CS112 Assignment 2
Viet Hoang Tran Duong
CS112 Spring 2019

Problem 1:

Link to code

a. **Your original data-generating equation**

The independent variable is the year of working experience. The dependent variable is the amount of debt each person is having, in thousand dollars (1000$). The positive amount of debt means they are in debt, while the negative amounts of debt mean they are not in debt and represents their total assets.

The outlier, in this case, is Elon Musk with 27 years of experience and have assets of 21.5 billion USD: points(27, -21500)

The original data generating equation:

```
################### Start R ###################
set.seed(23022019)
independent <- runif(999, 0, 30)
dependent <- c()
for (i in independent) {
  value = 0.5*i + rnorm(1, mean = 0, sd = 30) - 1
  dependent <- c(dependent, value)
}

independent_extra <- c(independent, 27)
dependent_extra <- c(dependent, -21500)
################### End R ###################
```

b. **Regression results for the original 999 (copy/paste the "summary" output)**

```
################### Start R ###################
lm(formula = dependent ~ independent)

Residuals:
    Min      1Q  Median      3Q     Max
-82.731 -21.058  -1.031  20.077 104.200

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.2851     1.9066   0.150 0.881180
independent   0.4178     0.1128   3.705 0.000223 ***
```

```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 30.11 on 997 degrees of freedom
Multiple R-squared:  0.01358,	Adjusted R-squared:  0.01259
F-statistic: 13.72 on 1 and 997 DF,  p-value: 0.0002233
################### End R ###################
```

c. **Regression results with the outlier included (copy/paste “summary” output)**

```
################### Start R ###################lm(formula = dependent_extra ~ 
independent_extra)

Residuals:
    Min      1Q  Median      3Q     Max 
-21444.1   -10.6    22.2    50.3   182.3 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)         33.279     43.085   0.772    0.440
independent_extra   -3.301      2.547  -1.296    0.195

Residual standard error: 680.5 on 998 degrees of freedom
Multiple R-squared:  0.001681,	Adjusted R-squared:  0.0006804 
F-statistic:  1.68 on 1 and 998 DF,  p-value: 0.1952
################### End R ###################
```

d. **A properly-labeled data visualization that shows the regression line based on the original 999 points, and another differentiated regression line (on the same axes) based on 1000 points**.

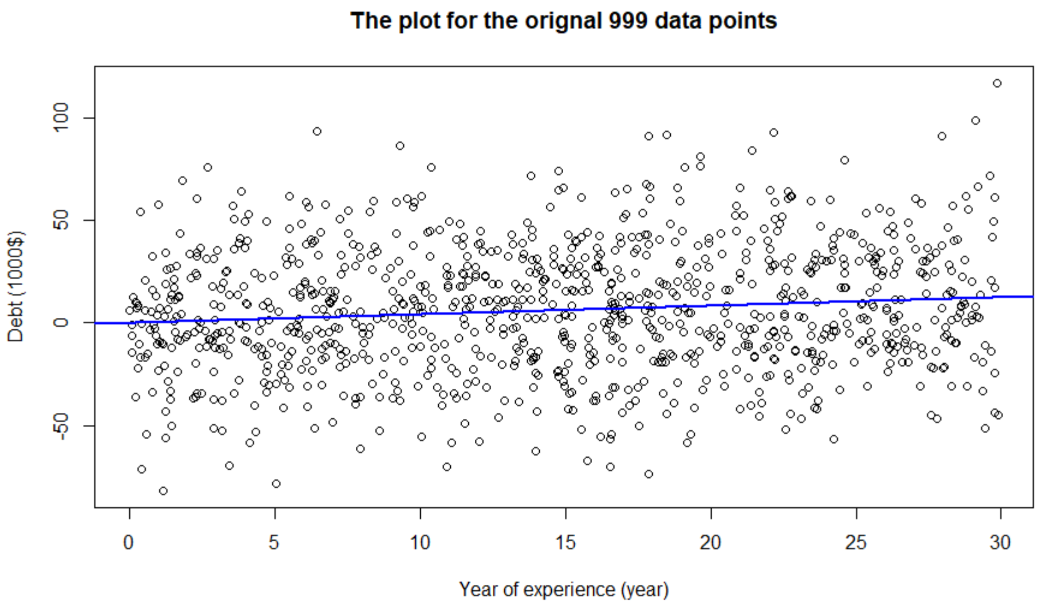

Figure 1.1: The Plot for the original 999 data points. The blue line is the line of best fit for the given data.

