# EC 711 Problem Set 2

Samuel Messer

February 16, 2024

*Collaborated with: Erin Eidschun, Rachel Vogt*

## Question 1: Effect of 401(k) on Asset Accumulation

### 1.

We consider

$$\mathbb{E}\left[\frac{(1-D)Y}{1-P(X)}\right]$$

Note that this expectation is taken with respect to $X$. By the law of iterated expectations (LIE), we can take expectations of the terms in the numerators given $X$ and the expression will be equivalent. In addition, the numerator will only be non-zero if $D = 0$, so we can condition $Y$ on $1 - D = 1$, or equivalently $D = 0$. This is the standard "trick" we employed in lecture several times. Then we can write this as

$$\mathbb{E}\left[\frac{(1-D)Y}{1-P(X)}\right] = \mathbb{E}\left[\frac{(1-D)\mathbb{E}\left[Y|X,1-D\right]}{1-P(X)}\right] = \mathbb{E}\left[\frac{\mathbb{E}\left[1-D|X\right]\mathbb{E}\left[Y|X,D=0\right]}{1-P(X)}\right]$$

Note that the first term in the numerator is exactly $1 - \mathbb{E}\left[D|X\right]$, and $\mathbb{E}\left[D|X\right] = P(X)$ by definition, so we have

$$\mathbb{E}\left[\frac{\mathbb{E}\left[1-D|X\right]\mathbb{E}\left[Y|X,D=0\right]}{1-P(X)}\right] = \mathbb{E}\left[\frac{(1-P(X))\mathbb{E}\left[Y|X,D=0\right]}{1-P(X)}\right] = \mathbb{E}\left[\mathbb{E}\left[Y|X,D=0\right]\right]$$

### 2.

We estimate the average treatment effect of 401k eligibility on net total financial assets using 3 methods:

1. The non-parametric regression of Hahn (1998).
2. The propensity score re-weighting estimator of Hirano, Imbens, and Ridder (2005)
3. The doubly robust estimator of Robins, Rotnitzky, and Zhao (1994)

Results for all 3 methods are reported in table 1.

| Estimator | ATE |
|---|---|
| Non-parametric | 7929.00 |
| Propensity Score | 7910.29 |
| Doubly Robust | 7978.99 |

Table 1

All 3 methods yield very similar results. The estimated parameter here is the ATE: that is, conditional on the observables, the effect on net total financial assets for a person of becoming eligible for a 401(k). This has a that exact causal interpretation: given covariates (the low-dimensional specification seen in class), the average effect of being made eligible for a 401(k) on net total financial assets is $\$7,978.99$. The identification conditions here are the conditions SO1-SO3 seen in class:

1. $\mathbb{E}[\alpha_i|X_i, D_i] = \mathbb{E}[Y_{i0}|X_i, D_i] = \mathbb{E}[Y_{i0}|X_i] = \mathbb{E}[\alpha_i|X_i]$
2. $0 < P(X) < 1$
3. $\mathbb{E}[Y_{i1}|X_i, D_i] = \mathbb{E}[Y_{i1}|X_i]$

Conditions 1 and 3 are conditional random assignment conditions, which specify that conditional on the covariates, the outcomes $Y_{i0}$ and $Y_{i1}$ are mean independent of treatment status. In this context, these assumptions mean that conditional on our controls, net total financial assets are mean independent of 401(k) eligibility. This seems a reasonable assumption in this case because conditional on the controls we have included, it is hard to imagine some omitted variable that would affect both treatment status and net total financial assets. Of course, these assumptions are untestable because they are necessary for identification, and it is always possible that such a variable exists and we have just not thought of it.

Condition 2 is a common support condition, and is empirically verifiable (the fact that we get an estimate when applying this method confirms the common support assumption). In this context it means that the sets of treated and untreated individuals have exactly overlapping sets of covariates. If this condition does not hold, we will not get a propensity score re-weighting estimator because we divide by both $P(X)$ and $1 - P(X)$ in the estimation.

