R Code Sample [*]

Xiling (Celia) Zhu xiling@uchicago.edu

Aug 28, 2020

Contents

1 Background

In January 2012, the Cook County State's Attorney's Office established a program intended to reduce re-arrest among people on bail awaiting trial. The program ran through October 2013.

The objective of our analysis is to evaluate the effectiveness of the program. We start by cleaning data sets on demographics, cases, and academic performance. Next, we provide descriptive statistics for the study population and test their baseline equivalence. The final step is to evaluate whether participating in the program reduces the likelihood of re-arrest before disposition.

1.1 Load packages

```r
# Set working directory
knitr::opts_knit$set(root.dir = "~/Box/projects/ra_code_sample")
```

[*] Written by `Rmarkdown`. For source code, please see my github repository here

```
# Load packages
knitr::opts_chunk$set(echo = TRUE, warning = FALSE, message = FALSE)
pkg_list <- c(
  "tidyverse", "stringr", "testthat", "lubridate",
  "kableExtra", "stargazer", "fastDummies",
  "lmtest", "sandwich", "AER"
)
lapply(pkg_list, require, character.only = TRUE)

# Resolve the clash in `select` between `dplyr` and `MASS`
select <- dplyr::select
```

1.2 Read in data sets for cases, demographics, and grades

```
# Read in data set of case, demographics, and grades information.
filelist <- dir("data", pattern = "\\.csv$", full.names = TRUE)
namelist <- str_match(filelist, "data/(.*?).csv")
namelist <- namelist[, 2]
for (i in 1:length(filelist)) {
  assign(namelist[i], read_csv(filelist[i]))
}
```

1.3 Define functions for summary table, balance tests, inverse probability weighting (IPW), and bar plot.

```
# Function for summary statistics
# produce table for N, mean, sd, min, and max (output named `sum_tbl`)
summary_statistics <- function(df, varlist) {
  obs <- nrow(df)

  sum_tbl <- df %>%
    select(!!varlist) %>%
    skimr::skim() %>%
    as_tibble() %>%
    select(
      skim_variable, n_missing, numeric.mean,
      numeric.sd, numeric.p0, numeric.p100
    ) %>%
    left_join(label_tbl, by = c("skim_variable" = "covs")) %>%
    mutate(
      numeric.mean = round(numeric.mean, digits = 2),
      numeric.sd = round(numeric.sd, digits = 2),
      N = (obs - n_missing)
    ) %>%
    select(-skim_variable, -n_missing) %>%
    # Make N in the second column
    select(cov_labs, N, everything()) %>%
    # Clean the names of summary statistics
    rename(
      Variable = cov_labs,
      Mean = numeric.mean,
      "Standard deviation" = numeric.sd,
      Min = numeric.p0, Max = numeric.p100
    )

  output <- sum_tbl
}
```

```
# Test balance on prior_arrests, age, treat, black, asian, white, male, female.
# Regress each covariate on treatment variable with robust s.e.
balance_table <- function(df, varlist, treat_var, weight) {
  formula <- list()
  fit <- list() # homoskedastic results stored here
  ttest <- list() # to store robust s.e.

  for (i in 1:length(varlist)) {
    formula[[i]] <- paste0(varlist[i], " ~", treat_var)
    if (missing(weight)) {
      fit[[i]] <- lm(formula[[i]], data = df)
    } else {
      fit[[i]] <- lm(formula[[i]], data = df, weights = weight)
    }
    # Calculate robust s.e.
    ttest[[i]] <- coeftest(fit[[i]],
      vcov = vcovHC(fit[[i]], type = "HC1")
    )
  }

  # Create a tibble to store statistics in balance test
  bal_test <- tibble(
    "var" = character(), "mean0" = numeric(), "mean1" = numeric(),
    "diff" = numeric(), "se_diff" = numeric(), "df" = numeric(), "p" = numeric()
  )

  for (i in 1:length(varlist)) {
    # fill in names of covariates
    bal_test[i, 1] <- varlist[i]
    # fill in mean in unenrolled groups = the coefficient of constant term
    bal_test[i, 2] <- ttest[[i]][1, 1]
    # fill in mean in enrolled groups
    # = the coefficient of constant term + coefficient of `treat`
    bal_test[i, 3] <- ttest[[i]][1, 1] + ttest[[i]][2, 1]
    # fill in the difference in means = coefficient of `treat`
    bal_test[i, 4] <- ttest[[i]][2, 1]
    # fill in the se in the difference
    bal_test[i, 5] <- ttest[[i]][2, 2]
    # fill in degree of freedom
    bal_test[i, 6] <- fit[[i]]$df.residual
  }

  # Calculate pvalue
  bal_test <- bal_test %>%
    mutate(p = 2 * pt(-abs(diff / se_diff), df))

  # Format the balance table
  bal_tbl <- bal_test %>%
    # format numbers
    mutate_if(is.numeric, round, digits = 2) %>%
    # Clean the name of statistics
    rename(
      "Mean in unenrolled group" = mean0,
      "Mean in enrolled group" = mean1,
      "Difference in means" = diff,
      "P value" = p
    ) %>%
    select(-se_diff, -df)

  output <- bal_tbl
}
```

```
ipw <- function(df, formula, wt) {
  eval(bquote(
    lm(
      formula,
      df,
      weights = .(as.name(wt))
    )
  ))
}
```

```
# Define filling color for three racial groups
fill_pal <- c("#800000E6", "#800000B3", "#80000080")

# Calculate mean and confidence interval
barplot_errorbar <- function(df, treat_var, grouping_var, plot_var) {
  eval(bquote(
    df %>%
      group_by(.(as.name(treat_var)), .(as.name(grouping_var))) %>%
      summarise(
        avg = mean(.(as.name(plot_var))),
        sd = sd(.(as.name(plot_var))),
        n = n()
      ) %>%
      # Calculate 95% confidence interval
      mutate(ci = 1.96 * (sd / sqrt(n))) %>%
      ungroup() %>%
      mutate(
        plot_id = c(1, 2, 3, 5, 6, 7)
      ) %>%
      # Plot bar graph with error bar
      ggplot() +
      # Bar for averages
      geom_col(
        aes(x = plot_id, y = avg, fill = .(as.name(grouping_var))),
        position = "dodge"
      ) +
      # Error bar for 95% CI
      geom_errorbar(
        aes(x = plot_id, ymin = avg - ci, ymax = avg + ci),
        width = 0.2, color = "#666666", alpha = 0.9, size = 1
      ) +
      # Fill by racial groups
      scale_fill_manual(
        values = fill_pal, name = "Race",
        labels = c("Asian", "Black", "White")
      ) +
      # Unenrolled groups on the left side, enrolled on the right
      scale_x_continuous(breaks = c(2, 6), label = c("Unenrolled", "Enrolled")) +
      # Make titles in bold, reduce font size for portrait pdf
      theme(
        plot.title = element_text(size = 11, face = "bold"),
        plot.subtitle = element_text(size = 9),
        # Set background and grid to a minimalistic exhibition
        panel.background = element_rect(
          fill = "white", colour = "white",
          size = 0.5, linetype = "solid"
        ),
        panel.grid.major.y = element_line(
          size = 0.1, linetype = "solid",
          colour = "grey"
        ),
        panel.grid.minor = element_blank()
      )
  ))
}
```