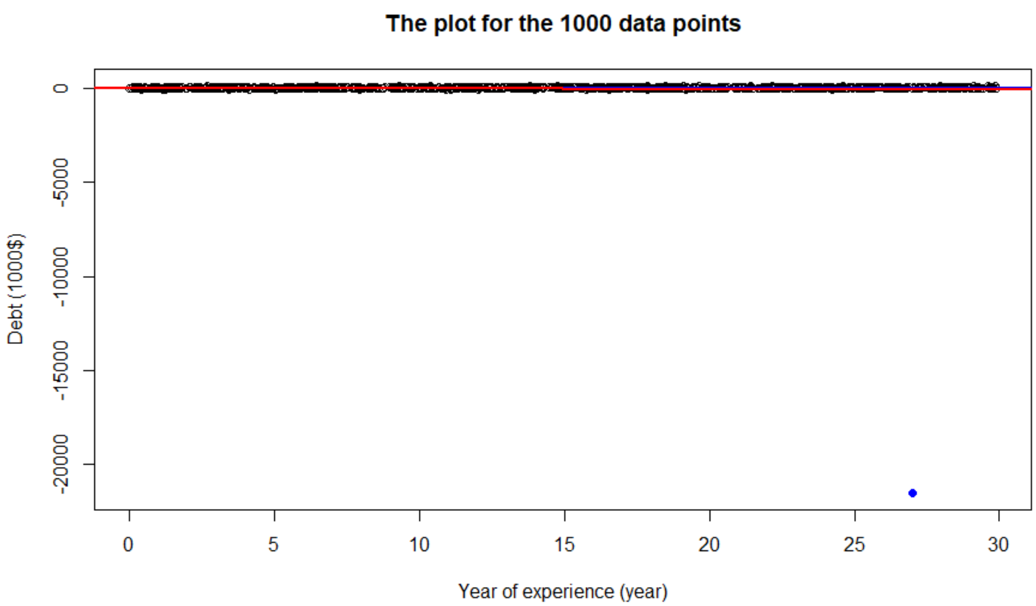

Figure 1.2: The Plot for the 1000 data points (999 original data points and an outlier). The blue line represents the line of best fit for the original 999 data points. The red line represents the line of best fit for the 1000 data points. The blue point is the outlier.

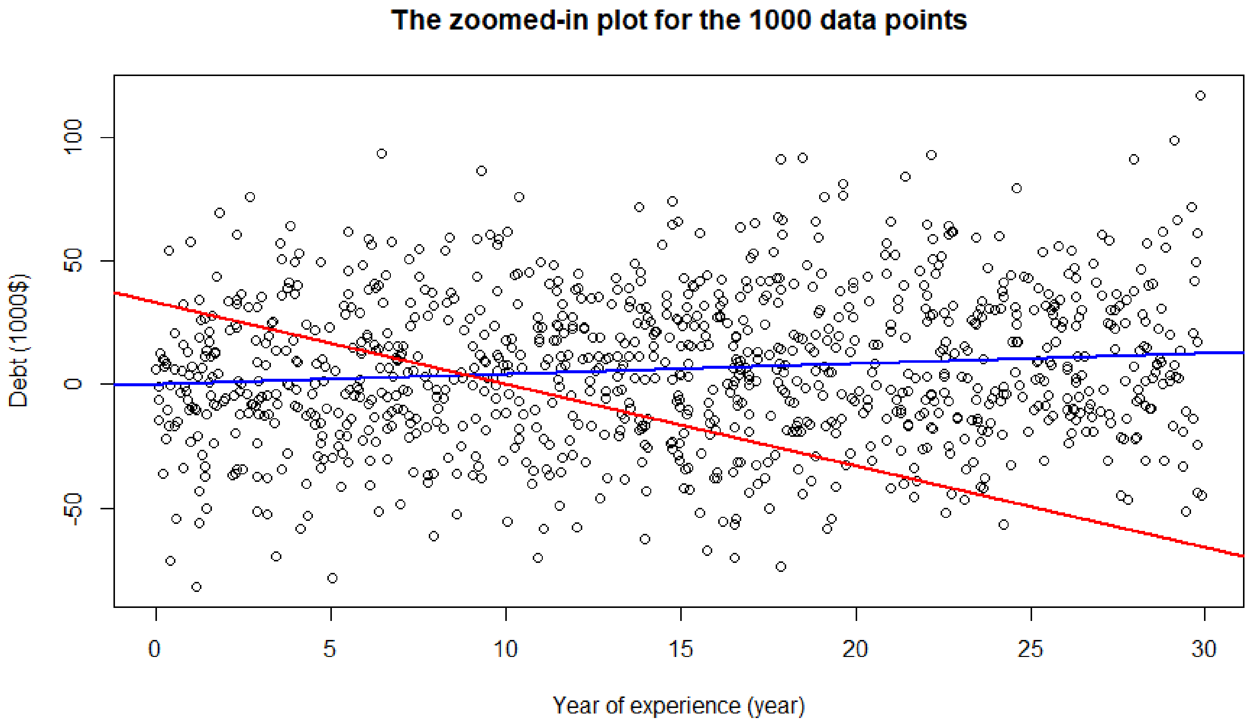

Figure 1.3: The zoomed-in plot for the 1000 data points (999 original data points and an outlier). It is zoomed-in by neglecting the extrapolation point (27, -21500) to see the before (999 points) and after (1000 points) trend. The blue line represents the line of best fit for the original 999 data points. The red line represents the line of best fit for the 1000 data points

e. **No more than 3 sentences that would serve as a caption for your figure if it were to be included in an econometrics textbook to illustrate the dangers of extrapolation.**

<u>Description</u>: There is an outlier in the lower right corner which is a high-level point (27, -21500) and it pulls the regression line around: from originally positive slope with 999 data points to a negative slope when we include the outlier.
This outlier pulls the regression line around, making the regression line fails to reflect the actual state of the data (positive slope when fitting the 999 data points - the majority of the data).
We can handle such situation by ignoring the outlier (which might lead to missing out on the critical element or relationship), include it (and potentially bias the regression line), or the best approach is to collect more data, especially data in the neighborhood of the outlier, and also the data about the outlier itself.

