestOwn, Data2016$RestOwn)
PrivExpComplete = c(Data2015$PrivExp, Data2016$PrivExp)

complete = cbind(YearComplete, RestOwnComplete, PrivExpComplete)
complete = as.data.frame(complete)
colnames(complete) = c("Year", "RestOwn", "PrivExp")

# Create a new variable that shows if there have been tax reductions

complete$TaxReduc = NA
complete$TaxReduc[complete$Year == 2015] = 0
complete$TaxReduc[complete$Year == 2016] = 1

summary(complete)
```

```
##       Year         RestOwn       PrivExp         TaxReduc
##  Min.   :2015   Min.   :0.0   Min.   : 7511   Min.   :0.0
##  1st Qu.:2015   1st Qu.:0.0   1st Qu.:22954   1st Qu.:0.0
##  Median :2016   Median :0.5   Median :26655   Median :0.5
##  Mean   :2016   Mean   :0.5   Mean   :26628   Mean   :0.5
##  3rd Qu.:2016   3rd Qu.:1.0   3rd Qu.:30326   3rd Qu.:1.0
##  Max.   :2016   Max.   :1.0   Max.   :44663   Max.   :1.0
```

```r
did_model = lm(PrivExp ~ RestOwn + TaxReduc + RestOwn * TaxReduc, data = complete)

summary(did_model)
```

```
##
## Call:
## lm(formula = PrivExp ~ RestOwn + TaxReduc + RestOwn * TaxReduc,
##     data = complete)
```

```
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -17160  -3509    132   3397  16744
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       24689.3      219.4 112.521  < 2e-16 ***
## RestOwn             -18.0      316.7  -0.057    0.955
## TaxReduc           2759.2      316.7   8.712  < 2e-16 ***
## RestOwn:TaxReduc   2185.3      447.9   4.879 1.15e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5004 on 1996 degrees of freedom
## Multiple R-squared:  0.1491,Adjusted R-squared:  0.1478
## F-statistic: 116.6 on 3 and 1996 DF,  p-value: < 2.2e-16
```

Interpretation: the regression model we used to estimate the differences-in-differences shows that—if the parallel trends assumption holds—there is a difference of 2185.3 Z-Dollars in the private expenditures of restaurant owners versus all other small business owners. Given that the only exogenous change we are aware of is the change in tax policy, we would attribute this difference to the change in tax policy. The difference is statistically significant at a level of $p < 0.001$ and thus at all common levels of statistical significance.

The results appear to confirm that a tax reduction for small business owners could have a positive impact on their private expenditures. Insofar, our statistical results would lead to a recommendation of the tax reduction if the government wants to achieve an increase in the private expenditures of small business owners.

Problem 3

a) Differences-in-differences is a powerful tool for causal inference if its assumptions are true. The most crucial assumption in the simple diff-in-diff framework is the parallel trends assumption. Why is this the case? We compare the characteristics of two groups over time. In the vast majority of cases, the characteristic of interest is not static but moves. The

parallel trends assumption means *that the characteristic of interest does not trend for the two groups in a systematically different way.* Note that there may still be some degree of random error for individual observations but as long as we have a sufficiently large number of observations drawn at random from the population, this random error will not bias the results. The parallel trends assumption is such an important part of diff-in-diff because we *approximate* (however, we do never perfectly meet the requirements of) a randomized experiment in which we have one control group and one treatment group.

Most importantly, if the parallel trends assumption is true for the population and individual deviations are caused by random error only, the group not affected by the treatment can *plausibly function as (near-)counterfactual* to the group that is affected by it.

In other words, when the parallel trends assumption is true, the comparison of group differences at time t-1 and t reveals the effect of the exogeneous treatment that only affects one of the two groups because both groups are subject to the same intertemporal dynamics in all other respects. Accordingly, the following mathematical expression captures the magnitude of the difference:

$$\delta = (\bar{x}_{t2} - \bar{y}_{t2}) - (\bar{x}_{t1} - \bar{y}_{t1})$$

Where x is the treatment group and y is the control group and $t1$ and $t2$ indicate time points 1 and 2 respectively.

b) Without any additional model features or assumptions, when the parallel trends assumption is violated, the difference in differences could be caused by diverging intertemporal dynamics in the characteristic of interest. This means that the effect cannot be effectively reduced to the treatment by which only one group is affected. Because we cannot distinguish between the impact of the treatment and the impact of other unobserved factors, we can no longer make an inference about the magnitude of the treatment effect. Note that this also means that we have to be very confident in the parallel trends assumption to make causal claims based on a simple diff-in-diff framework.

c) If we have three time points per group that we analyze, we can estimate time trends that are group-specific. As long as the time trend of the treatment group is correctly captured by at least two points in time before the treatment, we can still make claims about the effect that the treatment had. We can technically catpure such a trend through an interaction of unit dummies and time variables. Therefore, these individual time trends allow us to relax

the parallel trends assumption. We merely need a stable trend in the analyzed groups.