EDS241: FINAL

Briana Barajas

02/23/2024

Make sure to read through the setup in markdown. Remember to write out interpretations and report your results in writing/table/plot forms.

1 Part 1: RCTs, treatment ignorability (selection on observables), propensity scores *(15 points total)*

Setup

This exercise is inspired by Costello et al. 2008 article in science "Can Catch Shares Prevent Fisheries Collapse", which we also discussed in class (lecture 5). "Inspired" means that the data `final_fisheries_data.csv` are synthetically generated to simplify things for our purposes. It contains the variables on 11,135 fisheries (only cross sectional, no time observations): These fisheries were either regulated by an Individual Transferable Quota (ITQ) for all years between 1990 and 2012 or in none of those years. Variables in the dataset include:

The outcome and treatment variables are:

`COLL_SHARE` = share of years a fishery is collapsed between 1990 and 2012 (collapse defined as harvest being more than 10% below maximum recorded harvest).

`ITQ` = dummy variable indicating 'treatment' with an ITQ (equal to 1 if the fishery has been regulated by an ITQ and 0 otherwise).

The control variables are:

`MET1`, `MET2`, ...`MET6` = Dummy variables indicating to which Marine Ecosystem Type (MET) the fishery belongs to (coral reefs, kelp forests, seagrass meadows, open ocean, deep sea, mangrove forests). This type does not change over the relevant time period and does not depend on human influence.

`IND_SR` = Index of species richness in 1980 with values between 0 and 100 indicating the biodiversity with respect to species in the fishery. Bounds of 0 and 100 are the lowest and highest observed values of species diversity across all fisheries in 1980, respectively.

`COMM_VAL` = Commercial value of fisheries in 1980 in million US-$

The basic question of interest is **"What is the average treatment effect of implementing an ITQ in the time period from 1990 to 2012 on the share of years with a collapse.** It is likely that the probability a fishery is selected for an ITQ depends on the pre-treatment characteristics given. It is also quite likely that the pre-treatment characteristics have an effect on the share of collapse for each fishery, i.e. our outcome variable of interest.

```
## Load Data
fish <- read_csv(here("final", "data", "final_fisheries_data.csv")) %>%
  clean_names()
```

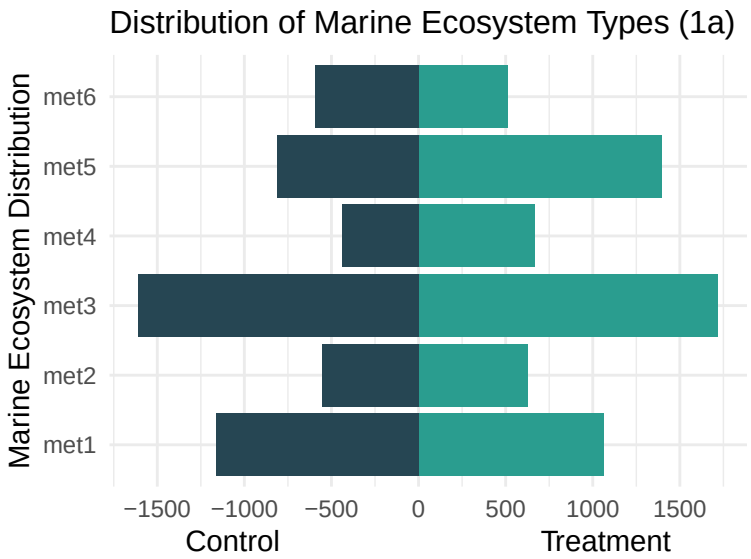
```r
# pivot data (turn met into single column)
fish_pivot <- fish %>%
  pivot_longer(cols = 1:6, names_to = "ecosystem_type", values_to = "value") %>%
  mutate(ecosystem_type = as.factor(ecosystem_type)) %>%
  filter(value == 1)
```