<u>Addition: Context description for the given variable and scenario above</u>: Consider the original 999 points as ordinary US citizens: we can infer that the population has more debt over time, approximately 0.4178 (thousand dollars) more in debt per year. However, with the appearance of the outlier (Elon Musk) and his enormous assets (27 year of experience and -21.5 billion USD in debt), the trend was reversed from positive slope to negative slope, implying overall, the populations are experiencing 3.301 (thousand dollars) less in debt per year. Because of the extrapolation, the linear model fails to reflect the actual state of the economy: people are getting less debt in contrast to the fact that the majority of people are getting more debt, which leads to giving a false hope that people's lives are improving, and the economy is getting better.

Problem 2:

[Link to code]

	educ	re74	re75
Median	10	0	0
The 75% quartile	11	139.4247	650.0963

Table 2.1: The table represents the median and the 75% quantile of the variables that we would hold as constant.

- **the 95% interval of expected values for re78, for every unit (i.e., each age 17-55, spanning the age range in the data set).**

	17	18	19	20	21	22	23	24	25	26	27	28	29
lower 95% quantile	3078.389	3173.783	3242.595	3278.062	3268.087	3237.138	3182.118	3115.756	3032.031	2945.722	2860.035	2791.894	2727.854
median	4491.669	4396.981	4304.024	4215.736	4137.946	4066.754	4013.562	3963.405	3920.226	3885.437	3861.353	3842.450	3831.742
upper 95% quantile	5868.670	5588.441	5333.686	5137.182	4986.386	4891.372	4850.935	4839.871	4834.658	4844.317	4868.970	4902.398	4946.549

30	31	32	33	34	35	36	37	38	39	40	41	42
2672.854	2642.184	2613.603	2591.433	2584.267	2579.096	2587.632	2576.784	2567.684	2535.282	2518.127	2483.915	2454.503
3828.701	3834.497	3846.516	3869.468	3904.537	3941.026	3984.080	4038.908	4100.947	4168.369	4244.148	4332.294	4429.953
4981.487	5024.904	5074.865	5128.584	5189.460	5281.288	5369.592	5473.786	5587.827	5739.647	5914.851	6131.455	6365.895

43	44	45	46	47	48	49	50	51	52	53	54	55
2388.110	2308.012	2201.455	2104.050	2008.479	1863.431	1735.003	1568.107	1405.971	1211.962	1015.198	805.4072	597.079
4539.446	4649.984	4774.912	4902.840	5043.791	5191.000	5333.326	5498.928	5662.906	5853.044	6043.702	6233.6041	6428.202
6611.194	6905.963	7237.517	7578.977	7952.823	8367.493	8790.186	9271.319	9779.838	10326.815	10913.978	11501.0255	12115.519

Table 2.2: The table represents the lower and upper quartile, median of the expected values for the re78 for each age if holding other variables at their median.

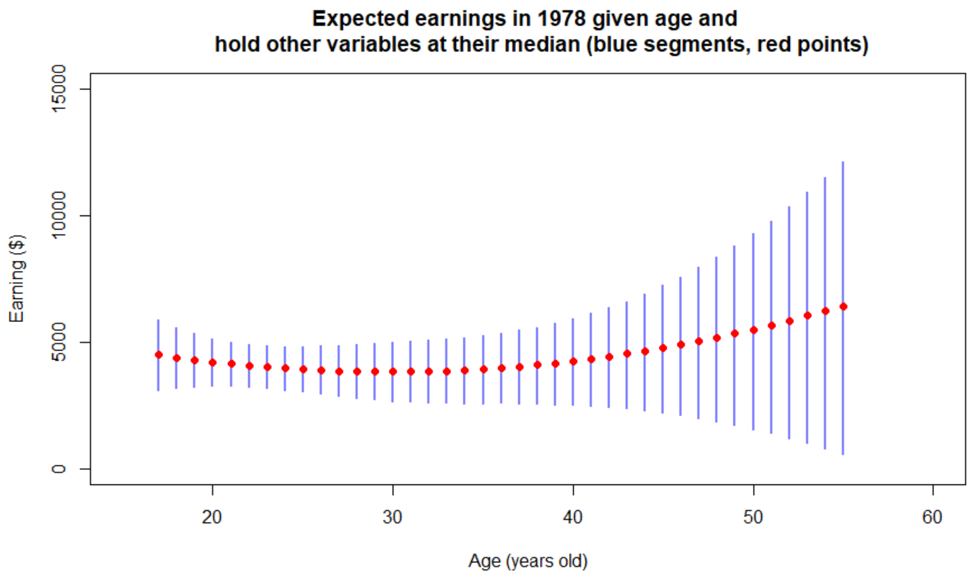

Figure 2.2: The plot represents the bounds of the 95% intervals of expected values of re78 (earnings) for the different ages if holding other variables at their median. The red point is the median of the interval.

