Data task

Camilo Pedraza Jimenez

febrero 23, 2024 | 16:51:22 | CET

```
#Load used packages
library(pacman)
pacman::p_load(tidyverse, ggplot2, vtable, rempsyc, knitr)
```

Task 1

Load the Lalonde data set (see attachment) into either Stata or R. The data belongs to a study that looked at the effectiveness of a job training program (the treatment) on the real earnings of an individual, a couple years after completion of the program. It consists of a number of demographic variables as well as a treatment indicator and the real earnings in the year 1978 (the response).

```
lalonde <- read.csv("lalonde.csv")
```

Task 2

Produce a table with summary statistics (incl. mean, standard deviation) for the variables in the dataset.

```
lalonde %>%
  select(!nr) %>% #The variable nr includes only the number of the row, so it
      is not necessary for the summary statistics
  st(., simple.kable = T) #Function st() from the package vtable. simple.kable
       = T option to export the table to PDF format
```

Taking the first 5 variables as examples for the 445 observations in the dataset, we can see that:

- **Age:** The average age is 25 with a standard deviation of 7.1 years.
- **Education:** The average years of education is 10 with a standard deviation of 1.8 years.
- **Black:** As a dummy variable, since the mean is above 0.5 (0.83), it is possible to say that the majority of people on the dataset are black.
- **Hispanic:** As a dummy variable, since the mean is bellow 0.5 (0.088), it is possible to say that the majority of people on the dataset are not Hispanic.
- **Married:** As a dummy variable, since the mean is below 0.5 (0.17), it is possible to say that the majority of people on the dataset are not Married.

