# POLS 207

Problem Set 3*

*Villaseñor-Derbez, J.C.*

Problem 1: NSW experiment and observational comparison

```
# Load packages
suppressPackageStartupMessages({
  library(startR)
  library(Matching)
  library(ebal)
  library(foreign)
  library(here)
  library(tidyverse)
})

# Load data
nsw_dat <- read.dta("nsw_exper.dta")
```

### Using the experimental data, obtain a simple unbiased estimate of the ATE of NSW on 1978 earnings, its SE and a 95% confidence interval.

Because this is experimental data, the difference in means is an unbiased estimate of the ATE.

```
# Simple unbiased estimate of the ATE
mT <- mean(nsw_dat$re78[nsw_dat$nsw == 1], na.rm = T)
mC <- mean(nsw_dat$re78[nsw_dat$nsw == 0], na.rm = T)

ATE <- mT - mC
```

```
# Calculate standard errors
# Get variances
sigma_y1 <- var(nsw_dat$re78[nsw_dat$nsw == 1], na.rm = T)
sigma_y0 <- var(nsw_dat$re78[nsw_dat$nsw == 0], na.rm = T)

# Get sample sizes
N1 <- sum(nsw_dat$nsw == 1, na.rm = T)
N0 <- sum(nsw_dat$nsw == 0, na.rm = T)

# Calculate Standard Errors
SE <- sqrt((sigma_y1 / N1) + (sigma_y0 / N0))
```

```
# Calculate 95% CIs
CI_h <- ATE + (1.96 * SE)
CI_l <- ATE - (1.96 * SE)
```

The $\hat{ATE} = 1794.343$, with $SE = 670.99$ and $CI_{95} = (479.18, 3109.49)$.

*Available on GitHUb: https://github.com/jcvdav/POLS207/blob/master/ps3/ps3.pdf

**With the experimental data, use OLS to estimate the ATE controlling for age, education, race, ethnicity, marital status, and employment and earnings in 1974 and 1975. Report the estimate and its SE. Compare it to the one obtained in (a), explaining the reason for any similarities or differences.**

The table below shows the coefficient estimates for the stated regression. The $\hat{ATE}_{OLS} = 1682.58$. The previous estimate was slightly higher ($\hat{ATE} = 1794.343$), but the OLS estimate is within the 95% CIs calculated before.

```
lm(re78 ~ nsw + age + educ + black + hisp + married + re74 + re75, data = nsw_dat) %>%
  stargazer::stargazer(.,
                       se = estimatr::starprep(.,
                                               se_type = "HC2"),
                       t.auto = T,
                       p.auto = T,
                       header = F,
                       title = "Estimate of the Sample Average Treatment Effect.",
                       single.row = T
  )
```

Table 1: Estimate of the Sample Average Treatment Effect.

| | *Dependent variable:* |
|---|---|
| | re78 |
| nsw | 1,682.588** (658.486) |
| age | 55.771 (39.927) |
| educ | 405.883** (158.113) |
| black | −2,169.781** (1,016.918) |
| hisp | 157.926 (1,381.290) |
| married | −140.271 (876.130) |
| re74 | 0.083 (0.111) |
| re75 | 0.052 (0.128) |
| Constant | 621.739 (2,402.376) |
| Observations | 445 |
| $R^2$ | 0.055 |
| Adjusted $R^2$ | 0.037 |
| Residual Std. Error | 6,506.044 (df = 436) |
| F Statistic | 3.161*** (df = 8; 436) |
| *Note:* | *p<0.1; **p<0.05; ***p<0.01 |

**File `nsw_psid_withtreated.dta` contains treated units taken from the experiment, but control units replaced by the non-experimental sample from the PSID.**

```
psid_dat <- read.dta("nsw_psid_withtreated.dta")
```

**Check the covariate balance in this merged dataset. Decide on a few sensible balance statistics and report them in a table.**

```
balance <- psid_dat %>%
  drop_na() %>%
  MatchBalance(
    nsw ~ age + educ + black + hisp + married + re74 + re75 + u74 + u75,
    data = .,
    print.level = 0
  ) %>%
  baltest.collect(
    var.names = c(
      "age",
      "educ",
      "black",
      "hisp",
      "married",
      "re74",
      "re75",
      "u74",
      "u75"),
    after = F
  ) %>%
  as_tibble(rownames = "Covariate") %>%
  dplyr::select(-contains("qq"), -contains("pooled"))

knitr::kable(
  balance,
  booktabs = T,
  col.names = c(
    "Covariate",
    "Mean (Treatment)",
    "Mean (Control)",
    "Standardized difference",
    "Variance ratio",
    "T p-value",
    "KS p-value"
  ),
  caption = "Pre-matching balance of covariates."
)
```

