Final

Wanlin Ji

6/6/2018

Problem 1 Treatment Effects

(a) Without further information, my best(least wrong, to be precise) guess would be the mean of true distribution of income of undocumented immigrants in California.

Proof.

Let g denote our guess. Firstly, we define Mean Squared Error as our error function, where $MSE = E[(Y-g)^2] = \frac{1}{n}\sum_{n=1}^{n}(Y_i - g)^2$.

Secondly, we want g to minimize our MSE. Obviously, we have

$$\begin{aligned} E[(Y-g)^2] &= E[((Y-E[Y]) + (E[Y]-g))^2] \\ &= E[(Y-E[Y])^2] + E[(E[Y]-g)^2] + 2E[(Y-E[Y])(E[Y]-g)] \\ &= E[(Y-E[Y])^2] + E[(E[Y]-g)^2] + 2(E(Y) - E[E[Y]])(E[Y]-g) \\ &= E[(Y-E[Y])^2] + E[(E[Y]-g)^2] + 0 \end{aligned} \tag{1}$$

Only by taking $g = E[Y]$ can we get the second non-negative term to 0 and minimize the MSE. Thus $\arg\max_g E[(Y-g)^2] = E[Y]$.

(b) It would be $E[Y|x]$, where x denotes the known proportion of that interviewee. In this case, $MSE = E[Y - h(X)]^2$, and by transforming it we would get

$$\begin{aligned} E[Y-h(X)]^2 &= E[(Y-E[Y|X])(h(X)-E[Y|X])^2] \\ &= E[(Y-E[Y|X])^2 + (h(X)-E[Y|X])^2 - 2(Y-E[Y|X])(h(X)-E[Y|X])] \\ &= E[(Y-E[Y|X])^2] + E[(h(X)-E[Y|X])^2] - 2E[E[(Y-E[Y|X])(h(X)-E[Y|X])|X]] \\ &= E[(Y-E[Y|X])^2] + E[(h(X)-E[Y|X])^2] - 2E[(E[Y|X]-E[Y|X])(h(X)-E[Y|X])] \\ &= E[(Y-E[Y|X])^2] + E[(h(X)-E[Y|X])^2] - 0 \end{aligned} \tag{2}$$

Note that Law of Iterated Expectations is used in expanding the last term in (2). And only by taking $h(X) = E[Y|X]$ can we get the second non-negative term to 0 and minimize the MSE. Let $X = x$, then we get $E[Y|x]$.

(c) Without considering X, the difference in means can be defined and decomposed as follows,

$$\begin{aligned} E[Y|D=1] - E[Y|D=0] &= E[Y_1|D=1] - E[Y_0|D=0] \\ &= E[Y_1 - Y_0|D=1] + \{E[Y_0|D=1] - E[Y_0|D=0]\} \end{aligned} \tag{3}$$

The first term is causal effect of implementing Prop 187, and the bias term is in curly braces. The overall difference in expected incomes in this context can be interpreted as the average treatment effect on those who had Prop 187, plus the difference on untreated income between those units who had Prop 187 and those didn't.

(d) Yes, it is identified. Given assumption (1) and (2), we have $F(Y_{oi}, Y_{1i}|X_i = x, D_i = 1) = F(Y_{oi}, Y_{1i}|X_i = x, D_i = 0)$

$$\begin{aligned}\tau(x) &= E_x[Y_i(1) - Y_i(0)|X = x] \\ &= E_x[Y_i(1)|X = x] - E_x[Y_i(0)|X = x] \\ &= E[Y_i(1)] - E[Y_i(0)] because of the mean independence assumption, introduced by the assumption (1) \\ &= \frac{1}{n_1}\sum_{i=1}^{n_1} Y_i(1) - \frac{1}{n_2}\sum_{i=1}^{n_2} Y_i(0)\end{aligned} \tag{4}$$

(e) The estimation is described in (d). Assumption (2) indicates that it might not be equal probability of D=1 and D=0, which leads to different sample sizes for n_1 and n_2 in (d), making the estimator not effective. But this has no influence on the unbiasedness and consistency.

(f) Given the results in (d), the expected difference in observed responses to the two treatments at X=x is equal to ATE, that is $E[Y_i(1)|X = x] - E[Y_i(0)|X = x] = E[Y_i(1) - Y_i(0)|X = x]$

(g) This will lead to omitted variable bias and a biased and inconsistent τ.