Question (a) Pretreatment Ecosystem Characteristic Comparison, Visual *(3 pts)*

(a) Compare the distributions of pre-treatment ecosystem characteristics (i.e. MET1, MET2, „, MET6) between the treated and the control groups by drawing back to back histograms [2 pts]. Write one sentence discussing the (dis)similarity between the two groups [1pt].

```r
# create subset of treated and untreated
fish_treat <- fish_pivot %>% filter(itq == 1)
fish_control <- fish_pivot %>% filter(itq == 0)

# plot to compare distribution of ecosystem types
ggplot() +
  geom_col(data = fish_treat, aes(x = ecosystem_type, y = value), fill = "#2a9d8f") +
  geom_col(data = fish_control, aes(x = ecosystem_type, y = -value), fill = "#264653") +
  scale_y_continuous(n.breaks = 10) +
  coord_flip() +
  theme_minimal() +
  labs(y = "Control                                    Treatment",
       x = "Marine Ecosystem Distribution",
       title = "Distribution of Marine Ecosystem Types (1a)") +
  theme(plot.title = element_text(size = 12))
```



ANS 1A: I decided to pivot the data frame so marine ecosystem type would be a single column, making it easier to compare the treatment and control groups. The distribution for ecosystem types in the control and treatment groups (where treatment is `itq`) are not equal. For example, in `met5` and `met4` there appear to be more observations for the treatment group.

Table 1: Mean differences of Treatment vs. Control (1b)

Variable	Mean Treated	Mean Control	P-Value
ind_sr	57.38515	48.55968	0
comm_val	117.22839	84.87908	0

Question (b) Pretreatment Ecosystem Characteristic Comparison, Mean differences *3 pts)*

(b) Do a test on mean differences between the treated and control groups for the species richness index (`IND_SR`) and commercial value (`COMM_VAL`) variables. Interpret the results (estimated difference and significance) [2 pts] and make a conclusion regarding the similarity between the groups [1pt].

```
## Mean Differences (remember to use prop.test or t.test when applicable)

# calculate differences in mean for sp richness
mean_diff_ind <- broom::tidy(
  t.test(ind_sr ~ itq, data = fish)) %>%
  select(estimate1, estimate2, p.value) %>%
  mutate(Variable = "ind_sr")

# calculate differences in mean for commercial value
mean_diff_comm <- broom::tidy(
  t.test(comm_val ~ itq, data = fish)) %>%
  select(estimate1, estimate2, p.value) %>%
  mutate(Variable = "comm_val")

# bind result tables
t_test_results <- rbind(mean_diff_ind, mean_diff_comm) %>%
  relocate(Variable)

# print results
t_test_results %>%
  kbl(col.names = c("Variable", "Mean Treated", "Mean Control", "P-Value"),
      caption = "Mean differences of Treatment vs. Control (1b)") %>%
  kable_paper()
```

ANS 1B: A t-test was used to compare the difference in the means for the continuous variables, and both tests resulted in a p-value that was less than 0.05. Given a significance level (α) of 0.05, we can reject the null hypothesis that the difference in means between the control group and treatment group equals zero. This is true for both `ind_sr` and `comm_val`.

Question (c) Treatment Ignorability *(1 pt)*

(c) Based on your results from (a) and (b), do you see a problem with just comparing the outcome variable means between treated and untreated fisheries?

ANS 1C: Based on the histogram from *Question 1a*, there appears to be differences in the marine ecosystem type (`met`) of the control group versus the treatment. This would cause bias when statistically estimating whether or not regulation by an Individual Transferable Quota (ITQ) altered the share of years a fishery collapsed (`coll_share`). Similarly the t-test results above demonstrate that the difference in the

mean species richness index (`ind_sr`) and commercial value (`comm_val`) are statistically significant. This adds to the potential for bias, since pre-treatment characteristics vary for the treated and untreated.

