Causal Inference
Problem Set 4

Due Monday May 14th

Problem 1

Consider the linear model $Y_i = \tau_X D_i + \beta X_i + \epsilon_i$; i.e., a model in which the treatment effect varies by strata of X (heterogenous treatment effects).

(a) Write down an expression for the $\hat{\tau}_{OLS}$ obtained by regressing Y on D and X.

$$\hat{\theta}_{OLS} = (Z'Z)^{-1}Z'Y,$$

with $\hat{\theta}_{OLS} = \begin{pmatrix} \hat{\tau}_{OLS} \\ \hat{\beta}_{OLS} \end{pmatrix}$, and $Z = \begin{bmatrix} D & X \end{bmatrix}$

Recall the least squares normal equation:
$Z'Z\hat{\theta} = Z'y$

But now consider Z as comprised of D and X,

$$Z'Z = \begin{bmatrix} D' \\ X' \end{bmatrix} \begin{bmatrix} D & X \end{bmatrix}$$

And so the least squares normal equation becomes the following set of equations:

$$\begin{bmatrix} D'D & D'X \\ X'D & X'X \end{bmatrix} \begin{bmatrix} \hat{\tau} \\ \hat{\beta} \end{bmatrix} = \begin{bmatrix} D'y \\ X'y \end{bmatrix}$$

This yields the following equation:

$$D'D\hat{\tau} + D'X\hat{\beta} = D'y$$

which can be rearranged to solve for $\hat{\tau}$:

$$\hat{\tau} = (D'D)^{-1}D'y - (D'D)^{-1}D'X\hat{\beta}$$

(b) Write down an expression for the $\hat{\tau}_{OLS}$ obtained from the bivariate regression of Y on $\tilde{D}$, where $\tilde{D}$ is defined as $M_X D$ (i.e. the residuals from regressing D on X). Your expression should be a function only of Y and $\tilde{D}$.

$$\hat{\tau}_{OLS} = (\tilde{D}'\tilde{D})^{-1}\tilde{D}'Y,$$

with $\tilde{D} = M_X D$, the residuals from regressing D on X.

Furthermore, because $\tilde{D}$ has mean 0 (by construction),

$$\tilde{D}'\tilde{D} = N \cdot Var(\tilde{D})$$

and

$$\tilde{D}'Y = N \cdot Cov(Y, \tilde{D})$$

and thus

$$\hat{\tau}_{OLS} = \frac{N \cdot Cov(Y, \tilde{D})}{N \cdot Var(\tilde{D})} = \frac{Cov(Y, \tilde{D})}{Var(\tilde{D})}$$

(c) Show that the expressions from parts (a) and (b) are equivalent. (Hint: Recall the FWL theorem.)

This is shown by the FWL theorem.

From the least squares normal equations, we have:

$$X'D\hat{\tau} + X'X\hat{\beta} = X'y$$

which yields for $\hat{\beta}$:

$$\hat{\beta} = (X'X)^{-1}X'y - (X'X)^{-1}X'D\hat{\tau}$$

As shown above, the least squares normal equations also yield:

$$D'D\hat{\tau} + D'X\hat{\beta} = D'y$$

Combining the two equations (i.e. plugging in for $\hat{\beta}$), we get:

$$D'X(X'X)^{-1}X'y - D'X(X'X)^{-1}X'D\hat{\tau} + D'D\hat{\tau} = D'y$$

which after some manipulation, becomes:

$$\hat{\tau} = [D'(I - X(X'X)^{-1}X')D]^{-1}[D'(I - X(X'X)^{-1}X')y]$$

where $(I - X(X'X)^{-1}X')$ is the residual-maker matrix for the column space of X—i.e. when multiplied by another vector or matrix, it will yield the residuals for regression of that vector or each of the columns of the other matrix on X. M_X is also idempotent. Hence,

$$\hat{\tau} = [D'(I - X(X'X)^{-1}X')D]^{-1}[D'(I - X(X'X)^{-1}X')y]$$

$$= [D'M_XD]^{-1}[D'M_Xy] = [D'M_X'M_XD]^{-1}[D'M_X'y] = (\tilde{D}'\tilde{D})^{-1}\tilde{D}'y$$

All of this is to say that the OLS estimator of τ ($\hat{\tau}$) in the complete regression containing both D and X, is the same as the estimate of τ for a regression of just y on $\tilde{D}$ (i.e. in the latter case, D has had the variance it shared with X already partialed out, and thus D has become $\tilde{D}$).

(d) Beginning with the expression for $\hat{\tau}_{OLS}$ from part (b), derive an expression for $\hat{\tau}_{OLS}$ in terms of a weighted sum of the τ_X's (i.e. the stratum-specific treatment effects). If you have not yet done so, it will be helpful to write your expression from part (b) in non-matrix terms, which should be simple given its bivariate form. You will then need to algebraically manipulate this expression in order to obtain a weighted sum of the τ_X's. What are the weights?

$$\begin{aligned}
\hat{\tau}_{OLS} &= \frac{Cov(Y, \tilde{D})}{Var(\tilde{D})} \\
&= \frac{Cov(Y, D - E[D|X])}{E(D - E[D|X])^2} \\
&= \frac{Cov(\hat{Y}, D - E[D|X])}{E(D - E[D|X])^2} \\
&= \frac{Cov(E[Y|X, D], D - E[D|X])}{E(D - E[D|X])^2} \\
&= \frac{E\{E[Y|X, D](D - E[D|X])\}}{E(D - E[D|X])^2} \\
&= \frac{E\{(E[Y|X, D = 0] + \tau_x D)(D - E[D|X])\}}{E(D - E[D|X])^2} \\
&= \frac{E[\tau_x(D - E[D|X])^2]}{E[(D - E[D|X]^2)]} = \frac{E[\tau_x E[(D - E[D|X])^2|X]]}{E[E[(D - E[D|X]^2)|X]]} = \frac{E[\tau_x \sigma^2_{D,X}]}{E[\sigma^2_{D,X}]} \\
&= \frac{\sum_x \tau_x [Pr[D = 1|X = x](1 - Pr[D = 1|X = x])]Pr[X = x]}{\sum_x [Pr[D = 1|X = x](1 - Pr[D = 1|X = x])]Pr[X = x]}
\end{aligned}$$

(e) Recall the weights used in the subclassification estimator, which is unbiased for the true τ_{ATE}. Are these weights equal to or different than those implied by OLS? If so, how?

They are different. In the subclassification estimator the weights are proportional to the share of

units in each stratum, while OLS weights by the marginal distribution of X and the conditional variance of $Var[D|X]$ in each stratum.

Problem 2

Use the data `dataQ2.csv` to estimate the ATE using:

i) Regression assuming the following model: $y_i = \alpha + \tau D_i + \epsilon_i$

ii) Regression assuming the following model: $y_i = \alpha + \tau D_i + \beta x_i + \epsilon_i$

iii) Subclassification

iv) Matching, using one match per treated unit (with replacement), setting `ties = TRUE`

v) Matching, using one match per treated unit (with replacement), setting `ties = FALSE`

In these data, selection on the observables is satisfied. With the insights from Problem 1, answer the following questions.

(a) Present your estimates of the ATE with standard errors for the five approaches.

```
> dat <- read.csv("dataQ2.csv")
> head(dat)
  D X           Y
1 0 0  0.01013758
2 1 0  9.44139306
3 1 0  9.57398593
4 0 0 -0.23094179
5 1 0 10.32727881
6 1 0 10.10340707
>
>
> # Regression 1 ------------------------------------------------------------
>
> mod1 <- lm(Y ~ D, data=dat)
> coeftest(mod1, vcov=vcovHC(mod1,"HC2"))

t test of coefficients:

             Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)  0.707241   0.034477 20.5136   <2e-16 ***
D           -2.220397   2.384330 -0.9312    0.352
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

>
> ATE.reg1 <- coeftest(mod1, vcov=vcovHC(mod1,"HC2"))[2,1]
> ATE.reg1
[1] -2.220397
>
> # Regression 2 ------------------------------------------------------------
>
> mod2 <- lm(Y ~ D + X, data=dat)
> coeftest(mod2, vcov=vcovHC(mod2,"HC2"))

t test of coefficients:

            Estimate Std. Error t value  Pr(>|t|)
(Intercept)  3.61116    0.65762  5.4913 5.064e-08 ***
D           -4.24285    2.62957 -1.6135    0.1069
X           -8.26897    1.86158 -4.4419 9.913e-06 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

>
> ATE.reg2 <- coeftest(mod2, vcov=vcovHC(mod2,"HC2"))[2,1]
> ATE.reg2
[1] -4.242854
>
> # Subclassification -------------------------------------------------------
> ATE.sub <-
+   (mean(dat$Y[dat$D==1 & dat$X==0]) - mean(dat$Y[dat$D==0 & dat$X==0]))*
(sum(dat$X==0)/nrow(dat)) +
+   (mean(dat$Y[dat$D==1 & dat$X==1]) - mean(dat$Y[dat$D==0 & dat$X==1]))*
(sum(dat$X==1)/nrow(dat))
> ATE.sub
[1] -23.36667
>
```

```
>
> #Variance of ATE within X=0
>
> var.ATEhat.X0 <-
+   var(dat$Y[dat$D==1 & dat$X==0])/length(dat$Y[dat$D==1 & dat$X==0]) +
+   var(dat$Y[dat$D==0 & dat$X==0])/length(dat$Y[dat$D==0 & dat$X==0])
>
> #Variance of ATE within X=1
>
> var.ATEhat.X1 <-
+   var(dat$Y[dat$D==1 & dat$X==1])/length(dat$Y[dat$D==1 & dat$X==1]) +
+   var(dat$Y[dat$D==0 & dat$X==1])/length(dat$Y[dat$D==0 & dat$X==1])
>
> #Variance of overall ATE
> N <- nrow(dat)
> NX1 <- sum(dat$X == 1)
> NX0 <- sum(dat$X == 0)
>
>
> var.ATEhat.sub <- (NX0/N)^2 * var.ATEhat.X0 + (NX1/N)^2 * var.ATEhat.X1
> SE.ATEhat.sub <- sqrt(var.ATEhat.sub)
> SE.ATEhat.sub
[1] 0.02178867
> # Matching ----------------------------------------------------------------
>
>
> mout1 <- Match(dat$Y,dat$D,dat$X,estimand = "ATE", M = 1, ties=TRUE)
> summary(mout1)

Estimate...  -23.367
AI SE......  3.4647
T-stat.....  -6.7441
p.val......  1.5393e-11

Original number of observations..............  1000
Original number of treated obs...............  197
Matched number of observations...............  1000
Matched number of observations  (unweighted).  195236
```

```
>
> mout2 <- Match(dat$Y,dat$D,dat$X,estimand = "ATE", M = 1, ties=FALSE)
> summary(mout2)

Estimate...  -23.365
SE.........  1.6002
T-stat.....  -14.601
p.val......  < 2.22e-16

Original number of observations..............  1000
Original number of treated obs...............  197
Matched number of observations...............  1000
Matched number of observations  (unweighted).  1000
```

