BT2101 GA2 Group 67 Submission

Lo Zhi Hao

2022-10-07

1 Econometric Analysis using R

Please use the jtrain2 and jtrain3 datasets from the Wooldridge package in R to answer this question. The dataset jtrain2 is the outcome of a job training experiment, where training status was randomlyassigned. The file jtrain3 contains observational data, where individuals largely determine if they wouldlike to participate in job training. These two datasets cover the same time period. Please carefully read the dataset documents before your answer these questions.

```
## Setting up the environment for further studies

## install.packages("wooldridge")
## install.packages("dplyr")
## install.packages("MASS")
## install.packages("ggplot2")

library(wooldridge)
library(dplyr)
library(MASS)
library(knitr)
library(corrplot)
library(ggplot2)

## documentation for MASS is at https://cran.r-project.org/web/packages/MASS/MASS.pdf

## Downloading the dataset

data('jtrain2')
data('jtrain3')
## ?jtrain2
## ?jtrain3
```

```
## for jtrain2
str(jtrain2)
```

```
## 'data.frame':    445 obs. of  19 variables:
##  $ train   : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ age     : int  37 22 30 27 33 22 23 32 22 33 ...
##  $ educ    : int  11 9 12 11 8 9 12 11 16 12 ...
##  $ black   : int  1 0 1 1 1 1 1 1 1 0 ...
##  $ hisp    : int  0 1 0 0 0 0 0 0 0 0 ...
##  $ married : int  1 0 0 0 0 0 0 0 0 1 ...
##  $ nodegree: int  1 1 0 1 1 1 0 1 0 0 ...
##  $ mosinex : int  13 13 13 13 13 13 6 6 14 13 ...
##  $ re74    : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ re75    : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ re78    : num  9.93 3.6 24.91 7.51 0.29 ...
##  $ unem74  : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ unem75  : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ unem78  : int  0 0 0 0 0 0 1 0 0 0 ...
##  $ lre74   : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ lre75   : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ lre78   : num  2.3 1.28 3.22 2.02 -1.24 ...
##  $ agesq   : int  1369 484 900 729 1089 484 529 1024 484 1089 ...
##  $ mostrn  : int  13 13 13 13 13 13 6 6 14 13 ...
##  - attr(*, "time.stamp")= chr "25 Jun 2011 23:03"
```

```
head(jtrain2)
```

```
##   train age educ black hisp married nodegree mosinex re74 re75     re78 unem74
## 1     1  37   11     1    0       1        1      13    0    0  9.93005      1
## 2     1  22    9     0    1       0        1      13    0    0  3.59589      1
## 3     1  30   12     1    0       0        0      13    0    0 24.90950      1
## 4     1  27   11     1    0       0        1      13    0    0  7.50615      1
## 5     1  33    8     1    0       0        1      13    0    0  0.28979      1
## 6     1  22    9     1    0       0        1      13    0    0  4.05649      1
##   unem75 unem78 lre74 lre75     lre78 agesq mostrn
## 1      1      0     0     0  2.295566  1369     13
## 2      1      0     0     0  1.279792   484     13
## 3      1      0     0     0  3.215249   900     13
## 4      1      0     0     0  2.015723   729     13
## 5      1      0     0     0 -1.238599  1089     13
## 6      1      0     0     0  1.400318   484     13
```

```
## for jtrain3
str(jtrain3)
```

```
## 'data.frame':    2675 obs. of  20 variables:
##  $ train  : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ age    : int  37 30 27 33 22 23 32 22 19 21 ...
##  $ educ   : int  11 12 11 8 9 12 11 16 9 13 ...
##  $ black  : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ hisp   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ married: int  1 0 0 0 0 0 0 0 0 0 ...
##  $ re74   : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ re75   : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ unem75 : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ unem74 : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ re78   : num  9.93 24.91 7.51 0.29 4.06 ...
##  $ agesq  : int  1369 900 729 1089 484 529 1024 484 361 441 ...
##  $ trre74 : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ trre75 : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ trun74 : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ trun75 : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ avgre  : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ travgre: num  0 0 0 0 0 0 0 0 0 0 ...
##  $ unem78 : int  0 0 0 0 0 1 0 0 0 0 ...
##  $ em78   : int  1 1 1 1 1 0 1 1 1 1 ...
##  - attr(*, "time.stamp")= chr "25 Jun 2011 23:03"
```

```
head(jtrain3)
```

```
##   train age educ black hisp married re74 re75 unem75 unem74     re78 agesq
## 1     1  37   11     1    0       1    0    0      1      1  9.93005  1369
## 2     1  30   12     1    0       0    0    0      1      1 24.90950   900
## 3     1  27   11     1    0       0    0    0      1      1  7.50615   729
## 4     1  33    8     1    0       0    0    0      1      1  0.28979  1089
## 5     1  22    9     1    0       0    0    0      1      1  4.05649   484
## 6     1  23   12     1    0       0    0    0      1      1  0.00000   529
##   trre74 trre75 trun74 trun75 avgre travgre unem78 em78
## 1      0      0      1      1     0       0      0    1
## 2      0      0      1      1     0       0      0    1
## 3      0      0      1      1     0       0      0    1
## 4      0      0      1      1     0       0      0    1
## 5      0      0      1      1     0       0      0    1
## 6      0      0      1      1     0       0      1    0
```