- **the 95% interval of expected values for re78, for every unit, using simulation**

	17	18	19	20	21	22	23	24	25	26	27	28	29
lower 95% quantile	3212.555	3308.165	3381.361	3418.779	3430.236	3423.335	3399.310	3347.223	3291.071	3224.424	3154.141	3095.853	3047.929
median	4720.133	4622.199	4531.850	4454.559	4382.708	4319.103	4269.274	4228.606	4188.081	4161.683	4137.558	4124.866	4118.244
upper 95% quantile	6261.755	5965.374	5707.330	5493.305	5339.088	5214.546	5149.172	5111.799	5095.070	5104.314	5129.938	5153.056	5189.594

30	31	32	33	34	35	36	37	38	39	40	41	42
2672.854	2642.184	2613.603	2591.433	2584.267	2579.096	2587.632	2576.784	2567.684	2535.282	2518.127	2483.915	2454.503
3828.701	3834.497	3846.516	3869.468	3904.537	3941.026	3984.080	4038.908	4100.947	4168.369	4244.148	4332.294	4429.953
4981.487	5024.904	5074.865	5128.584	5189.460	5281.288	5369.592	5473.786	5587.827	5739.647	5914.851	6131.455	6365.895

43	44	45	46	47	48	49	50	51	52	53	54	55
2388.110	2308.012	2201.455	2104.050	2008.479	1863.431	1735.003	1568.107	1405.971	1211.962	1015.198	805.4072	597.079
4539.446	4649.984	4774.912	4902.840	5043.791	5191.000	5333.326	5498.928	5662.906	5853.044	6043.702	6233.6041	6428.202
6611.194	6905.963	7237.517	7578.977	7952.823	8367.493	8790.186	9271.319	9779.838	10326.815	10913.978	11501.0255	12115.519

Table 2.3: The table represents the lower and upper quartile, median of the expected values for the re78 for each age if holding other variables at their 75% quantile.

Figure 2.2: The plot represents the bounds of the 95% intervals of expected values of re78 (earnings) for the different ages if holding other variables at their 75% quantile. The blue point is the median of the interval.

- **the 95% prediction interval for re78, for every unit (i.e., each age, spanning the age range in the data set), using simulation.**

	17	18	19	20	21	22	23	24	25	26	27	28	29
lower 95% quantile	-6488.301	-6547.319	-6452.738	-6996.559	-6791.051	-6717.278	-6943.599	-7022.418	-6954.418	-7058.309	-7194.839	-7024.594	-7128.052
median	4572.662	4346.347	4247.964	4213.474	4259.935	3984.338	3952.821	3897.405	3862.550	3884.207	3916.315	3841.865	3851.168
upper 95% quantile	15546.380	15149.651	15073.288	15034.119	14916.343	14800.345	15050.086	15005.208	14587.517	14830.611	14785.016	14700.715	14366.520

30	31	32	33	34	35	36	37	38	39	40	41	42
-7088.716	-7339.462	-6647.994	-6935.922	-7252.409	-7051.983	-7070.709	-6681.098	-6838.242	-6764.643	-6755.751	-6718.080	-6728.827
3836.529	3728.841	3880.952	3788.478	3802.453	3934.223	4060.850	4031.663	4153.682	4170.030	4289.768	4290.425	4525.665
14652.317	14586.191	14800.283	14804.324	14594.870	14969.075	15002.132	15129.879	15124.165	15119.512	15521.105	15667.533	15282.125

43	44	45	46	47	48	49	50	51	52	53	54	55
-6773.830	-6808.825	-6358.169	-6485.434	-6453.748	-6238.239	-6054.383	-6119.081	-6233.774	-5924.358	-5894.832	-5738.078	-5972.334
4505.464	4623.292	4766.307	4832.655	4966.671	5077.860	5465.905	5492.322	5608.542	5821.581	5991.597	6199.585	6399.304
15572.525	15876.184	16094.364	16049.915	16553.986	16360.800	17185.377	16843.276	17449.668	17545.523	17382.201	18243.588	18874.916

Table 2.4: The table represents the lower and upper quartile, median of the predicted values for the re78 for each age if holding other variables at their median.