Question (d) Propensity Scores *(2 pts)*

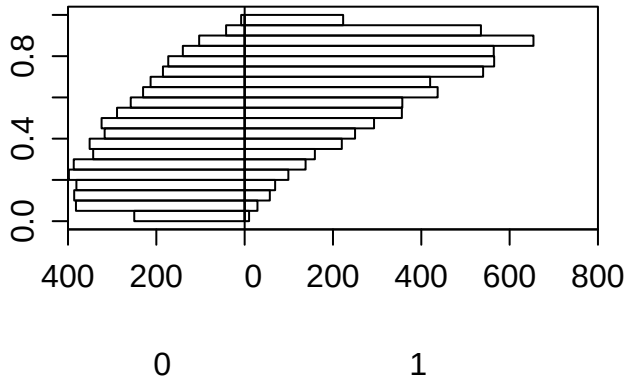
(d) Estimate the propensity scores (probability of being treated) using a logit model, assume that all covariates are relevant and should be included in the estimation [0.5 pt]. Draw separate histograms (back to back) of the propensity scores for the treated and the untreated group [0.5 pt]. Comment on the overlap, do you have any concerns? Why/why not? [1]

```
## ..........Propensity Score Estimates.............

propensity_scores <- glm(itq ~ met1 + met2 +met3 + met4 + met5 + met6 +
                           ind_sr + comm_val, data = fish, family = binomial())

fish$psvalue <- predict(propensity_scores, type = "response")

histbackback(split(fish$psvalue, fish$itq))
```



ANS 1D: There is a decent amount of overlap for the center values, so matching should work. I am a bit concerned about the observations on the more extreme ends. For example, there are many observations in the treated group with a value around 1, but there are not many counterfactuals.

Question (e) ATT with Nearest Neighbor Matching *(3 pts: 2 pt estimate, 1 pt interpretation)*

(e) Use the propensity scores from (d) to estimate the Average Treatment Effect on the Treated (ATT) with a nearest neighbor matching estimator. Interpret the result (just the size of the estimate)

```
## ..........Nearest Neighbor Matching.................

# find matches using nearest-neighbors method
match_itq <- matchit(itq ~  met1 + met2 +met3 + met4 + met5 + met6 +
```

```
                      ind_sr + comm_val, data = fish,
                    method = "nearest", ratio = 1)

# store results
match_itq_df <- match.data(match_itq)

# calculate the difference by subgroup
subgroup_diff <- match_itq_df %>%
  group_by(subclass) %>%
  mutate(diff = mean(coll_share[itq == 1]) - mean(coll_share[itq == 0]),
         .groups = "drop")

## .................Estimate ATT.....................

# calculate the mean for all differences of aTT
ATT <- mean(subgroup_diff$diff)

paste("Estimate of ATT using matched dataset:", ATT)
```

```
## [1] "Estimate of ATT using matched dataset: -0.0713262272262593"
```

ANS 1E: The average treatment effect, -0.0713, indicates the difference in the proportion of years that a fishery was collapsed (between 1990-2012). On average, fisheries in the treated group were collapsed for fewer years than the control group.

Question (f) ATE with WLS *(3 pts: 1 pt estimate, 1 pt interpretation)*

(f) Estimate the Average Treatment Effect (ATE) using the weighted least squares on the full sample. Interpret the estimated size and conclude if it is significantly different from zero from a statistical perspective.

```
## ..................WLS Matching.....................

# isolate variables needed to calculate WLS estimator
PS <- fish$psvalue
D <- fish$itq

# calculate weights
fish <- fish %>%
  mutate(wgt = (D/PS + (1-D)/(1-PS)))

fish$wgt = ifelse(D == 1,
                  (1 / PS),
                  (1 / (1 - PS)))

## .................Estimate ATE.....................

# calculate WLS estimate without controls
fish_wls <- lm(coll_share ~ itq,
               data = fish, weights = wgt)

# calculate WLS estimate with controls
```

Table 2: WLS Without Controls (1f)

term	estimate	std.error	statistic	p.value
(Intercept)	0.2578	0.0010	271.2388	0
itq	-0.0763	0.0013	-56.6355	0