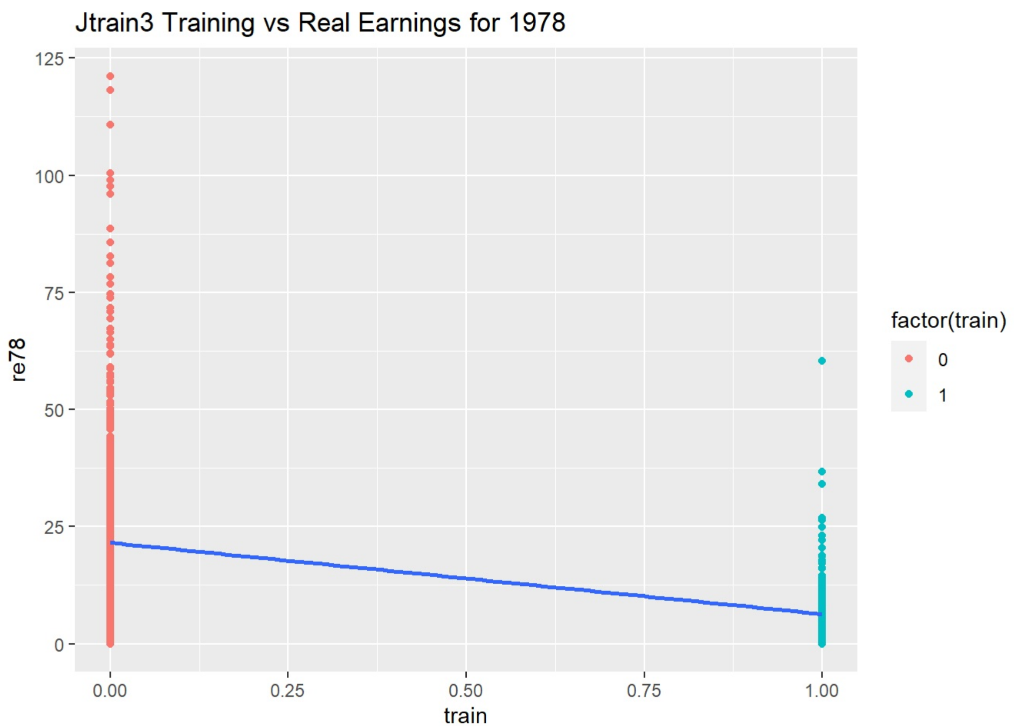
(a) Use jtrain2 and jtrain3 to plot the variable re78 (i.e., DV) against train (i.e., IV) and compare their distributions and slope (i.e., β) of the simple regression lines. (Hint: to visualize the regression line) Explain potential reasons for the different slopes.

```
## for jtrain2

ggplot(jtrain2, aes(x = train, y = re78)) +
    geom_point(aes(colour = factor(train))) +
  geom_smooth(method = "lm", se = FALSE) +
  labs(
    title = paste("Jtrain2 Training vs Real Earnings for 1978")
)
```

```
## `geom_smooth()` using formula 'y ~ x'
```


```
## for jtrain3

ggplot(jtrain3, aes(x = train, y = re78)) +
    geom_point(aes(colour = factor(train))) +
  geom_smooth(method = "lm", se = FALSE) +
    labs(
    title = paste("Jtrain3 Training vs Real Earnings for 1978")
)
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```
## searching for potential underlying problems

par(mfrow = c(1, 2))

## Key statistics for JTRAIN2

df1 <- data.frame(mean(jtrain2$re78), sd(jtrain2$re78))
colnames(df1) <- c("Mean for jtrain2 Recorded Earnings in 1978", "SD for jtrain2 Recorded Earnings in 1978")
kable(df1)
```

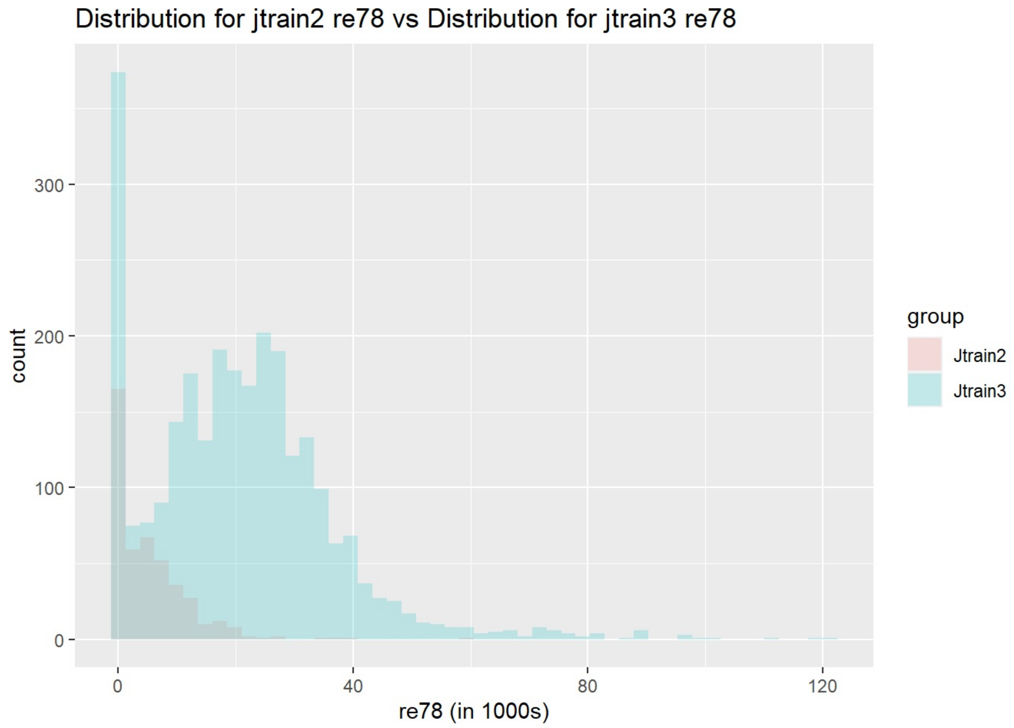
Mean for jtrain2 Recorded Earnings in 1978	SD for jtrain2 Recorded Earnings in 1978