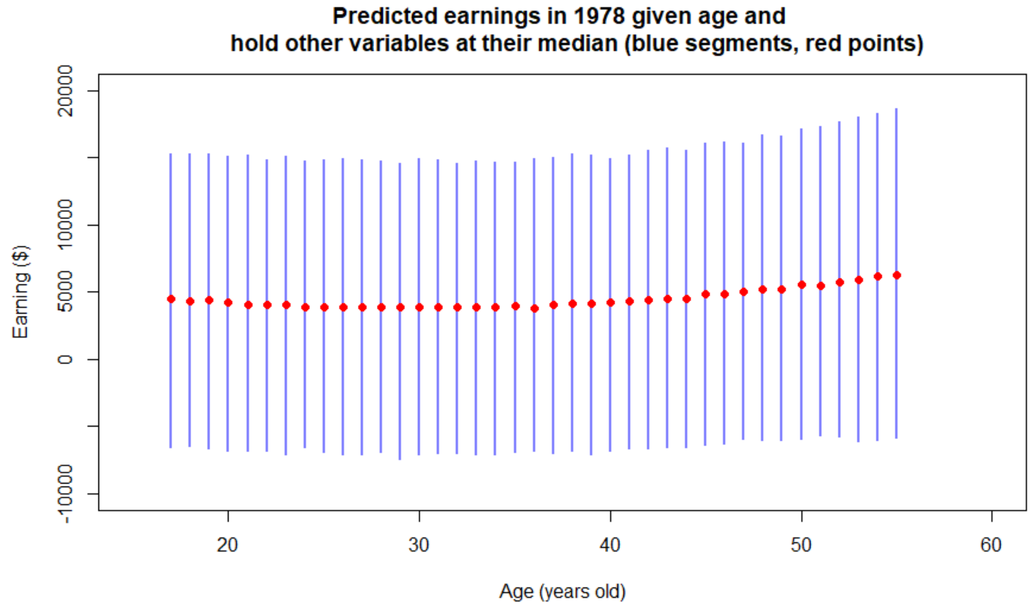

Figure 2.3: The plot represents the bounds of the 95% intervals of predicted values of re78 (earnings) for the different ages if holding other variables at their median. The red point is the median of the interval.

- **the 95% prediction interval for re78, for every unit, using simulation**

	17	18	19	20	21	22	23	24	25	26	27	28	29
lower 95% quantile	-6396.858	-6203.892	-6515.127	-6510.259	-6432.981	-6550.021	-6677.282	-6891.355	-6559.466	-6575.827	-6816.277	-6771.344	-6926.323
median	4662.859	4645.974	4524.224	4450.186	4372.537	4332.138	4188.292	4170.901	4253.662	4250.507	4038.068	4123.905	3999.768
upper 95% quantile	15781.441	15384.999	15396.128	15272.729	15448.366	15382.476	15274.143	15084.462	15008.100	14993.791	15218.702	15328.187	15069.154

30	31	32	33	34	35	36	37	38	39	40	41	42
-6642.574	-6908.641	-6667.732	-6948.622	-6541.060	-6726.731	-6713.172	-6769.134	-6427.478	-6510.269	-6313.253	-6204.352	-6217.998
4142.719	4125.778	4167.716	4207.391	4274.532	4219.794	4304.263	4406.198	4470.266	4552.070	4606.001	4722.707	4784.847
15075.070	15014.993	15189.171	15139.745	15060.695	14828.712	15323.478	15380.663	15279.721	15579.956	15525.375	15708.320	15775.696

43	44	45	46	47	48	49	50	51	52	53	54	55
-6184.155	-6063.113	-5739.589	-5959.258	-5989.542	-5771.966	-5837.785	-5773.275	-5734.607	-5423.873	-5403.049	-5767.567	-5964.431
4897.940	4880.698	5002.721	5427.399	5366.797	5642.818	5743.799	5836.613	6090.520	6253.904	6403.223	6664.661	6711.842
16225.460	16131.555	16056.847	16416.800	16791.051	17288.287	17134.138	17277.642	17887.619	18155.395	18241.977	18901.887	19095.628

Table 2.5: The table represents the lower and upper quartile, median of the predicted values for the re78 for each age if hold other variables at their 75% quantile.

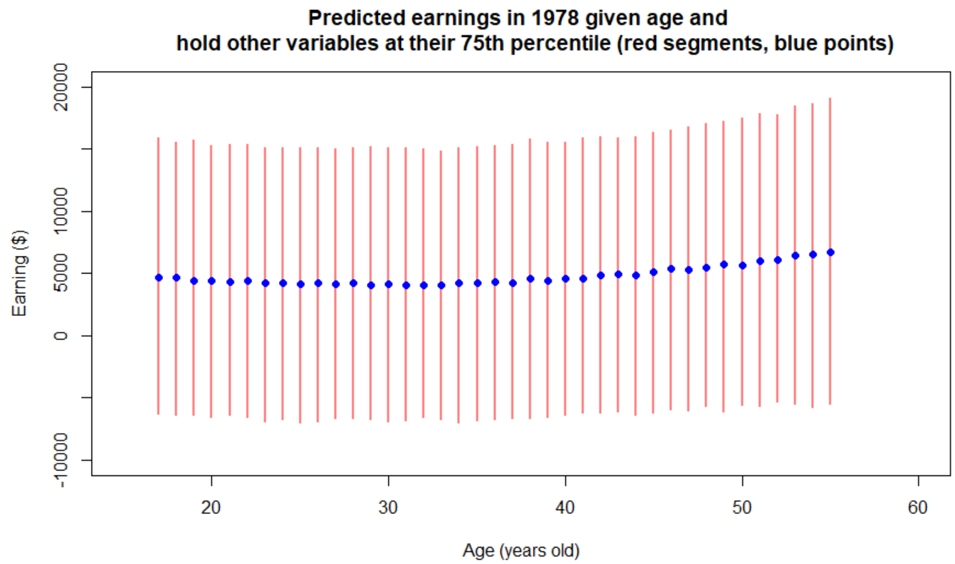

Figure 2.4: The plot represents the bounds of the 95% intervals of predicted values of re78 (earnings) for the different ages if holding other variables at their 75% interval. The blue point is the median of the interval.

