PSET 9

Sarah Van Alsten

3/30/2020

1. Card and Krueger

a. The difference-in-difference estimator would be defined as the difference in observed outcomes for the treated at time 1 versus 0 minus the difference in observed outcomes for the control group at time 1 versus 0: $(\beta - \delta) - (\alpha - \gamma)$. The estimator represents the estimated effect of the treatment on the treated group as the portion of the total change in the treatment group that is left unexplained by typical (ie control) changes over time.

b. $\alpha = E[Y_{i0}(0)|D_i = 1]$ $\beta = E[Y_{i1}(1)|D_i = 1]$ $\delta = E[Y_{i1}(0)|D_i = 0]$ $\gamma = E[Y_{i0}(0)|D_i = 0]$

Where Y_{i0} denotes observation Y_i at time 0, and Y_{i1} denotes observation Y_i at time 1.

The D.I.D. estimator then becomes:

$$(E[Y_{i1}(1)|D_i = 1] - E[Y_{i1}(0)|D_i = 0]) - (E[Y_{i0}(0)|D_i = 1] - E[Y_{i0}(0)|D_i = 0])$$

Which can also be written as:

$$(E[Y_{i1}(1)|D_i = 1] - E[Y_{i0}(0)|D_i = 1]) - (E[Y_{i1}(0)|D_i = 0] - E[Y_{i0}(0)|D_i = 0])$$

The parallel trends assumption is needed to identify the ATT as the difference-in-difference estimator: we need to be able to use the trend in the control group to impute the potential outcomes under control at time 1 for the treated group and this can only be done if the magnitude of change is expected to be the same for the two groups (which would mean $(E[Y_{i1}(0)|D_i = 1] - E[Y_{i0}(0)|D_i = 1]) = (E[Y_{i1}(0)|D_i = 0] - E[Y_{i0}(0)|D_i = 0])$. Assuming parallel trends, the estimator can be expressed as:

$$(E[Y_{i1}(1)|D_i = 1] - E[Y_{i0}(0)|D_i = 1]) - (E[Y_{i1}(0)|D_i = 1] - E[Y_{i0}(0)|D_i = 1])$$

$$= E[Y_{i1}(1)|D_i = 1] - E[Y_{i1}(0)|D_i = 1]$$

And this expression is exactly what we want to estimate.

c. No, we cannot find evidence for this assumption in the data. We only have data for two timepoints: right before the intervention and in the period after, so we cannot evaluate whether trends paralleled in the months or years before the intervention.

d.

```
ck.summary <- ck %>%
  group_by(state) %>% #group by d; 0 = PA, 1 = NJ
  summarise(February = mean(emp.pre, na.rm = T),
            November = mean(emp.post, na.rm = T))

ck.summary
```

```
## # A tibble: 2 x 3
##   state        February November
##   <fct>           <dbl>    <dbl>
## 1 New Jersey       20.4     21.0
## 2 Pennsylvania     23.3     21.2
```

```
#var/cov matrix
ck$nj_pre <- ifelse(ck$state == "New Jersey", ck$emp.pre, NA)
ck$pa_pre <- ifelse(ck$state != "New Jersey", ck$emp.pre, NA)
ck$nj_post <- ifelse(ck$state == "New Jersey", ck$emp.post, NA)
ck$pa_post <- ifelse(ck$state != "New Jersey", ck$emp.post, NA)

cov(ck[, c("nj_pre", "nj_post", "pa_pre", "pa_post")],
    use = "pairwise.complete.obs")
```

```
##           nj_pre  nj_post   pa_pre  pa_post
## nj_pre  82.92359 50.72446       NA       NA
## nj_post 50.72446 86.36029       NA       NA
## pa_pre        NA       NA 140.5714 48.30230
## pa_post       NA       NA  48.3023 68.50429
```

The diagonals of the matrix represent the variances of each parameter (namely, the variances in unemployment (Y) at time 0/time 1 for NJ/PA) and are calculated as:

$$\frac{\sum(Y_{dt} - \overline{Y_{dt}})^2}{n_{dt} - 1} \text{ for d = 0,1; t = 0,1}$$

The covariances of nj_pre/nj_post and pa_pre/pa_post represent the correspondence in the unemployment numbers at a firm for a given NJ or PA firm between February (X) and November(Y) periods and are calculated as:

$$\frac{\sum(Y_d - \overline{Y_d})^2(X_d - \overline{X_d})^2}{n_d - 1} \text{ for d = 0,1}$$

```
#estimate of the ATT
att <- (ck.summary[1,3] - ck.summary[1,2]) - (ck.summary[2,3] - ck.summary[2,2])
att[1,1]
```

```
## [1] 2.753606
```