5.300765	6.631493

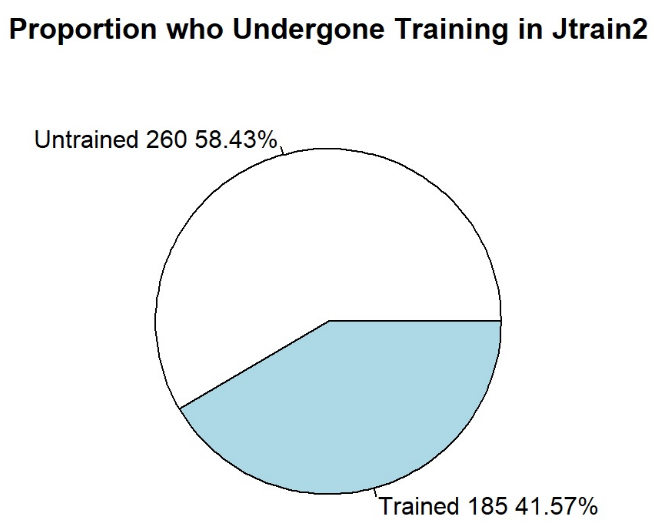
```
## Key statistics for JTRAIN3

df2 <- data.frame(mean(jtrain3$re78), sd(jtrain3$re78))
colnames(df2) <- c("Mean for jtrain3 Recorded Earnings in 1978", "SD for jtrain3 Recorded Earnings in 1978")
kable(df2)
```

Mean for jtrain3 Recorded Earnings in 1978	SD for jtrain3 Recorded Earnings in 1978
20.50238	15.63252

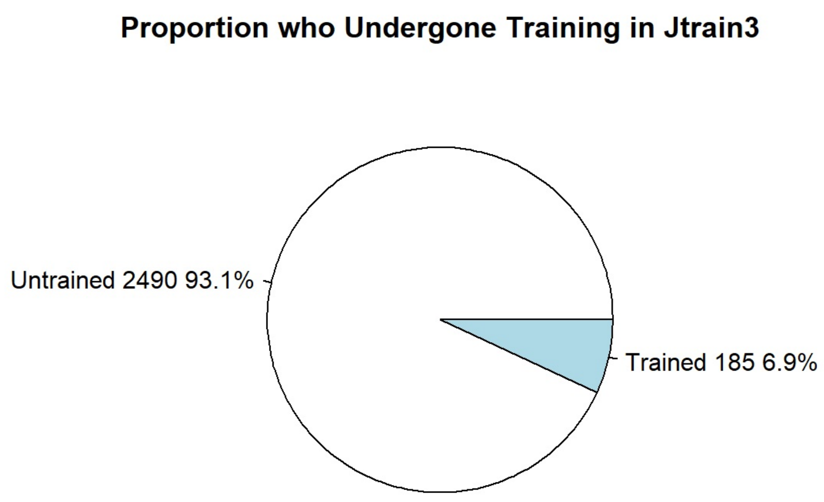
```
## plotting out the distribution

data1 <- data.frame(values = c(jtrain2$re78, jtrain3$re78),
                   group = c(rep("Jtrain2", nrow(jtrain2)),
                             rep("Jtrain3", nrow(jtrain3))))

ggplot(data1, aes(x = values, fill = group)) +
    geom_histogram(position = "identity", alpha = 0.2, bins = 50) +
    labs(
    title = paste("Distribution for jtrain2 re78 vs Distribution for jtrain3 re78"), x = "re78 (in 1000s)"
)
```




```
## Proportion of training and non-training individuals

## JTRAIN2
train.pop1 <- jtrain2 %>% count(train)
pie(train.pop1$n, labels = c("Untrained 260 58.43%", "Trained 185 41.57%"), main = "Proportion who Undergone Trai
ning in Jtrain2")
```




```
## JTRAIN3

train.pop2 <- jtrain3 %>% count(train)
pie(train.pop2$n, labels = c("Untrained 2490 93.1%", "Trained 185 6.9%"), main = "Proportion who Undergone Traini
ng in Jtrain3")
```



From the scatterplot that was plotted above, as well as the regression line that was plotted, we can notice a positive slope for Jtrain2 (slope = 1.7943) and a negative slope for Jtrain3 (slope = -15.2048).

By classifying the data according to whether the employee undergone training or not, we can also notice that the proportion of employees undergone training in Jtrain2 is significantly higher than in Jtrain3 (43.57% of Jtrain2 employees undergone training while only 6.9% of Jtrain3 employees undergone training).

Moreover, by plotting out the scatterplots and calculating the means for recorded earnings in each dataset, we can also notice that the mean earnings of Jtrain2 dataset is significantly lower than Jtrain3 (5.300765 (in thousands)) in Jtrain2 compared to 20.50238 (in thousands)) in Jtrain3.

This might be because of the sample selection for both datasets being different from each other. For jtrain2, it is the outcome from a job training experiment, and hence the targeted employees in the sample are low earners and targeted to receive a training. As such, it might not be an accurate representative of the entire population. Meanwhile, for jtrain3, it contains larger amount of data, and the proportion of men taken job training is also lower at 6.9%, which suggests it might represent a random sample from the population of men working in 1978.