(b) Why are the estimates different (or the same) across the approaches?

(c) Under what conditions would each approach be an unbiased estimate of the ATE? What method(s) deliver an unbiased estimate of the ATE in this case?

(d) Reproduce (exactly) the estimate of the ATE from the second regression by calculating and averaging (with appropriate weights) the within-stratum treatment effects.

```
> # Relationship between Regression 2 and Subclassification -----------------
>
> tau0 <- mean(dat$Y[dat$D == 1 & dat$X == 0]) - mean(dat$Y[dat$D == 0 & dat$X == 0])
> tau1 <- mean(dat$Y[dat$D == 1 & dat$X == 1]) - mean(dat$Y[dat$D == 0 & dat$X == 1])
>
> NX1 <- sum(dat$X == 1)
> NX0 <- sum(dat$X == 0)
>
> p.X0 <- NX0/N
> p.X1 <- NX1/N
>
> varD.X0 <- var(dat$D[dat$X == 0]) * (NX0 - 1)/NX0
> varD.X1 <- var(dat$D[dat$X == 1]) * (NX1 - 1)/NX1
>
> (tau0*p.X0*varD.X0 + tau1*p.X1*varD.X1)/(p.X0*varD.X0 + p.X1*varD.X1)
[1] -4.242854
> ATE.reg2
```

```
[1] -4.242854
```

Problem 3

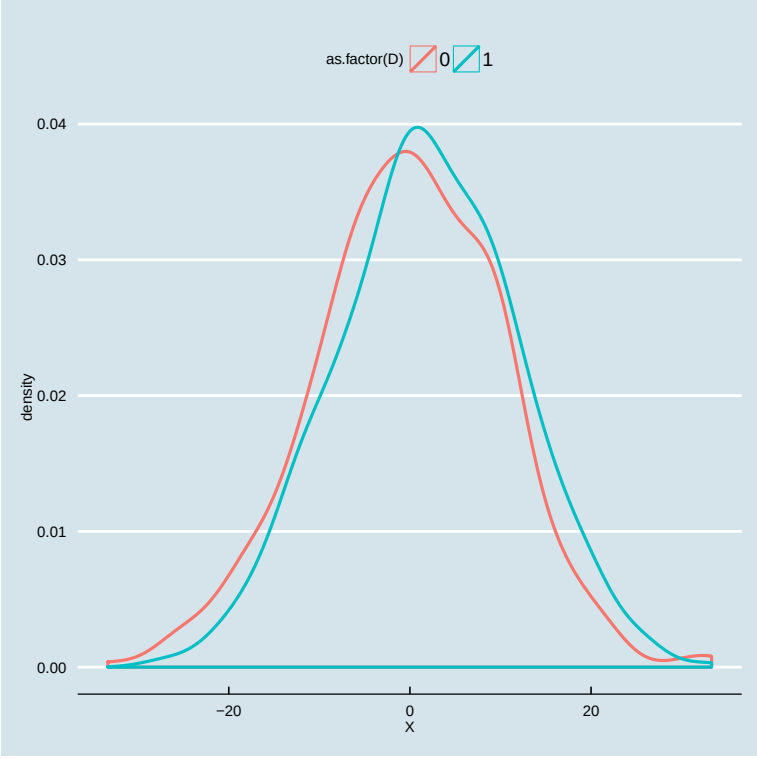
This question will make use of two datasets, `simdata1.csv` and `simdata2.csv`. Each dataset is simulated and contains a treatment vector D, a response variable vector Y, and a covariate vector X. The assumption of selection on the observables is also met by both datasets, and the response variable was generated according to the following model: $y_i = \alpha + \tau D_i + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \epsilon_i$, with $\epsilon \sim \mathcal{N}(0, 0.25)$, $\alpha = 0$, $\tau = 10$, $\beta_1 = 5$, $\beta_2 = 1$, and $\beta_3 = -0.05$. However, the joint distribution of D and X is different in each dataset.

Because the data are simulated, we happen to know the true relationship between Y and D and X. However, in real-world situations, we will not know such true relationships. Even if we know that we must condition on X to achieve ignorability of the treatment, for instance, we will probably not know the functional form by which Y relates to X; you can think of the specification used for the simulated DGP in this problem as some arbitrary, complicated functional form that we are unlikely to figure out in the real world. We will explore the consequences of this for regression and matching estimators under different overlap conditions.

(a) Using `simdata1.csv`, report balance statistics for X across the treatment and control groups, and overlay its density for the treatment and control groups.

```
> library(Matching)
> source("baltestcollect.r")
>
> #Part A
> w <- read.csv("simdata1.csv")
>
> mb <- MatchBalance(D ~ X,data=w)

***** (V1) X *****
before matching:
mean treatment........ 1.8772
mean control.......... -0.21898
std mean diff......... 21.101

mean raw eQQ diff..... 2.1013
med  raw eQQ diff..... 2.001
```

```
max  raw eQQ diff..... 6.3195

mean eCDF diff........ 0.05669
med  eCDF diff........ 0.060611
max  eCDF diff........ 0.094363

var ratio (Tr/Co)..... 0.94415
T-test p-value........ 0.0010793
KS Bootstrap p-value.. 0.016
KS Naive p-value...... 0.023709
KS Statistic.......... 0.094363

> out.bef <- baltest.collect(matchbal.out=mb,var.names="X",after=FALSE)
Minimum P value from T-Tests is 0.001079306
Minimum P value from KS-Tests is 0.016
Max qq.max.diff 0.09436273
> round(out.bef,3)
  mean.Tr mean.Co  sdiff sdiff.pooled var.ratio T pval KS pval qqmeandiff qqmeddiff qqmax
X   1.877  -0.219 21.101       29.292     0.944  0.001   0.016      0.057     0.061     0
```



(b) Estimate τ using: (1) OLS regression with a bivariate model (i.e. $y_i = \alpha + \tau D_i + u_i$); (2) OLS regression with a model that controls for the covariate (i.e. $y_i = \alpha + \tau D_i + \beta_1 x_i + u_i$); (3) OLS

regression that specifies the true model (i.e. $y_i = \alpha + \tau D_i + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + u_i$); and (4) a nearest-neighbor (with replacement) matching estimator. What do you notice?

