# POLI 171: Problem Set 4: Regression and Matching

Due Tuesday, February 27, at the beginning of class

For this assignment, you may collaborate with one other student. You may also complete this assignment on your own, if you prefer. If you collaborate, you and your co-author will turn in a single document, and you will each receive the same grade. You and your co-author are expected to work together on *all* parts of this assignment; do not simply divide the questions between yourselves.

Please type your answers and submit them as a hardcopy. Please show your work by including your `R` code and output in your answers.

## The National Supported Work Program

The National Supported Work (NSW) program was a temporary employment program created in the late-1970s. It was designed to provide low-wage, short-term employment to four broad categories of people who were having trouble finding well-paying jobs on the regular labor market: poor single mothers, recovering drug addicts, ex-convicts who had recently been released from prison, and teenagers who had dropped out of high-school.

For this problem set, you will use three datasets:

- `Experimental`: The dataset from a pilot experiment used to evaluate the effectiveness of the NSW program. Every individual in this dataset was **qualified** to participate in the NSW and **applied** to participate in the NSW. Approximately 40% of these applicants were randomly selected to participate in the NSW, and the remaining 60% were excluded from the program. For this dataset, treatment-assignment was random. The observations are individuals.

- `ObservationalCross`: A non-experimental dataset. This dataset includes all of the treated individuals who are in the `Experimental` dataset, but the data on the non-treated individuals comes from a survey of the broader population that was carried out at the same time. For this dataset, treatment-assignment was **not** random, and many of the non-treated individuals did not qualify for the NSW. The observations are individuals.

- `ObservationalPanel`: A non-experimental dataset in panel format. This dataset is based on the `ObservationalCross` dataset and it includes all of the same individuals, but it has fewer variables and it was transformed so that the observations are individual-years instead of just individuals. Each individual appears in the `ObservationalPanel` dataset twice: once for 1974 and once for 1978.

The `Experimental` and `ObservationalCross` datasets include the following variables:

| Variable | Description |
|---|---|
| `NSW` | Indicates whether the individual participated in the NSW. This is the treatment variable. |
| `Age` | The individual's age in 1978 |
| `Education` | The individual's education, in years of schooling |
| `Black` | = 1 if the individual identifies as African American, and = 0 otherwise |
| `Hispanic` | = 1 if the individual identifies as Hispanic, and = 0 otherwise |
| `Income74` | The individual's inflation-adjusted income in 1974 |
| `Income75` | The individual's inflation-adjusted income in 1975 |
| `Unemployed74` | = 1 if the individual was unemployed in 1974, and = 0 otherwise |
| `Unemployed75` | = 1 if the individual was unemployed in 1975, and = 0 otherwise |
| `Income78` | The individual's inflation-adjusted income in 1974. This is the outcome variable. |

The treatment variable is `NSW` and the outcome variable is `Income78`. All other variables are "pre-treatment variables."

The `ObservationalPanel` dataset includes the following variables:

| Variable | Description |
|---|---|
| `ID` | The individual's ID number |
| `Year` | The year |
| `NSW` | Indicates whether the individual participated in the NSW in the given year. This is the treatment variable. |
| `Income` | The individual's inflation-adjusted income in the given year |

In this problem set, you will estimate the effect of participation in the NSW on income using the various methods that we have covered in this course so far. For question 1, you will use the `Experimental` dataset to calculate an unbiased estimate of the treatment effect; you will refer back to this unbiased estimate throughout the remainder of the problem set. For question 2, you will use the `ObservationalCross` dataset, and you will estimate the treatment effect by controlling for pre-treatment variables in a regression. For question 3, you will use the `ObservationalCross` dataset and you will estimate the treatment effect by matching on pre-treatment variables. For question 4, you will use the `ObservationalPanel` dataset, and you will estimate the treatment effect using a fixed effects regression. **Make sure that you use the correct dataset for each question!**

## Question 1: The "True" Treatment Effect (2 points)

For this first question, you will use the `Experimental` dataset to calculate the treatment effect and test for balance. The methods that you need for this question are the same ones that you used for Problem Set 3. You can assume that all subjects complied with their treatment assignment.

a) Using the experimental dataset, calculate the effect of participation in the NSW on 1978 income. Interpret this effect in words. Is it statistically significant?

b) Verify that the treatment and control groups are balanced on the two pre-treatment income variables (`Income74` and `Income75`). There is no need to make a full balance table this time; just report the results of the difference-in-means test or the regression that you used and interpret the results. Are the treatment and control groups balanced on these variables?

