# POLS 207

Problem Set 2*

*Villaseñor-Derbez, J.C.*

## Problem 1: Analyzing Experimental Data

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

# Load data
olken_dat <- read.csv(here("ps2", "OlkenData.csv"))
```

### Estimate the average treatment effect in this new dataset, using the difference in means estimator

```
olken_dat %>%
  group_by(treat_invite) %>%
  summarize(mean_est = mean(pct_missing, na.rm = T)) %>%
  ungroup() %>%
  mutate(treat_invite = ifelse(treat_invite == 0, "control", "treatment")) %>%
  spread(treat_invite, mean_est) %>%
  mutate(diff_means = treatment - control) %>%
  knitr::kable(
    booktabs = T,
    col.names = c(
      "Non-treatment",
      "Treatment",
      "Difference in means"),
    caption = "Difference in means estimate for % budget missing for projects where villagers were invi
  )
```

Table 1: Difference in means estimate for % budget missing for projects where villagers were invited to public hearings.

| Non-treatment | Treatment | Difference in means |
|---|---|---|
| 0.2521056 | 0.2289582 | -0.0231474 |

*Available on GitHUb: https://github.com/jcvdav/POLS207/blob/master/ps2/ps2.pdf

## Derive a simple estimator for the standard error of the above difference-in-means

Both $Y_1$ and $Y_0$ are independent from each other and describe two different random variables, which have means $\hat{Y_1}$ and $\hat{Y_0}$. Each of them has a variance around them, given by $\mathbb{V}(\hat{Y_1})$ and $\mathbb{V}(\hat{Y_0})$. The variance around the difference in means is then given by:

$$\mathbb{V}(\hat{Y_1} - \hat{Y_0})$$

Since the variance of a variable is given by the expectation of the squared difference between the value of a variable and the expectation of this value (*i.e.* $\mathbb{V}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2]$), we can rewrite the above as:

$$\mathbb{V}(\bar{Y_1} - \bar{Y_0}) = \mathbb{E}[((\bar{Y_1} - \bar{Y_0}) - \mathbb{E}[(\bar{Y_1} - \bar{Y_0})])^2]$$

We can expand the terms in the expectation on the right hand side and obtain:

$$\mathbb{V}(\bar{Y_1} - \bar{Y_0}) = \mathbb{E}[(\bar{Y_1} - \bar{Y_0} - \mathbb{E}[\bar{Y_0}] + \mathbb{E}[\bar{Y_1}])^2]$$

Re-grouping and factorizing a $-1$ from the second term gives us:

$$\mathbb{V}(\bar{Y_1} - \bar{Y_0}) = \mathbb{E}[((\bar{Y_1} - \mathbb{E}[\bar{Y_1}]) + (\bar{Y_0} - \mathbb{E}[\bar{Y_0}]))^2]$$

The first term $((\bar{Y_1} - \mathbb{E}[\bar{Y_1}]))$ contains the deviations from the expectation for $\bar{Y_1}$, and the second term contains the deviations from the expectation for $\bar{Y_0}$.

We can expand the squared term and obtain:

$$\mathbb{V}(\bar{Y_1} - \bar{Y_0}) = \mathbb{E}[(\bar{Y_1} - \mathbb{E}[\bar{Y_1}])^2 + 2(\bar{Y_1} - \mathbb{E}[\bar{Y_1}])(\bar{Y_0} - \mathbb{E}[\bar{Y_0}]) + (\bar{Y_0} - \mathbb{E}[\bar{Y_0}])^2]$$

We can expand the outer-most expectation and obtain:

$$\mathbb{V}(\bar{Y_1} - \bar{Y_0}) = \mathbb{E}[(\bar{Y_1} - \mathbb{E}[\bar{Y_1}])^2] + \mathbb{E}[(\bar{Y_0} - \mathbb{E}[\bar{Y_0}])^2] + 2(\bar{Y_1} - \mathbb{E}[\bar{Y_1}])(\bar{Y_0} - \mathbb{E}[\bar{Y_0}])$$

Per the deffinition of the variance, we would obtain that this is just:

$$\mathbb{V}(\bar{Y_1} - \bar{Y_0}) = \mathbb{V}(\bar{Y_1}) + \mathbb{V}(\bar{Y_0}) + 2(\bar{Y_1} - \mathbb{E}[\bar{Y_1}])(\bar{Y_0} - \mathbb{E}[\bar{Y_0}])$$

The last term on the right is just the covariance times a constant (2), which then gives us:

$$\mathbb{V}(\bar{Y_1} - \bar{Y_0}) = \mathbb{V}(\bar{Y_1}) + \mathbb{V}(\bar{Y_0}) + 2\text{cov}(\bar{Y_1}, \bar{Y_0})$$

Since we assumed $\bar{Y_1}$ and $\bar{Y_0}$ to be independent from each other, we would expect $\text{cov}(\bar{Y_1}, \bar{Y_0}) = 0$. Therefore, the standard error of the difference in means estimator is given by:

$$SE = \sqrt{\frac{\sigma^2_{Y1}}{N_1} + \frac{\sigma^2_{Y2}}{N_2}}$$

## Use the data to estimate the standard error you derived in [the previous exercise]