2 Data Cleaning

2.1 Clean demographic data

The demographic data were extracted from a system that inconsistently coded gender. Recode it so that males are consistently coded as “M” and females are consistently coded as “F”.

```
# Inconsistent encoding
demo %>%
  filter(gender != "M" & gender != "F") %>%
  distinct(gender)
```

```
## # A tibble: 2 x 1
##   gender
##   <chr>
## 1 male
## 2 female
```

```
# Clean "male", "female" encoding
demo_clean <- demo %>%
  mutate(
    gender = str_replace(gender, "^male$", "M"),
    gender = str_replace(gender, "^female$", "F")
  )

# Test that if there are inconsistent encoding
test_that(
  "Gender is consistently coded",
  expect_equal(
    0,
    nrow(demo_clean %>%
      filter(gender != "M" & gender != "F"))
  )
)
```

2.2 Clean arrests data

Merge the case (data on arrests is named as “case”) and demo data sets together so that each row in the case data set also contains the demographics of the defendant. Keep in mind that the populations in the case and demo data may not be 100% aligned.

```
# person_id is the primary key in this join. Test if they contain NAs
test_na <- function(df, x) {
  test_that(
    "No missing value in primary key person_id",
    expect_equal(
      0,
      nrow(df %>%
        filter(is.na(x)))
    )
```

```r
  )
}
test_na(demo_clean, "person_id")
test_na(case, "person_id")

# Check if person_id is the unique identifier in demo_clean
# test_that(
#  "person_id is the unique identifier in demo",
#  expect_equal(
#    nrow(demo_clean),
#    nrow(demo_clean %>%
#           distinct(person_id)
#  )
# )
# )

# The test failed. person_id is not the unique identifier in demo.
# Extract duplicate person_id and see if they are in the case
# Remove duplicate rows
demo_clean <- unique(demo_clean)

# Check how many observations in cases do not have a match in demo
anti_join(case, demo_clean, by = "person_id") %>%
  nrow()
```

```
## [1] 0
```

```r
# Check that `person_id` is the only variable that demo and case data share.
test_that(
  "no other shared variables excpet person_id",
  expect_equal(
    "person_id",
    intersect(names(demo), names(case))
  )
)

case_demo <- left_join(case, demo_clean, by = "person_id")

# Check that if the numbers of rows before and after join can match
test_that(
  "demo and case are merged correctly",
  expect_equal(
    nrow(case),
    nrow(case_demo)
  )
)
```

While the program was mostly rolled out to defendants in Chicago, the State's Attorney's Office also ran a pilot serving a small number of individuals arrested in other parts of Cook County. For the purpose of this analysis, restrict the data to only individuals who were arrested in Chicago.

```r
# It's possible to have "Chicago" inconsistently capitalized
case_demo_chi <- case_demo %>%
  filter(endsWith(address, "CHICAGO") |
    endsWith(address, "Chicago") |
    endsWith(address, "chicago"))
```