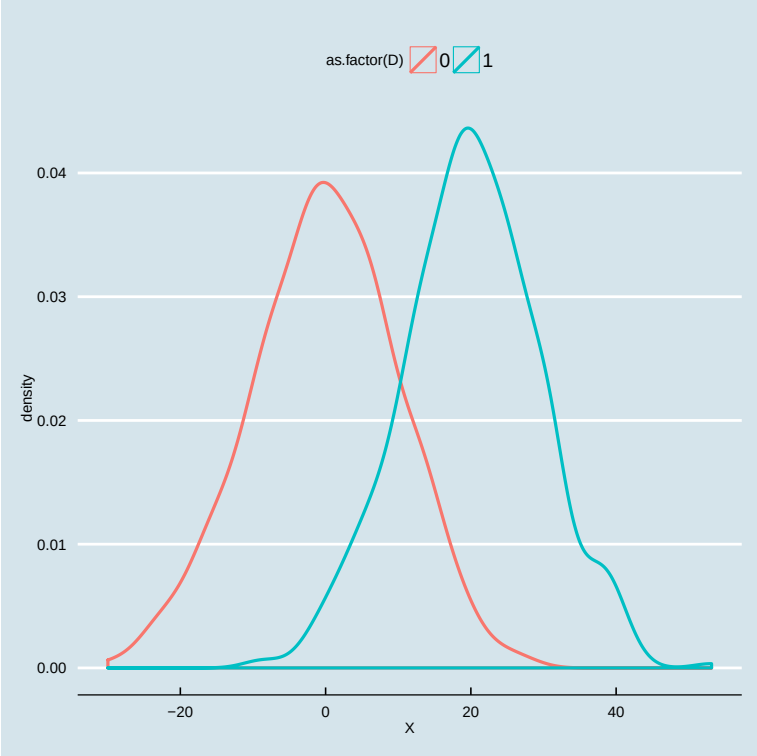
```
> reg1 <- lm(Y ~ D, data=w)
> reg2 <- lm(Y ~ D + X, data=w)
> reg3 <- lm(Y ~ D + X + I(X^2) + I(X^3), data=w)
>
> coeftest(reg1,vcov=vcovHC(reg1,"HC2"))

t test of coefficients:

            Estimate Std. Error t value Pr(>|t|)
(Intercept)  110.587     11.558  9.5679   <2e-16 ***
D            -16.062     13.346 -1.2035   0.2291
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

> coeftest(reg2,vcov=vcovHC(reg2,"HC2"))

t test of coefficients:

            Estimate Std. Error t value  Pr(>|t|)
(Intercept) 108.5263     9.7543 11.1260 < 2.2e-16 ***
D             3.6665    10.8582  0.3377    0.7357
X            -9.4115     1.3163 -7.1499 1.678e-12 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

> coeftest(reg3,vcov=vcovHC(reg3,"HC2"))

t test of coefficients:

               Estimate  Std. Error     t value Pr(>|t|)
(Intercept)  5.0229e-03  2.4855e-02      0.2021   0.8399
D            9.9842e+00  3.1047e-02    321.5809   <2e-16 ***
X            5.0033e+00  2.2575e-03   2216.2643   <2e-16 ***
I(X^2)       9.9986e-01  9.5000e-05  10524.8631   <2e-16 ***
I(X^3)      -5.0009e-02  4.6395e-06 -10779.1176   <2e-16 ***
```

```
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

>
>
> outmatchB <- Match(Y=w$Y, Tr=w$D,X= w$X,M=1,
+                    BiasAdjust=T, Weight=1,estimand="ATE")
> summary(outmatchB)

Estimate...  7.9433
AI SE......  1.8307
T-stat.....  4.339
p.val......  1.4312e-05

Original number of observations..............  1000
Original number of treated obs...............  531
Matched number of observations...............  1000
Matched number of observations  (unweighted).  1514
```

(c) Repeat (a) and (b) with `simdata2.csv`, and comment on the differences. Can the matching approach balance treatment and control groups on x? How are the results affected for each estimation given these new data? Why?

```
> w <- read.csv("simdata2.csv")
>
> mb <- MatchBalance(D ~ X,data=w)

***** (V1) X *****
before matching:
mean treatment........ 19.993
mean control.......... -0.18612
std mean diff......... 215.57

mean raw eQQ diff..... 20.232
med  raw eQQ diff..... 20.125
max  raw eQQ diff..... 25.066

mean eCDF diff........ 0.42999
med  eCDF diff........ 0.47196
```

```
max  eCDF diff........ 0.71204

var ratio (Tr/Co)..... 0.86809
T-test p-value........ < 2.22e-16
KS Bootstrap p-value.. < 2.22e-16
KS Naive p-value...... 0
KS Statistic.......... 0.71204

> out.bef <- baltest.collect(matchbal.out=mb,var.names="X",after=FALSE)
Minimum P value from T-Tests is 0
Minimum P value from KS-Tests is 0
Max qq.max.diff 0.7120402
> round(out.bef,3)
  mean.Tr mean.Co   sdiff sdiff.pooled var.ratio T pval KS pval qqmeandiff qqmeddiff qqmax
X  19.993  -0.186 215.574       203.74     0.868      0       0       0.43     0.472     0
```




```
> reg1C <- lm(Y ~ D, data=w)
> reg2C <- lm(Y ~ D + X, data=w)
> reg3C <- lm(Y ~ D + X + I(X^2) + I(X^3), data=w)
>
> coeftest(reg1C,vcov=vcovHC(reg1C,"HC2"))
```

```
t test of coefficients:

             Estimate Std. Error t value  Pr(>|t|)
(Intercept)  108.1525     9.3855 11.5233 < 2.2e-16 ***
D           -171.5208    19.7706 -8.6755 < 2.2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

> coeftest(reg2C,vcov=vcovHC(reg2C,"HC2"))

t test of coefficients:

            Estimate Std. Error  t value  Pr(>|t|)
(Intercept) 104.7158     8.7378  11.9843 < 2.2e-16 ***
D           201.0845    29.3862   6.8428 1.354e-11 ***
X           -18.4652     1.8394 -10.0389 < 2.2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

> coeftest(reg3C,vcov=vcovHC(reg3C,"HC2"))

t test of coefficients:

               Estimate  Std. Error     t value Pr(>|t|)
(Intercept) -9.8128e-04  2.5282e-02     -0.0388    0.969
D            1.0028e+01  4.7081e-02    213.0013   <2e-16 ***
X            4.9998e+00  2.2038e-03   2268.7443   <2e-16 ***
I(X^2)       1.0001e+00  1.1917e-04   8392.8322   <2e-16 ***
I(X^3)      -5.0006e-02  3.3450e-06 -14949.7715   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

>

> outmatchC <- Match(Y=w$Y, Tr=w$D,X= w$X,M=1,
+                    BiasAdjust=T, Weight=1, estimand="ATE")
> summary(outmatchC)
```

```
Estimate...  -58.358
AI SE......  32.283
T-stat.....  -1.8077
p.val......  0.070653

Original number of observations..............  1000
Original number of treated obs...............  481
Matched number of observations...............  1000
Matched number of observations  (unweighted).  1246
```

Problem 4

Now, you will replicate results in the following article:

> Card, D. and A. B. Krueger (1994), "Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania," *American Economic Review*, vol. 84, 772-793.

You'll want to download the original paper and data (`card_krueger.dta`) from Piazza. Card and Krueger are interested in estimating the impact of minimum wage on teenage employment. Conventional economic wisdom states that raises in minimum wages hurt employment, especially teenage employment since teenage wages are often set at the minimum wage. Such is the main argument of those who oppose raising minimum wages. However, empirical analysis has failed to find evidence of employment responses to raises in minimum wages. In 1992, New Jersey's minimum wage increased from \$4.25 to \$5.05 while the minimum wage in Pennsylvania remained at \$4.25. The authors used data on employment at fast-food establishments in New Jersey and Pennsylvania before and after the increase in the minimum wage to measure the impact of the increase in minimum wage on teenage employment.

All variables in the dataset are listed below.

Variable	Obs	Unique	Mean	Min	Max	Label
co_owned	410	2	.3439024	0	1	1 if company owned
southj	410	2	.2268293	0	1	1 if in southern NJ
centralj	410	2	.1536585	0	1	1 if in central NJ
pa1	410	2	.0878049	0	1	1 if in PA, northeast suburbs of Philadelphia
pa2	410	2	.104878	0	1	1 if in PA, Easton etc
wage_st	390	32	4.615641	4.25	5.75	Starting wage ($/hr) Before
hrsopen	410	23	14.43902	7	24	Hours Open Weekday Before
wage_st2	389	24	4.996273	4.25	6.25	Starting wage ($/hr) After
hrsopen2	399	23	14.46554	8	24	Hours Open Weekday After
emptot	398	103	20.99887	5	85	FTE Employment Before
emptot2	396	90	21.05429	0	60.5	FTE Employment After
nj	410	2	.8073171	0	1	1 if NJ; 0 if Pa
pa	410	2	.1926829	0	1	1 if Pa; 0 if NJ
bk	410	2	.4170732	0	1	1 if Burger King
kfc	410	2	.195122	0	1	1 if KFC
roys	410	2	.2414634	0	1	1 if Roy Rogers
wendys	410	2	.1463415	0	1	1 if Wendys
pmeal	387	154	3.290439	2.28	5.86	Price of Full Meal Before
pmeal2	376	164	3.341463	2.14	5.17	Price of Full Meal After
closed	410	2	.0146341	0	1	Closed Permanently After

As many of you experienced, the tables in Card and Krueger are not readily replicable. Most of the time you'll get very nearly the same result, but there are many sources of minor discrepancies including missing variables, unclear and inconsistent treatment of null and missing values, and ambiguity over how standard errors are computed. The lesson here is more that things don't always replicate perfectly.