```r
# Get variances
sigma_y1 <- var(olken_dat$pct_missing[olken_dat$treat_invite == 1], na.rm = T)
sigma_y0 <- var(olken_dat$pct_missing[olken_dat$treat_invite == 0], na.rm = T)

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

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

The standard error is 0.0301966.

**Check the covariate balance in this dataset on all covariates (all variables that are not the treatment assignment or the outcome vectors).**

```r
balance <- olken_dat %>%
  drop_na() %>%
  dplyr::select(-pct_missing) %>%
  MatchBalance(
    treat_invite ~ head_edu + mosques + pct_poor + total_budget,
    data = .,
    print.level = 0
  ) %>%
  baltest.collect(
    var.names = c(
      "head_edu",
      "mosques",
      "pct_poor",
      "total_budget"),
    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."
)
```

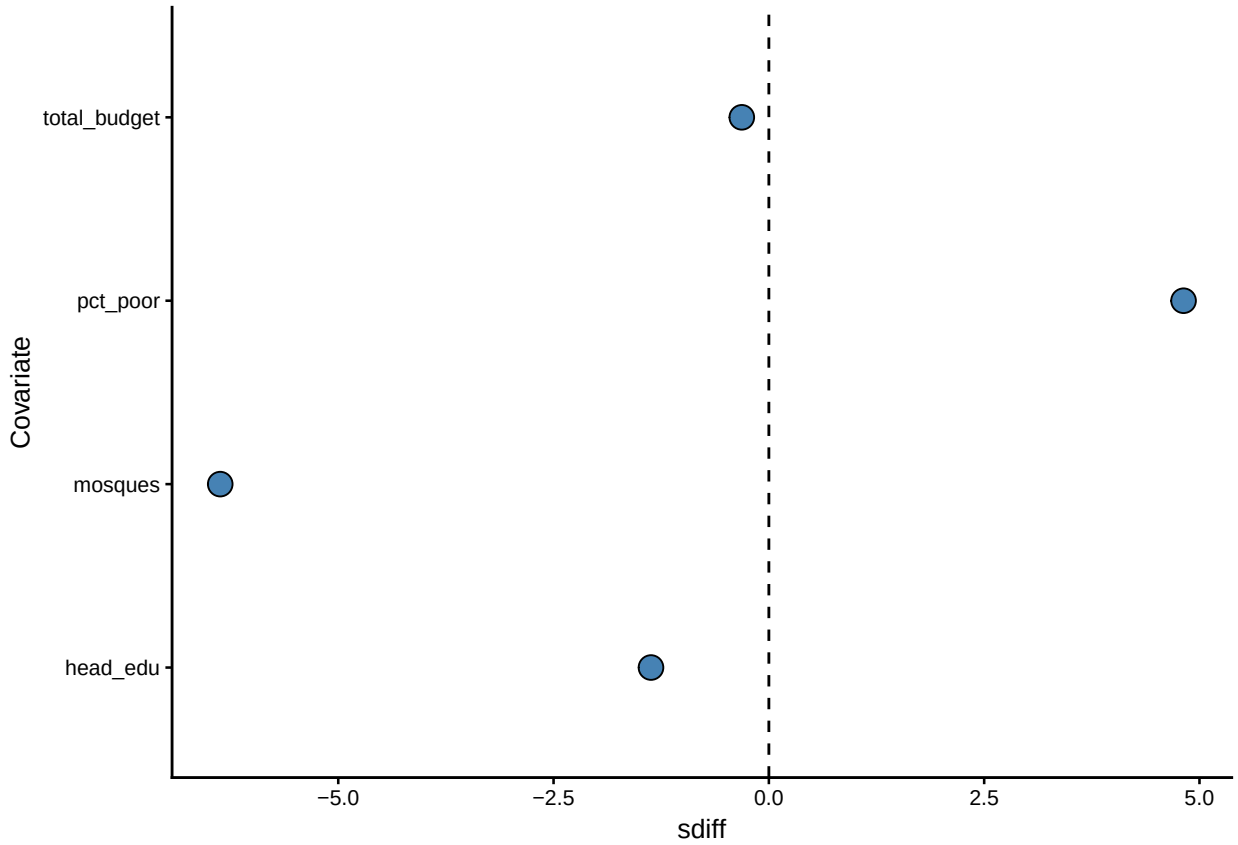



Figure 1: Pre-matching standardized difference in means of for covariates meassured for treadetd and untreated villages with projects.