## Question 2: Regression (4 points)

For this question, you will use the `ObservationalCross` dataset, and you will estimate the treatment effect using a regression on this non-experimental data.

a) Using the `ObservationalCross` dataset, regress `Income78` on just `NSW`. Interpret the estimated treatment effect in words. How does it compare to the treatment effect that you calculated using the experimental data in Question 1?

b) You suspect that the reason why the regression approach misestimated the treatment effect was because treatment status is correlated with pre-treatment variables that exert an independent effect on Income in 1978. Test this hypothesis by evaluating the balance between the treatment and control groups on the pre-treatment income variables (`Income74` and `Income75`). Are the treatment and control groups balanced on these variables?

c) You decide to take into account the pre-treatment differences between the treatment and control groups by controlling for pre-treatment variables in your regression. Regress `Income78` on `NSW` and **all** of the pre-treatment variables in the `ObservationalCross` dataset (i.e., control for all of the pre-treatment variables). Interpret the estimated treatment effect of participation in the NSW in words. How does it compare to the results that you found in Question 1(a) and Question 2(a)?

d) Based on the description of the NSW program at the beginning of this question prompt and the summary of the variables that are in this dataset, do you think that your regression in part (c) is controlling for all of the relevant pre-treatment variables? If not, name one individual-level variable that is not included in this dataset but is nevertheless relevant for explaining both treatment assignment and 1978 income.

## Question 3: Propensity-Score Matching (4 points)

For this question, you will continue to use the `ObservationalCross` dataset, but you will estimate the treatment effect using propensity-score matching.

a) Use the `matchit()` function to perform propensity-score matching on **all** of the pre-treatment variables. Allow the algorithm to drop unmatchable individuals from both the treatment and control groups by specifying the option "`discard="both"`." Ignore the error message. What percentage of the treated individuals do you end up keeping?

b) Are the matched treatment and control groups balanced on the variables that you matched on? Discuss any major imbalances between the treatment and control groups.

c) Estimate the treatment effect using the matched data. Begin by extracting the matched dataset by using the `match.data()` function. Then use this extracted matched dataset to regress `Income78` on `NSW`. If you like, you can control for the pre-treatment variables in this regression. Interpret the treatment effect in words. How does it compare to the treatment effect that you calculated using the experimental data in Question 1?

d) You are concerned by how poor a job your matching algorithm did at producing balance on the pre-treatment income variables, `Income74` and `Income75`. You think that these variables are more relevant to explaining post-treatment income than most of the other variables that you matched on, and you decide that you would rather ensure that your groups are balanced on just these two variables. Perform propensity-score matching again, but this time match on **only** `Income74` and `Income75`; ignore all of the other pre-treatment variables. Are the matched treatment and control groups balanced on `Income74` and `Income75` (the variables that you matched on)?

e) Repeat part (c) of this question using the new matched dataset that you created in part (d). Calculate the treatment effect on this new matched dataset. Interpret the new estimate of the treatment effect in words. How does it compare to the results that you found in Question 1 and Question 3(c)?

f) Compare the performance of your matching algorithm from parts (d-e) with the performance of your matching algorithm from parts (a-c). Which version of propensity-score matching got closest to the "true" treatment effect that you calculated in Question 1? What does this tell you about the advantages and disadvantages of matching on as many pre-treatment variables as possible?

## Question 4: Fixed Effects (4 points)

For this question, you will use the `ObservationalPanel` dataset and a fixed effects regression to estimate the treatment effect. Individuals are identified by the `ID` variable, and years are identified by the `Year` variable.

a) Use a fixed effects regression to calculate the treatment effect of participation in the NSW on income. Regress `Income` on `NSW`, and include fixed effects for each individual and each year (make sure to include **both** individual fixed effects **and** year fixed effects!). Don't worry if it takes your computer a few minutes to complete the calculations. Do **not** use the `coeftest()` function this time (it will take too long). Instead, extract just the coefficients using `coef(mod)`, where `mod` is the object that stores the fixed effects regression results. You do **not** need to report all of the coefficients in your write-up; report only the coefficient for the `NSW` variable. Interpret the estimated treatment effect of participation in NSW in words. How does it compare to the results that you found in Question 1a and Question 2c?

b) The `ObservationalPanel` and `ObservationalCross` datasets include the exact same individuals; the `ObservationalCross` dataset includes all of the information that is in the `ObservationalPanel` dataset, plus a lot of additional information. Why, then, did the fixed effects regression that you calculated in part (a) do a better job at estimating the treatment effect than the controlled regression that you calculated in question 2c?