However, there is an important point here regarding standard errors for diff-in-diff computations. Consider the mean full-time-employment (FTE) in four groups: PA in wave 1, PA in wave 2, NJ in wave 1, and NJ in wave 2. You can compute the mean of FTE (as emptot or emptot2) in each of these. What you want for diff-in-diff is either of the following:

- $\alpha = (E[FTE|wave = 2, PA = 1] - E[FTE|wave = 1, PA = 1]) - (E[FTE|wave = 2, PA = 0] - E[FTE|wave = 1, PA = 0])$

- $\alpha = (E[FTE|wave = 2, PA = 1] - E[FTE|wave = 2, PA = 0]) - (E[FTE|wave = 1, PA = 1] - E[FTE|wave = 1, PA = 0])$

In general, either choice would work. Ordinarily, you know how to compute SEs for all four of these means, for the two differences (either between state or between periods), and for the single double-difference. However the methods for computing SEs of a difference in means typically rely on zero covariance between the two means. E.g. $var(E[FTE|wave = 2, PA = 1] - E[FTE|wave = 1, PA = 1]) = var(E[FTE|wave = 2, PA = 1]) + var(E[FTE|wave = 2, PA = 1])$, only if $cov(E[FTE|wave = 2, PA = 0], E[FTE|wave = 2, PA = 1]) = 0$. This, unfortunately, is not the case as it has been in past examples, since the two samples used to get these means are actually the same. That is, sampling variation that make this mean higher in wave 1 may also make the mean higher in wave 2.

The simple solution is to first compute within-store changes between wave 1 and wave 2. This way, we don't have to guess about the variance of a difference: we can actually compute the difference first and then take its sample variance! Once you have the within-store changes between wave 1 and wave 2, you

can compute the mean within-store variation from PA, the mean within-store variation from NJ, and SEs for these each of these. Then you can compare these two means...and when you compute SEs for that mean difference in means, you have two independent samples and so you CAN use the usual trick. That all sounds very confusing, I know, so take a look at some sample code:

```
FTE_diff_within=emptot2-emptot  #within-store change
FTE_diff_within_PA=mean(FTE_diff_within[pa==1],na.rm=T)  #for PA stores
FTE_diff_within_NJ=mean(FTE_diff_within[pa==0],na.rm=T)  #for NJ stores

#Now every element of FTE_diff_within is already a difference.
#And recall that var(\bar(y))=var(y)/N.
var_FTE_diff_within_PA=var(FTE_diff_within[pa==1], na.rm=T)/sum(!is.na(FTE_diff_within[pa==1]
var_FTE_diff_within_NJ=var(FTE_diff_within[pa==0], na.rm=T)/sum(!is.na(FTE_diff_within[pa==0]

Diff_in_Diff=FTE_diff_within_PA-FTE_diff_within_NJ
SE_diff_in_diff=sqrt(var_FTE_diff_within_PA+var_FTE_diff_within_NJ)
```

Unfortunately this doesn't quite replicate the results in the paper, but it is a perfectly reasonable way to estimate the DID effect and its standard error.

Another important point on this question regards the identification assumption and its credibility. Remember that DID is only valid if you think that in the absence of the treatment the trends would have been the same between the two groups. Is that valid here? It is hard to tell since the authors didn't provide one of the most useful tests: a comparison of the trends in multiple pre-treatment periods. These days if you're going to publish a DID paper, it would typically need to show that in the several periods preceding the treatment onset, the two groups do have nearly identical trends. If they don't, then one has little reason to expect that they would be identical post-treatment in had the treatment not been administered. Another important consideration in this study is possible violation of the SUTVA.

1. Replicate Table 2 of Card and Krueger (1994) using t-tests, assuming unequal variance. Can you successfully replicate it? If not, explain why you think you cannot.

```
> library(foreign)
>
>
# d<-read.dta("...card_krueger.dta")
> covars1<-c("bk","kfc","roys","wendys","co_owned")
>
> d.nj<-d[d$nj==1,]
> d.pa<-d[d$nj==0,]
>
```

```
> data.sets<-list(d, d.nj, d.pa)
>
> means<-NA
> means.nj<-NA
> means.pa<-NA
>
> means.list<-list(means, means.nj, means.pa)
>
> for(i in 1:length(data.sets)){
+  for(j in 1:length(covars1)){
+
+  means.list[[i]][j]<-mean(data.sets[[i]][,covars1[j]], na.rm=T)
+
+  }
+ }
>
> t1<-cbind.data.frame(var=covars1, all=means.list[[1]],
nj=means.list[[2]], pa=means.list[[3]])
> t1
       var       all        nj        pa
1       bk 0.4170732 0.4108761 0.4430380
2      kfc 0.1951220 0.2054381 0.1518987
3     roys 0.2414634 0.2477341 0.2151899
4   wendys 0.1463415 0.1359517 0.1898734
5 co_owned 0.3439024 0.3413897 0.3544304
>
>
> p<-NA
> diff<-NA
>
> for(i in 1:length(covars1)){
+
+  t<-t.test(d[d$nj==1, covars1[i]], d[d$nj==0, covars1[i]] , na.rm=T)
+  p[i]<-t$p.value
+  diff[i]<-t$estimate[1] - t$estimate[2]
+
+ }
>
```

```
> p
[1] 0.6073881 0.2499542 0.5345289 0.2662397 0.8286651
>
> t1$diff<-diff
> t1$p<-p
>
>
> ##means in wave 1
>
>
> d$wage_before_425<-NA
> d$wage_before_425[d$wage_st<=4.25]<-1
> d$wage_before_425[d$wage_st>4.25]<-0
> table(d$wage_before_425)

  0   1
263 127
>
> covars2<-c("emptot","wage_st","wage_before_425","pmeal","hrsopen")
>
> d.nj<-d[d$nj==1,]
> d.pa<-d[d$nj==0,]
>
> data.sets<-list(d, d.nj, d.pa)
>
> means<-NA
> means.nj<-NA
> means.pa<-NA
>
> means.list<-list(means, means.nj, means.pa)
>
> for(i in 1:length(data.sets)){
+  for(j in 1:length(covars2)){
+
+  means.list[[i]][j]<-mean(data.sets[[i]][,covars2[j]], na.rm=T)
+
+
+
```

```
+  }
+ }
>
> t2<-cbind.data.frame(var=covars2, all=means.list[[1]],
nj=means.list[[2]], pa=means.list[[3]])
> t2
              var       all         nj         pa
1          emptot 20.998869 20.4394081 23.3311688
2         wage_st  4.615641  4.6121337  4.6301316
3 wage_before_425  0.325641  0.3216561  0.3421053
4           pmeal  3.290439  3.3510611  3.0423684
5         hrsopen 14.439024 14.4184290 14.5253165
>
> p<-NA
>
> for(i in 1:length(covars2)){
+
+  t<-t.test(d[d$nj==1, covars2[i]], d[d$nj==0, covars2[i]] , na.rm=T)
+  p[i]<-t$p.value
+  diff[i]<-t$estimate[1] - t$estimate[2]
+
+ }
>
> p
[1] 0.0479034119 0.6888144093 0.7372920289 0.0001291547 0.7704945298
>
> t2$diff<-diff
> t2$p<-p
>
> ##means in wave 2
>
> d$wage_after_425<-NA
> d$wage_after_425[d$wage_st2<=4.25]<-1
> d$wage_after_425[d$wage_st2>4.25]<-0
> table(d$wage_after_425)

  0   1
369  20
```

```
>
> d$wage_after_505<-NA
> d$wage_after_505[d$wage_st2<=5.05]<-1
> d$wage_after_505[d$wage_st2>5.05]<-0
> table(d$wage_after_505)

  0   1
320  69
>
>
> covars3<-c("emptot2","wage_st2","wage_before_425","pmeal2","hrsopen2")
>
> d.nj<-d[d$nj==1,]
> d.pa<-d[d$nj==0,]
>
> data.sets<-list(d, d.nj, d.pa)
>
> means<-NA
> means.nj<-NA
> means.pa<-NA
>
> means.list<-list(means, means.nj, means.pa)
>
> for(i in 1:length(data.sets)){
+  for(j in 1:length(covars1)){
+
+  means.list[[i]][j]<-mean(data.sets[[i]][,covars3[j]], na.rm=T)
+
+  }
+ }
>
> t3<-cbind.data.frame(var=covars3, all=means.list[[1]],
nj=means.list[[2]], pa=means.list[[3]])
> t3
             var       all         nj         pa
1         emptot2 21.054293 21.0274295 21.1655844
2        wage_st2  4.996273  5.0808492  4.6174648
3 wage_before_425  0.325641  0.3216561  0.3421053
```