To estimate the standard error of the difference in difference, we can use the formula for the variance of a difference: Var(a-b) = var(a) + var(b) - 2(cov(a,b)) Therfore, for the estimator

$$\begin{aligned}&(\beta - \delta) - (\alpha - \gamma)\\ &= (\beta - \alpha) - (\delta - \gamma)\end{aligned}$$

We can do the following:

$$Var(\beta - \alpha) = Var(\beta) + Var(\alpha) - 2cov(\beta, \alpha)$$
$$Var(\delta - \gamma) = Var(\delta) + Var(\gamma) - 2cov(\delta, \gamma)$$

And the variance of the difference:

$$\begin{aligned}&Var((\beta - \alpha) - (\delta - \gamma))\\ &= Var(\beta) + Var(\alpha) - 2cov(\beta, \alpha) + Var(\delta) + Var(\gamma) - 2cov(\delta, \gamma) - 2cov((\beta - \alpha), (\delta - \gamma))\end{aligned}$$

Because of the bilinear property of covariance, the last term in the equation can be separated out further:

$$
\begin{aligned}
&Var((\beta - \alpha) - (\delta - \gamma)) \\
&= Var(\beta) + Var(\alpha) - 2cov(\beta, \alpha) + Var(\delta) + Var(\gamma) - 2cov(\delta, \gamma) - 2cov((\beta - \alpha), (\delta - \gamma)) \\
&= Var(\beta) + Var(\alpha) - 2cov(\beta, \alpha) + Var(\delta) + Var(\gamma) - 2cov(\delta, \gamma) - 2(cov((\beta, \delta) - cov(\beta, \gamma) - cov(\alpha, \delta) + cov(\alpha, \gamma)) \\
&= Var(\beta) + Var(\alpha) - 2cov(\beta, \alpha) + Var(\delta) + Var(\gamma) - 2cov(\delta, \gamma) - 2cov(\beta, \delta) + 2cov(\beta, \gamma) + 2cov(\alpha, \delta) - 2cov(\alpha, \gamma)
\end{aligned}
$$

Assuming independence between the pre-period employment in one state and the post-period employment in the other state, we can assume that $cov(\alpha, \delta)$ and $cov(\beta, \gamma)$ are 0:

$$= Var(\beta) + Var(\alpha) - 2cov(\beta, \alpha) + Var(\delta) + Var(\gamma) - 2cov(\delta, \gamma) - 2cov(\beta, \delta) - 2cov(\alpha, \gamma)$$

We assume there is 0 covariance between units which simplifies the above to

$$Var(\beta) + Var(\alpha) - 2cov(\beta, \alpha) + Var(\delta) + Var(\gamma) - 2cov(\delta, \gamma)$$

```
#using the vars and covs estimated above, the variance of the DID is:
82.92359 + 86.36029 - 2*50.72446 +140.5714 + 8.50429 - 2*48.3023
```

```
## [1] 120.306
```

```
#and the se is
sqrt(82.92359 + 86.36029 - 2*50.72446 +140.5714 + 8.50429 - 2*48.3023)
```

```
## [1] 10.96841
```

```
#the 2 sided pvalue is
pnorm(q = att[1,1],
      sd = sqrt(82.92359 + 86.36029 - 2*50.72446 +140.5714 + 8.50429 - 2*48.3023),
      lower.tail = F) *2
```

```
## [1] 0.8017764
```

The ATE(se) is 2.75 (10.97) and is not significant (p = 0.80).

Question 2. Malesky et al.

a.

```
df <- as.data.frame(rbind(c("$\\gamma + \\theta + \\beta + \\alpha + \\delta$",
                            "$\\gamma + \\alpha + \\delta$", "$\\beta + \\theta$"),
                          c("$\\beta + \\alpha + \\delta$","$\\alpha + \\delta$","$\\beta$"),
                          c("$\\gamma + \\theta$","$\\gamma$","$\\theta$")))

names(df) <- c("After", "Before", "After Minus Before")
rownames(df) <- c("Treated", "Control", "Treated Minus Control")
knitr::kable(df, escape = FALSE, caption = "Meaning of Parameters",
             format = "latex", booktabs = TRUE)
```

b.