Table 2: Pre-matching balance of covariates.

| Covariate | Mean (Treatment) | Mean (Control) | Standardized difference | Variance ratio | T p-value | KS p-value |
|---|---|---|---|---|---|---|
| head_edu | 11.5466238 | 11.583851 | -1.3687731 | 0.9986533 | 0.8880239 | 0.898 |
| mosques | 1.4195176 | 1.472887 | -6.3705719 | 1.0294290 | 0.5081709 | 0.708 |
| pct_poor | 0.4106243 | 0.400366 | 4.8150808 | 1.0054616 | 0.6196482 | 0.764 |
| total_budget | 83.1643437 | 83.354209 | -0.3142903 | 1.9821467 | 0.9685551 | 0.364 |

```
ggplot(data = balance, mapping = aes(x = Covariate, y = sdiff)) +
  geom_point(size = 4, shape = 21, fill = "steelblue", color = "black") +
  coord_flip() +
  ggtheme_plot() +
  geom_hline(yintercept = 0, linetype = "dashed")
```

**Now use regression to estimate the SATE (sample average treatment effect). Is this estimate different from the difference-in-means estimate?**

```
lm(pct_missing ~ treat_invite, data = olken_dat) %>%
  stargazer::stargazer(.,
                       se = estimatr::starprep(.,
                                               se_type = "HC1"),
                       t.auto = T,
                       p.auto = T,
                       header = F,
                       title = "Estimate of the Sample Average Treatment Effect.",
                       single.row = T
  )
```

Table 3: Estimate of the Sample Average Treatment Effect.

| | *Dependent variable:* |
|---|---|
| | pct_missing |
| treat_invite | −0.023 (0.033) |
| Constant | 0.252*** (0.026) |
| Observations | 477 |
| $R^2$ | 0.001 |
| Adjusted $R^2$ | −0.001 |
| Residual Std. Error | 0.344 (df = 475) |
| F Statistic | 0.486 (df = 1; 475) |
| *Note:* | *p<0.1; **p<0.05; ***p<0.01 |

**Using your answer for (b), conduct a t-test of the null hypothesis that SAT E = 0. You may use a normal approximation for the cutoff value.**

```
bar_y1 <- mean(olken_dat$pct_missing[olken_dat$treat_invite == 1], na.rm = T)
bar_y0 <- mean(olken_dat$pct_missing[olken_dat$treat_invite == 0], na.rm = T)

dif_means <- bar_y1 - bar_y0

t_score <- abs(dif_means / SE)

df <- N1 + N0 - 2 #N1 and N0 were calculated in b)

p_value <- 2 * pt(t_score, df = df, lower.tail = F) #calculate two-tailed
```

Student's t-test for unparied samples suggests that there are no differences in percent missing budget between treated and untreated groups ($t(565) = 0.7665$; $p = 0.4436$).

## Is the standard error of the OLS estimate different than the standard error of the differencein-means estimate? Why or why not?

The standard error of the OLS estimate is different to the standard error that I calculated in the difference in means. In the OLS estimation, I used heteroskedastic-robust standard errors. Furthermore, the SE's used in my derivation assumed that $\text{cov}(\bar{Y}_1, \bar{Y}_0) = 0$, and ignored sampling with / without replacement issue.

## Re-estimate SAT E using three additional regression models

### One in which you include all pre-treatment covariates as additional linear predictors

```
m1 <- lm(pct_missing ~ ., data = olken_dat)
```

### Another in which you include arbitrary functions of the covariates

```
m2 <- lm(pct_missing ~ treat_invite + mosques * pct_poor + total_budget + I(total_budget^ 2),
         data = olken_dat)
```

### A third in which you interact the treatment variable with a demeaned covariate ($X_i - \bar{X}$)

```
olken_dat_demean <- olken_dat %>%
  mutate_at(.vars = vars(head_edu, mosques, pct_poor, total_budget),
            .funs = function(x){x - mean(x, na.rm = T)})

m3 <- lm(pct_missing ~ treat_invite * total_budget,
         data = olken_dat_demean)
```