Problem 2

```
library(foreign)
library(plm)

house=read.dta('~/Desktop/CausalInf/FinalExam/house.dta')
#transfer data into panel frame
house_data=pdata.frame(house, index = c("id", "post"))
# select data with treated equals 0
house0<-house_data [house$treated==0,]
 # select data with treated equals 1
house1<-house_data [house$treated==1,]
```

(a)

```
# Regression in (a)
lma=plm(rent~inc+famsiz+phoenix+post+phoenix*post+hed+femh+race,data=house0, model="pooling")
summary(lma)
```

Coefficients	Estimate	Std. Error	t-value	Pr(>|t|)
(Intercept)	4.5267e+01	4.8722e+00	9.2907	< 2.2e-16 ***
inc	2.7249e-03	5.1507e-04	5.2903	1.419e-07 ***
famsiz	5.2084e+00	6.5466e-01	7.9558	3.689e-15 ***
phoenix	1.1649e+01	3.7478e+00	3.1083	0.00192 **
post1	5.2629e+00	2.3528e+00	2.2368	0.02546 *
hed	3.2851e+00	3.8083e-01	8.6260	< 2.2e-16 ***
femh	9.2948e+00	2.2722e+00	4.0907	4.551e-05 ***
race	-1.5613e+01	2.5375e+00	-6.1529	9.960e-10 ***
phoenix:post1	2.2020e+01	5.2488e+00	4.1954	2.899e-05 ***

The coefficients before inc is 2.7249e-03 with SE as 5.1507e-04, that is very significant at the level of 0.001. It suggests that by controling all other variables, one additional dollar in annual income will let the average montly rents paid significantly marginally increase 2.7249e-03.

(b)

```
# Regresion in (b)
lmb=plm(rent~inc+famsiz+phoenix+post+phoenix*post+hed+femh+race,data=house0, model="fd")
summary(lmb)
```

Coefficients	Estimate	Std. Error	t-value	Pr(>|t|)
inc	9.2293e-04	5.9592e-04	1.5488	0.1219
famsiz	9.4002e+00	1.8343e+00	5.1248	3.886e-07 ***
hed	2.7777e+00	1.6969e+00	1.6369	0.1021
femh	3.7586e+00	5.6096e+00	0.6700	0.5031
phoenix:post1	2.9501e+01	3.4821e+00	8.4724	< 2.2e-16 ***

After the first-differences, the coefficient before inc is 9.2293e-04 with a p-value as 0.1219, has been not significant in 0.1 level. This suggests that when controling all other variables, one additional dollar in annual income will make the average montly rents paid marginalling incease 9.2293e-04, thought not significantly. This result is different from the ones in model in (a).

The reason is that data under out topic has individual effect. The pooling model in (a) can't solve this issue, so this leads to inconsistent results in coefficients estimation. But the result after first-difference has

consistency. Thus the method and result used in (b) is more reliable.

(c)

```
lmc=plm(rent~inc+famsiz+phoenix+post+phoenix*post+hed+femh+race,data=house0,model="within",index=c("id"
summary(lmc)
```

Coefficients	Estimate	Std. Error	t-value	Pr(>|t|)
inc	-0.00069523	0.00065712	-1.0580	0.29044
famsiz	10.54520432	1.81036640	5.8249	8.823e-09 ***
post1	9.80579464	1.82479565	5.3736	1.062e-07 ***
hed	2.94064958	1.66342912	1.7678	0.07754 .
femh	-1.10397918	5.57202167	-0.1981	0.84300
phoenix:post1	20.81238311	3.77647151	5.5111	5.069e-08 ***

Our coefficient before inc is negative -0.00069523, with p-value as 0.29044, not significant in the level of 0.1.

This suggests that by controling other variables, one additional dollar in annual income will let the average monthly rents paid decrease 0.00069523, which is not significant. Not like any results from (a) and (b)

(d)

```
predict(lma)
predict(lmb)
predict(lmc)
```

I intended to find a function that can predict the interval, but failed to get such a function. Thus I used the point estimate instead.

(e) The beta1 is the difference between first year and second year by controling the teared and its interaction term., this should be positive. Beta2 is the influence of treated by controling the post and its interaction terms, which should be 0. Beta3 is the difference on treatment group before the treatment and after the treatment, which is expected to be positive.

(f)

```
lm_e=plm(rent~post+treated+treated*post,data=house_data,model="pooling")
summary(lm_e)
```

Coefficients	Estimate	Std. Error	t-value	Pr(>|t|)
(Intercept)	117.4852	1.7088	68.7518	< 2.2e-16 ***
post1	10.0071	2.4167	4.1409	3.601e-05 ***
treated	-5.4980	3.0913	-1.7786	0.0754592 .
post1:treated	14.4545	4.3717	3.3064	0.0009615 ***

This result is different from (d), beta3 here really express the treatment effect.

(g) BetaA should be smaller than BetaB, because A uses all the data, and the no treatment in first year will tend to lower the treatment effect.

