ECON312 Problem Set 3: question 3

Futing Chen, Hongfan Chen, Will Parker

04/22/2020

Contents


```
library(AER)
library(broom)
library(knitr)
library(haven)
library(dplyr)
library(ggplot2)
```

Problem 3

a) Discuss instrument exogeneity, exclusion and monotonicity.

- *Exclusion restriction* requires that any effect of the lottery outcome (Z) on earnings (Y) must be through attending medical school (D), i.e., $Y_{d,0} = Y_{d,1},\ d \in \{0,1\},\ \forall i$. This assumption may fail.The try-again scheme implies that some lottery winners are those who lost the lottery before and hence put off attending the medical school and entering the job market. Winning lottery may thus be correlated with earnings by reducing work experience.
- *Random assignment* assumption requires that the lottery outcomes are independent of potential outcomes and potential treatment status, i.e., $(Y_{11}, Y_{10}, Y_{01}, Y_{00}, D_1, D_0) \perp Z\ \forall i$. This assumption fails if lottery winners and losers differed in important unobserved characteristics. Since the probability of winning the lottery depended on high school grades, (D_1, D_0) would not be independent of Z if the lottery winners were mostly those with high learning ability and motivation and hence were more likely to take the treatment than the lottery losers otherwise would have done, making the distribution of D_1 different across lottery winners and losers. Clearly, the potential Y would not be independent of Z either, if the lottery winners, with higher ability, earned more incomes by attending medical school than the lottery losers would otherwise have earned.
- *Instrument exogeneity* is the combinition of exclusion restriction and random assignment. If one of the two assumptions fails, instrument exogeneity will fail.
- In this setting, *monotonicity* means $D_1 \geqslant D_0\ \forall i$. This holds if there were no defiers, i.e., no one attended the medical school when he did not win the lottery and chose not to attend the school when he won the lottery. Since only those who apply for medical studies were assigned lotteries and were unlikely to forgo the chance if they won the lottery, monotonicity is reasonable.

b) Assess instrument relevance.

```
lottery <- read_dta("~/Desktop/UChicago/Empirical Analysis/EA III/part A/ps/ps3/lottery.dta")
first_stage <- lm(d~z, data=lottery)
kable(tidy(first_stage), digits=4, align='c',caption="First Stage")
```

Table 1: First Stage

term	estimate	std.error	statistic	p.value
(Intercept)	0.4101	0.0162	25.3159	0
z	0.5203	0.0195	26.7012	0

The coefficient on z in the first stage is nonzero and highly statistically significant ($t = 26.7,\ p < .000$). Thus, we can conclude that Z satisfies instrument relevance.

c) Estimate the return to attending medical school on earnings in 2007 using IV, and interpret the results.

```
iv<-ivreg(lnw~d|z, data=lottery)
kable(tidy(iv), digits=4, align='c',caption="Results of IV Regression")
```

Table 2: Results of IV Regression

term	estimate	std.error	statistic	p.value
(Intercept)	3.0106	0.0407	73.9813	0e+00
d	0.1871	0.0505	3.7076	2e-04

```
d_iv<-coef(iv)[2]
100*(exp(d_iv)-1)
```

```
##        d
## 20.57689
```

The IV estimate of d is 0.1871 ($t = 3.708, p < .000$). This result suggests that, on average, attending medical schools increased earnings in 2007 by 20.58% among the compliers, i.e., the applicants who attended medical schools after winning the lottery and did not attend medical schools after losing the lottery.

d) Count the number of compliers, and compare them to the population of applicants in terms of gender.

The share of compliers is the wald first-stage:

$$P(D_1 = 1, D_0 = 0) = E[D|Z = 1] - E[D|Z = 0]$$

The relative likelihood a complier's gender X is $x \in 0, 1$ compared to the population is

$$\frac{P(X = x|D_1 > D_0)}{P(X = x)} = \frac{E[D|Z = 1, X = x] - E[D|Z = 0, X = x]}{E[D|Z = 1] - E[D|Z = 0]}$$

```
#The size of compliers
z_firststage<-coef(first_stage)[2]
num_complier<-round(dim(lottery)[1]*z_firststage)

#The relative likelihood a complier is female
z_f <- coef(lm(d~z, data=lottery[lottery$female==1,]))[2]
complier_f<-z_f/z_firststage

#The relative likelihood a complier is male
z_m <- coef(lm(d~z, data=lottery[lottery$female==0,]))[2]
complier_m<-z_m/z_firststage

results <- tibble("quantity of interest"=c("number of compliers",
                                           "relative likelihood a complier is female",
                                           "relative likelihood a complier is male"),
                  "estimate"=c(num_complier,complier_f,complier_m))
kable(results,digits=3)
```

quantity of interest	estimate
number of compliers	768.000
relative likelihood a complier is female	0.969
relative likelihood a complier is male	1.050