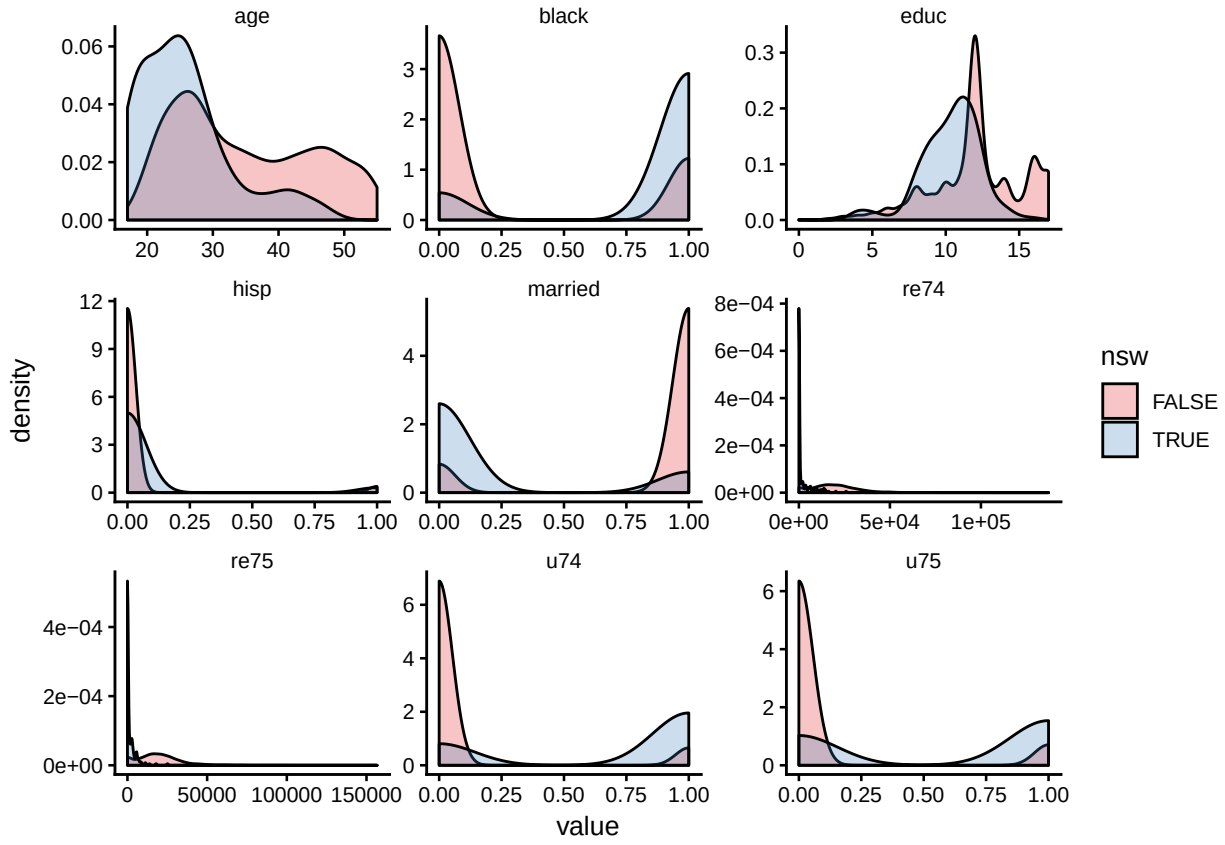
Table 2: Pre-matching balance of covariates.