Table 1: Meaning of Parameters

	After	Before	After Minus Before
Treated	$\gamma + \theta + \beta + \alpha + \delta$	$\gamma + \alpha + \delta$	$\beta + \theta$
Control	$\beta + \alpha + \delta$	$\alpha + \delta$	β
Treated Minus Control	$\gamma + \theta$	γ	θ

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Fri, Apr 03, 2020 - 10:36:10 AM

	Dependent variable:					
	Infrastructure	Agriculture	Health	Education	Communications	HH Business
	(1)	(2)	(3)	(4)	(5)	(6)
Year = 2010	0.224***	0.042	−0.014	0.075***	−0.046**	−0.011
	(0.053)	(0.043)	(0.017)	(0.028)	(0.022)	(0.032)
Treatment	−0.269**	0.074	−0.013	0.057	−0.197**	−0.034
	(0.115)	(0.114)	(0.022)	(0.090)	(0.084)	(0.160)
Region 2	0.116	−0.327**	0.057	0.113	−0.447***	−0.062
	(0.111)	(0.157)	(0.040)	(0.069)	(0.132)	(0.192)
Region 3	0.041	−0.621***	0.349***	0.360***	−0.693***	−0.418**
	(0.168)	(0.166)	(0.119)	(0.114)	(0.080)	(0.174)
Region 4	0.216	−0.003	0.036	0.080	−0.250***	−0.032
	(0.176)	(0.132)	(0.028)	(0.081)	(0.084)	(0.169)
Region 5	0.248**	−0.340***	0.035	−0.156	0.012	0.314
	(0.113)	(0.116)	(0.029)	(0.128)	(0.075)	(0.246)
Region 7	0.631***	−1.152***	−0.062**	−0.080	−0.031	0.020
	(0.120)	(0.180)	(0.028)	(0.109)	(0.060)	(0.166)
Region 8	−0.006	−0.530***	0.003	−0.261***	−0.097*	−0.120
	(0.133)	(0.147)	(0.027)	(0.080)	(0.056)	(0.156)
Log Area	0.170***	0.102*	−0.078***	0.231***	0.032	0.368***
	(0.060)	(0.058)	(0.023)	(0.045)	(0.046)	(0.073)
Log Population Density	0.313***	0.061	−0.130***	0.200***	0.089*	0.454***
	(0.052)	(0.053)	(0.020)	(0.041)	(0.047)	(0.072)
City	0.126	−0.001	0.030*	0.236**	−0.022	0.189
	(0.103)	(0.077)	(0.018)	(0.092)	(0.041)	(0.190)
Treatment*Year	0.225*	−0.003	0.123***	0.091	0.152**	0.007
	(0.129)	(0.109)	(0.033)	(0.091)	(0.076)	(0.100)
Constant	1.039***	2.281***	1.019***	−0.017	1.705***	−1.253**
	(0.372)	(0.399)	(0.157)	(0.316)	(0.354)	(0.562)
Observations	4,126	4,126	4,126	4,126	4,126	4,126
R^2	0.099	0.124	0.139	0.039	0.131	0.116
Adjusted R^2	0.097	0.122	0.137	0.037	0.128	0.113
Residual Std. Error (df = 4113)	1.006	0.889	0.386	0.853	0.664	1.021
F Statistic (df = 12; 4113)	37.855***	48.560***	55.488***	14.094***	51.633***	44.861***

Note: *p<0.1; **p<0.05; ***p<0.01

For all indicies except agriculture, recentralization had a postive effect on outcomes, with effects only being significant in infrastructure, health, and communications. The effect of recentralization was strongest in infrastructure, as the policy change led to an increase of .225 in the infrastructure index compared to increases of .123 and .152 in health and communications, respectively.

c. The key assumption to identify the causal effect of recentralization is [conditional] parallel trends, which implies that the trends in indicies for treated and non-treated groups will be the same during

the pretreatment period after controlling for the covariates of region, log area, log population density, and city. These confounders also have to act additively and be time-invariant, meaning that there aren't multiplicative effects of the confounders and that confounders themselves do not change over time. We also assume no omitted variable bias when we condition on these covariates, and that none of them open up a backdoor path/are colliders. In the context of the experiment, one confounder that could violate this assumption would be log population- people may be moving into/out of an area during the time in which the intervention took place. This is an issue in that it could change the area's population, and in that migration effects could also account for observed differences rather than the intervention alone (individuals who stand to gain from recentralization might move to areas being recentralized and drive up the observed effect on various indicies). Looking back to part a, time invariance is critical because if it does not hold, then the delta's for the control and treatment group do not cancel out and the estimate of theta is biased.