The result shows that the compliers are more likely to be males and less likely to be females than the average applicant in the sample.

e) Is the IV estimate an estimate of the ATT? Explain why or why not.

$$ATT = E[Y_1 - Y_0|D=1] = \underbrace{E[Y_1 - Y_0|D_1 > D_0]}_{LATE} P(D_1 > D_0|D=1) + \underbrace{E[Y_1 - Y_0|D_0 = 1]}_{\text{effect on always-takers}} P(D_0 = 1|D=1)$$

where the two probabilities sum to one and

$$P(D_1 > D_0|D=1) = \frac{P(Z=1)(E[D|Z=1] - E[D|Z=0])}{P(D=1)}$$

Hence, the IV estimate is ATT if $P(D_1 > D_0|D=1) = 1$ or, equivalently, there are no always-takers. The latter is unlikely to be true, because for applicants who lost the lottery in 1989, some of them might win in 1990 and then attend medical schools. We can also check from data by estimating $P(D_1 > D_0|D=1)$:

```
complier_t <- mean(lottery$z)*z_firststage/mean(lottery$d)
kable(complier_t, col.names = "P(D1>D0|D=1)")
```

	P(D1>D0|D=1)
z	0.4671748

Because 0.467 is far smaller than 1, the IV estimate is unlikely to estimate ATT.

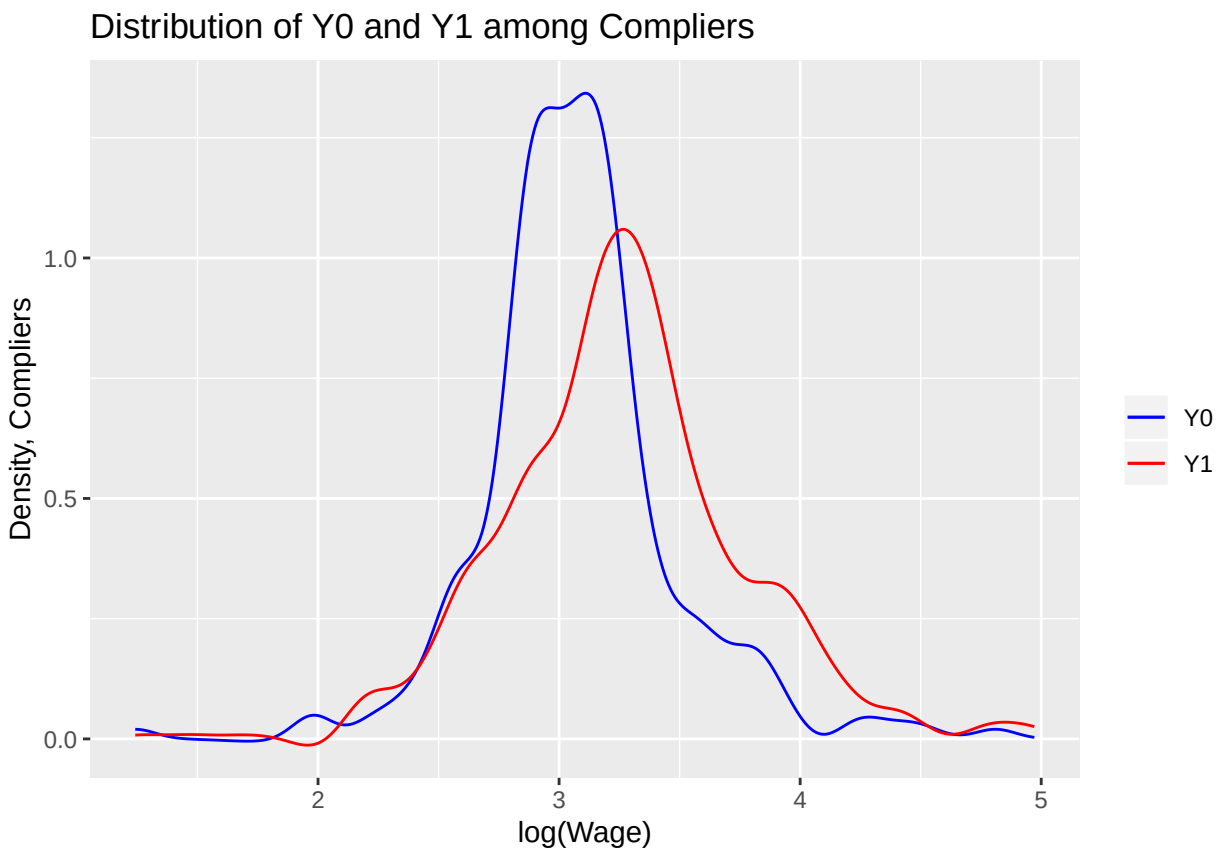
f) Estimate the mean and distribution of Y0 and Y1 for compliers.

```
lottery <- mutate(lottery, Y1=lnw*d, Y0=lnw*(1-d),md=1-d)
mean_Y1 <- coef(ivreg(Y1~d|z, data=lottery))[2]
mean_Y0 <- coef(ivreg(Y0~md|z, data=lottery))[2]
kable(tibble("quantity of interest"=c("E[Y1|C=c]",
                                      "E[Y0|C=c]"),
             "estimate"=c(mean_Y1,mean_Y0)))
```