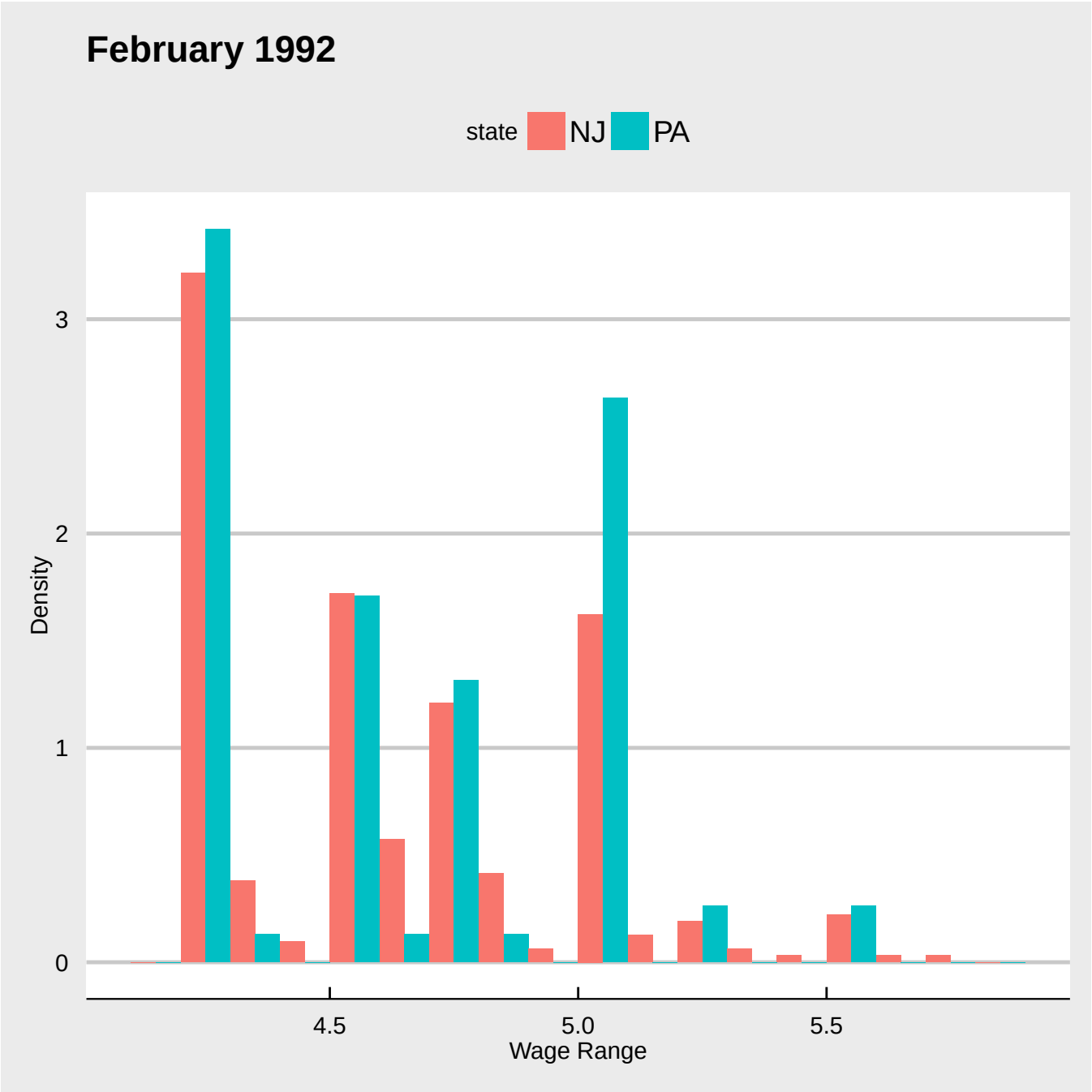
```
4          pmeal2  3.341463  3.4147541  3.0266197
5        hrsopen2 14.465539 14.4197819 14.6538462
>
> p<-NA
>
> for(i in 1:length(covars3)){
+
+  t<-t.test(d[d$nj==1, covars3[i]], d[d$nj==0, covars3[i]] , na.rm=T)
+  p[i]<-t$p.value
+  diff[i]<-t$estimate[1] - t$estimate[2]
+
+ }
>
> p
[1] 8.981531e-01 8.417388e-17 7.372920e-01 1.486448e-06 5.175281e-01
>
> t3$diff<-diff
> t3$p<-p
>
> results<-rbind.data.frame(t1, t2, t3)
>
> for(i in 2:ncol(results)){
+
+  results[,i]<-round(results[,i], digits=2)
+
+ }
> results
              var   all    nj    pa  diff    p
1              bk  0.42  0.41  0.44 -0.03 0.61
2             kfc  0.20  0.21  0.15  0.05 0.25
3            roys  0.24  0.25  0.22  0.03 0.53
4          wendys  0.15  0.14  0.19 -0.05 0.27
5        co_owned  0.34  0.34  0.35 -0.01 0.83
6          emptot 21.00 20.44 23.33 -2.89 0.05
7         wage_st  4.62  4.61  4.63 -0.02 0.69
8  wage_before_425  0.33  0.32  0.34 -0.02 0.74
9           pmeal  3.29  3.35  3.04  0.31 0.00
10        hrsopen 14.44 14.42 14.53 -0.11 0.77
```

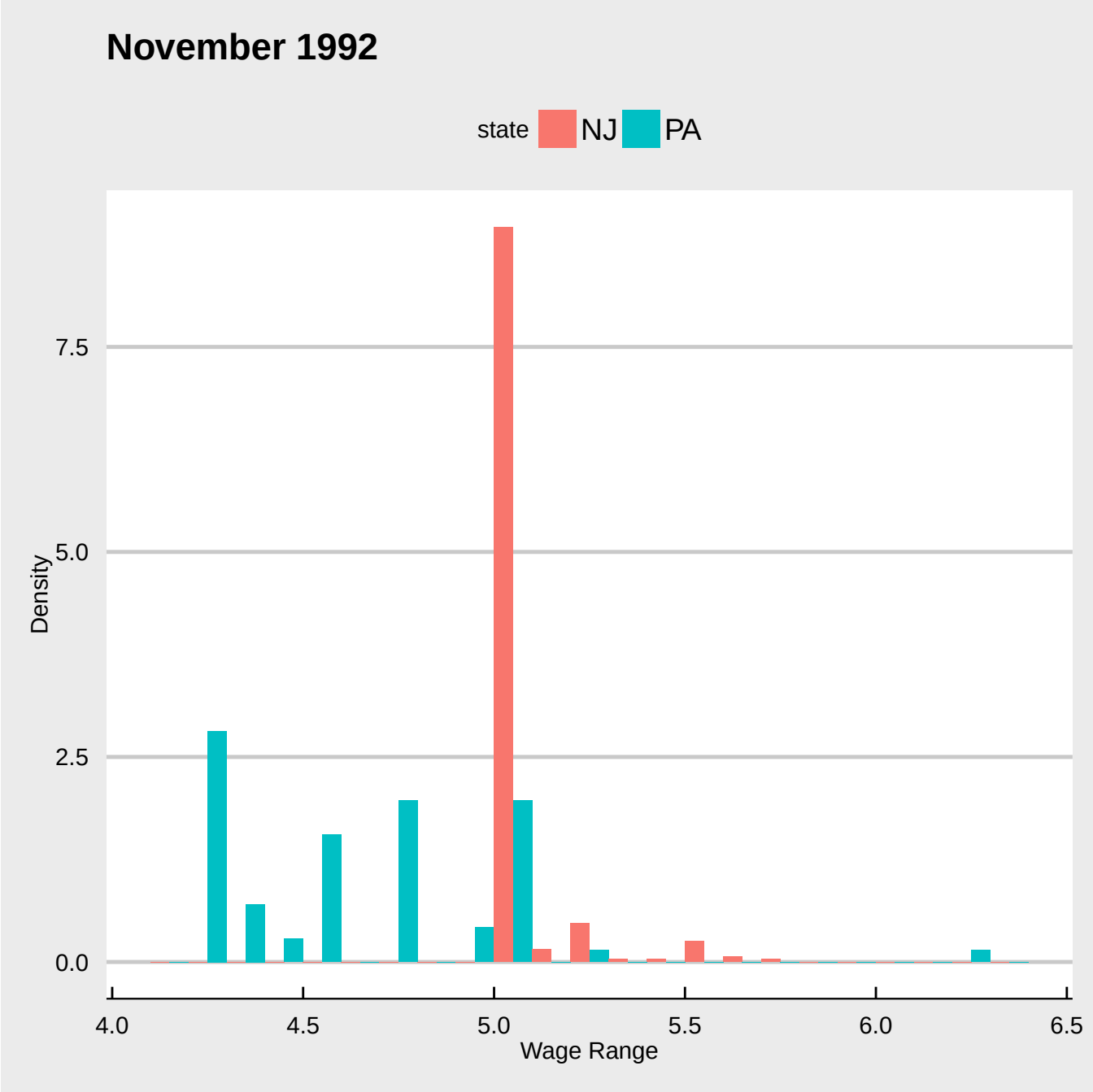
```
11          emptot2 21.05 21.03 21.17 -0.14 0.90
12        wage_st2  5.00  5.08  4.62  0.46 0.00
13 wage_before_425  0.33  0.32  0.34 -0.02 0.74
14           pmeal2  3.34  3.41  3.03  0.39 0.00
15        hrsopen2 14.47 14.42 14.65 -0.23 0.52
```

2. Based on the results you report in your replicated Table 2, what is the major finding with regard to FTE before and after the rise of the minimum wage?

   For NJ, FTE increased from period 1 to period 2, while in PA FTE decreased. As a result, there is a positive difference in differences. If we made the parallel trends assumption (i.e. that NJ would have followed the same trajectory as PA if the treatment had not occurred), we could interpret this as a positive causal effect of the wage increase on FTE.

3. Replicate Figure 1. Explain what is going on after the minimum wage legislation.


February 1992
state
NJ
PA
Density
3
2
1
0
4.5
5.0
5.5
Wage Range



The implications of these plots are clear. The distributions of starting wages in NJ and PA were virtually indistinguishable prior to the minimum wage increase in NJ. After the increase, the restaurants in NJ that had been paying less than the new minimum wage began reporting starting wages at the level of the new minimum wage, with the distribution of restaurants that had previously been paying above the new minimum wage remaining about the same. This indicates compliance with the new legislation. Meanwhile, in wave 2, the distribution in PA remains similar to wave 1.