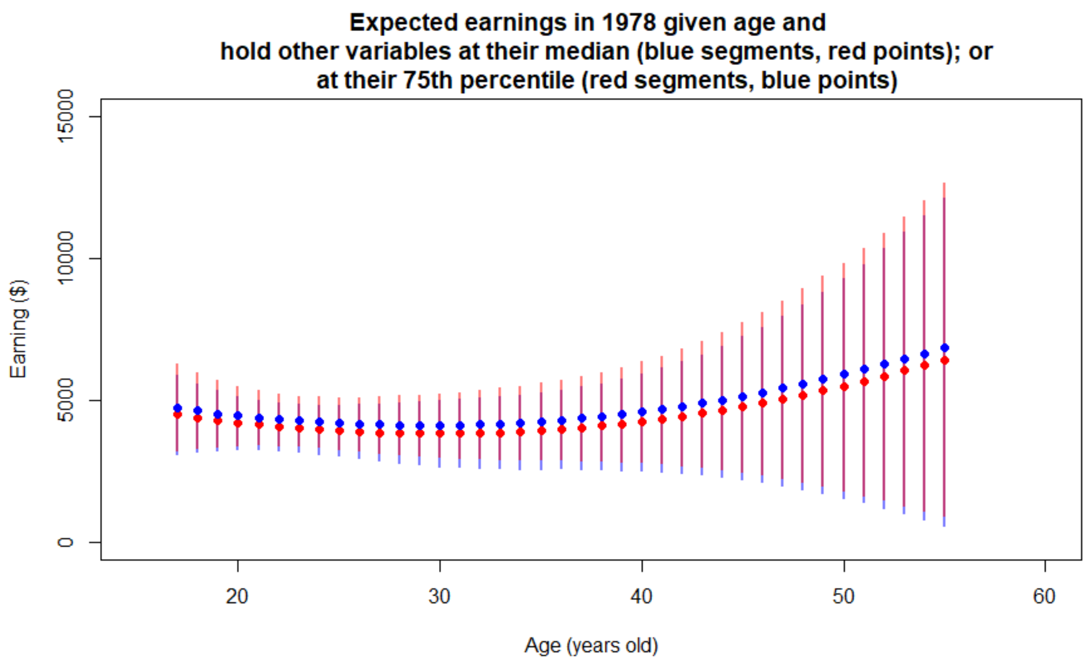

Figure 2.5: The plot represents the bounds of the 95% intervals of expected values of re78 (earnings) for the different ages if holding other variables at their median (blue segments, red points) and 75% quantile (red segments, blue points). The points are the median of the given range. We can see that if we hold other variables at the 75% quantile, the expected values seem to be higher than the expected value when we hold other variables at their median.

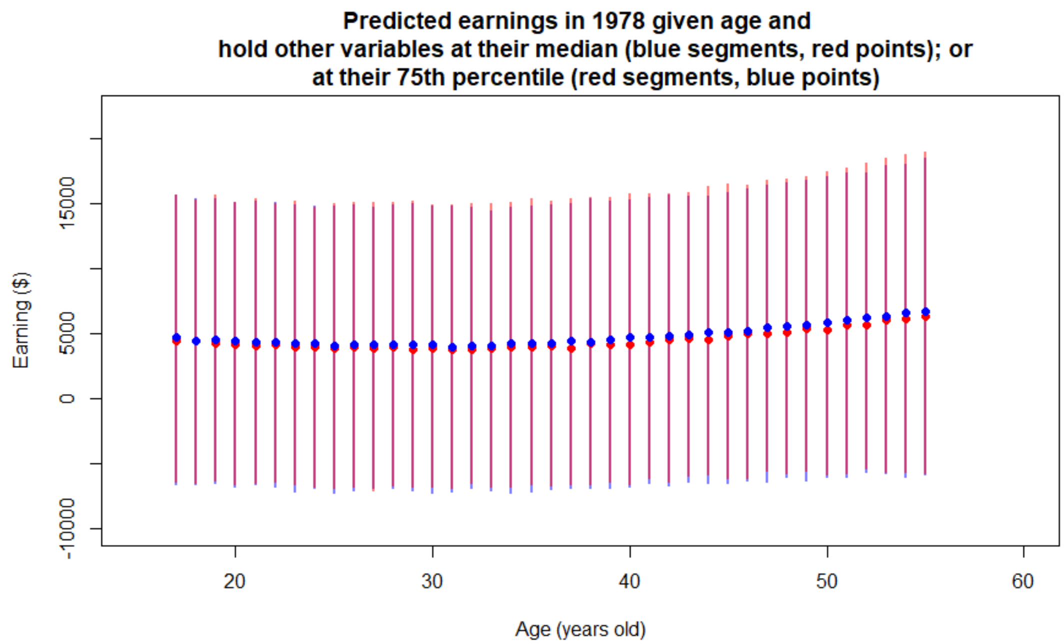

Figure 2.6: The plot represents the bounds of the 95% intervals of predicted values of re78 (earnings) for the different ages if holding other variables at their median (blue segments, red points) and 75% quantile (red segments, blue points). The points are the median of the given range. We can see that if we hold other variables at the 75% quantile, the predicted values seem to be higher than the expected value when we hold other variables at their median. Also, the 95% interval of the predicted values are wider than of the expected values.