(b) Using jtrain2, run a simple regression of re78 on train. Interpret your regression outputs.

```r
## running the regression

linear.model <- lm(re78 ~ train, jtrain2)
summary(linear.model)
```

```
##
## Call:
## lm(formula = re78 ~ train, data = jtrain2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.349 -4.555 -1.829  2.917 53.959
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.5548     0.4080  11.162  < 2e-16 ***
## train         1.7943     0.6329   2.835  0.00479 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.58 on 443 degrees of freedom
## Multiple R-squared:  0.01782,    Adjusted R-squared:  0.01561
## F-statistic: 8.039 on 1 and 443 DF,  p-value: 0.004788
```

```r
# linear.model2 <- lm(re78 ~ train, jtrain3)
# summary(linear.model2)
```

The relationship is as follows:

`re78` = 4.5548 + 1.7943 x `train`

Multiple R squared is 0.01782, and Adjusted R squared is 0.01561.

If an employee is not assigned to job training, he is expected to earn an average of 4.5548 x 1000 = $4554.8 in real earnings in 1978. If a person is in job training, the real earnings of him in 1978 is associated with a 1.7943 thousands (1794.3) increase (a nontrivial amount). Hence, after the employee had been assigned to job training, he is expected to earn an average of (4.5548 + 1.7943) x 1000 = $6349.1 in real earnings in 1978. The two coefficients are statistically significant, which suggests that we can confidently conclude that these variables are statistically different from 0.

(c) Using jtrain2, now adds variables re74, re75, educ, age, black, and hisp as control variables to the regression in question (b). Will the estimated effect of job training on re78 change much? Explain, that is, why or why not?

```r
## running new regression model

linear.model2 <- lm(re78 ~ train + re74 + re75 + educ + age + black + hisp, jtrain2)
summary(linear.model2)
```

```
##
## Call:
## lm(formula = re78 ~ train + re74 + re75 + educ + age + black +
##     hisp, data = jtrain2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.890 -4.424 -1.661  3.012 54.113
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.67407    2.42272   0.278  0.78097
## train        1.68005    0.63086   2.663  0.00803 **
## re74         0.08331    0.07653   1.089  0.27694
## re75         0.04677    0.13068   0.358  0.72062
## educ         0.40360    0.17485   2.308  0.02145 *
## age          0.05435    0.04382   1.240  0.21560
## black       -2.18007    1.15550  -1.887  0.05987 .
## hisp         0.14356    1.54092   0.093  0.92582
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.499 on 437 degrees of freedom
## Multiple R-squared:  0.05476,    Adjusted R-squared:  0.03962
## F-statistic: 3.617 on 7 and 437 DF,  p-value: 0.0008396
```

After adding in all the control variables (re74, re75, educ, age, black, hist), the coefficient of train in the linear model reduces from 1.79 to 1.68. This is not a huge change from section (b) as estimated.

This is because in order to be a randomised controlled experiment, the training programme has to be assigned randomly to employees without taking into account any of the other variables. As such, they should be roughly uncorrelated to the other independent variables. Thus, the estimated effect of adding other control variables into the regression is expected to be small, and it is proven by the results of the model shown.

(d) Using jtrain3, following the same logic of comparison between univariate and multivariate linear regression (i.e., question (b) & (c)), will a multivariate regression give different results? Explain, that is, why or why not?

```
# running same regression for jtrain3
par(mfrow = c(1, 2))

linear.model3 <- lm(re78 ~ train, jtrain3)
summary(linear.model3)
```

```
##
## Call:
## lm(formula = re78 ~ train, data = jtrain3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -21.554  -9.732  -0.866   7.705  99.620
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  21.5539     0.3036   70.98   <2e-16 ***
## train       -15.2048     1.1546  -13.17   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.15 on 2673 degrees of freedom
## Multiple R-squared:  0.06092,    Adjusted R-squared:  0.06057
## F-statistic: 173.4 on 1 and 2673 DF,  p-value: < 2.2e-16
```

```
linear.model4 <- lm(re78 ~ train + re74 + re75 + educ + age + black + hisp, jtrain3)
summary(linear.model4)
```