d.

```
#create parallel trend plots

#combine 2006 placebo data with real 2008-2010 data
mal.all <- mal.p %>%
  filter(year == "2006") %>%
  #add the real data
  rbind(mal)

mal.all %>%
  #restructure data so index + its corresponding value are 2 separate columns
  pivot_longer(cols = c("index1", "index2", "index3", "index4", "index5", "index6"),
               names_to = "index_type", values_to = "value") %>%
  #give indices informative names
  mutate(index_type = ifelse(index_type == "index1", "Infrastructure",
                             ifelse(index_type == "index2", "Agriculture",
                                    ifelse(index_type == "index3", "Health Care",
                                           ifelse(index_type == "index4", "Education",
                                                  ifelse(index_type == "index5", "Communications",
                                                         "Household Business")))))) %>%
  #plot
  ggplot(aes(x = year, y = value, group = factor(treatment), color = factor(treatment)))+
  #the mean = point, 95% CI = error bar, line connecting
  stat_summary(fun = mean, geom = "point") +
  stat_summary(fun.data = mean_cl_normal, geom = "errorbar", width = .2) +
  stat_summary(fun= mean, geom = "line") +
  #add vertical line indicating when tx occurred
  geom_vline(xintercept = 2009, linetype = "dashed") +
  theme_bw() +
  #split apart by index and let y-axis vary
  facet_wrap(~index_type, scales = "free_y") +
  labs(x = "Year", y = "Index Value", color = "Treatment")
```

Based on the plots, it looks like there were parallel trends in agriculture, education, and household business. Trends in healthcare and communications clearly diverge between 2008 and 2009. There also seems to be a sharper increase in infrastructure for treated units than control units between 2008 and 2009, but this differnce in slopes is less apparent than for healthcare and communications, in which slopes are of sufficiently different magnitude that the trends for treatment and control "cross" prior to intervention.

e.