Problem 3:
Link to code

The selected regression is the linear model. We have the following graph:

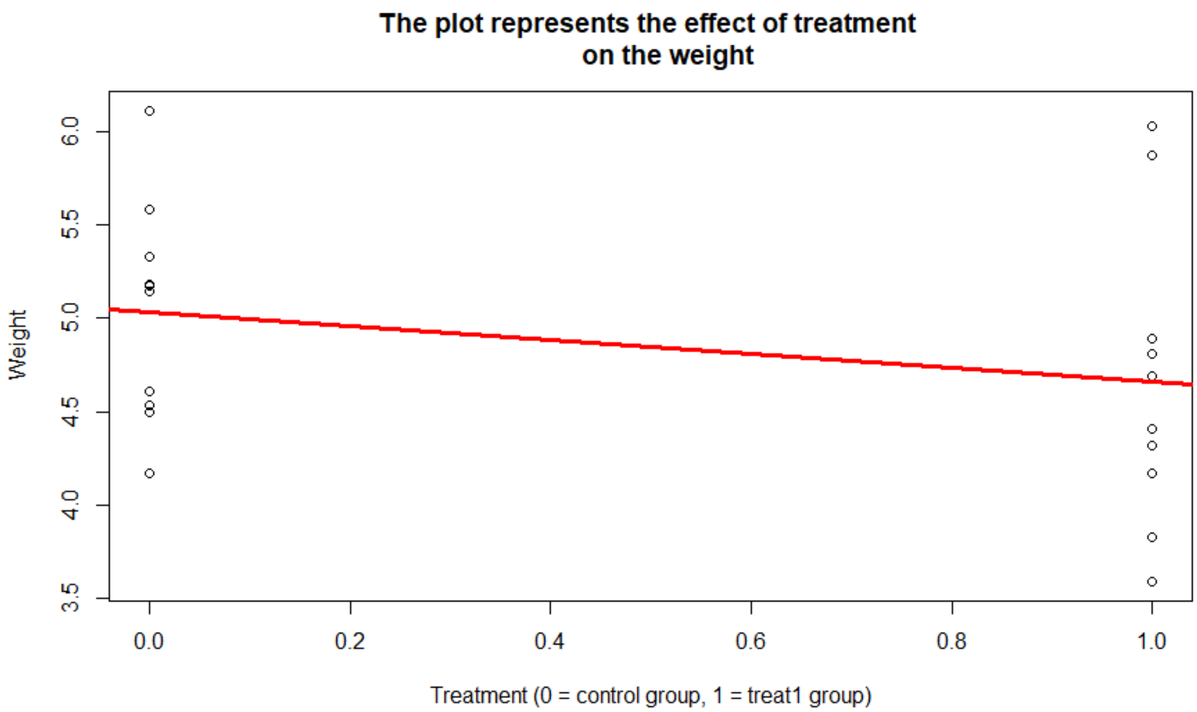

Figure 3.1: The figure represents the line of best fit represents the effect of the treatment on the weight.

	Lower 95% confidence interval boundaries	**Mean (the coefficients)**	**Upper 95% confidence interval boundaries**
Linear model	-1.02530	-0.371	0.2833003
Bootstrap-sample results	-0.9490250	-0.3735098	0.2401223

Table 3.1: The table represents the bounds on the two 95% confidence intervals from the linear model and the bootstrap-sample results.