Create an age variable equal to the defendant's age at the time of arrest for each case.

```r
case_demo_chi <- case_demo_chi %>%
  mutate(age = round((as_date(arrest_date) - as_date(bdate)) / 365.25, 1)) %>%
  mutate(age = as.double(age))
```

2.3 Clean grades data

The State's Attorney is interested in pursuing a partnership with the Chicago Public Schools to investigate the relationship between high school achievement and criminal justice outcomes in early adulthood. To that end, the State's Attorney's Office has requested 9th and 10th grade course grade data from defendants between the ages of 18 and 24. These data are included in grades.csv. Please construct measures for 9th and 10th grade GPA for this target population. When constructing GPA, please use a 4 point scale, where: A=4, B=3, C=2, D=1, and F=0.

```r
# Calculate GPA
grades_clean <- grades %>%
  mutate_at(
    vars(starts_with("gr")),
    funs(case_when(
      . == "A" ~ 4,
      . == "B" ~ 3,
      . == "C" ~ 2,
      . == "D" ~ 1,
      . == "F" ~ 0
    ))
  )

# GPAs for 9th and 10th grades
for (i in 9:10) {
  grades_clean[, paste0("gpa", i)] <- rowMeans(
    select_at(grades_clean, vars(starts_with(paste0("gr", i)))),
    na.rm = TRUE
  )
}

# Keep person_id and gpa for 9th and 10th graders
grades_clean <- grades_clean %>%
  select_at(vars(contains("person_id"), starts_with("gpa")))

# Save cleaned grades data
```

```
saveRDS(grades_clean, file = "r/processed/grades_clean.csv")
```

3 Statistical Analysis

Determine if the program should be continued/expanded by estimating the program's effect on **re-arrests prior to disposition**. Because we only have grades data for young adults, do not use these data to inform your statistical analysis.

```
# Check distinct values of race.
case_demo_chi %>%
  distinct(race)
```

```
## # A tibble: 3 x 1
##   race
##   <chr>
## 1 WHITE
## 2 BLACK
## 3 ASIAN
```

```
# Create dummies for statistical analysis
analysis_data <- case_demo_chi %>%
  dummy_cols(select_columns = c("race", "gender")) %>%
  rename(
    asian = race_ASIAN, black = race_BLACK, white = race_WHITE,
    female = gender_F, male = gender_M
  )

# Check that the study population has 25,000 subjects.
test_that(
  "The study population has 25,000 subjects",
  expect_equal(
    25000,
    nrow(analysis_data)
  )
)

# From now on I will use analysis_data for statistical analysis.
# Save the cleaned data.
saveRDS(analysis_data, "r/processed/analysis_data.rds")
```

3.1 Summary statistics of study population

The study population are predominantly male, with only 20% cases having female defendants. As to race, 73% of the cases involve Black defendants, only 22% are white, and 5% are Asian. On average, the study population has around 4 prior arrests before the program rolled out, and their average age is approximately 30.

See table 1 for summary statistics.

```r
# Create a crosswalk for covariates and its labels
covs <-
  c("asian", "black", "white", "male", "female", "prior_arrests", "age")
cov_labs <-
  c("Asian", "Black", "White", "Male", "Female", "Number of prior arrests", "Age")
label_tbl <- tibble(covs, cov_labs)

# Use the function `summary_statistics()` to get summary table
summary_tbl <- summary_statistics(analysis_data, covs)

# Exhibit latex output
kbl(summary_tbl, "latex",
  caption = "Summary Statistics",
  booktabs = TRUE,
  align = "l"
) %>%
  kable_styling(
    latex_options = "hold_position",
    full_width = TRUE
  )
```

3.2 Balance tests for demographic characteristics

The enrolled and unenrolled groups are not balanced at the baseline.

The average numbers of prior arrests are significantly different in the two groups. Cases with more prior arrests are more likely to be enrolled into the program.

Their age is also imbalanced. Cases with older defendants are more likely to be enrolled into the program.

The imbalance are not due to random coincidence. The F-test indicates that these covariates didn't pass joint orthogonality, either.

The imbalance at baseline signals the problem of selection.

See table 2 for balance test.

```r
# Test balance on prior_arrests, age, treat, black, asian, white, male, female.
# Use function balance_table() for balance test of individual orthogonality
bal_tbl <- balance_table(analysis_data, covs, "treat") %>%
  # Format the variable names
  left_join(label_tbl, by = c("var" = "covs")) %>%
  select(-var) %>%
  rename(Variable = cov_labs) %>%
  select(Variable, everything())

# F-test for joint orthogonality
# Check perfect collinearity
alias(
  lm(treat ~ black + asian + white + male + female + prior_arrests + age,
    data = analysis_data
  )
)
```

```
## Model :
## treat ~ black + asian + white + male + female + prior_arrests +
##     age
##
## Complete :
##        (Intercept) black asian male prior_arrests age
## white   1          -1    -1     0    0             0
## female  1           0     0    -1    0             0
```

```
# Use female and Asian as base group,
# because our study population are predominately male and Black.
f_fit <- lm(
  treat ~ black + white + male + prior_arrests + age,
  data = analysis_data
)
ftest <- linearHypothesis(
  f_fit,
  c("black = 0", "white = 0", "male = 0", "prior_arrests = 0", "age = 0"),
  white.adjust = "hc1"
) %>%
  filter(!is.na(Df)) %>%
  select(F, "Pr(>F)")

ftest_pvalue <- format(round(ftest$`Pr(>F)`, digits = 3), nsmall = 2)

# Assemble balance table (add F-test in balance table)
# Latex output
kbl(bal_tbl, "latex", caption = "Balance Test", booktabs = TRUE, align = "l") %>%
  kable_styling(latex_options = "hold_position", full_width = TRUE) %>%
  footnote(
    general = paste("F-test of joint orthogonality (P value)",
      ftest_pvalue,
      sep = " "
    ),
    footnote_as_chunk = TRUE
  )
```