(h)

```
lmh=plm(rent~post+treated+treated*post+phoenix+phoenix*post+treated*phoenix+treated*post*phoenix,data=h
summary(lmh)
```

Coefficients	Estimate	Std. Error	t-value	Pr(>|t|)
(Intercept)	112.8837	2.2098	51.0831	< 2.2e-16 ***
post1	7.3366	3.1251	2.3476	0.018989 *
treated	-6.5195	4.0357	-1.6155	0.106367
phoenix	10.6967	3.3692	3.1748	0.001522 **
post1:treated	4.8194	5.7073	0.8444	0.398530
post1:phoenix	6.2076	4.7648	1.3028	0.192785
treated:phoenix	1.9248	6.0785	0.3167	0.751538
post1:treated:phoenix	21.4133	8.5963	2.4910	0.012818 *

Beta0 is the average fixed effect, and beta1 is the time fixed effect, beta2 is the average difference between treatment group and control group, beta 3 is the marginal effect of phoenix, beta5 is the time difference of phoenix, beta 6 is the difference of phoenix between treatment group and control group, beta 7 is the difference between 2nd year treatment group and other samples after controling all previous factors. The real treatment effect should be beta3.

(i)

```
library(zoo)
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
library(lmtest)
coeftest(lmh, vcov=function(x) vcovHC(x, cluster="group", type="HC1"))
```

```
##
## t test of coefficients:
##
##                       Estimate Std. Error t value  Pr(>|t|)
## (Intercept)           112.8837     1.8285 61.7365 < 2.2e-16 ***
## post1                   7.3366     1.8523  3.9609 7.724e-05 ***
## treated                -6.5195     2.9876 -2.1822  0.029210 *
## phoenix                10.6967     3.2054  3.3371  0.000862 ***
## post1:treated           4.8194     3.3715  1.4295  0.153024
## post1:phoenix           6.2076     2.9680  2.0915  0.036608 *
## treated:phoenix         1.9248     5.2296  0.3681  0.712870
## post1:treated:phoenix  21.4133     5.6617  3.7821  0.000160 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Coefficients	Estimate	Std. Error	t-value	Pr(>|t|)
(Intercept)	112.8837	1.8285	61.7365	< 2.2e-16 ***
post1	7.3366	1.8523	3.9609	7.724e-05 ***
treated	-6.5195	2.9876	-2.1822	0.029210 *
phoenix	10.6967	3.2054	3.3371	0.000862 ***
post1:treated	4.8194	3.3715	1.4295	0.153024
post1:phoenix	6.2076	2.9680	2.0915	0.036608 *
treated:phoenix	1.9248	5.2296	0.3681	0.712870
post1:treated:phoenix	21.4133	5.6617	3.7821	0.000160 ***

Almost all coefficents are significant so we can trust the model.

Problem 3 LATE

It is important to know that some students went back home at the canvassing time, so there are some units assigned to treatment group but not contacted/treated, as the data shows.

(a)

```
library(foreign)

data3 <- read.dta("~/Desktop/CausalInf/FinalExam/beijing.dta")

# data quantification: transforming yes into 1 and no into 0
require(dplyr)
data3 <- data3 %>%
  mutate(turnout = ifelse(turnout == "no",0,1)) %>%
  mutate(treat2 = ifelse(treat2 == "no",0,1)) %>%
  mutate(contact = ifelse(contact == "no", 0,1))

# ITT/Reduced form, taking the assignment variable treat2 as the IV,
# and contact as treatment variable
summary(lm(turnout ~ treat2, data = data3))
```

As we can see the ITT is 0.13193, with the standard error 0.01422. The calculated mean difference corresponds to the ITT, meaning that given assigned treatment exerts a causal effect equal to 0.13193. And the standard error is used to explain whether or not the mean difference is also statistically significant among the wider population. We generally reject the null hypothesis (no difference within the wider population) at $p<0.05$.

The assumptions are

0. SUTVA
1. treat2 has 0 covariance with error term of treat2 on contact,
2. first stage effect of treat2 on contact is not 0, this assumption ensures that there is significant first stage
3. treat2 has 0 covariance with error term of contact on turnout; it says that the counterfactual outcome only depends on D, and once you know D you do not need to know z. This is the exclusion restriction, and it implies that we can write Y (z; D) = Y (D).

Yes, they are plausible.

Note that the experiment randomized for dorm rooms not for students. This violates the assumption as treatment assignment very likely to be clustered within dorm rooms. Even worse, several background characteristics of individuals are also very likely to be clustered the same way (age, gender, major field, being friends with each other, ect.). All of these factors are excellent sources of bias in a study of voting turnout. Also note that ITT estimates tend to be more conservative than LATE estimates because they include always takers, never takers, and defiers (this adds a lot of random noise); ITT is problematic, however, if always takers, never takers, and defiers aren't randomly distributed. Here, selection bias becomes immanent, if one group (say students in dorm room 5) are more likely to not comply given treatment if they (for instance because they are all part of a club opposed to the communist party). Ideally for ITT you want the random noise to remain random and not to be clustered in any way.

(b)