4. Replicate Table 4 by carefully generating relevant variables and running regressions. Interpret your coefficient estimates for the variables of "New Jersey dummy" and "Initial wage gap." Be sure to interpret the statistical and substantive significance.

Replication of Table 4:

Reduced-Form Models For Change in Employment

	Model				
	(i)	(ii)	(iii)	(iv)	(v)
New Jersey dummy	2.33 (1.19)	2.30 (1.20)	-	-	-
Initial wage gap	-	-	15.65 (6.08)	14.92 (6.21)	11.98 (7.42)
Controls for chain and ownership	no	yes	no	yes	yes
Controls for region	no	no	no	no	yes
Standard error of regression	8.79	8.79	8.76	8.76	8.75
Probability value for controls	-	0.34	-	0.43	0.39

Interpretation of coefficients in each regression:
Model 1: The minimum wage legislation is associated with, on average, 2.33 workers increase in FTE employment in NJ. This increase is substantively large and statistically significant at 0.1 level (but not at 0.05 level).

Model 2: The minimum wage legislation is associated with, on average, 2.30 workers increase in FTE employment in NJ, controlling for chain type and ownership. This increase is substantively large and statistically significant at 0.1 level (but not at 0.05 level).

Model 3: One percentage point increase in the initial wage gap is associated with, on average, 0.157 workers increase in FTE employment among stores that payed below the new minimum wage before the legislation. This increase is substantively large because if we move from the previous minimum wage \$4.25 to the new one \$5.05 (18.82% increase), we need to increase around 3 workers on average. This increase is also statistically significant at 0.05 significance level.

Model 4: One percentage point increase in the initial wage gap is associated with, on average, 0.149 workers increase in FTE employment among stores that payed below the new minimum wage before the legislation, controlling for chain type and the ownership. This increase is substantively large because if we move from the previous minimum wage \$4.25 to the new one \$5.05 (18.82% increase), we need to increase around 2.8 workers on average. This increase is also statistically significant at 0.05 significance level.

Model 5: One percentage point increase in the initial wage gap is associated with, on average, 0.119 workers increase in FTE employment among stores that payed below the new minimum

wage before the legislation, controlling for chain type, ownership, and regions. This increase is substantively large because if we move from the previous minimum wage $4.25 to the new one $5.05 (18.82% increase), we need to increase around 2.2 workers on average. This increase is not statistically significant at 0.05 or 0.1 significance level.

```
balanced2 <- card
balanced2$emptot2[balanced2$closed==1] <- 0
balanced2$wage_st2[balanced2$closed==1] <- 0
balanced2 <- balanced2[complete.cases(balanced2[c("emptot","emptot2", "wage_st", "wage_s

balanced2$gap<-NA
balanced2$gap<-ifelse(balanced2$pa==0 & balanced2$wage_st<5.05,(5.05-balanced2$wage_st)/

balanced2$outcome <- balanced2$emptot2 - balanced2$emptot

#reg 1
table4.out <- lm(outcome~nj, data=balanced2)
summary(table4.out)
#reg 2
table4.out2 <- lm(outcome~nj+kfc+roys+wendys+co_owned, data=balanced2)
summary(table4.out2)
#reg 3
table4.out3 <- lm(outcome~gap, data=balanced2)
summary(table4.out3)
#reg 4
table4.out4 <- lm(outcome~gap+kfc+roys+wendys+co_owned, data=balanced2)
summary(table4.out4)
#reg 5
table4.out5 <- lm(outcome~gap+kfc+roys+wendys+co_owned+southj+centralj+pa1+pa2, data=bal
summary(table4.out5)

#F TEST of controls
#reg 2
table4.out2 <- lm(outcome~nj+kfc+roys+wendys+co_owned, data=balanced2)
summary(table4.out2)
test2 <- lm(outcome~nj, data=balanced2)
lrtest(test2, table4.out2)
```

```
#reg 4
table4.out4 <- lm(outcome~gap+kfc+roys+wendys+co_owned, data=balanced2)
summary(table4.out4)
test4 <- lm(outcome~gap, data=balanced2)
lrtest(test4, table4.out4)

#reg 5
table4.out5 <- lm(outcome~gap+kfc+roys+wendys+co_owned+southj+centralj+pa1+pa2, data=bal
summary(table4.out5)
test5 <- lm(outcome~gap, data=balanced2)
lrtest(test5, table4.out5)
```

5. Explain the DID identification strategy in the context of this example. Do you think it is credible? Why or why not? Be sure to think hard about potential violations, and then decide what you think about how much you'd trust these results. How could you improve the study?

This approach relies on the assumption that, in the absence of treatment, NJ and PA would have experienced parallel trends over time in the outcome variable. It is difficult to asses the validity of this assumption without pre-treatment data observed at multiple points in time. It could be the case that some time-varying factor that differs between the two groups is responsible for the observed effect.

Using matching before DID basically combines the strengths of using a lagged DV and using a DID estimator. By matching on a lagged DV, we are narrowing (or even eliminating) baseline differences in outcomes between the two groups. This makes the parallel trends assumption more credible, since two groups with large baseline differences could have been different for a time-varying reason that we did not get to observe. We then apply the DID estimator to the matched data, thereby removing the influence of any remaining unobserved time invariant factors.

Problem 5

We observe random samples of campaign contributions from two groups of voters (call them red and blue) in two different periods. We don't necessarily observe any given person in both periods (we can think of this as the repeated cross-section case discussed in class). We sample and observe $N_{g,t}$ individuals from group $g \in \{red, blue\}$ at time $t \in \{0, 1\}$. Let G_i be the group individual i in the sample belongs to, with the convention that $G_i = 1$ means the individual is in the red group, and T_i be the period in which we observe i's contributions.

As in a difference-in-difference scenario, assume that in the second period, the red group received a

treatment (for example, a change in contribution rules for the red party). In class, we assumed a very particular model of the dependence between outcomes, time periods, treatment status, and unobservables *in the absence of treatment*. Instead, suppose we posit a more general model. Individual i in the sample has an unobservable determinant of contributions in the absence of treatment. Call this unobservable determinant ϵ_i, and think of it as the strength of partisan sentiment. In keeping with the potential outcomes framework, we let the relationship between partisan sentiment, time period, and potential outcomes (contributions) *in the absence of treatment* be as general as possible:

$$Y_{0i} = h_g(\epsilon_i, T_i) \tag{1}$$

That is to say, each group of voters $g \in \{red, blue\}$ might have a different function linking time and sentiment to contributions in the absence of treatment, but we are willing to assume that the function is the same for everyone in a given group (for instance, you might suppose that blue workers' contributions increase less rapidly as their partisan sentiment rises than contributions for red workers). Since, in the DID setup, people were not randomized into the two groups, we have to presume that the distribution of the unobservables differs across groups. In other words, the distribution of ability ϵ_i in group g at time t has a cdf $F_{g,t}(\epsilon_i)$ and density $f_{g,t}(\epsilon_i)$.

We leave the outcome under treatment (which is only observable for the red group in $t = 1$) unrestricted; that is, $Y_{1i}(t)$ is the period t outcome after having been treated, but we don't impose any conditions on it, other that its being independent across i's. In other words, we don't write down a function for the Y_{1i}s the way we did for the Y_{0i}s in equation (1).

(a) Recall that the ATT in the DID framework is expressed as

$$ATT = E[Y_{1i}(1) - Y_{0i}(1)\big|D_i = 1] = E[Y_{1i}(1)\big|D_i = 1] - E[Y_{0i}(1)\big|D_i = 1]$$

Using the notation described above, express the ATT terms of the h's, f's, ϵ's, and counterfactual outcome under treatment. (Hint: recall the definition of expected value: $E(X) = \int X f(x)dx$. This problem simply asks you to rewrite the second expectation in the ATT expression as an integral. The first expectation in the ATT expression can be kept as is.)

$$\begin{aligned} ATT &= E\left(Y_{1i}(1) - Y_{0i}(1)\middle|D_i = 1\right) \\ &= E\left(Y_{1i}(1) - Y_{0i}(1)\middle|G_i = 1, T_i = 1\right) \\ &= E\left(Y_{1i}(1)\middle|G_i = 1, T_i = 1\right) - \int h_1(\epsilon_i, 1) f_{1,1}(\epsilon_i) d\epsilon_i \end{aligned}$$

(b) The usual DID estimator is

$$\hat{\beta} = \{\bar{Y}_{red,1} - \bar{Y}_{red,0}\} - \{\bar{Y}_{blue,1} - \bar{Y}_{blue,0}\}$$

We could obtain this directly from the means themselves, or by running a regression of Y_i on a constant, T_i, G_i and $T_i \times G_i$, as we saw in lecture. Express the estimator in terms of the functions, sample sizes, and random variables given—i.e., rewrite the latter three $\bar{Y}$s in terms of equation (1), but keep the first $\bar{Y}$ (outcome under treatment) unrestricted.

$$\hat{\beta} = \left\{ \frac{1}{N_{1,1}} \sum_{\substack{j: \\ G(j)=1, \\ T(j)=1}} Y_{1j}(1) - \frac{1}{N_{1,0}} \sum_{\substack{j: \\ G(j)=1, \\ T(j)=0}} h_1(\epsilon_j, 0) \right\}$$
$$- \left\{ \frac{1}{N_{0,1}} \sum_{\substack{j: \\ G(j)=0, \\ T(j)=1}} h_0(\epsilon_j, 1) - \frac{1}{N_{0,0}} \sum_{\substack{j: \\ G(j)=0, \\ T(j)=0}} h_0(\epsilon_j, 0) \right\}$$

(c) What is the probability limit of this estimator? Your answer should in terms of the h's and $f_{g,t}$'s for all terms referring to outcomes in the absence of treatment. What do you have to assume about the sample sizes $N_{g,t}$ to derive this plim?

$$plim(\hat{\beta}) = \left\{ E(Y_{1i}(1)|G(i)=1, T(i)=1) - \int h_1(\epsilon_j, 0) f_{1,0}(\epsilon_j) d\epsilon_j \right\}$$
$$- \left\{ \int h_0(\epsilon_j, 1) f_{0,1}(\epsilon_j) d\epsilon_j - \int h_0(\epsilon_j, 0) f_{0,0}(\epsilon_j) d\epsilon_j \right\}$$

Have to assume that $N_{g,t} \to \infty$ to apply the WLLN.