3.3 Visualize age and number of prior arrests by enrollment status and race

Number of prior arrests and age are not balanced across enrolled and unenrolled groups. See figure 1 and 2.

```
# Use pre-defined function to plot average number of prior arrests
# by enrollment status and race
```

```
n_arrests_plot <- barplot_errorbar(
  analysis_data,
  "treat", "race", "prior_arrests"
)
n_arrests_plot +
  labs(
    x = "Enrollment status",
    y = "Average number of prior arrests",
    title = "Number of Prior Arrests Imbalanced between Enrolled and Unenrolled Group",
    subtitle = "Average number of prior arrests and 95% confidence interval by race and enroll
    caption = "Source: Cook County State's Attorney's Office"
  )
```

```
# Save plot
ggsave("r/output/barplot_arrests.png")
```

```
# Use pre-defined function to plot average age by enrollment status and race
n_arrests_plot <- barplot_errorbar(
  analysis_data,
  "treat", "race", "age"
)
n_arrests_plot +
  labs(
    x = "Enrollment status",
    y = "Average age",
    title = "Age Imbalanced between Enrolled and Unenrolled Group",
    subtitle = "Average age and 95% confidence interval by race and enrollment status",
    caption = "Source: Cook County State's Attorney's Office"
  )
```

```
# Save plot
ggsave("r/output/barplot_age.png")
```

3.4 Estimate the effect of the program on reducing the likelihood of re-arrest before disposition

One difficulty in estimating the effect of the program is that I don't have enough information about the program: if program was an randomized controlled trial and if so, how was the compliance, or if it was an observational study.

1. OLS (or Linear Probability Model)

We start with the OLS model.

$$rearrest_{ic} = \tau treat_{ic} + \beta X_{ic} + \epsilon_{ic}$$

where i is the individual, c is the case, X_{ic} is the vector for race, gender, age, and number of prior arrests. We could have added individual fixed effects u_i to control for those individual-invariant characteristics. But by examining the unique values of `person_id`, we can conclude that for most defendants, they only have one or two cases. Individual fixed effect is not desirable here.

It can correctly estimate the treatment effect given that 1) there is no selection on unobservables and we've controlled all observables that could be selected upon; and 2) how people are self-selected based on those variables can be approximated by a linear function. But these conditions are unlikely to be true.

That said, suppose the program was a randomized trial, the treatment was administered on the case level, not on the individual level. Hence, we don't cluster standard errors here.

```r
# 14353 unique person_id, with 25000 unique case_id.
analysis_data %>%
  distinct(person_id) %>%
  nrow()
```

```
## [1] 14353
```

```r
# Specification 1: OLS
# female and asian as base group
ols <- lm(
  re_arrest ~ treat + prior_arrests + male + black + white + age,
  data = analysis_data
)
ols$rse <- sqrt(diag(vcovHC(ols, type = "HC1")))
```

2. Logit Model

Assume the program was not an experimental study, we can improve our **prediction** on the likelihood by using a logit model instead of a linear probability model, which was implemented in specification 1. But the results given by logit specification is for prediction, not for causal inference.

$$Pr(rearrest = 1|X_{ic}) = \frac{exp^{X'_{ic}\beta}}{1 + e^{X'_{ic}\beta}}$$

```r
# Specification 2: logit
# female and asian as base group
logit <- mfx::logitmfx(
  re_arrest ~ treat + prior_arrests + male + black + white + age,
  data = analysis_data,
  robust = TRUE,
  clustervar1 = "person_id"
)
```

3. Inverse Probability Weighting (IPW)

To estimate the causal effect via propensity score matching (more specifically, inverse probability weighting (IPW)), given that the program was **observational**, not experimental, we need to assume that there is **no selection on unobservables** and we've controlled all observables that could be selected upon.

One improvement from OLS or logit model is that we can relax the assumption on the functional form.