```
# Start manually compute the latedata3$turnout=ifelse(data3$turnout=="yes",1,0)
data31=data3[data3$treat2==1,]
data30=data3[data3$treat2==0,]
v1=mean(data31$turnout,na.rm=TRUE)-mean(data30$turnout)
v2=mean(data31$contact)-mean(data30$contact)
LATE=v1/v2
```

```
LATE

# Se
se = 0.01422/0.8859
se
```

LATE = 0.1489, with the SE 0.01605147.

The assumptions are

0. SUTVA
1. treat2 has 0 covariance with error term of treat2 on contact,
2. first stage effect of treat2 on contact is not 0, this assumption ensures that there is significant first stage
3. treat2 has 0 covariance with error term of contact on turnout; it says that the counterfactual outcome only depends on D, and once you know D you do not need to know z. This is the exclusion restriction, and it implies that we can write Y (z; D) = Y (D).

Yes, they are plausible.

(c) The complier ratio is 0.8859. The pattern of compliance means basically create a cross tab of assigned encouragement and received treatment. To be done?

```
# Compliers Ratio
C <- mean(data3[data3$treat2==1,]$contact) - mean(data3[data3$treat2==0,]$contact)
C
```

Problem 4

(a)

i. ATE

$\tau_{ATE} = E[response_1 - response_0] = \int\int\int E[response_1 - response_0|cov_1, cov_2, cov_3]dP(cov_1)dP(cov_2)dP(cov_3)$
$= \int\int\int (E[response_1|cov_1, cov_2, cov_3, D = 1] - E[response_0|cov_1, cov_2, cov_3], D = 0)dP(cov_1)dP(cov_2)dP(cov_3)$

ii. ATT

$\tau_{ATT} = E[response_1 - response_0|D = 1] = \int\int\int (E[response_1|cov_1, cov_2, cov_3, D = 1] - E[response_0|cov_1, cov_2, cov_3], D = 0)dP(cov_1|D = 1)dP(cov_2|D = 1)dP(cov_3|D = 1)$

iii. ATC

$\tau_{ATC} = E[response_1 - response_0|D = 0] = \int\int\int (E[response_1|cov_1, cov_2, cov_3, D = 1] - E[response_0|cov_1, cov_2, cov_3], D = 0)dP(cov_1|D = 0)dP(cov_2|D = 0)dP(cov_3|D = 0)$

(b)

```
data4 <- read.csv("~/Desktop/CausalInf/FinalExam/Exam2018sim.csv")

# Use probit regresion to produce the propensity score and named them est.prop
psout <- glm(treatment ~ cov1 + cov2 + cov3, family = binomial(link = "probit"), data=data4)
summary(psout)
data4$est.prop <- psout$fitted
```

(c)

```
# i. Match the units without replacement
library(Matching)
# xvars <- c("cov1", "cov2", "cov3")
m <- Match(Y=data4$response, Tr=data4$treatment,X= data4$est.prop, M=1 , replace = FALSE)

# ii. Balance check on covariates
#mb <- MatchBalance(treatment ~ cov1 + cov2 + cov3, data=data4 , match.out=m)

# ii. Balance check on prop score
mb <- MatchBalance(treatment ~ est.prop, data=data4 , match.out=m)
```

ii.

No, the balance is not improved as we see just similiar results before and after matching.

iii.

Yes, but not much.

(d)

```
summary(lm(response~treatment, data = data4))
summary(lm(response~treatment + cov1 + cov2 + cov3, data = data4))
summary(lm(response~treatment + est.prop, data = data4))

data5 <- data4
```

The results from i

Coefficients	Estimate	Std. Error	t-value	Pr(>|t|)
(Intercept)	-0.40408	0.05890	-6.861	7.67e-12 ***
treatment	3.23690	0.07399	43.748	< 2e-16 ***

The results from ii

Coefficients	Estimate	Std. Error	t-value	Pr(>|t|)
(Intercept)	-0.02469	0.05395	-0.458	0.647
treatment	2.61419	0.07021	37.234	<2e-16 ***
cov1	0.38779	0.03267	11.870	<2e-16 ***
cov2	0.84564	0.03068	27.559	<2e-16 ***
cov3	0.95362	0.03236	29.466	<2e-16 ***

The results from iii

Coefficients	Estimate	Std. Error	t-value	Pr(>|t|)
(Intercept)	-2.47852	0.11050	-22.43	<2e-16 ***
treatment	2.48852	0.07859	31.66	<2e-16 ***
est.prop	4.02550	0.18453	21.82	<2e-16 ***