## 3.

We estimate the average treatment effect of 401k participation on net total financial assets using 3 methods:

1. The non-parametric regression of Hahn (1998).
2. The propensity score re-weighting estimator of Hirano, Imbens, and Ridder (2005)
3. The doubly robust estimator of Robins, Rotnitzky, and Zhao (1994)

Results for all 3 methods are reported in table 2.

| Estimator | ATE |
|---|---|
| Non-parametric | 11458.16 |
| Propensity Score | 11604.89 |
| Doubly Robust | 11518.57 |

Table 2

All 3 methods yield very similar results. However, the causal interpretation of this estimate is more suspect. The identification conditions here are the conditions SO1-SO3 seen in class:

1. $\mathbb{E}[\alpha_i|X_i, D_i] = \mathbb{E}[Y_{i0}|X_i, D_i] = \mathbb{E}[Y_{i0}|X_i] = \mathbb{E}[\alpha_i|X_i]$
2. $0 < P(X) < 1$
3. $\mathbb{E}[Y_{i1}|X_i, D_i] = \mathbb{E}[Y_{i1}|X_i]$

Given these conditions, the identified parameter is the ATE: conditional on the observables, the effect of participating in the 401(k) on net total assets is $\$11,518.57$. However, one could argue that the above identification conditions do not hold.

Conditions 1 and 3 are conditional random assignment conditions, which specify that conditional on the covariates, the outcomes $Y_{i0}$ and $Y_{i1}$ are mean independent of treatment status. In this context, these assumptions mean that conditional on our controls, net total financial assets are mean independent of 401(k) participation. This does not seem a reasonable assumption in this context because we do not control for 401(k) eligibility. Thus, people who participate in 401(k) plans may have different unobserved characteristics (financial planning goals, etc.) that are correlated with their participation and outcome. We could fix this issue by using an instrumental variables approach as in Abadie (2003), but we leave that analysis for another time.

Condition 2 is a common support condition, and is empirically verifiable (the fact that we get an estimate when applying this method confirms the common support assumption). In this context it means that the sets of participators and non-participators have fully overlapping sets of covariates. This is confirmed by the fact that we get an estimate, as in the process we divide by both $P(X)$ and $1 - P(X)$.

# Question 2: Head Start (Ludwig and Miller, 2007)

The original paper by Ludwig and Miller studies the effect of Head Start on child outcomes using a regression discontinuity framework. The 300 poorest counties in the US received help filling out applications for Head Start funding, generating a sharp discontinuity in Head Start participation at the poverty rate of the 300th poorest county. One main result is that the mortality rate for 5-9 year olds on the treated side of the cutoff falls. The result is robust to several different specifications, both parametric and non-parametric. However, there are several key methods developed for RD designs in the intervening years that can be applied to check the robustness of their results.

In particular, Ludwig and Miller state "Unfortunately there is currently no widely agreed-upon method for selection of optimal bandwidths in the nonparametric RD context." However, improvements have been made on this front. In addition, the rise of randomization inference, where counties near the cutoff are treated as if they were randomly selected into treatment. In what follows, I draw heavily on the work of Cattaneo, Tituinik, and Vazquez-Bare (2017), who revisit the work of Ludwig and Miller with a more robust approach.

We begin with optimal bandwidth setting using the rdrobust command in R. Using a local linear specification, we get the results in table 3.

| | Mortality, age 5-9 | Mortality, age 25+ | Injuries, age 25+ |
|---|---|---|---|
| Estimate | −2.409 | 2.033 | 0.052 |
| Robust SE | (1.206) | (5.977) | (5.378) |
| P-value | 0.042 | 0.866 | 0.731 |
| Optimal Bandwidth | 6.811 | 8.054 | 6.206 |

Table 3: Local linear specification of RD, with optimal bandwidth setting.