Still, the estimate based upon IPW is not entirely valid, because we omit important characteristics such as grades, household income, neighborhood, etc.. However, this is less restrictive and therefore more plausible than the other two models.

```r
# Get propensity score
# female and Asian be base group
pscore_estimate <- glm(treat ~ prior_arrests + male + black + white + age,
  data = analysis_data,
  family = binomial
)
pscore <- predict(pscore_estimate, type = "response")
analysis_data$pscore <- pscore

# Inverse probability weighting (IPW)
# Weight for average treatment effect and average treatment effect on the treated
analysis_data <- analysis_data %>%
  mutate(
    ate_weight = if_else(
      treat == 1, 1 / pscore, 1 / (1 - pscore)
    ),
    atet_weight = if_else(
      treat == 1, 1, pscore / (1 - pscore)
    )
  )

# Common support
pscore_min <- analysis_data %>%
  filter(treat == 1) %>%
  summarise(min = min(pscore))
pscore_min <- pscore_min$min

pscore_max <- analysis_data %>%
  filter(treat == 0) %>%
  summarise(max = max(pscore))
pscore_max <- pscore_max$max

# An indicator on/off common support
analysis_data <- analysis_data %>%
  mutate(support = if_else(pscore >= pscore_min & pscore <= pscore_max, 1, 0))

# Data with observation on the common support
analysis_data_cs <- analysis_data %>%
  filter(support == 1)
```

Before we estimate average treatment effect (ATE) and average treatment effect on the treated (ATET) of this program, we first assess the validity of propensity score by 1) the distribution of the propensity score, 2) summary statistics of observations within the common support, and 3) the balance of the sample within common support, weighted by the IPW for ATE.

See figure 3 for the distribution of propensity score.

```r
# Facet label names
treat.labs <- c("Unenrolled", "Enrolled")
names(treat.labs) <- c("0", "1")

# Plot faceted histogram
ggplot() +
  geom_histogram(
    data = analysis_data, binwidth = 0.05,
    fill = "#800000B3", aes(x = pscore)
  ) +
  geom_vline(xintercept = pscore_max, color = "grey", size = 0.5, linetype = 2) +
  geom_vline(xintercept = pscore_min, color = "grey", size = 0.5, linetype = 2) +
  facet_grid(rows = vars(treat), labeller = labeller(treat = treat.labs)) +
  labs(
    x = "Propensity score",
    y = "Frequency",
    title = "Histogram of Propensity Scores by Enrollemnt Status",
    caption = "Source: Cook County State's Attorney's Office \n(Observation bounded by vertical
  ) +
  theme(
    plot.title = element_text(size = 11, face = "bold"),
    plot.subtitle = element_text(size = 9),
    panel.background = element_rect(
      fill = "white", colour = "white",
      size = 0.5, linetype = "solid"
    ),
    panel.grid.major.y = element_line(
      size = 0.1, linetype = "solid",
      colour = "grey"
    ),
    panel.grid.minor = element_blank()
  )
```

See table 3 for the summary statistics of observations within the common support.

```r
# Use the function `summary_statistics` to get summary table
summary_tbl_cs <- summary_statistics(analysis_data_cs, covs)

# Exhibit output
kbl(summary_tbl_cs,
  caption = "Summary Statistics (observations within common support)",
  booktabs = TRUE, align = "l"
) %>%
  kable_styling(latex_options = "hold_position", full_width = TRUE)
```

See table 4 for the balance test of observation within common support, weighted by inverse probability for ATE.

```
# 3) balance tests with weights, using data within common support----
# Use function balance_table() for balance test of individual orthogonality
weight <- analysis_data_cs$ate_weight
bal_tbl_ipw <- balance_table(analysis_data_cs, covs, "treat", weight) %>%
  # Format the variable names
  left_join(label_tbl, by = c("var" = "covs")) %>%
  select(-var) %>%
  rename(Variable = cov_labs) %>%
  select(Variable, everything())

# F-test for joint orthogonality
# use female and asian as base group
f_fit_ipw <- lm(treat ~ black + white + male + prior_arrests + age,
  weights = ate_weight,
  data = analysis_data_cs
)
ftest_ipw <- linearHypothesis(f_fit_ipw,
  c("black = 0", "white = 0", "male = 0", "prior_arrests = 0", "age = 0"),
  white.adjust = "hc1"
) %>%
  filter(!is.na(Df)) %>%
  select(F, "Pr(>F)")

ftest_pvalue_ipw <- format(round(ftest_ipw$`Pr(>F)`, digits = 3), nsmall = 2)

# Assemble balance table
# Exhibit latex output
kbl(bal_tbl_ipw, "latex",
  caption = "Balance Test Weighted by Inverse Probability for ATE",
  booktabs = TRUE, align = "l"
) %>%
  kable_styling(latex_options = "hold_position", full_width = TRUE) %>%
  footnote(
    general = paste("F-test of joint orthogonality (P value): ", ftest_pvalue_ipw),
    footnote_as_chunk = TRUE
  )
```

Estimate ATE and ATET with inverse probability weighting.

```
ipw_ate <- ipw(analysis_data_cs, "re_arrest ~ treat", "ate_weight")
```

```
ipw_atet <- ipw(analysis_data_cs, "re_arrest ~ treat", "atet_weight")
```