quantity of interest	estimate
E[Y1|C=c]	3.264167
E[Y0|C=c]	3.077049

```
# estimate the share of compliers, never-takers and always-takers
pc <- z_firststage
pn <- mean(lottery$md[lottery$z==1])
pa <- mean(lottery$d[lottery$z==0])
# estimate pdf for each cell and counterfactual distributions
dist <- lottery %>% group_by(z,d) %>%
  summarise (pdf=list(density(lnw, from = min(.$lnw), to = max(.$lnw))))
x   <- dist$pdf[[1]]$x
```

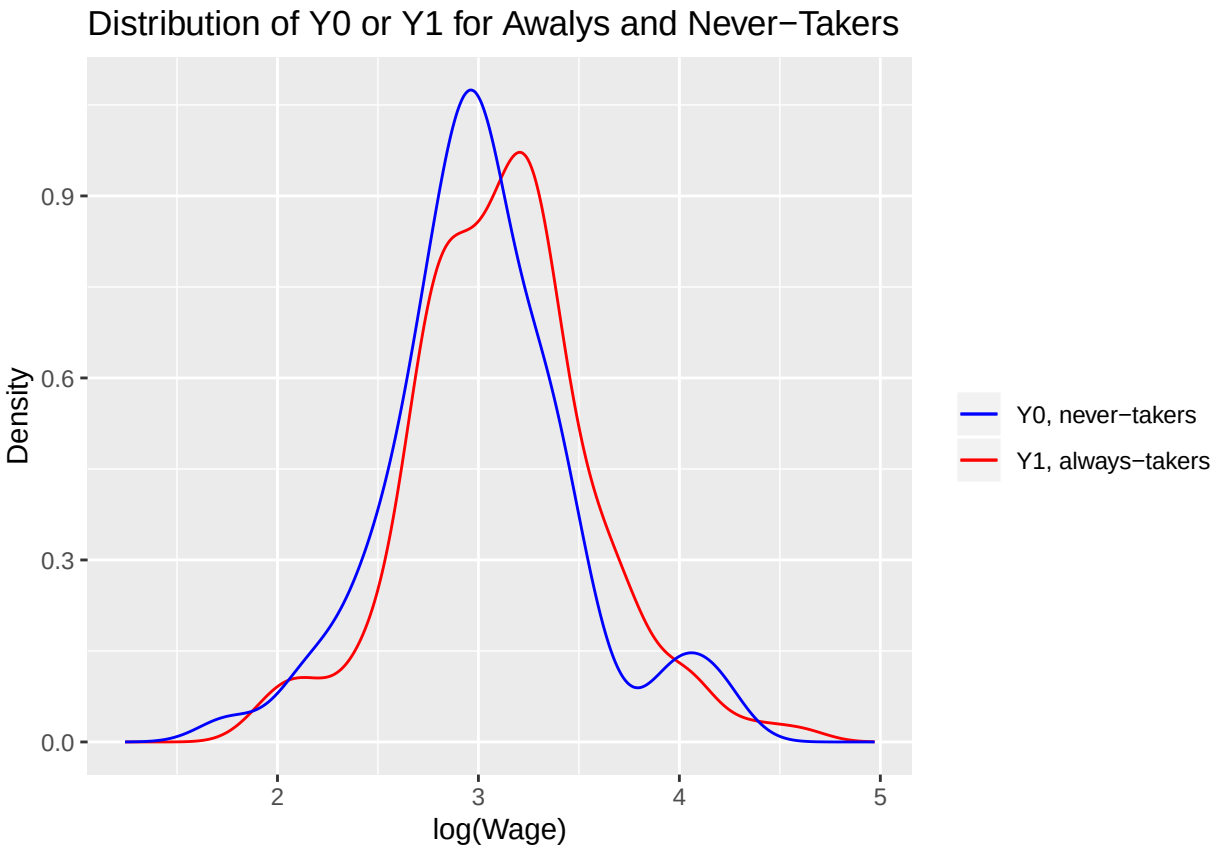
```r
f00 <- dist$pdf[[1]]$y
f01 <- dist$pdf[[2]]$y
f10 <- dist$pdf[[3]]$y
f11 <- dist$pdf[[4]]$y
dist <- data.frame(x, f00, f01, f10, f11)
dist <- mutate(dist,
               gc0 = f00*(pc+pn)/pc-f10*pn/pc,
               gc1 = f11*(pc+pa)/pc-f01*pa/pc)
# plot
ggplot(dist, aes(x)) +
  geom_line(aes(y=gc0, color = "Y0")) +
  geom_line(aes(y=gc1, color = "Y1")) +
  xlab("log(Wage)") + ylab("Density, Compliers") +
  ggtitle("Distribution of Y0 and Y1 among Compliers") +
  scale_color_manual("",values = c("blue","red"))
```