This suggests that with optimal bandwidth setting, such that balance holds between treatment and control groups, we find a significant effect of Head Start on mortality. We also find no significant effects on outcomes for people who would not be affected by Head Start, which is reassuring. This is in line with the results in Ludwig and Miller, and we find an optimal bandwidth lower than the lowest they considered, 9. We also consider a local quadratic specification, with results in table 4.

| | Mortality, age 5-9 | Mortality, age 25+ | Injuries, age 25+ |
|---|---|---|---|
| Estimate | −3.474 | 2.243 | −1.989 |
| Robust SE | (1.368) | (7.343) | (6.151) |
| P-value | 0.009 | 0.751 | 0.731 |
| Optimal Bandwidth | 7.578 | 10.098 | 7.897 |

Table 4: Local quadratic specification of RD, with optimal bandwidth setting.

In the local quadratic specification, we once again find a significant effect on mortality but not the placebo controls. In addition, in both specifications, the standard errors and p-values are robust.

Finally, we consider a local randomization strategy. We first determine the appropriate window using a covariate balance test. We want the widest window with a p-value above 0.15, which is generally the level at which balance holds (or at least the profession's agreed upon number). Doing this gives us an optimal window of $\pm 1.3$. this differs from the level in Cattaneo et al, which is 1.1. For consistency, we follow that number instead of the one we arrived at through optimal window selection. The results we get using randomization inference are displayed in table 5

| | p(0) | p(1) |
|---|---|---|
| Estimate−2.280 | −2.515 | |
| Robust P-value0.011 | 0.005 | |

Table 5: Randomization inference with a bandwidth of ±1.1.

The column headers refer to the order of polynomial fitted on each side of the discontinuity. When running with window ±1.3 as I found, we actually find an insignificant effect when using a local linear model. Thus, we choose to follow the Cattaneo et al paper. As we see, the result in Ludwig and Miller holds up when using randomization inference as well. We get negative significant estimates of the effect of Head Start on mortality. Even with more robust methods, the original analysis in Ludwig and Miller holds.

# Appendix: Code

The following are the two R scripts used to produce the analysis in this problem set. Significant parts of question 2 come directly from the replication package of Cattaneo et al.