(d) Does this equal the expression for ATT derived in (a)?

No.

(e) Without assuming anything else about the f's and h's, state a necessary and sufficient condition such that plim($\hat{\beta}$) =ATT. (Hint: this is not a difficult part of the question; your assumption should simply equate a term from the ATT expression in (b) with terms from the plim expression in (c).)

$$\int h_1(\epsilon_i, 1) f_{1,1}(\epsilon_i) d\epsilon_i = \int h_1(\epsilon_j, 0) f_{1,0}(\epsilon_j) d\epsilon_j + \int h_0(\epsilon_j, 1) f_{0,1}(\epsilon_j) d\epsilon_j - \int h_0(\epsilon_j, 0) f_{0,0}(\epsilon_j) d\epsilon_j$$

or

$$\int h_1(\epsilon_i, 1) f_{1,1}(\epsilon_i) d\epsilon_i - \int h_1(\epsilon_j, 0) f_{1,0}(\epsilon_j) d\epsilon_j = \int h_0(\epsilon_j, 1) f_{0,1}(\epsilon_j) d\epsilon_j - \int h_0(\epsilon_j, 0) f_{0,0}(\epsilon_j) d\epsilon_j$$

(f) What did we call this condition (assumption) in class? In this general case in which we don't restrict the form of the h function, does this condition have a behavioral (structural) interpretation? If so, what is it?

This is the parallel trends assumption. Without any further restrictions on the $h(\cdot)$s, this condition does not have a clear behavioral meaning.

(g) Suppose you don't want to assume the condition above directly because you are not clear about what it implies about the h's and f's. Rather, you want to understand what would have to be true about the h's and f's in order for the condition to be satisfied, so that you can estimate the ATT with the DID estimator. You start by making the plausibly uncontroversial assumption that the functions $h_1(\cdot)$ and $h_0(\cdot)$ are the same—that is, that the mapping between partisan sentiment, time, and contributions (in the absence of treatment) is the same across groups. Call this assumption (A1). Can you now conclude that $\text{plim}(\hat{\beta}) = \text{ATT}$? In other words, is the condition from (d) now satisfied? If so, prove your result. If not, state why this new assumption does not suffice.

The condition is now:

$$\int h(\epsilon_i, 1) f_{1,1}(\epsilon_i) d\epsilon_i - \int h(\epsilon_j, 0) f_{1,0}(\epsilon_j) d\epsilon_j = \int h(\epsilon_j, 1) f_{0,1}(\epsilon_j) d\epsilon_j - \int h(\epsilon_j, 0) f_{0,0}(\epsilon_j) d\epsilon_j$$

It looks as if the first LHS term might cancel with the first RHS term, and as if the second LHS term might cancel with the second RHS term. Rearranging to make this clearer:

$$\int h(\epsilon_i, 1) f_{1,1}(\epsilon_i) d\epsilon_i - \int h(\epsilon_j, 1) f_{0,1}(\epsilon_j) d\epsilon_j = \int h(\epsilon_j, 0) f_{1,0}(\epsilon_j) d\epsilon_j - \int h(\epsilon_j, 0) f_{0,0}(\epsilon_j) d\epsilon_j$$

Clearly, this condition is not sufficient. Even with $h_1(\cdot) = h_0(\cdot)$, the densities (fs) of unobservables still differ across groups. You can think of each side of the last expression above as the continuous version of a difference of weighted sums; since the weights f are different, the weighted sums (that is, integrals) won't cancel as we'd hoped.

(h) Suppose, instead, that you choose to make simplifying assumptions about the f's (distributions of unobservable ability in the samples), without restricting the h's. You start by assuming that $f_{g,1}(\epsilon) = f_{g,0}(\epsilon)$ for all g. (For notational convenience, we can now rewrite the fs with only one subscript, for group.) Call this assumption (A2). Explain what this assumption means substantively for the example of red and blue workers. Using this assumption, can you now conclude that $\text{plim}(\hat{\beta}) =$

ATT? Why or why not?

Substantively, this assumption means that the distribution of ability in each group does not change over time (across periods). Now, rewriting our original condition from (e) under assumption (A2):

$$\int h_1(\epsilon_i, 1) f_1(\epsilon_i) d\epsilon_i - \int h_1(\epsilon_j, 0) f_1(\epsilon_j) d\epsilon_j = \int h_0(\epsilon_j, 1) f_0(\epsilon_j) d\epsilon_j - \int h_0(\epsilon_j, 0) f_0(\epsilon_j) d\epsilon_j$$

where the single subscript on f denotes group. Now, it looks as if the LHS terms might cancel and the RHS terms might cancel—but of course they don't, since $h_g(\epsilon_j, 1) \neq h_g(\epsilon_j, 0)$.

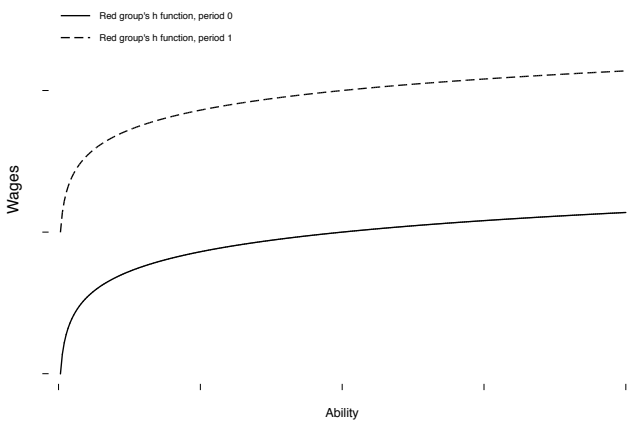
(i) Now consider the special case in which $h_g(\epsilon_j, 1) = h_g(\epsilon_j, 0) + k$ for each group. What does this condition mean substantively? Under this condition together with (A2), does our condition hold? Can we use the DID estimator to get at the ATT?

Substantively, this assumption has two components:

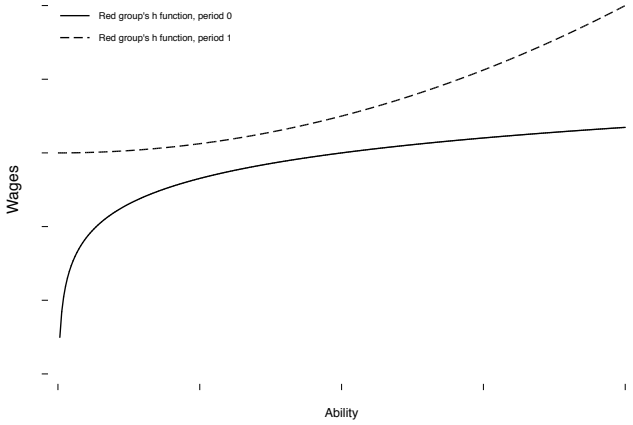
(1) The first is that, within each group, the hs—which, recall, are the functions mapping ability to wages—differ only by a constant. Figure 1 shows examples of functions that do and do not meet this requirement.

Figure 1: Illustration of the assumption that $h_0(\epsilon_j, 1) = h_0(\epsilon_j, 0) + k$

(a) Meets the assumption



(b) Violates the assumption



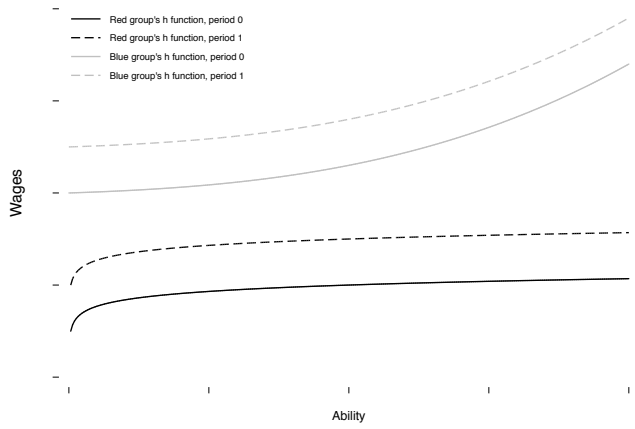
(2) The second is that the constant that shifts the wage function across time (k) is the same for both groups. Figure 2 illustrates functions consistent with and violating this requirement.