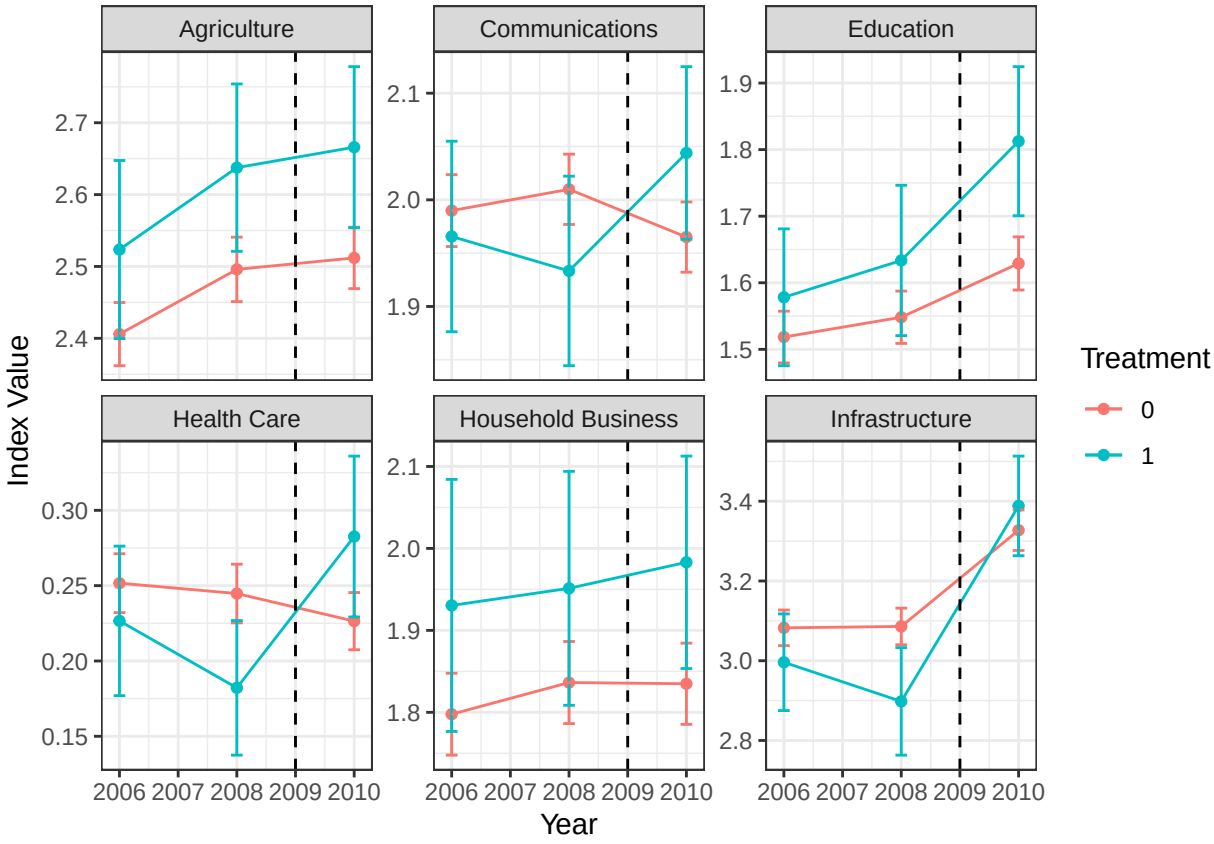

Figure 1: Trends in 6 indicies pre and post treatment

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Fri, Apr 03, 2020 - 10:36:17 AM

	Dependent variable:					
	Infrastructure	Agriculture	Health	Education	Communications	HH Business
	(1)	(2)	(3)	(4)	(5)	(6)
Year = 2008	−0.0002	0.083**	−0.007	0.022	0.025	0.032
	(0.031)	(0.036)	(0.016)	(0.028)	(0.022)	(0.033)
Treatment	−0.158	−0.041	0.021	−0.002	−0.146*	−0.033
	(0.135)	(0.141)	(0.031)	(0.101)	(0.085)	(0.162)
Region 2	0.0001	−0.359**	0.092**	0.050	−0.486***	−0.155
	(0.115)	(0.165)	(0.039)	(0.106)	(0.135)	(0.212)
Region 3	−0.022	−0.543***	0.384**	0.238*	−0.740***	−0.451**
	(0.233)	(0.188)	(0.151)	(0.134)	(0.076)	(0.195)
Region 4	0.006	−0.042	0.017	−0.045	−0.235**	0.019
	(0.136)	(0.180)	(0.031)	(0.102)	(0.099)	(0.163)
Region 5	0.087	−0.429***	0.050*	−0.218	0.008	0.217
	(0.105)	(0.109)	(0.030)	(0.137)	(0.089)	(0.265)
Region 7	0.444***	−1.076***	−0.015	−0.112	−0.004	0.056
	(0.101)	(0.163)	(0.032)	(0.142)	(0.063)	(0.155)
Region 8	−0.155	−0.670***	0.002	−0.363***	−0.058	−0.206
	(0.124)	(0.177)	(0.029)	(0.105)	(0.060)	(0.147)
Log Area	0.108*	0.121	−0.089***	0.234***	−0.024	0.408***
	(0.065)	(0.077)	(0.021)	(0.048)	(0.051)	(0.083)
Log Population Density	0.198***	0.109	−0.132***	0.202***	0.041	0.501***
	(0.064)	(0.067)	(0.020)	(0.041)	(0.049)	(0.086)
City	0.095	0.101	0.035	0.187	−0.006	0.148
	(0.205)	(0.112)	(0.032)	(0.146)	(0.055)	(0.124)
Treatment*Year	−0.114*	0.080	−0.033	0.033	−0.051	−0.024
	(0.069)	(0.144)	(0.033)	(0.050)	(0.051)	(0.074)
Constant	1.866***	1.964***	1.058***	0.022	2.050***	−1.565**
	(0.454)	(0.513)	(0.152)	(0.311)	(0.371)	(0.647)
Observations	4,220	4,220	4,220	4,220	4,220	4,220
R^2	0.049	0.116	0.139	0.035	0.129	0.123
Adjusted R^2	0.046	0.114	0.137	0.032	0.127	0.120
Residual Std. Error (df = 4207)	0.978	0.919	0.393	0.846	0.683	1.043
F Statistic (df = 12; 4207)	17.999***	46.132***	56.717***	12.538***	52.044***	49.069***
Note:						*p<0.1; **p<0.05; ***p<0.01

This suggests that there were no significant violations in parallel trends (all interaction p-values < 0.05) for any of the 6 indicies measured, though it should be noted that there may be slight differences in the slopes for the treatment vs. control group in the infrastructure index ($p < 0.1$). The difference between these results and those shown in c may be accounted for by the addition of covariates in the OLS models here versus

the unconditional means estimated in c. In this case, *conditional* parallel trends hold, whereas *unconditional* seem not to have. Given that the conditional parallel trends assumption holds, the DID estimates in part b can be interpreted as ATEs, meaning that the increases in indices observed does appear to be a result of recentralization.