# Modeling and Representation of Data - Team Project 1

27th September, 2021

## Introduction

-add
-your
-intro
-here
-and
-here
-and
-here

### Data Pre-processing

We study a subset of of the data collected by the NSW demonstration. It was originally organized in Lalonde, R. J. (1986), which is our main reference for this analysis. In the data provided, the treatment group includes male participants for which 1974 earnings can be obtained, and the control group includes all the unemployed males in 1976 whose income in 1975 was below the poverty level.

Refer to appendix for the data dictionary. A summary of the continuous variables can be found in Table 1. The following transformations are applied to the data:

- Combine 'black' and 'Hispanic' into a 'race' variable, where 0 indicates 'other' races, 1 indicates black men and 2 indicates Hispanic men.
- Create a new variable 'growth', which is the difference between the real earnings in 1978 and the real earnings in 1974.

We have 614 records in our final data, with 429 participants in the control group and 185 in the treatment group. Large number of zero values are seen in each of the income columns (details in Appendix), but given that the experiment was run on disadvantaged workers lacking job skills, we conclude that this may be by design. Outliers are also seen in the data: in re78 for example, the maximum value is more than 5 times the value of the third quartile. We make note of this, and decide to deal with it in our exploratory analysis.

Table 1: Summary Statistics

| Statistic | N | Mean | St. Dev. | Min | Pctl(25) | Pctl(75) | Max |
|---|---|---|---|---|---|---|---|
| age | 614 | 27.363 | 9.881 | 16 | 20 | 32 | 55 |
| educ | 614 | 10.269 | 2.628 | 0 | 9 | 12 | 18 |
| re74 | 614 | 4,557.547 | 6,477.964 | 0 | 0 | 7,888.5 | 35,040 |
| re75 | 614 | 2,184.938 | 3,295.679 | 0 | 0 | 3,249.0 | 25,142 |
| re78 | 614 | 6,792.834 | 7,470.731 | 0.000 | 238.283 | 10,893.590 | 60,307.930 |
| growth | 614 | 2,235.288 | 8,033.464 | −25,256.800 | −1,240.175 | 6,459.640 | 60,307.930 |