With this assumption, starting with (A2), we go from:

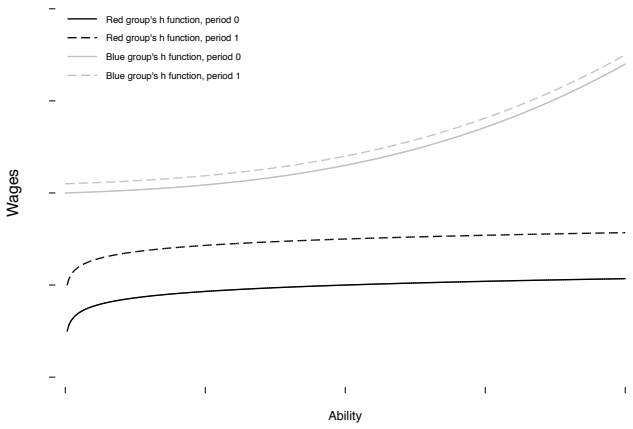
$$\int h_1(\epsilon_i, 1) f_1(\epsilon_i) d\epsilon_i - \int h_1(\epsilon_j, 0) f_1(\epsilon_j) d\epsilon_j = \int h_0(\epsilon_j, 1) f_0(\epsilon_j) d\epsilon_j - \int h_0(\epsilon_j, 0) f_0(\epsilon_j) d\epsilon_j$$

Figure 2: Illustration of the assumption that $h_g(\epsilon_j, 1) = h_g(\epsilon_j, 0) + k$

(a) Meets the assumption



(b) Violates the assumption



to

$$\int \Big[h_1(\epsilon_i, 0) + k\Big] f_1(\epsilon_i) d\epsilon_i - \int h_1(\epsilon_j, 0) f_1(\epsilon_j) d\epsilon_j = \int \Big[h_0(\epsilon_j, 0) + k\Big] f_0(\epsilon_j) d\epsilon_j - \int h_0(\epsilon_j, 0) f_0(\epsilon_j) d\epsilon_j$$

$$k \int f_1(\epsilon_i) d\epsilon_i = k \int f_0(\epsilon_j) d\epsilon_j$$

which of course is true since the densities integrate to one.

(j) What if you combined (A2)—that the distribution of (unobserved) partisan sentiment strength is constant within group over time—with (A1), that the functions $h_1(\cdot)$ and $h_0(\cdot)$ are the same? Can you conclude, with (A1) and (A2) together (but without the assumption in (i)) that the necessary and sufficient condition for $\text{plim}(\hat{\beta}) = \text{ATT}$ holds? If so, prove your result. If not, state why not, and give one additional assumption about the f's that is needed to be able to conclude that $\text{plim}(\hat{\beta}) = \text{ATT}$. Does this additional assumption strike you as reasonable?

Rewriting our original condition from (e) under assumptions (A1) and (A2) together:

$$\int h(\epsilon_i, 1) f_1(\epsilon_i) d\epsilon_i - \int h(\epsilon_j, 0) f_1(\epsilon_j) d\epsilon_j = \int h(\epsilon_j, 1) f_0(\epsilon_j) d\epsilon_j - \int h(\epsilon_j, 0) f_0(\epsilon_j) d\epsilon_j$$

This still isn't enough! Without the assumption in (i), the terms on each side don't cancel. And if we rearrange, we see that:

$$\int h(\epsilon_i, 1) f_1(\epsilon_i) d\epsilon_i - \int h(\epsilon_j, 1) f_0(\epsilon_j) d\epsilon_j = \int h(\epsilon_j, 0) f_1(\epsilon_j) d\epsilon_j - \int h(\epsilon_j, 0) f_0(\epsilon_j) d\epsilon_j$$

In order to make the LHS terms cancel in this rearranged expression, we would need the additional assumption that $f_1(\epsilon_j) = f_0(\epsilon_j)$. Substantively, this says that the time-invariant distribution of observables ϵ_j is identical across groups—which is tantamount to assuming random assignment, in

which case we wouldn't need DID at all!

(k) A researcher suggests that the assumption $f_{red,1}(\epsilon) = f_{blue,1}(\epsilon)$ (call it (A3)), added to (A1), would let you estimate ATT consistently with $\{\bar{Y}_{red,1} - \bar{Y}_{blue,1}\}$. Show this.

Recall that:

$$ATT = E\left(Y_{1i}(1)\middle|G_i = 1, T_i = 1\right) - \int h_1(\epsilon_i, 1) f_{1,1}(\epsilon_i) d\epsilon_i$$

and that in the probability limit,

$$\bar{Y}_{blue,1} = \int h_0(\epsilon_j, 1) f_{0,1}(\epsilon_j) d\epsilon_j$$

Given assumption (A1)

$$\int h_0(\epsilon_j, 1) f_{0,1}(\epsilon_j) d\epsilon_j = \int h_1(\epsilon_j, 1) f_{0,1}(\epsilon_j) d\epsilon_j$$

and then adding assumption (A3)

$$\int h_0(\epsilon_j, 1) f_{0,1}(\epsilon_j) d\epsilon_j = \int h_1(\epsilon_j, 1) f_{0,1}(\epsilon_j) d\epsilon_j = \int h_1(\epsilon_j, 1) f_{1,1}(\epsilon_j) d\epsilon_j$$

Thus, with these two assumptions:

$$\begin{aligned}
&plim\{\bar{Y}_{red,1} - \bar{Y}_{blue,1}\} \\
&= E\left(Y_{1i}(1)\middle|G_i = 1, T_i = 1\right) - \int h_0(\epsilon_i, 1) f_{0,1}(\epsilon_i) d\epsilon_i \\
&= E\left(Y_{1i}(1)\middle|G_i = 1, T_i = 1\right) - \int h_1(\epsilon_i, 1) f_{1,1}(\epsilon_i) d\epsilon_i \\
&= ATT
\end{aligned}$$

Substantively, this assumption implies as-good-as-randomized treatment assignment, which in turn implies that ATT=ATE, which we can estimate with a simple difference in means.

(l) Explain, in social scientific terms, the meaning of assumption (A3). Also explain why (A3) is a much more problematic assumption, in behavioral terms, than (A2).

A3 means that the distribution of ϵ (partisan sentiment) is the same across the red and blue groups in the second period. This is a more problematic assumption than A2 (the distribution of partisan sentiment is constant across time within each group) because it requires the distribution of a covari-

ate (i.e. partisan sentiment) to be the same across treatment groups in the post-treatment period. This equates random assignment.

(m) Taking a slightly different approach, suppose we assume that 1) the h's are the same for both groups and 2) h is separable in time effects and the strength of partisan sentiment:

$$h(\epsilon_i, t) = h(t) + g(\epsilon_i)$$

Call this assumption (A1′) (a stronger version of (A1)). Would (A1′) and (A2) together make the DID estimator consistent? What other assumption would you need?

Rewriting our condition under (A1′) and (A2).

$$\begin{aligned}&\int \Big[h(1) + g(\epsilon_i)\Big] f_1(\epsilon_i) d\epsilon_i - \int \Big[h(0) + g(\epsilon_j)\Big] f_1(\epsilon_j) d\epsilon_j \\ &= \int \Big[h(1) + g(\epsilon_j)\Big] f_0(\epsilon_j) d\epsilon_j - \int \Big[h(0) + g(\epsilon_j)\Big] f_0(\epsilon_j) d\epsilon_j\end{aligned}$$

$$h(1) - h(0) = h(1) - h(0)$$

Thus in this case, plim($\hat{\beta}$)=ATT under (A1′) and (A2) without additional assumptions.

(n) Suppose, in addition to (A1′), we assumed (A4), $g(\epsilon) = \epsilon$, i.e., sentiment comes in linearly into the wage function in the absence of treatment. What minimal condition could you assume about the distributions of the ϵ's for (A1′) and (A4) together to justify DID? Is this weaker or stronger than (A2)? Than (A3)?

Rewriting our condition under (A1′) and (A4).

$$\begin{aligned}&\int \Big[h(1) + \epsilon_j\Big] f_{1,1}(\epsilon_i) d\epsilon_i - \int \Big[h(0) + \epsilon_j\Big] f_{1,0}(\epsilon_j) d\epsilon_j \\ &= \int \Big[h(1) + \epsilon_j\Big] f_{0,1}(\epsilon_j) d\epsilon_j - \int \Big[h(0) + \epsilon_j\Big] f_{0,0}(\epsilon_j) d\epsilon_j\end{aligned}$$

$$\int \epsilon_j f_{1,1}(\epsilon_i) d\epsilon_i - \int \epsilon_j f_{1,0}(\epsilon_j) d\epsilon_j = \int \epsilon_j f_{0,1}(\epsilon_j) d\epsilon_j - \int \epsilon_j f_{0,0}(\epsilon_j) d\epsilon_j$$

Thus, what we need is that the means of the the unobserved distribution of ability change (across periods) by the same amount in each group. This is much weaker than (A3), which requires the distribution of unobservables to be the same in each group in the second period.