| Covariate | Mean (Treatment) | Mean (Control) | Standardized difference | Variance ratio | T p-value | KS p-value |
|---|---|---|---|---|---|---|
| age | 25.8162162 | 3.485060e+01 | -126.26641 | 0.4696319 | 0.0000000 | 0 |
| educ | 10.3459459 | 1.211687e+01 | -88.07705 | 0.4254861 | 0.0000000 | 0 |
| black | 0.8432432 | 2.506024e-01 | 162.56425 | 0.7073935 | 0.0000000 | NA |
| hisp | 0.0594595 | 3.253010e-02 | 11.35663 | 1.7858904 | 0.1317327 | NA |
| married | 0.1891892 | 8.662651e-01 | -172.40603 | 1.3307596 | 0.0000000 | NA |
| re74 | 2095.5740112 | 1.942875e+04 | -354.70658 | 0.1328502 | 0.0000000 | 0 |
| re75 | 1532.0556149 | 1.906334e+04 | -544.57642 | 0.0560566 | 0.0000000 | 0 |
| u74 | 0.7081081 | 8.634540e-02 | 136.39138 | 2.6331760 | 0.0000000 | NA |
| u75 | 0.6000000 | 1.000000e-01 | 101.78586 | 2.6800826 | 0.0000000 | NA |

**How do the treatment and control group differ?**

The treatment group has younger people, with less education, a greater percentage of black and hispanics, and most people are single. The treatment group has lower average real earnings. The figure below shows the density distributions for each covariate in the dataset.

```
psid_dat %>%
  select(-c(re78, u78)) %>%
  gather(variable, value, -c(nsw)) %>%
  mutate(nsw = nsw == 1) %>%
  ggplot(aes(x = value, fill = nsw, group = nsw)) +
  geom_density(alpha = 0.25) +
  facet_wrap(~variable, scales = "free") +
  ggtheme_plot() +
  scale_fill_brewer(palette = "Set1")
```




**Among the observed covariates, what seem to be the most important factors that determine selection into the program (the "treatment")?**

The data suggest that your real earnings for 1974 and 1975 and ethnicity are prediuctors of selection into the program.

**Estimate the (naive) ATE of the program on 1978 earnings without adjusting for any of the covariates. Report the estimate and a standard error.**

```
# Simple unbiased estimate of the ATE for psid data
mTpsid <- mean(psid_dat$re78[psid_dat$nsw == 1], na.rm = T)
mCpsid <- mean(psid_dat$re78[psid_dat$nsw == 0], na.rm = T)

ATEpsid <- mTpsid - mCpsid
```

```
# Calculate standard errors
# Get variances
sigma_y1psid <- var(psid_dat$re78[psid_dat$nsw == 1], na.rm = T)
sigma_y0psid <- var(psid_dat$re78[psid_dat$nsw == 0], na.rm = T)

# Get sample sizes
N1psid <- sum(psid_dat$nsw == 1, na.rm = T)
N0psid <- sum(psid_dat$nsw == 0, na.rm = T)

# Calculate Standard Errors
SEpsid <- sqrt((sigma_y1psid / N1) + (sigma_y0psid / N0))
```

In this case, the naive ATE estimate for 1978 earnigns appears to be $ATE = -15204.78$, suggesting that the program reduces the average earnings. The standard error is now $SE1124.821$

**Repeat (b) using the non-experimental data. Does the estimate of the ATE change? Why or why not?**

```
lm(re78 ~ nsw + age + educ + black + hisp + married + re74 + re75, data = psid_dat) %>%
  stargazer::stargazer(.,
                       se = estimatr::starprep(.,
                                               se_type = "HC2"),
                       t.auto = T,
                       p.auto = T,
                       header = F,
                       title = "Estimate of the Sample Average Treatment Effect using the non-experimen
                       single.row = T
  )
```

**Using the non-experimental data, condition (only) on the marital status of individuals, and manually compute the subclassification estimator of the ATT.**

Table 3: Estimate of the Sample Average Treatment Effect using the non-experimental data.

| | *Dependent variable:* |
|---|---|
| | re78 |
| nsw | 859.769 (768.537) |
| age | $-81.537^{***}$ (20.719) |
| educ | $528.024^{***}$ (88.743) |
| black | $-542.706$ (442.463) |
| hisp | $2{,}165.572^{*}$ (1,227.276) |
| married | $1{,}220.269^{**}$ (496.886) |
| re74 | $0.278^{***}$ (0.062) |
| re75 | $0.568^{***}$ (0.067) |
| Constant | 776.729 (1,489.169) |
| Observations | 2,675 |
| $R^2$ | 0.586 |
| Adjusted $R^2$ | 0.585 |
| Residual Std. Error | 10,070.410 (df = 2666) |
| F Statistic | $472.194^{***}$ (df = 8; 2666) |
| *Note:* | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 |