Table 3: WLS With Controls (1f)

term	estimate	std.error	statistic	p.value
(Intercept)	0.3748	0.0030	123.7914	0
itq	-0.0767	0.0010	-77.4356	0
ind_sr	-0.0015	0.0000	-35.8043	0
comm_val	0.0004	0.0000	33.9440	0
met1	-0.1120	0.0019	-58.1740	0
met2	-0.1286	0.0022	-58.2007	0
met3	-0.0969	0.0018	-53.5442	0
met4	-0.0635	0.0022	-28.3526	0
met5	-0.0323	0.0019	-16.8212	0
met6	NA	NA	NA	NA

```
fish_wls_ctrl <- lm(coll_share ~ itq + ind_sr + comm_val + met1 + met2 +
                      met3 + met4 + met5 + met6, data = fish,
                    weights = wgt)

# store and print results to view ATE
wls_results <- broom::tidy(fish_wls)
wls_ctrl_results <- broom::tidy(fish_wls_ctrl)

wls_results %>%
  kbl(caption = "WLS Without Controls (1f)", digits = 4)
```

```
wls_ctrl_results %>%
  kbl(caption = "WLS With Controls (1f)", digits = 4)
```

ANS 1F: The estimated average treatment effect (ATE) when using weighted least squares with all control variables is -0.0766 years. This ATE indicates that, on average, there is a decrease of 0.0766 in the proportion of years that a fishery is collapsed. The low p-value (less than 0.05) indicates that the difference in `coll_share` between fisheries with and without individual transferable quotas is significant.

2 Part 2 Difference in Difference Estimation *(10 points total + 3pts extra credit)*

Here we return for a final time to the dataset from Gertler, Martinez, and Rubio-Codina (2012) and use a different way of estimating the **effect of the Mexican conditional cash transfer on the value of animal holdings** of recipients. We'll use the panel data from assignment 2, where you have both the pre-program and post-program observations. See Template for dataset preparation instructions.

Data Preparation

*Note: You will need to install the packages plm and dplyr (included in template preamble). Again, you can find a description of the variables at the bottom of PDF and HERE.

Prepare Data: Load the new data (progresa_pre_1997.csv) and the follow-up data (progresa_post_1999.csv) into R. Note that we created a time denoting variable (with the same name, 'year') in BOTH datasets. Again, you will create a panel dataset by appending the data (i.e. binding the dataset row-wise together creating a single dataset). We want to examine the same outcome variable as before, value of family animal holdings (vani). You will use the full dataset for each estimate. NOTE: you should not change any NAs from the TREATED column in your analysis, as we expect that spillover was likely in this program. NAs will be excluded from your calculations/estimations.

Question (a) DiD Estimator, ATE *(5 pts: 3 pts estimate, 2 pts interpretation)*

(a) Calculate the DiD estimator of the treatment effect (ATE) of the program on the value of animal holdings (vani) "manually" i.e. based on group mean values without running a regression. Report and interpret the result (Note: no significance test or standard errors is possible, so you do not need to report these values).

```
## ........Calculate means to estimate ATE.............

# calculate mean vani for treatment and control, both years
diff <- progresa_full %>%
  filter(treatment == 1 | treatment == 0) %>% # remove NAs for treatment

  group_by(year, treatment) %>%
  summarise(mean_vani = mean(vani, na.rm = TRUE)) %>% # calculate mean vani for year+treatment

  ungroup() %>%
  group_by(treatment) %>%
  mutate(diff_1999_97 = mean_vani - dplyr::lag(mean_vani))  #calculate mean differences


## ................Compute the DiD.....................

# calculate did
did <- diff %>% ungroup() %>%
  drop_na(diff_1999_97) %>%
  mutate(did = diff_1999_97 - dplyr::lag(diff_1999_97))

glue::glue("Difference-in-difference estimator:", as.numeric(did[2,5]))
```

```
## Difference-in-difference estimator:287.904957673713
```

ANS 2A: The average treatment effect, standardized for changes over time using the difference-in-difference estimator, is 287.9 pesos. This indicates that, on average, households that underwent treatment saw an increase of 287.9 pesos in their value of animal holdings (`vani`).

Question (b) Difference in Difference using OLS *(5 pts)*

(b) Now set up an OLS-regression using group mean values to estimate the same ATE. Interpret the estimated treatment effect [3 pts]. Also interpret the coefficients on the time dummy and the group dummy variable (see interpretation done in class in lecture 9) [2 pts].