## Question 1

```
# Samuel Messer
# EC711 Problem Set 2
# Worked closely on code with Erin Eidschun
library(tidyverse) # Data manipulation
library(readstata13) # To read in stata data file
library(xtable) # To make output convenient

# Clear environment
rm(list = ls())

# Read in and clean data
sipp_raw <- read.dta13("data/sipp1991.dta")

# Split income into 7 quantiles
break_points = as.numeric(quantile(sipp_raw$inc, probs = ((1:6) / 7) ))
sipp <- sipp_raw %>%
  mutate(inc_bin = as.factor(findInterval(inc, break_points)),
         inc2 = inc^2, # Quadratic in income
         age2 = age^2, age3 = age^3, # Cubic in age
         fsize2 = fsize^2, # Quadratic in fsize
         educ2 = educ^2) # Quadratic in educ

# Generate low-dim control matrix
# Need this to calculate fitted values later
ld_control <- model.matrix(~ marr + twoearn + db + pira + hown +
                             fsize + fsize2 + educ + educ2 + age + age2 + age3 +
                             inc_bin + inc_bin:inc + inc_bin:inc2, data = sipp)

# Remove raw data
rm(sipp_raw)

### Eligibility as treatment

## Non-parametric
# Covariate control formula
ctrl_form = net_tfa ~ marr + twoearn + db + pira + hown +
                      fsize + fsize2 + educ + educ2 + age + age2 + age3 +
                      inc_bin + inc_bin:inc + inc_bin:inc2

# Fit OLS model only on treated
elg_tr <- sipp[sipp$e401 == 1,]
mod_elg_tr <- lm(ctrl_form, data = elg_tr)

# Fit OLS model only on untreated
elg_untr <- sipp[sipp$e401 == 0,]
mod_elg_untr <- lm(ctrl_form, data = elg_untr)
```

```
# Calculate fitted Y vals
elg_coef = mod_elg_tr$coefficients - mod_elg_untr$coefficients
y_hat_np_elg = as.matrix(ld_control) %*% as.matrix(elg_coef)

ATE_np_elg = mean(y_hat_np_elg)

## Propensity score reweighting
mod_elg_ps = glm(e401 ~ marr + twoearn + db + pira + hown +
                 fsize + fsize2 + educ + educ2 + age + age2 + age3 +
                 inc_bin + inc_bin:inc + inc_bin:inc2, data = sipp,
                 family = binomial(link = "logit"))

# Fitted propensity scores
ps = fitted(mod_elg_ps)

# Check if common support holds
min(ps) > 0
max(ps) < 1

# Calculate ATE according to HIR
ATE_ps_elg = mean((sipp$net_tfa * sipp$e401 / ps) - (sipp$net_tfa * (1 - sipp$e401) / (1 - ps)))


## Doubly robust estimator
# We calculated the expectations for the adjustment already in the np estimator
# Treated
tr_fit = as.matrix(ld_control) %*% as.matrix(mod_elg_tr$coefficients)

# Untreated
untr_fit = as.matrix(ld_control) %*% as.matrix(mod_elg_untr$coefficients)

# Adjustment term in dr estimator
dr_elg_adj = mean((sipp$e401 * tr_fit) / ps - (1 - sipp$e401) * untr_fit / (1 - ps))
ATE_dr_elg = ATE_np_elg + ATE_ps_elg  - dr_elg_adj


print(xtable(data.frame(Estimator = c("Non-parametric", "Propensity Score", "Doubly Robust"),
                        ATE = c(ATE_np_elg, ATE_ps_elg, ATE_dr_elg))),
      include.rownames = F)

### Participation as treatment

## Non-parametric
# Fit OLS model only on treated
par_tr <- sipp[sipp$p401 == 1,]
mod_par_tr <- lm(ctrl_form, data = par_tr)

# Fit OLS model only on untreated
par_untr <- sipp[sipp$p401 == 0,]
mod_par_untr <- lm(ctrl_form, data = par_untr)

# Plug into framework to predict Y
par_coef = mod_par_tr$coefficients - mod_par_untr$coefficients
y_hat_np_par = as.matrix(ld_control) %*% as.matrix(par_coef)
```

```
ATE_np_par = mean(y_hat_np_par)

## Propensity score reweighting
mod_par_ps = glm(p401 ~ marr + twoearn + db + pira + hown +
                   fsize + fsize2 + educ + educ2 + age + age2 + age3 +
                   inc_bin + inc_bin:inc + inc_bin:inc2, data = sipp,
                 family = binomial(link = "logit"))

# Fitted propensity scores
ps = fitted(mod_par_ps)

# Check if common support holds
min(ps) > 0
max(ps) < 1

# Calculate ATE according to HIR
ATE_ps_par = mean((sipp$net_tfa * sipp$p401 / ps) - (sipp$net_tfa * (1 - sipp$p401) / (1 - ps)))


## Doubly robust estimator
# We calculated the expectations for the adjustment already in the np estimator
# Treated
tr_fit = as.matrix(ld_control) %*% as.matrix(mod_par_tr$coefficients)

# Untreated
untr_fit = as.matrix(ld_control) %*% as.matrix(mod_par_untr$coefficients)

# Adjustment term in dr estimator
dr_par_adj = mean((sipp$p401 * tr_fit) / ps - (1 - sipp$p401) * untr_fit / (1 - ps))
ATE_dr_par = ATE_np_par + ATE_ps_par  - dr_par_adj


print(xtable(data.frame(Estimator = c("Non-parametric", "Propensity Score", "Doubly Robust"),
                        ATE = c(ATE_np_par, ATE_ps_par, ATE_dr_par))),
      include.rownames = F)
```