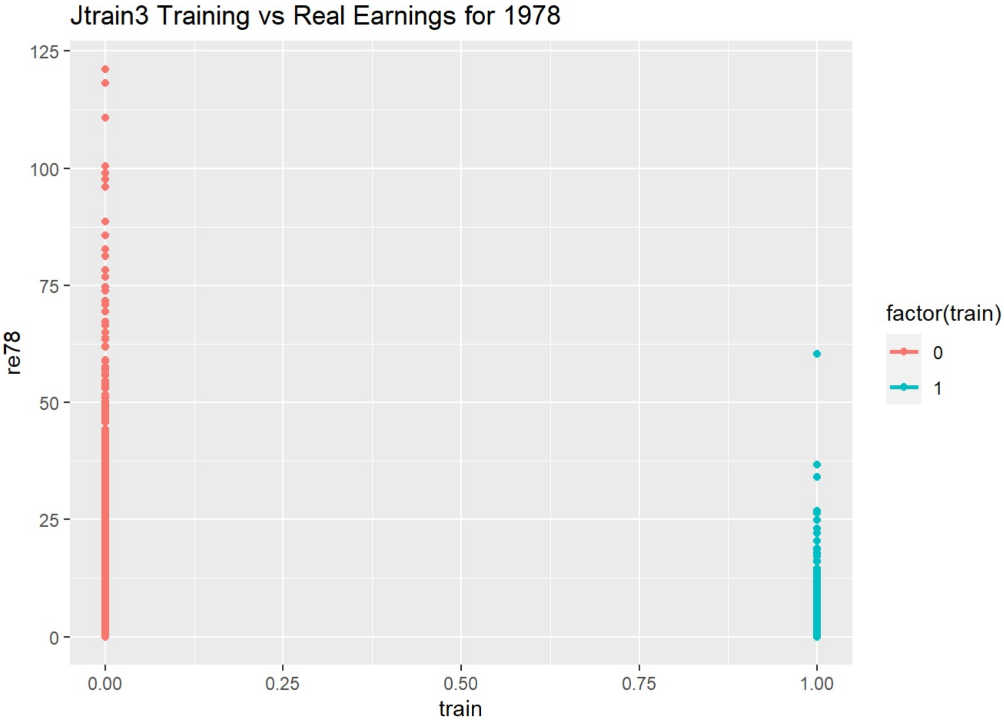
```
##
## Call:
## lm(formula = re78 ~ train + re74 + re75 + educ + age + black +
##     hisp, data = jtrain3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -65.246  -4.355  -0.465   3.770 110.950
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.64755    1.30093   1.266 0.205465
## train        0.21323    0.85339   0.250 0.802716
## re74         0.28098    0.02790  10.071  < 2e-16 ***
## re75         0.56929    0.02757  20.648  < 2e-16 ***
## educ         0.52006    0.07522   6.914 5.89e-12 ***
## age         -0.07507    0.02047  -3.667 0.000251 ***
## black       -0.64771    0.49193  -1.317 0.188056
## hisp         2.20261    1.09279   2.016 0.043944 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.08 on 2667 degrees of freedom
## Multiple R-squared:  0.5856, Adjusted R-squared:  0.5845
## F-statistic: 538.4 on 7 and 2667 DF,  p-value: < 2.2e-16
```

```
## Proving and understanding the underlying concepts behind the coefficients

jtrain3.trained <- jtrain3 %>% filter(train == 1)
jtrain3.non_trained <- jtrain3 %>% filter(train == 0)

df3 <- data.frame(mean(jtrain3.trained$re78), mean(jtrain3.non_trained$re78))
colnames(df3) <- c("Mean for jtrain3 (Trained) Recorded Earnings in 1978", "Mean for jtrain3 (Non-Trained) Record
ed Earnings in 1978")
kable(df3)
```

Mean for jtrain3 (Trained) Recorded Earnings in 1978	Mean for jtrain3 (Non-Trained) Recorded Earnings in 1978
6.349145	21.55392

```
ggplot(jtrain3, aes(x = train, y = re78, colour = factor(train))) +
    geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
    labs(
    title = paste("Jtrain3 Training vs Real Earnings for 1978")
)
```

```
## `geom_smooth()` using formula 'y ~ x'
```



When we run the univariated linear regression model, if an employee is not assigned to job training, he is expected to earn an average of 21.5539 x 1000 = $21, 554 in real earnings in 1978. After the employee had been assigned to job training, he is expected to earn an average of (21.5539 - 15.2048) x 1000 = $6349.1 in real earnings in 1978. The two coefficients are statistically significant, which suggests that we can confidently conclude that these variables are statistically different from 0.

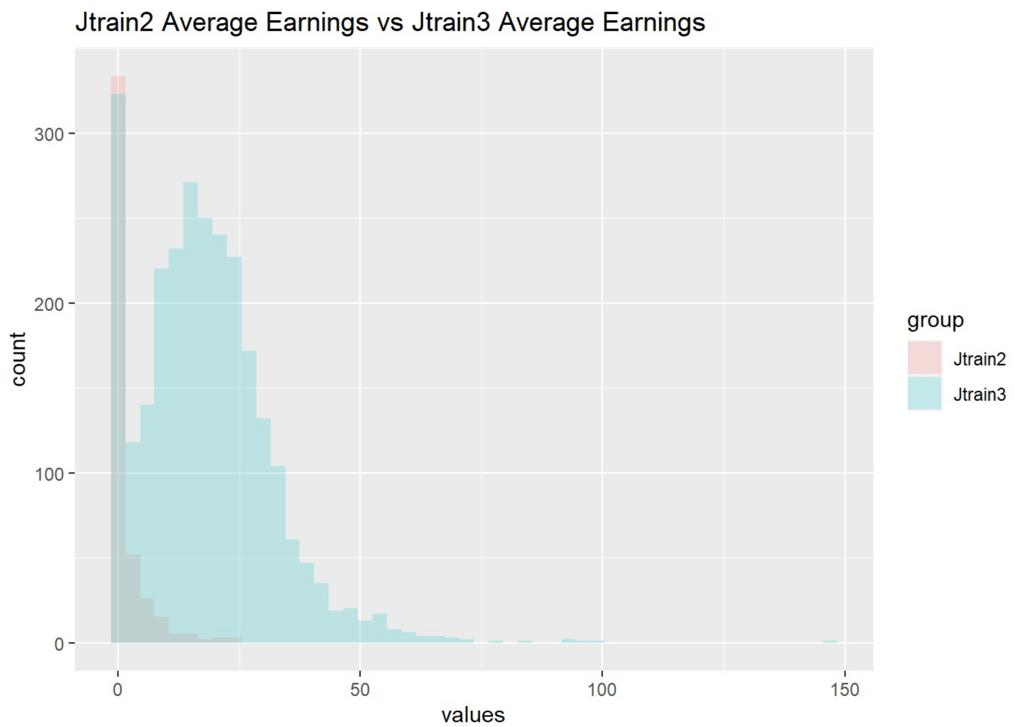
Something to take note about is that the coefficient of job training is -15.2048, which is a huge negative coefficient. This is hard to believe, as it is counter intuition for employees who undergone job training to have decreased wages despite the increase in skills and abilities they gain from the training. Further analysis of the sample suggests that those who undergone job training are from a lower income group. This suggests that the job training for employees in this dataset is most likely not going to be randomly selected, but instead is choosen to join the job training on purpose.