### Report the treatment effect estimates and their robust standard errors

```
models <- list(m1, m2, m3)

stargazer::stargazer(models,
                     se = estimatr::starprep(models,
                                             se_type = "HC1"),
                     t.auto = T,
                     p.auto = T,
                     header = F,
                     title = "Estimate of the Sample Average Treatment Effect.",
                     single.row = T
)
```

Table 4: Estimate of the Sample Average Treatment Effect.

| | *Dependent variable:* | | |
|---|---|---|---|
| | pct_missing | | |
| | (1) | (2) | (3) |
| treat_invite | −0.026 (0.033) | −0.020 (0.033) | −0.023 (0.033) |
| head_edu | −0.006 (0.006) | | |
| mosques | −0.048** (0.019) | −0.095** (0.041) | |
| pct_poor | −0.118 (0.075) | −0.320** (0.154) | |
| total_budget | 0.001* (0.0003) | 0.001** (0.001) | 0.001 (0.001) |
| I(total_budget^2) | | −0.00000 (0.00000) | |
| mosques:pct_poor | | 0.134 (0.090) | |
| treat_invite:total_budget | | | −0.0001 (0.001) |
| Constant | 0.390*** (0.087) | 0.354*** (0.079) | 0.251*** (0.027) |
| Observations | 472 | 474 | 477 |
| $R^2$ | 0.029 | 0.034 | 0.008 |
| Adjusted $R^2$ | 0.019 | 0.022 | 0.001 |
| Residual Std. Error | 0.341 (df = 466) | 0.341 (df = 467) | 0.343 (df = 473) |
| F Statistic | 2.823** (df = 5; 466) | 2.756** (df = 6; 467) | 1.220 (df = 3; 473) |
| *Note:* | | | *p<0.1; **p<0.05; ***p<0.01 |

# Problem 2: Randomization Inference

### is the sharp-null, and how does it compare to the null hypothesis we tested in the previous question?

The sharp-null states that the treatment effect is 0 for all units. This means that, unlike the previous example, the treated and untreated potential outcomes are the same for each observation.

### Why is the sharp null a convenient choice? I.e., what special property of the sharp null allows us to obtain a distribution of outcomes for a test statistic under this null?

The sharp null is a non-parametric test. This means that we de not make any assumptions about the distribution of the variable of interest. By the central limit theorem, we can then use a normal distribution (or approximate one with a t distribution).

### Write your own function in `R` that takes as arguments a vector of outcomes, Y , and the original treatment assignment vector, D, and produces:

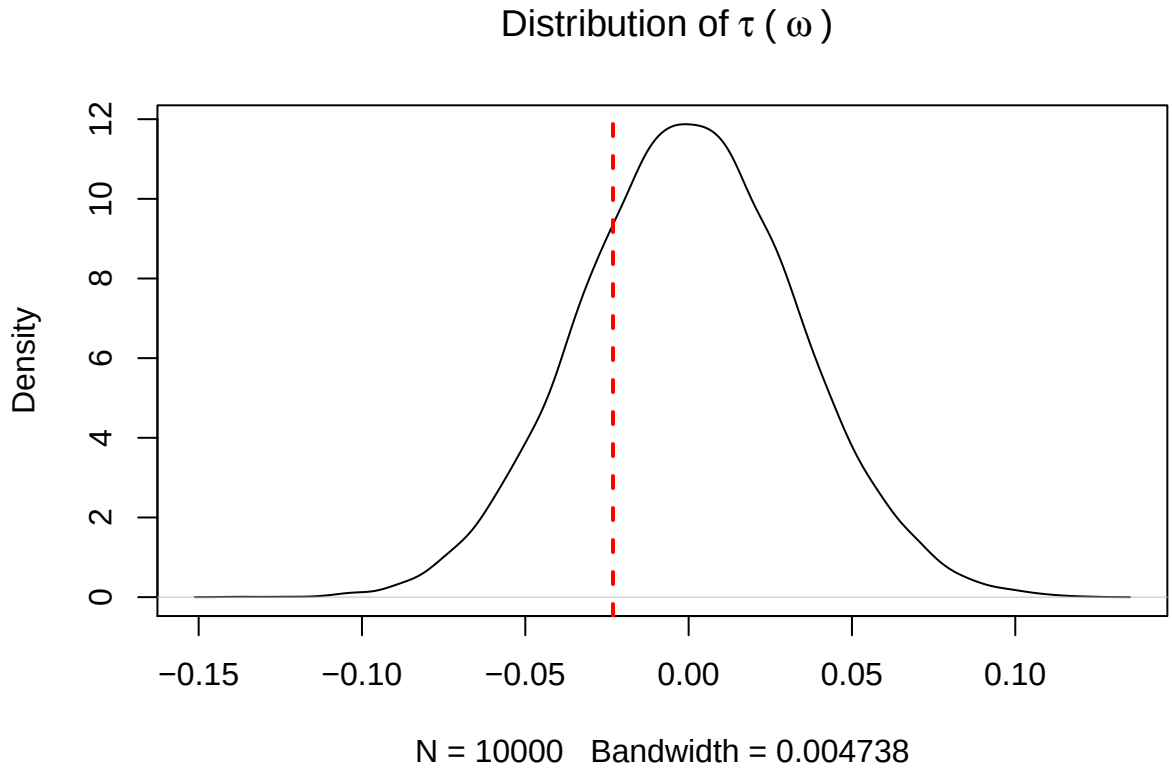
- (i) a plot showing the distribution of the difference in means statistic under the sharp-null
- (ii) a vertical line representing the observed difference in means relative to this distribution
- (iii) a p-value for the difference in means statistic against the sharp-null.

```r
sharp_null_fxn <-
  function(y = NULL, D = NULL, n_perms = 10000, two_sided = FALSE, seed = 42, ...){
    # Run checks
    ## Did the user specify all parameters?
    if(is.null(y))
    {stop("You did not specify a vector of outcomes.")
    }
    if(is.null(D)){
      stop("You did not specify a vector of treatments.")
    }
    n_obs <- length(y)
    # Are parameters the correct size?
    if(n_obs != length(D)){
      stop(paste0("y and D have different lengths(", n_obs," and ", length(D),")."))
    }
    # Is D the correct class?
    if(!is.logical(D)){
      stop("D must be a logical vector with TRUE or FALSE indicating treatment or control.")
    }
    # Is the suggested number of permutations large enough?
    if((n_perms / n_obs) < 10){
      warning("Your number of permutations might not be high enough")
    }

    # Get treated and control units
    treated <- y[D]
    not_treated <- y[!D]
    # Calculate true difference in means
    true_diff_in_means <- mean(treated, na.rm = T) - mean(not_treated, na.rm = T)

    # Set a random seed for reprodcibility
    set.seed(seed)
    # Create empty vector to save the omega-estimated
    omega_diff_in_means <- rep(NA, length = n_obs) #safer than numeric(length = n_obs)
    # iterare over n_perms
    for(omega in 1:n_perms){
      # Generate a random treatment assignment vector.
      # Sample without replacement to obtain same proportion of treated and untreated.
      D_omega <- sample(x = D, size = n_obs, replace = FALSE)
      # Get treated and not treated based on D_omega
      treated_omega <- y[D_omega]
      not_treated_omega <- y[!D_omega]
      # Calculate difference in means
      omega_diff_in_means[omega] <-
        mean(treated_omega, na.rm = T) - mean(not_treated_omega, na.rm = T)
    }
    # Plot the density of the iterated estimates
    plot(density(omega_diff_in_means), main = quo("Distribution of"~tau~"("~omega~")"))
    # Add a line with the true difference in means
    abline(v = true_diff_in_means, col = "red", lwd = 2, lty = "dashed")
    # Calculate probabilities
    if(two_sided){
      p <- sum(abs(omega_diff_in_means) >= abs(true_diff_in_means)) / n_perms
```

```
    } else {
      p <- sum(omega_diff_in_means >= true_diff_in_means) / n_perms
    }

    return(p)
  }
```

**Apply your function to the data from the Olken experiment in the previous section. How do your results compare to the results under the t-test or regression? Which do you trust, and how do you interpret any differences?**

```
# Create vector of outcomes
y <- olken_dat$pct_missing

# Create treatment vector
D <- olken_dat$treat_invite == 1

# Run the sharp null test
sharp_null <- sharp_null_fxn(y = y, D = D)
```




With the previous run, 75.75% of the $\hat{\tau}(\omega)$ had a value equal to or greater than $\hat{\tau}_{ATE}$. In terms of our one-sided hypothesis, this means that only 24.25% of the random treatment assignment vectors produced an effect size larger than or equal to our estimated effect. I can re-run the function asking for the two-sided (`sharp_null_fxn(y = y, D = D, two_sided = T)`), in which case $p = 0.4869$. The t-test and the regression

suggested that there was a 0.44 and 0.48 chance of randomly obtaining a value equal to or larger than the difference in means estimate. The interpretation is a bit different, given that the null hypothesis assumes no change on any unit.