# Review on Causality

Your Name

## The data

This data was collected to evaluate the National Supported Work (NSW) Demonstration project in Lalonde (1986).

The NSWD evaluation employed a randomized experimental design, which is a key strength in causal inference. In a randomized controlled trial (RCT), participants are randomly assigned to either the treatment group (those receiving the intervention) or the control group (those not receiving the intervention). Randomization helps ensure that any observed differences in outcomes between the groups can be attributed to the intervention itself rather than pre-existing differences in participant characteristics.

In the case of the NSWD, eligible individuals were randomly assigned to either the program group (those receiving supported work) or the control group (those not receiving the program). This random assignment aimed to create comparable groups, allowing researchers to isolate the causal impact of the intervention.

Literature:

*Lalonde, R. (1986) Evaluating the Econometric Evaluations of Training Programs, American Economic Review, 76, 604-620.*

| Variable | Description |
|---|---|
| id | Matched pair id, 1, 1, 2, 2, . . . , 185, 185. |
| z | z=1 for treated, z=0 for control |
| age | Age in years |
| edu | Education in years |
| black | 1=black, 0=other |
| hisp | 1=Hispanic, 0=other |
| married | 1=married, 0=other |
| nodegree | 1=no High School degree, 0=other |
| re74 | Earnings in 1974, a covariate |
| re75 | Earnings in 1975, a covariate |
| re78 | Earnings in 1978, an outcome |

## 1. Descriptive statistics

First, load the data and ensure it is clean, if necessary. Rename variables and convert appropriate binary indicators into factors for analysis. Assess the balance between treatment and control groups for the variables `black`, `married`, `nondegree`, and `age`.

```
# load data

nsw <- read.csv(file="data/nsw.csv")
```

```r
# clean data
nsw$treat <- "Control"
nsw$treat[nsw$z==1] <- "Treated"


## proportion of black people
table(nsw$treat, nsw$black)
```

```
##
##              0   1
##   Control  27 158
##   Treated  29 156
```

```r
## proportion of college educated
sort(unique(nsw$edu))
```

```
##  [1]  3  4  5  6  7  8  9 10 11 12 13 14 15 16
```

```r
nsw$college <- "no"
nsw$college[nsw$edu>12] <- "yes"
table(nsw$treat, nsw$college)
```

```
##
##             no yes
##   Control 178   7
##   Treated 170  15
```