g) What can you say about Y0 and Y1 for always- and never-takers?

For always-takers, $D_1 = D_0 = 1$, so $Y = Y_1$. Similarly, for never-takers, $Y = Y0$. Their distributions are:

```r
ggplot(dist, aes(x)) +
  geom_line(aes(y=f01, color = "Y1, always-takers")) +
  geom_line(aes(y=f10, color = "Y0, never-takers")) +
  xlab("log(Wage)") + ylab("Density") +
  ggtitle("Distribution of Y0 or Y1 for Awalys and Never-Takers") +
  scale_color_manual("",values = c("blue","red"))
```



The plot shows that the potential earnings of always-takers are in general higher than the potential earnings of never-takers.

h) Estimate lottery category×year specific LATEs and combine these in one estimate. Compare this to the specification where you control for lottery category×year dummies and also interact the instrument with these dummies.

```
# Estimate covariate-specific LATEs
LATEs <- lottery %>% group_by(year,lotcateg) %>%
  do(LATE = coef(lm(lnw~d|z,data=.))[2]) %>%
  mutate(LATE=as.numeric(LATE))

kable(tibble("year"=c(1988,1989),
             "Category 3"=LATEs$LATE[LATEs$lotcateg==3],
             "Category 4"=LATEs$LATE[LATEs$lotcateg==4],
             "Category 5"=LATEs$LATE[LATEs$lotcateg==5],
             "Category 6"=LATEs$LATE[LATEs$lotcateg==6]),
      digits=4, align='c', caption="Lottery Category × Year Specific LATEs")
```

Table 6: Lottery Category × Year Specific LATEs

year	Category 3	Category 4	Category 5	Category 6
1988	-0.3915	0.1160	0.1394	0.0910
1989	1.1366	0.0515	0.1671	0.0642