When we run the multivariated linear regression, if an employee is not assigned to job training, as well as holding other independent variables at 0 (0 for re74, re75, educ, age, black and hist), he is expected to earn an average of 1.64755 x 1000 = $1,647.55 in real earnings in 1978. After the employee had been assigned to job training, he is expected to earn an average of (1.64755 + 0.21323) x 1000 = $1,860.78 in real earnings in 1978. The two coefficients have large p-values, which suggests that we have insufficient statistical evidence to suggests that the coefficients are significantly different from 0. As such, the effect of training on the overall recorded earnings on 1978 is small, positive and statistically insignificant.

Compared to the original univariated linear regression, the coefficient for train in the multivariated linear regression is more believable, as the small and insignificant statistics (0.213) is a more convincing representation of the effect on recorded earnings, as training provides employees with more value adding skills but will not change the recorded earnings significantly in a short period of time.

(e) Define avgRe = (re74 + re75) / 2. Create a graph to compare the distribution of avgRe across jtrain2 and jtrain3. Do you agree that these data sets can represent the same populations in 1978? Explain. In addition, using at least two statistics except average, present your intuitive evidence (i.e., not a two-sample t-test) and analysis regarding this representativity issue.

```
# mutating and creating a new column

jtrain2 <- jtrain2 %>% mutate(avgRe = (re74 + re75) / 2)
# str(jtrain2)

jtrain3 <- jtrain3 %>% mutate(avgRe = (re74 + re75) / 2)
# str(jtrain3)

## Creating the graph

data <- data.frame(values = c(jtrain2$avgRe, jtrain3$avgRe),
                   group = c(rep("Jtrain2", nrow(jtrain2)),
                             rep("Jtrain3", nrow(jtrain3))))

ggplot(data, aes(x = values, fill = group)) +
    geom_histogram(position = "identity", alpha = 0.2, bins = 50) +
    labs(
    title = paste("Jtrain2 Average Earnings vs Jtrain3 Average Earnings")
)
```




```
## Two statistics to represent the intuitive evidence

## 1ST statistics - interquartile range

summary(jtrain2$avgRe)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.000   0.000   1.740   1.492  24.376
```

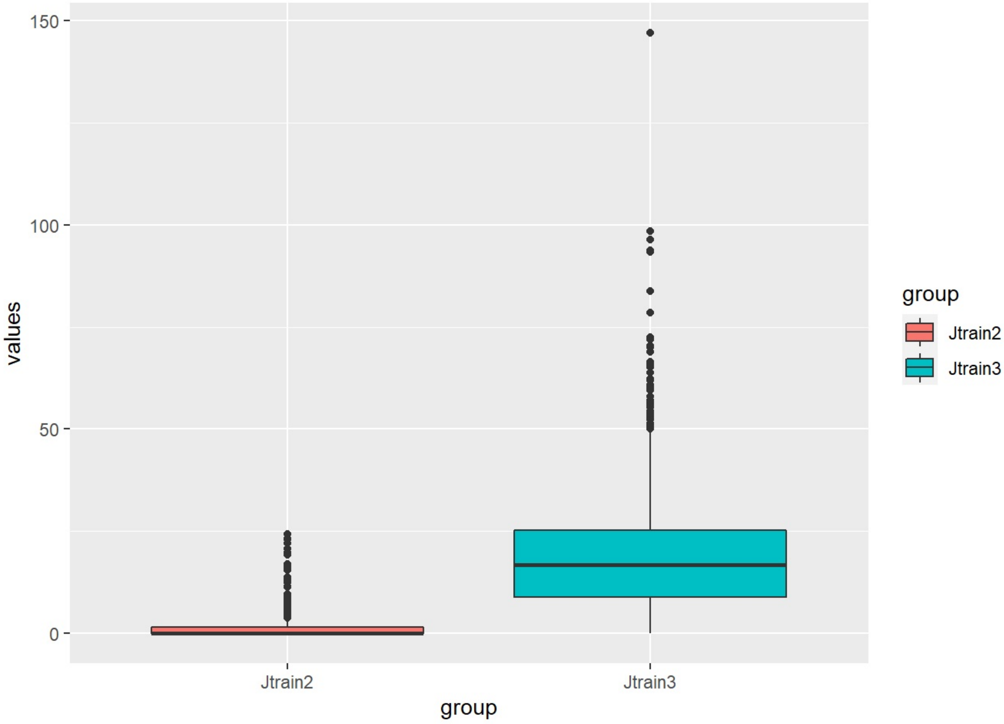
```
summary(jtrain3$avgRe)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   8.829  16.873  18.040  25.257 146.901
```

```
IQR_jtrain2 <- "[0, 1.492]"
IQR_jtrain3 <- "[8.829, 25.257]"

IQR <- data.frame(IQR_jtrain2, IQR_jtrain3)
kable(IQR)
```

IQR_jtrain2	IQR_jtrain3
[0, 1.492]	[8.829, 25.257]

```
ggplot(data, aes(x = group, y = values, fill = group)) + geom_boxplot()
```




```
## 2nd statistics - median and sd of both datasets

jtrain2.median <- median(jtrain2$avgRe)
jtrain2.sd <- sd(jtrain2$avgRe)

jtrain3.median <- median(jtrain3$avgRe)
jtrain3.sd <- sd(jtrain3$avgRe)

df7 <- data.frame(jtrain2.median, jtrain2.sd, jtrain3.median, jtrain3.sd)
kable(df7)
```

jtrain2.median	jtrain2.sd	jtrain3.median	jtrain3.sd
0	3.900095	16.87315	13.29345

No, we do not agree that the two datasets represent the same population. This is mainly because of the major difference in distribution of data in the two datasets, as shown visually in the histogram graphed above. This is further proven by looking into the documentation for both jtrain2 and jtrain3, where jtrain2 is the outcome of a job training experiment where participants are low earners who are targeted to get training, while jtrain3 is the outcome of a random sample from the population of men working in 1978.