See table 5

```
stargazer(ols, logit$fit, ipw_ate, ipw_atet,
  # retrieve marginal effects for logit
  coef = list(NULL, logit$mfxest[, 1], NULL, NULL),
  # retrieve robust s.e. for ols and logit model
  se = list(ols$rse, logit$mfxest[, 2], NULL, NULL),
  title = "Estimation Results", align = TRUE,
  dep.var.labels = rep("Rearrested before disposition", 2),
  covariate.labels = c(
    "Enrolled into program", "Number of prior arrests",
    "Male", "Black", "White", "Age", "Constant"
  ),
  notes = c(
    "Marginal effects reported in logit model;",
    "Robust s.e. reported in OLS model, clustered robust s.e.reported",
    "in logit model"
  ),
  notes.append = TRUE,
  notes.align = "l",
  model.numbers = FALSE,
  model.names = FALSE,
  keep.stat = "n",
  multicolumn = TRUE,
  column.labels = c("OLS", "Logit", "IPW (ATE)", "IPW(ATET)"),
  column.sep.width = "10pt",
  type = "latex",
  header = FALSE,
  float = TRUE,
  table.placement = "h",
  label = "tab:regression-results",
  out = "r/output/reg_results.tex"
)
```

3.5 Conclusion

Overall, the treatment significantly reduces the likelihood of re-arrest before disposition. With the information we have, we can conclude that the program is effective and should be expanded or continued, or should be furthered examined with an experiment.

However, please note that the causal inference has much room for improvement. Though the sample on the common support and weighted by the inverse probability for ATE is balanced, the distribution of propensity score is not ideal. Some cases not selected into the program still have relatively high propensity score. And the distribution didn't improve much when I added higher-ordered terms, like age squared.

It would have better performance if we can obtain more characteristics. To name a few, household income, and the neighborhoods defendants live in; and for younger defendants, we can also incorporate their academic performance and disciplinary incidents in school, and the school districts they live in.

Tables

Table 1: Summary Statistics

Variable	N	Mean	Standard deviation	Min	Max
Asian	25000	0.05	0.22	0.0	1.0
Black	25000	0.73	0.44	0.0	1.0
White	25000	0.22	0.41	0.0	1.0
Male	25000	0.80	0.40	0.0	1.0
Female	25000	0.20	0.40	0.0	1.0
Number of prior arrests	25000	3.80	2.14	0.0	16.0
Age	25000	30.34	7.80	9.5	70.1

Table 2: Balance Test

Variable	Mean in unenrolled group	Mean in enrolled group	Difference in means	P value
Asian	0.05	0.05	0.00	0.51
Black	0.73	0.73	0.00	0.45
White	0.22	0.22	-0.01	0.24
Male	0.80	0.80	0.00	0.34
Female	0.20	0.20	0.00	0.34
Number of prior arrests	3.15	4.38	1.23	0.00
Age	28.73	31.79	3.07	0.00

Note: F-test of joint orthogonality (P value) 0.00

Table 3: Summary Statistics (observations within common support)

Variable	N	Mean	Standard deviation	Min	Max
Asian	24980	0.05	0.22	0.0	1.0
Black	24980	0.73	0.44	0.0	1.0
White	24980	0.22	0.41	0.0	1.0
Male	24980	0.80	0.40	0.0	1.0
Female	24980	0.20	0.40	0.0	1.0
Number of prior arrests	24980	3.79	2.12	0.0	12.0
Age	24980	30.33	7.78	9.5	70.1

Table 4: Balance Test Weighted by Inverse Probability for ATE

Variable	Mean in unenrolled group	Mean in enrolled group	Difference in means	P value
Asian	0.05	0.05	0.00	0.90
Black	0.73	0.73	0.00	0.98
White	0.22	0.22	0.00	0.93
Male	0.80	0.80	0.00	0.91
Female	0.20	0.20	0.00	0.91
Number of prior arrests	3.80	3.79	-0.01	0.78
Age	30.39	30.33	-0.06	0.63