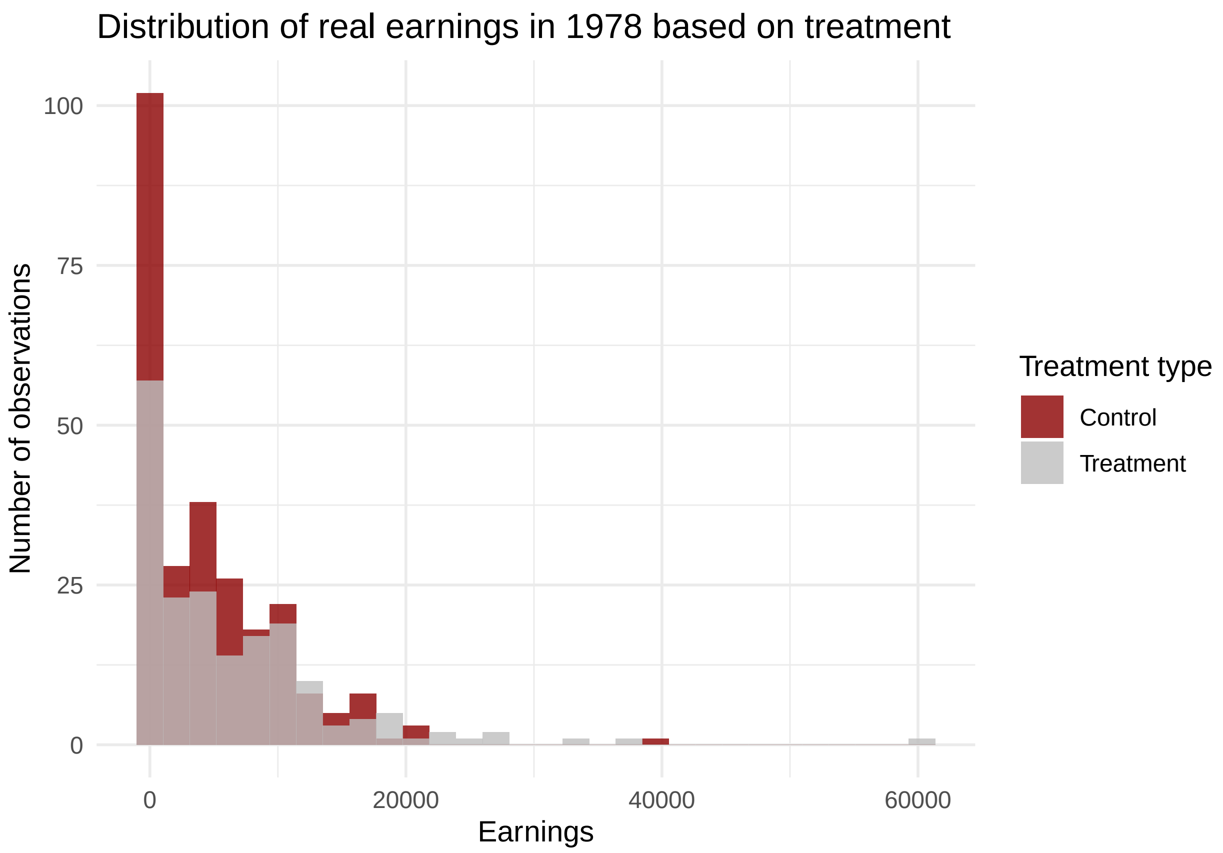
Table 1: Summary Statistics

Variable	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
age	445	25	7.1	17	20	28	55
educ	445	10	1.8	3	9	11	16
black	445	0.83	0.37	0	1	1	1
hisp	445	0.088	0.28	0	0	0	1
married	445	0.17	0.37	0	0	0	1
nodegr	445	0.78	0.41	0	1	1	1
re74	445	2102	5364	0	0	824	39571
re75	445	1377	3151	0	0	1221	25142
re78	445	5301	6631	0	0	8125	60308
u74	445	0.73	0.44	0	0	1	1
u75	445	0.65	0.48	0	0	1	1
treat	445	0.42	0.49	0	0	1	1

Task 3

Produce two histograms showing the distribution of real earnings 1978 (re78) for individuals in the treatment and in the control group separately.

```
ggplot(lalonde, aes(x = re78, fill = factor(treat))) +                      #
    Factor by treatment to get both histograms in one graph
  geom_histogram(position = "identity", alpha = 0.8, bins = 30) +           #
      alpha option to ajust the transparency
  scale_fill_manual(values = c("darkred", "grey"),
                    breaks = c(0, 1),
                    labels = c("Control", "Treatment")) +
  labs(title = "Distribution of real earnings in 1978 based on treatment",
       x = "Earnings",
       y = "Number of observations",
       fill = "Treatment type") +
  theme_minimal()
```



The distribution of the real earnings in 1978 is fairly even between the treated and the control group, since the distribution of the earnings for both groups stays mostly between 0 and 20000. Apart from the similarity of both distributions, it is possible to notice that the control group has more observation than the treatment group, since the higher frequencies can be found in the control group. It is possible to confirm that the control group is bigger by looking at the summary statistics, in which the mean of the treat variable is 0.42, which is lower than 0.5. Since the treat variable only takes the values 0 and 1, a mean lower than 0.5 would indicate that more control observations are found in the dataset. Finally, the highest earnings can be seen mostly for the treatment group, over 20000. The highest earnings for the control (40000) and the treatment (around 35000 to 60000) groups.

Task 4

Test whether the difference between treatment and control group is statistically significant at conventional levels. What do you find?

```
nice_t_test(lalonde,                #nice_t_test() function from the rempsyc
    package for a summary of the t-test
            response = "re78",
            group = "treat",
            warning = FALSE) %>%
  knitr::kable()                    #kable() function from the knitr package to
      get an organized table
```

```
## Using Welch t-test (base R's default; cf. https://doi.org/10.5334/irsp.82).
## For the Student t-test, use `var.equal = TRUE`.
##
```

Dependent Variable	t	df	p	d	CI_lower	CI_upper
re78	-2.674146	307.1325	0.007893	-0.2727154	-0.4619349	-0.0831909

Based on the p-value (0.007893), it is possible to argue that, with a level of confidence of 99%, we can reject the null hypothesis ($\mu_1 = \mu_2$ the mean of earnings for the control and treatment group are equal). This means that we have sufficient evidence to say that the mean of earning in 1978 between the two populations is different. Although with the histogram graphs, the distribution of earnings seemed similar, there could be two possible reasons that make the difference in the average earnings significant:

- The highest earnings of the treatment group.
- The high frequency of observations from the control group that have earnings around 0.