In addition, by calculating the interquartile range for both jtrain2 and jtrain3 and creating a boxplot to represent the range in a visual way, we can see that the interquartile range for both datasets are significantly different from each other, with jtrain2 having an interquartile range way below that of jtrain3. This provides further evidence that the two datasets are not from the same population, as if the two datasets are from a same distribution, the majority of data for real earnings (data between the 25th and 75th percentile of the distribution) will not be as different as shown in the interquartile range and boxplot analysis.

Similarly, by calculating the median and standard deviation for both jtrain2 and jtrain3, we realise that there is a significant difference between the median and standard deviation of both datasets. For example, the median and sd for jtrain2 are 0 and 3.900095 while median and sd for jtrain3 are 16.87315 and 13.29345. As such, this provides further evidence that the two datasets aer not from the same population due to the large difference in median and standard deviation values.

(f) Almost 96% of men in the data set jtrain2 have avgRe less than $10,000. Using only these men, run the regression below:

re78 on train, re74, re75, educ, age, black, hisp (1)
Please report the coefficient of train and its t-statistic. Is there any difference between the regression result of the full sample and that of this subsample of 96 % men? How to justify this difference, if any?

```
## filter out men with avgRe less than $10, 000

jtrain2.filtered <- jtrain2 %>% filter(avgRe <= 10)

linear.model5 <- lm(re78 ~ train + re74 + re75 + educ + age + black + hisp, jtrain2.filtered)
summary(linear.model5)
```

```
##
## Call:
## lm(formula = re78 ~ train + re74 + re75 + educ + age + black +
##     hisp, data = jtrain2.filtered)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -8.217 -4.349 -1.750  3.044 53.977
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.73691    2.44601   0.710   0.4780
## train        1.58303    0.63245   2.503   0.0127 *
## re74        -0.11676    0.12378  -0.943   0.3461
## re75         0.17321    0.18879   0.917   0.3594
## educ         0.35821    0.17591   2.036   0.0423 *
## age          0.04400    0.04388   1.003   0.3166
## black       -2.38353    1.16779  -2.041   0.0419 *
## hisp        -0.36940    1.55105  -0.238   0.8119
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.377 on 419 degrees of freedom
## Multiple R-squared:  0.04627,    Adjusted R-squared:  0.03033
## F-statistic: 2.904 on 7 and 419 DF,  p-value: 0.005643
```

```
df4 <- data.frame(nrow(jtrain2), nrow(jtrain2.filtered))
kable(df4)
```

nrow.jtrain2.	nrow.jtrain2.filtered.
445	427

The relationship is as follows:

`re78` = 1.73691 + 1.58303 x `train` - 0.11676 x `re74` + 0.17321 x `re75` + 0.35821 x `educ` + 0.044 x `age` - 2.38353 x `black` - 0.3694 x `hisp`

For `train` in the model, the coefficient of `train` is 1.58303 and the t-statistics for `train` is 2.503.

There is a difference between the coefficient (1.58303 compared to 1.68005 in original model) and t-statistics (2.503 compared to 2.663 in original model). This difference is expected, as we are dismissing the higher earners in the new jtrain2 model.

As such, the lower coefficient for train is expected, as we have fewer observations in the new model, and the remaining observations are all slightly lower earners as the higher earners are filtered out.

Meanwhile, despite the standard error remains similar (0.63245 compared to 0.63086 in original model), the t-statistics for train still decreases. This is due to the reduced number of extreme observations recorded in the data, as we removed the higher earners out from our dataset. As such, the remaining values will be closer to the mean of the real earnings, and according to the formula for t-statistics, t-statistics will be slightly lower than the original model.

(g) Run the same regression above for both jtrain3 and jtrain2, also using only men with avgRe <= 10. Regarding the sub-sample regressions, will the coefficients of train be different across these two datasets? How to justify this difference, if any?

```
## filter out men with avgRe less than $10, 000

## running on jtrain2

jtrain2.filtered <- jtrain2 %>% filter(avgRe <= 10)

linear.model6 <- lm(re78 ~ train + re74 + re75 + educ + age + black + hisp, jtrain2.filtered)
summary(linear.model6)
```

```
##
## Call:
## lm(formula = re78 ~ train + re74 + re75 + educ + age + black +
##     hisp, data = jtrain2.filtered)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -8.217 -4.349 -1.750  3.044 53.977
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.73691    2.44601   0.710   0.4780
## train        1.58303    0.63245   2.503   0.0127 *
## re74        -0.11676    0.12378  -0.943   0.3461
## re75         0.17321    0.18879   0.917   0.3594
## educ         0.35821    0.17591   2.036   0.0423 *
## age          0.04400    0.04388   1.003   0.3166
## black       -2.38353    1.16779  -2.041   0.0419 *
## hisp        -0.36940    1.55105  -0.238   0.8119
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.377 on 419 degrees of freedom
## Multiple R-squared:  0.04627,    Adjusted R-squared:  0.03033
## F-statistic: 2.904 on 7 and 419 DF,  p-value: 0.005643
```