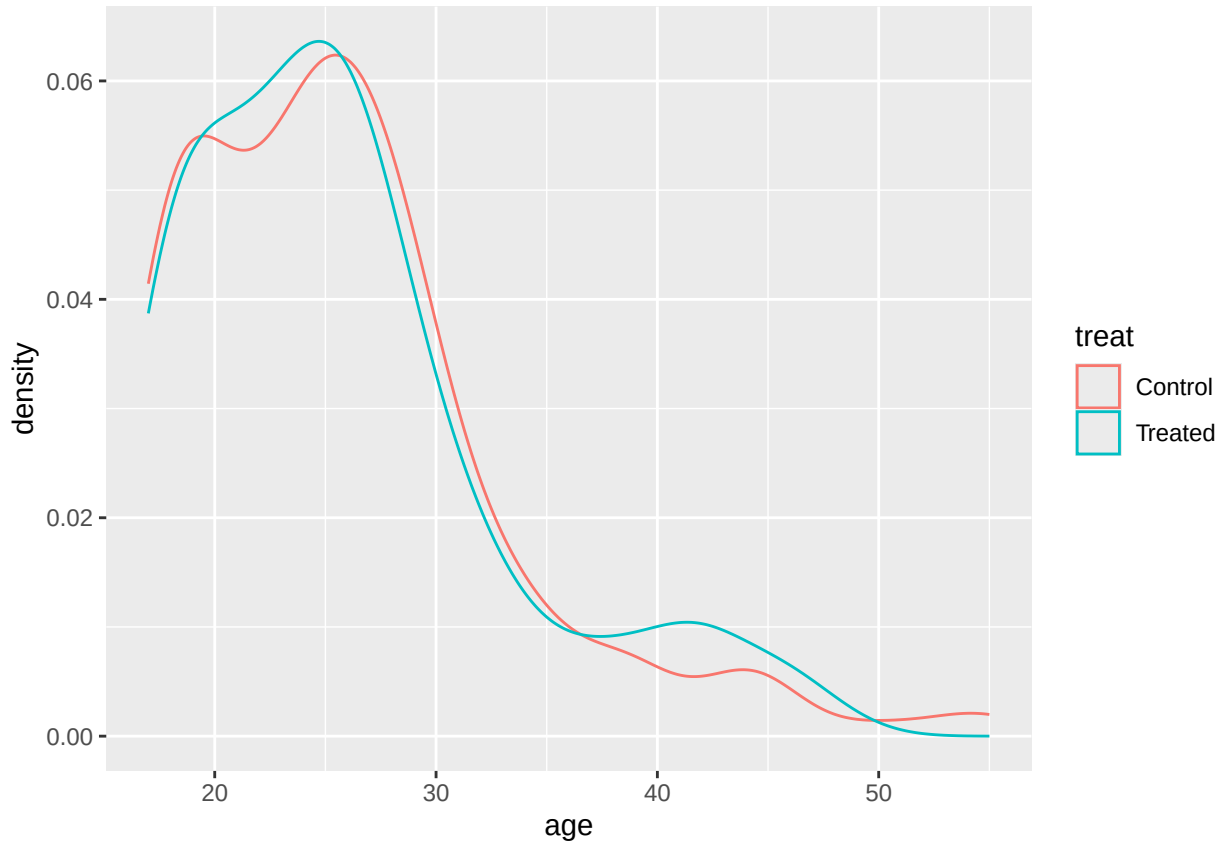
```r
## proportion of married
table(nsw$treat, nsw$married)
```

```
##
##              0   1
##   Control 148  37
##   Treated 150  35
```

```r
## balance on age
table(nsw$treat, nsw$age)
```

```
##
##           17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39
##   Control 14 14 14 10  4 12  7 12 17 10 11 13 10  4  6  3  2  4  2  1  0  2  4
##   Treated 13 13 12  9  9 11 12  9 18  9 16  6  9  3  7  1  5  0  3  0  2  3  0
##
##           40 41 42 43 44 45 46 48 50 54 55
##   Control  0  0  0  1  3  2  0  0  1  1  1
##   Treated  2  2  4  1  1  1  3  1  0  0  0
```

```r
ggplot(data=nsw, aes(x=age, group=treat))+
  geom_density(aes(color=treat))
```




Second, report the correlation matrix of all the quantiative variables using the function `cor()`.

```
round(
  cor(nsw[,-c(1,12:13)]), digits = 2
)
```

```
##              z   age   edu black  hisp married nodegree  re74  re75  re78
## z         1.00  0.01  0.04 -0.02  0.01   -0.01    -0.07  0.01  0.01  0.15
## age       0.01  1.00  0.00  0.08 -0.06    0.20    -0.09  0.01  0.05  0.07
## edu       0.04  0.00  1.00 -0.01 -0.09    0.07    -0.67  0.09  0.01  0.13
## black    -0.02  0.08 -0.01  1.00 -0.58    0.06     0.07 -0.03 -0.08 -0.14
## hisp      0.01 -0.06 -0.09 -0.58  1.00    0.00     0.04  0.01  0.09  0.08
## married  -0.01  0.20  0.07  0.06  0.00    1.00     0.00  0.18  0.28  0.04
## nodegree -0.07 -0.09 -0.67  0.07  0.04    0.00     1.00 -0.09  0.05 -0.12
## re74      0.01  0.01  0.09 -0.03  0.01    0.18    -0.09  1.00  0.70  0.09
## re75      0.01  0.05  0.01 -0.08  0.09    0.28     0.05  0.70  1.00  0.08
## re78      0.15  0.07  0.13 -0.14  0.08    0.04    -0.12  0.09  0.08  1.00
```