Table 4: OLS Model - Mean Data (2b)

term	estimate	std.error	statistic	p.value
(Intercept)	2848.2175	NaN	NaN	NaN
treatment	-237.6927	NaN	NaN	NaN
post_treat	-1156.7517	NaN	NaN	NaN
time_x_treated	287.9050	NaN	NaN	NaN

Hints: You will need to create a new dataframe with a variety of dummy variables to do this. The R example provided with the DiD module (and/or the excel file) should help.

```
## ..........Calculate means for estimate.............

# create df with vani mean and dummy variables
group_mean <- progresa_full %>%
  filter(treatment == 1 | treatment == 0) %>%

  # calculate group means
  group_by(year, treatment) %>%
  summarise(mean_vani = mean(vani, na.rm = TRUE)) %>%
  ungroup() %>%

  # add binary post-treatment column
  mutate(post_treat = case_when(
    year == 1999 ~ 1,
    year == 1997 ~ 0),

    # add time-treatment interaction column
    time_x_treated = treatment * post_treat)

## ................OLS Regression.....................

# run the OLS regression w/dummies
model1 <- lm(mean_vani ~ treatment + post_treat + time_x_treated, data = group_mean)

## ............Present results in table..............

model1_results <- broom::tidy(model1)

model1_results %>% kbl(caption = "OLS Model - Mean Data (2b)")
```

ANS 2B: The intercept value indicates that the average value of animal holdings (`vani`) prior to treatment is 2,848 pesos. The `post_treat` estimate can be interpreted as the change in `vani` between pre-treatment and post-treatment time. In this case, between 1997 and 1999 the value of animal holdings decreased by 1156.75 pesos on average. The `treatment` estimate indicates the mean difference in `vani` between the treated and untreated groups, holding all other variables constant. This demonstrates a decrease in `vani` based on treatment, but that would be an incorrect interpretation.

The value we're truly interested in for assessing the average treatment effect is the `time_x_treated` estimate. The `time_x_treated` term estimates the effect of the cash treatment on households, adjusting for time. Therefore, the treatment increased `vani` by an average of 287.9 pesos.

Table 5: OLS Model - Full Data (2c)

term	estimate	std.error	statistic	p.value
(Intercept)	2848.218	60.615	46.989	0.000
treatment	-237.693	80.453	-2.954	0.003
post_treat	-1156.752	85.722	-13.494	0.000
time_x_treated	287.905	113.778	2.530	0.011

3 Extra Credit: ATE with OLS using full dataset *(3 pts: 2 pts estimate, 1 pt interpretation)*

(c) Estimate the ATE with an OLS-regression based on the original units as observations (i.e. not with group mean values, you will need to use the entire dataset). Even though the specification is the same as in the regression with the group mean values above, you'll need to create new indicator variables for the treatment group and the post treatment time period as well as their interaction term. Verify that you get the same result as above. Now report also on the precision of the estimation and test whether the estimated coefficient is different from zero.

```
## .............Create dummy variables.................
progresa_full <- progresa_full %>%

  # add dummy variables to full df
  mutate(post_treat = case_when(
    year == 1999 ~ 1,
    year == 1997 ~ 0),

    time_x_treated = treatment * post_treat)

## .................OLS Regression.....................

## conduct regression using dummy variables
model2 <- lm(vani ~ treatment + post_treat + time_x_treated, data = progresa_full)

## ............Present results in table...............

# store and print model results
model2_results <- broom::tidy(model2)

model2_results %>% kbl(caption = "OLS Model - Full Data (2c)", digits = 3)
```

ANS 2C: The OLS regression with the full data returned the same results as the regression with the mean `vani` values. Using the p-values in the second regression, we can confirm that the interpretations above are statistically significant (p-value $<$ *alpha*). More specifically, the difference in `vani` between the treated and untreated, controlling for time is statistically significant. This corresponds to the `time_x_treated` variable, as discussed above.