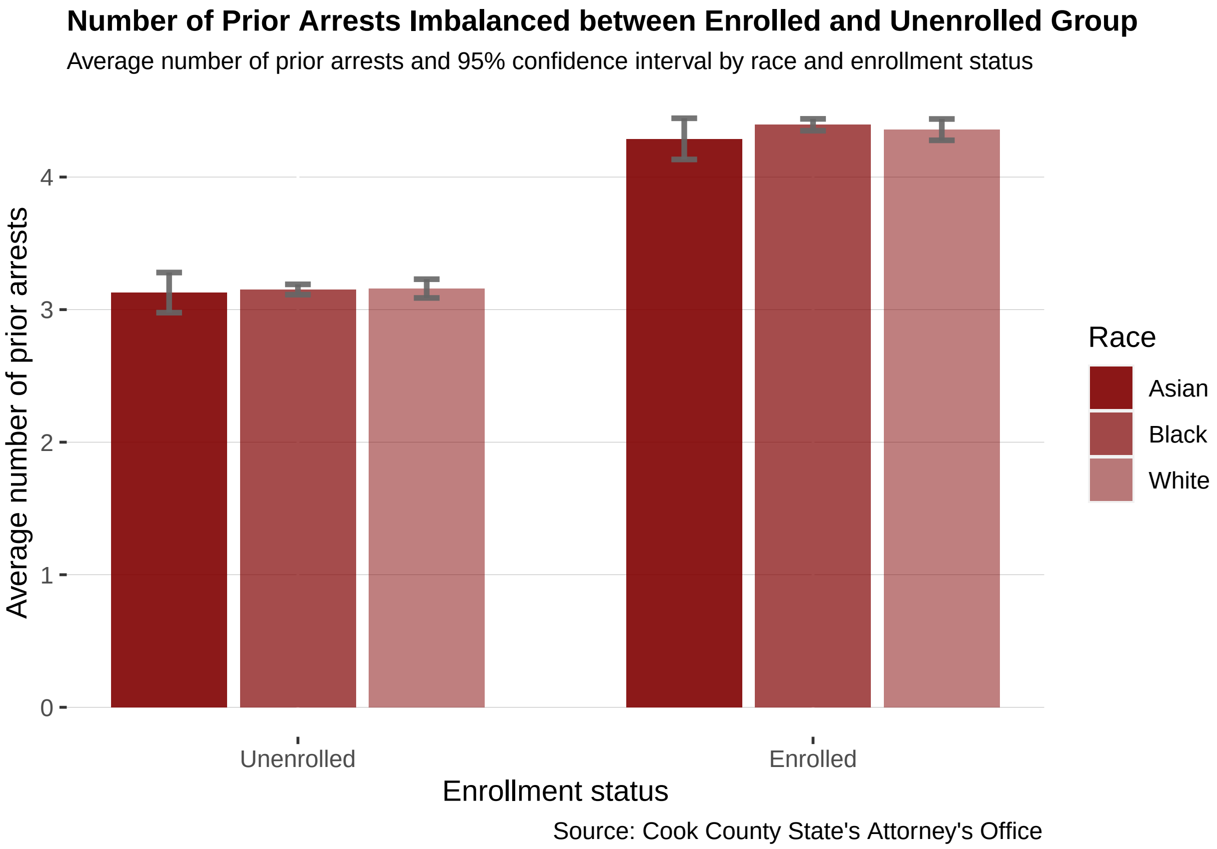
Note: F-test of joint orthogonality (P value): 0.998

Table 5: Estimation Results

	Dependent variable:			
	Rearrested before disposition			
	OLS	Logit	IPW (ATE)	IPW(ATET)
Enrolled into program	-0.015^{***}	-0.015^{***}	-0.017^{***}	-0.017^{***}
	(0.005)	(0.005)	(0.005)	(0.005)
Number of prior arrests	0.016^{***}	0.015^{***}		
	(0.002)	(0.001)		
Male	0.005	0.005		
	(0.006)	(0.006)		
Black	0.006	0.005		
	(0.012)	(0.012)		
White	0.003	0.003		
	(0.013)	(0.013)		
Age	0.004^{***}	0.004^{***}		
	(0.0005)	(0.0004)		
Constant	0.035^{**}		0.219^{***}	0.234^{***}
	(0.016)		(0.004)	(0.004)
Observations	25,000	25,000	24,980	24,980

Note: *p<0.1; **p<0.05; ***p<0.01
Marginal effects reported in logit model;
Robust s.e. reported in OLS model, clustered robust s.e.reported in logit model

Figures


Number of Prior Arrests Imbalanced between Enrolled and Unenrolled Group

Average number of prior arrests and 95% confidence interval by race and enrollment status