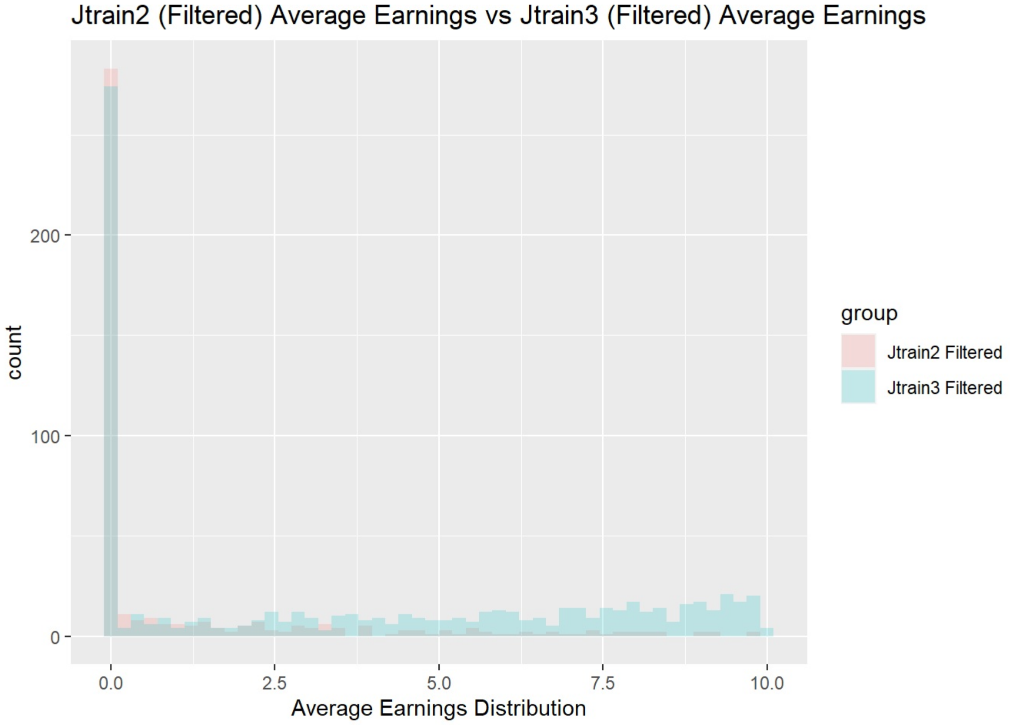
```
## running on jtrain3

jtrain3.filtered <- jtrain3 %>% filter(avgRe <= 10)

linear.model7 <- lm(re78 ~ train + re74 + re75 + educ + age + black + hisp, jtrain3.filtered)
summary(linear.model7)
```

```
##
## Call:
## lm(formula = re78 ~ train + re74 + re75 + educ + age + black +
##     hisp, data = jtrain3.filtered)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -18.673  -4.387  -1.751   2.804  60.545
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.44801    2.14136   1.610  0.10777
## train        1.84445    0.89311   2.065  0.03924 *
## re74         0.31311    0.06919   4.525    7e-06 ***
## re75         0.77435    0.07557  10.247  < 2e-16 ***
## educ         0.32831    0.11034   2.975  0.00302 **
## age         -0.08315    0.03068  -2.710  0.00688 **
## black       -1.97331    0.72072  -2.738  0.00633 **
## hisp        -1.10072    1.43184  -0.769  0.44228
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.962 on 757 degrees of freedom
## Multiple R-squared:  0.2344, Adjusted R-squared:  0.2273
## F-statistic: 33.11 on 7 and 757 DF,  p-value: < 2.2e-16
```

```
df5 <- data.frame(nrow(jtrain3), nrow(jtrain3.filtered))
kable(df5)
```

nrow.jtrain3.	nrow.jtrain3.filtered.
2675	765

```r
## showing distribution of average earnings for jtrain2 and jtrain3 after filtering

data <- data.frame(values = c(jtrain2.filtered$avgRe, jtrain3.filtered$avgRe),
                   group = c(rep("Jtrain2 Filtered", nrow(jtrain2.filtered)),
                             rep("Jtrain3 Filtered", nrow(jtrain3.filtered))))

ggplot(data, aes(x = values, fill = group)) +
    geom_histogram(position = "identity", alpha = 0.2, bins = 50) +
    labs(
    title = paste("Jtrain2 (Filtered) Average Earnings vs Jtrain3 (Filtered) Average Earnings")
) + xlab("Average Earnings Distribution")
```



For Jtrain2, the relationship is as follows:

`re78` = 1.73691 + 1.58303 x `train` - 0.11676 x `re74` + 0.17321 x `re75` + 0.35821 x `educ` + 0.044 x `age` - 2.38353 x `black` - 0.3694 x `hisp`

For Jtrain3, the relationship is as follows:

`re78` = 3.44801 + 1.84445 x `train` - 0.31311 x `re74` + 0.77435 x `re75` + 0.32831 x `educ` - 0.08315 x `age` - 1.97331 x `black` - 1.10072 x `hisp`

Yes, the subsample regressions will be different across these two datasets. This is because the underlying population for these two datasets are different, with jtrain2 being more likely to be a selected group of low earning and less skillful employees selected for job training while jtrain3 more likely to be a random selection from the overall job market, as shown by the histogram visualisation above. As such, the underlying population of the two datasets are different, and it justifies why the subsample regressions achieve significantly different results from each other.

As shown from the plot, despite both jtrain2 and jtrain3 average earnings are filtered to be below /$10,000, the distribution of average earnings for jtrain3 is still higher than that of jtrain2. This suggests a higher level of average earnings is received by data in jtrain3, and it reflects on the regression model as jtrain3 achieve higher coefficients for both earnings before training and after training. This suggests that the difference in underlying distribution is the main reason why the two regression models achieve different outcomes.