By problem 2 b), we can combine the covariate-specific LATEs into conditional LATE:

$$E[Y_1 - Y_0|D_1 > D_0] = \sum_x LATE(x) \frac{E[D|Z=1, X=x] - E[D|Z=0, X=x]}{E[D|Z=1] - E[D|Z=0]} P(X=x)$$

```
Px <- group_by(lottery, year, lotcateg)  %>%
  summarize(Px = n()/dim(lottery)[1])
z_x <-  group_by(lottery, year, lotcateg)  %>%
  do(z_x = coef(lm(d~z,data=.))[2]) %>%
  mutate(z_x=as.numeric(z_x))
LATE <- left_join(LATEs, z_x,  by=c('year', 'lotcateg')) %>%
  left_join(., Px,by=c('year', 'lotcateg'))
LATE_uncond <- sum(LATE$LATE*(LATE$z_x/z_firststage)*LATE$Px)
```

Alternatively, we can use Abadie's kappa to estimate unconditional LATE:

$$E[Y_1 - Y_0|D_1 > D_0] = \frac{1}{P(D_1 > D_0)} E[\kappa LATE(x)]$$

where $\kappa = 1 - \frac{D(1-Z)}{P(Z=0|X)} - \frac{(1-D)Z}{P(Z=1|X)}$.

```
# use Abadie's kappa to estimate E[Y1-Y0|D1>D0]
lottery <- group_by(lottery, year, lotcateg)  %>%
  mutate(P_Z1_X=mean(z),      # P(Z=1|X)
         P_Z0_X=1-mean(z))    # P(Z=0|X)
lottery <- left_join(lottery, LATEs, by=c('year', 'lotcateg'))
lottery <- mutate(lottery,
                  kappa=1-d*(1-z)/P_Z0_X-z*(1-d)/P_Z1_X,
                  kappa_LATE=kappa*LATE)  #kappa*covariate-specific LATE
kappa_LATE <- mean(lottery$kappa_LATE)
# compute unconditional LATE
LATE_uncond_k <- kappa_LATE/z_firststage
kable(tibble("method"=c("integrate over X", "Abadie's kappa"),
             "estimate"=c(LATE_uncond,LATE_uncond_k)),
      caption = "Unconditional LATE E[Y1-Y0|D1>D0]")
```

Table 7: Unconditional LATE E[Y1-Y0|D1>D0]

method	estimate
integrate over X	0.1128143
Abadie's kappa	0.1128143

Next, we estimate the following specification using 2SLS:

$$lnw = \alpha + \beta d + \alpha_x I\{X=x\} + \alpha_{xd} I\{X=x\} * d + e$$
$$d = \gamma + \pi z + \gamma_x I\{X=x\} + \gamma_{xz} I\{X=x\} * z + u$$

```
iv_c <- ivreg(lnw ~ d*factor(lotcateg)*factor(year) |
                z*factor(lotcateg)*factor(year) , data=lottery)
kable(tidy(iv_c), digits=4, align='c',caption="2SLS Estimation")
```

Table 8: 2SLS Estimation

term	estimate	std.error	statistic	p.value
(Intercept)	3.8725	0.9336	4.1479	0.0000
d	-0.8037	1.0038	-0.8006	0.4235
factor(lotcateg)4	-0.7353	0.9529	-0.7717	0.4404
factor(lotcateg)5	-0.7867	0.9380	-0.8387	0.4018
factor(lotcateg)6	-0.8511	0.9359	-0.9094	0.3633
factor(year)1989	-2.0686	1.1356	-1.8216	0.0687
d:factor(lotcateg)4	0.9232	1.0298	0.8965	0.3701
d:factor(lotcateg)5	0.9629	1.0108	0.9526	0.3409
d:factor(lotcateg)6	1.0132	1.0079	1.0053	0.3149
d:factor(year)1989	2.3679	1.2395	1.9103	0.0563
factor(lotcateg)4:factor(year)1989	2.0381	1.1878	1.7158	0.0864
factor(lotcateg)5:factor(year)1989	1.6897	1.1543	1.4639	0.1434
factor(lotcateg)6:factor(year)1989	2.0412	1.1396	1.7911	0.0735
d:factor(lotcateg)4:factor(year)1989	-2.4754	1.3013	-1.9022	0.0573
d:factor(lotcateg)5:factor(year)1989	-2.0084	1.2643	-1.5886	0.1124
d:factor(lotcateg)6:factor(year)1989	-2.4321	1.2463	-1.9514	0.0512

The 2SLS estimate is -0.8037 ($t = -0.801, p = 0.4235$). This value is statistically insignificant and is of opposite sign of LATE.