## Question 2

```
# Samuel Messer
# EC711 Problem Set 2
# This code draws heavily on the replication package from Cattaneo, Titiunik,
# and Vazquez-Bare (2017)
rm(list = ls())

library(readstata13)
library(rdlocrand)
library(rdrobust)
library(rddensity)
library(stargazer)

# Read data
headstart <- read.dta13("data/headstart.dta")
```

```
# Set cutoff poverty rate
cutoff = 59.1984
# Outcome variable of interest
Y = headstart$mort_age59_related_postHS

# Generate running variable
R = headstart$povrate60 - cutoff

# Generate treatment indicator
D = as.numeric(R >= 0)

# Placebo outcomes
# These people should not be affected by head start, so we would expect to see no effect here
Plac <- cbind(headstart$mort_age25plus_related_postHS, headstart$mort_age25plus_injuries_postHS)

#Initialize table to store results
local_linear <- array(NA, dim = c(4, 3))

# Local Linear Regression
# masspoints and stdvars are applied according to instructions in replication package
# Outcome of interest
tmp <- rdrobust(Y, R, p = 1, masspoints = "off", stdvars = "on")
local_linear[1, 1] = tmp$coef[1]
local_linear[2, 1] = tmp$se[1]
local_linear[3, 1] = tmp$pv[3]
local_linear[4, 1] = tmp$bws[1,1]

# Placebo outcomes
tmp <- rdrobust(Plac[, 1], R, p = 1, masspoints = "off", stdvars = "on")
local_linear[1, 2] = tmp$coef[1]
local_linear[2, 2] = tmp$se[1]
local_linear[3, 2] = tmp$pv[3]
local_linear[4, 2] = tmp$bws[1,1]

# Placebo outcome 2
tmp <- rdrobust(Plac[, 2], R, p = 1, masspoints = "off", stdvars = "on")
local_linear[1, 3] = tmp$coef[1]
local_linear[2, 3] = tmp$se[1]
local_linear[3, 3] = tmp$pv[3]
local_linear[4, 3] = tmp$bws[1,1]

round(local_linear,3)

stargazer(local_linear)

#Initialize table to store results
local_quadratic <- array(NA, dim = c(4, 3))

# Local quadratic Regression
# masspoints and stdvars are applied according to instructions in replication package
# Outcome of interest
tmp <- rdrobust(Y, R, p = 2, masspoints = "off", stdvars = "on")
local_quadratic[1, 1] = tmp$coef[1]
local_quadratic[2, 1] = tmp$se[1]
```

```
local_quadratic[3, 1] = tmp$pv[3]
local_quadratic[4, 1] = tmp$bws[1,1]

# Placebo outcomes
tmp <- rdrobust(Plac[, 1], R, p = 2, masspoints = "off", stdvars = "on")
local_quadratic[1, 2] = tmp$coef[1]
local_quadratic[2, 2] = tmp$se[1]
local_quadratic[3, 2] = tmp$pv[3]
local_quadratic[4, 2] = tmp$bws[1,1]

# Placebo outcome 2
tmp <- rdrobust(Plac[, 2], R, p = 2, masspoints = "off", stdvars = "on")
local_quadratic[1, 3] = tmp$coef[1]
local_quadratic[2, 3] = tmp$se[1]
local_quadratic[3, 3] = tmp$pv[3]
local_quadratic[4, 3] = tmp$bws[1,1]

round(local_quadratic,3)

stargazer(local_quadratic)

# Specify covariates for randomization window selection
cov_1960 <- cbind(headstart$census1960_pop,
                  headstart$census1960_pctsch1417,
                  headstart$census1960_pctsch534,
                  headstart$census1960_pctsch25plus,
                  headstart$census1960_pop1417,
                  headstart$census1960_pop534,
                  headstart$census1960_pop25plus,
                  headstart$census1960_pcturban,
                  headstart$census1960_pctblack)


cov_test <- cbind(headstart$mort_age59_related_preHS, cov_1960)
# Plot p-values for winselect test
tmp <- rdwinselect(R, cov_test, reps = 1000, statistic = "ksmirnov", wmin = .3 , wstep = .05, level = .

# Set optimal window
w = 1.1

# Initialize otput array
rand_inf <- array(NA, dim = c(4,2))

# Outcome

tmp <- rdrandinf(Y, R, wl = -w, wr = w, reps = 1000, quietly = TRUE)
rand_inf[1:4,1] <- c(0, tmp$window[2], tmp$obs.stat, tmp$p.value)


tmp <- rdrandinf(Y, R, wl = -w, wr = w, reps = 1000, p = 1, quietly = TRUE)
rand_inf[1:4,2] <- c(1, tmp$window[2], tmp$obs.stat, tmp$p.value)

round(rand_inf, 3)
```

```
stargazer(rand_inf)
```