Source: Cook County State's Attorney's Office

Figure 1: Average number of prior arrests by enrollment status and race

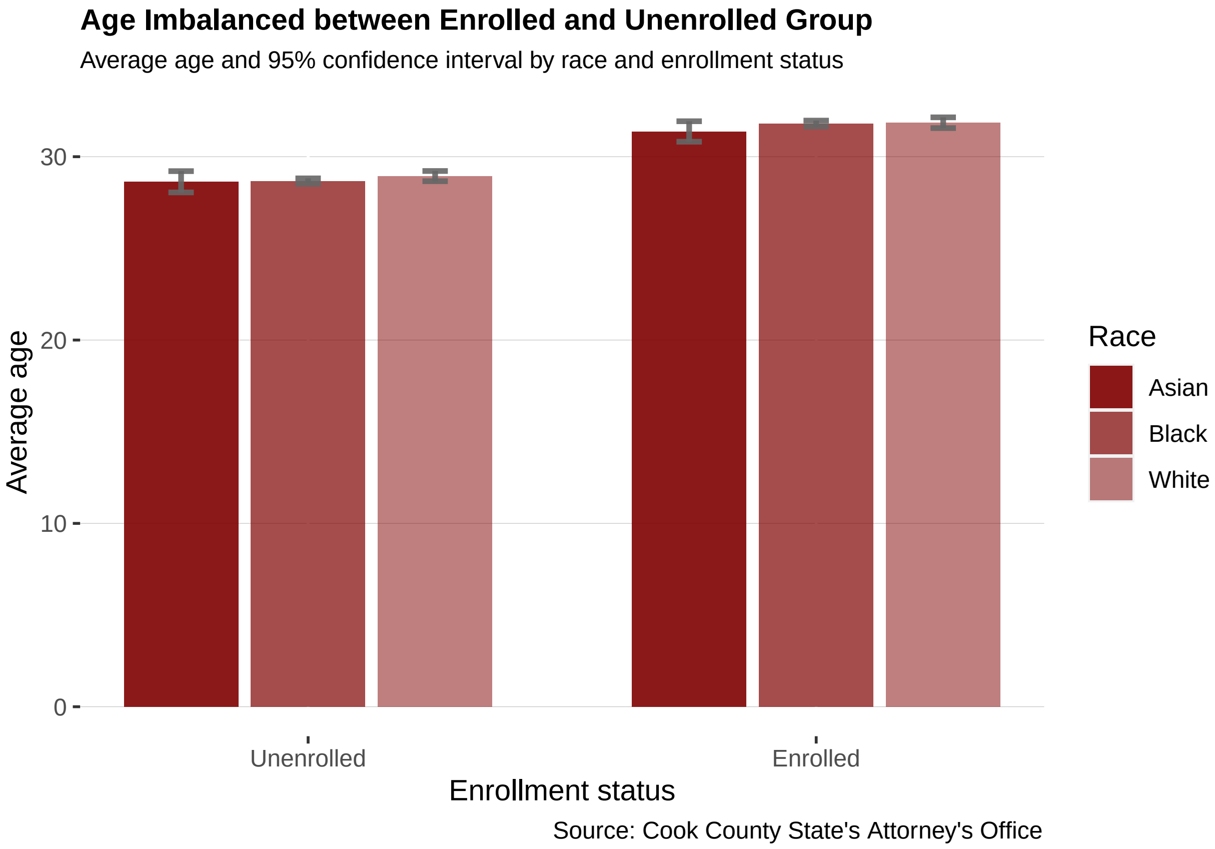
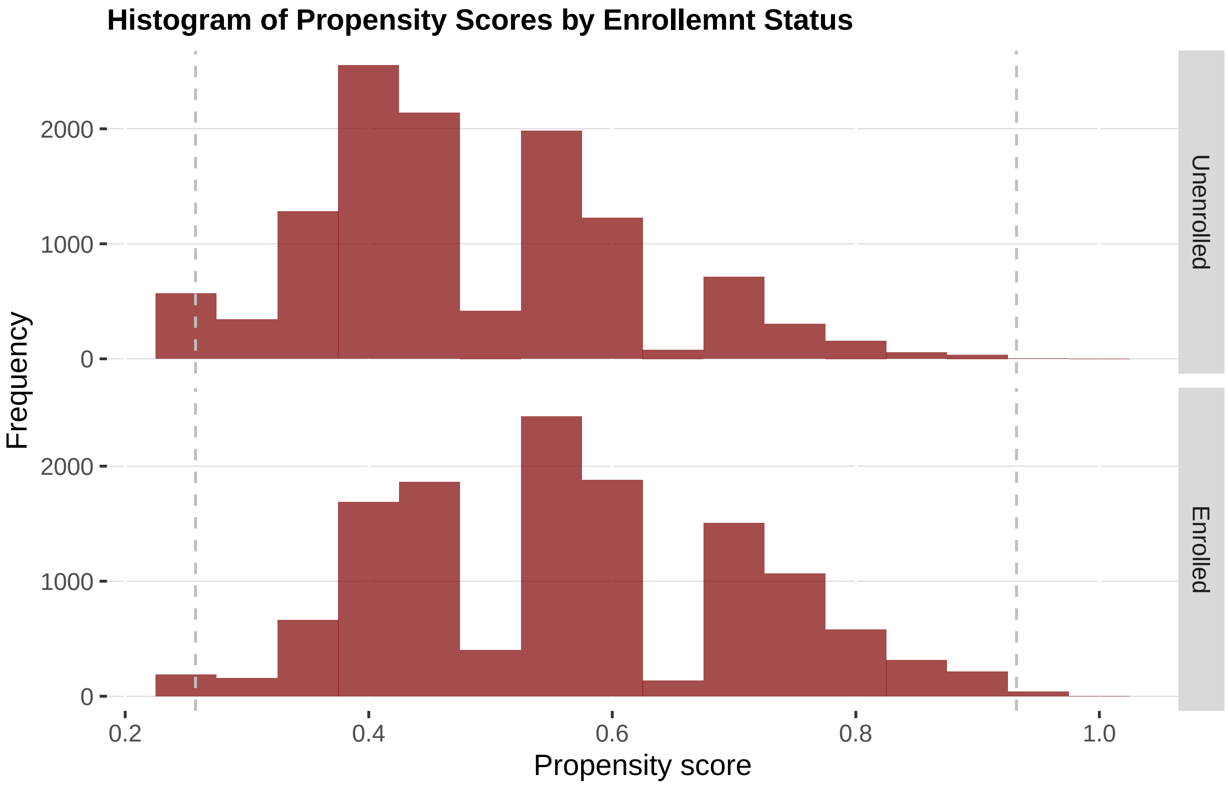


Figure 2: Average age by enrollment status and race



Source: Cook County State's Attorney's Office
(Observation bounded by vertical lines within common support)

Figure 3: Histogram of Propensity Score