Third, generate a table displaying descriptive statistics for all variables except `id`. Utilize the `stargazer()` function, specifying only numerical/quantitative variables and setting the `type=` argument to `text` for console printing or LaTeXfor output in the knitted PDF.

```
# Note: to display LaTeX tables from stargazer in your knitted PDF you must set your code chunk with th

stargazer(nsw[,-1],
```

```
          header = FALSE,
          type = "latex") # change to type = "text" to display in the console
```

Table 2:

| Statistic | N | Mean | St. Dev. | Min | Max |
|---|---|---|---|---|---|
| z | 370 | 0.500 | 0.501 | 0 | 1 |
| age | 370 | 25.759 | 7.202 | 17 | 55 |
| edu | 370 | 10.270 | 1.865 | 3 | 16 |
| black | 370 | 0.849 | 0.359 | 0 | 1 |
| hisp | 370 | 0.057 | 0.232 | 0 | 1 |
| married | 370 | 0.195 | 0.396 | 0 | 1 |
| nodegree | 370 | 0.738 | 0.440 | 0 | 1 |
| re74 | 370 | 2,052.511 | 4,945.299 | 0.000 | 35,040.070 |
| re75 | 370 | 1,508.557 | 3,308.116 | 0.000 | 25,142.240 |
| re78 | 370 | 5,328.255 | 6,643.759 | 0.000 | 60,307.930 |

## 2. Exploratory data analysis

Create a **histogram** or **density plot** to compare age distribution of both the treatment and control groups, and display a vertical line showing the age **median** for the treatment and control group to visualize the balance of this distribution. If you will use ggplot, you will have to use the functions `geom_density` and `geom_vline`.

Moreover, create a scatter plot between the `re78` and `edu`, and make sure to differentiate with colors those observations from the treatment and control groups. In ggplot, you will need to use the `color=` aesthetic and `geom_point`.

```
## visualize the age distribution by treatment group


# scatter plot
```

## 3. Estimation of policy effects

We want to evaluate the treatment effect on the outcome. Before estimating this quantity, double check whether the outcome `re78` is correlated or associated with the unbalanced variables. Then, estimate the following conditional expectations:

$$E[\ re78 \mid z = treated] \quad - \quad E[\ re78 \mid z = control]$$

Can we interpret this estimator as a causal effect? Why?

Re-estimate the quantity using linear regression. Fit one model with only the treatment variable and another controlling for all non-balanced variables. Assign the model outputs to objects and use the `stargazer` function to display the results in a nicely formatted table.

In particular you must fit the following two models:

$$re78_i = \alpha + \beta_1 Z_i + e_i \quad (1)$$
$$re78_i = \alpha + \beta_1 Z_i + \beta_2 edu_i + \beta_3 nodegree_i + e_i \quad (2)$$

```
# remember to set the code chunk option:
```

Given the temporal dimension of this experiment, estimate a differences-in-differences estimator using `re75` to represent participants' earnings before the intervention and `re78` to represent earnings after the intervention. Under what assumptions this esitmator identifies a causal effect? Recall:

$$\text{DiD} = [\bar{Y}(1)_{after} - \bar{Y}(0)_{after}] - [\bar{Y}(1)_{before} - \bar{Y}(0)_{before}]$$

```
# differences in means before the intervention:

# differences in means after the intervention:

# differences in differences:
```

What is the **internal** and **external validity** of this experiment results?

## 4. Parallel trends

The assumption of parallel trends cannot be fully verified because both the treatment and control groups should exhibit parallel trends in their unobservable factors. However, providing visual evidence that parallel trends hold in the outcome during the years preceding the intervention is valuable.

Create a line plot showing the average earnings of the control and treatment groups for the periods `74`, `75`, and `78`. Begin by extracting the columns `re74`, `re75`, and `re78` for the treatment group and transforming them from a wide to long data structure. This process will result in three variables: one for treatment, one for earnings, and one for year indicators. Utilize the `pivot_longer` function for this task.

Once the data is in long format, employ the `aggregate` function to estimate the mean earnings by year and treatment group indicators. Use this aggregated data to plot the trends in earnings for both the control and treatment groups.

You can find information on how to use these functions in my **Module 2** slides.