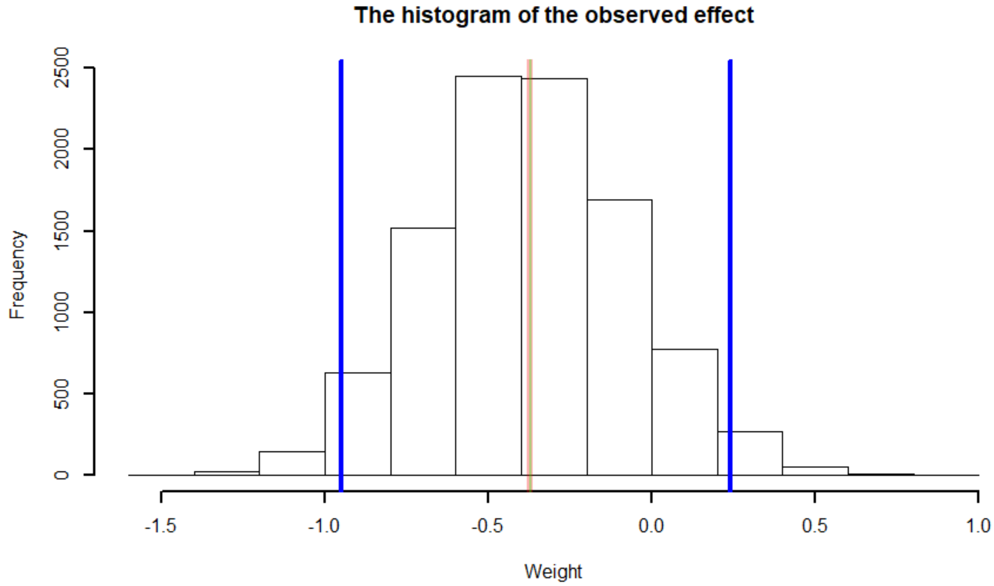

Figure 3.2: The histogram represents the distribution of the bootstrapped data. The blue lines are the 95% confidence interval of the values of the coefficient for treatment. The middle line is the overlapped of 2 lines: the red line is the mean of the bootstrapped values of the coefficients for the treatment variable, the green line is the real coefficient derived from the linear model.

Description: The mean, the lower boundaries, and the upper boundaries of the 95% confidence interval are approximately the same between the results from the linear model and the results from bootstrapped data, which implies that the bootstrap, as a resampling method, can be used to estimate the standard errors of the coefficients from a linear regression fit. The bootstrap is based roughly on the law of large numbers, which says, in short, that with enough data the empirical distribution will be a good approximation of the true distribution. Also, the histogram of the distribution of the bootstrap results is quite similar to be a normal distribution and should approximate the density of the true distribution.

Problem 4: Write your function (5 lines max) that takes Ys and predicted Ys as inputs, and outputs R^2.
Link to code

```
################### Start R ###################
find_r_squared <- function(y , predicted_y) {
  mean_y <- mean(y)
  SSError <- sum((y - predicted_y)^2)
  SSTotal <- sum((y - mean_y)^2)
  return(1 - SSError/SSTotal)
}

data(PlantGrowth)
data <- PlantGrowth

data2 <- data[-which(data$group == 'trt2'), ]
data2$group_2 <- as.numeric(data2$group == 'trt1')

find_r_squared(data2$weight, predict(lm(weight ~ group_2 , data = data2)))
summary(lm(weight ~ group_2 , data = data2))$r.squared

############### Output R ###############
find_r_squared(data2$weight, predict(lm(weight ~ group_2 , data = data2)))
[1] 0.0730776

summary(lm(weight ~ group_2 , data = data2))$r.squared
[1] 0.0730776
############### End Output R ###############

################### End R ###################
```

Problem 5:
[Link to code]

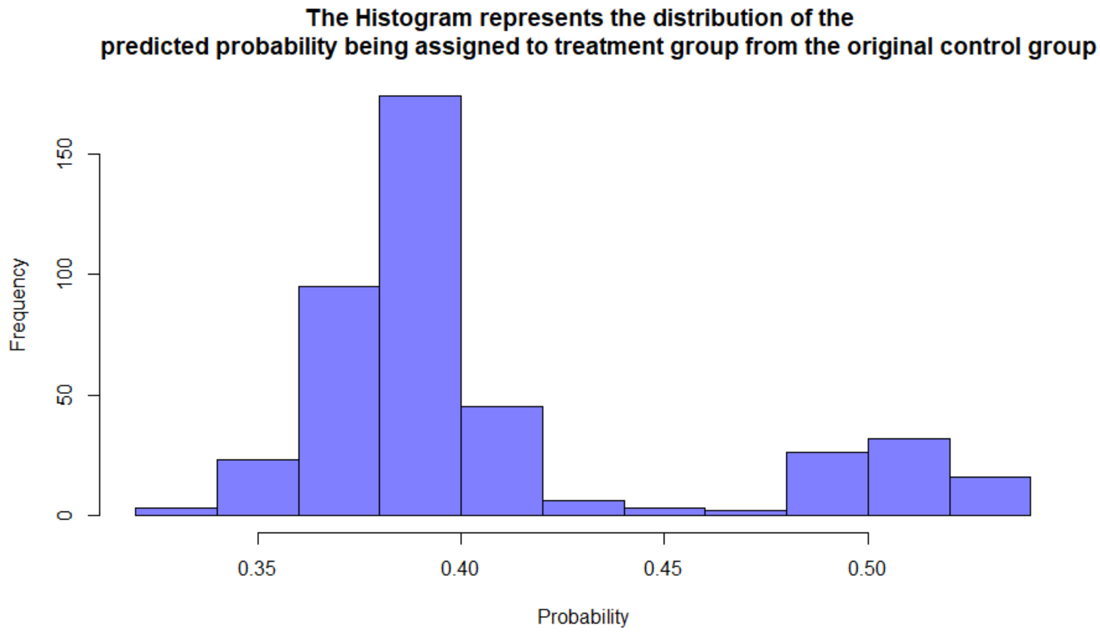

Figure 5.1: The histogram represents the distribution of the estimated probability of being assigned to the treatment group from the data initially from the control group.

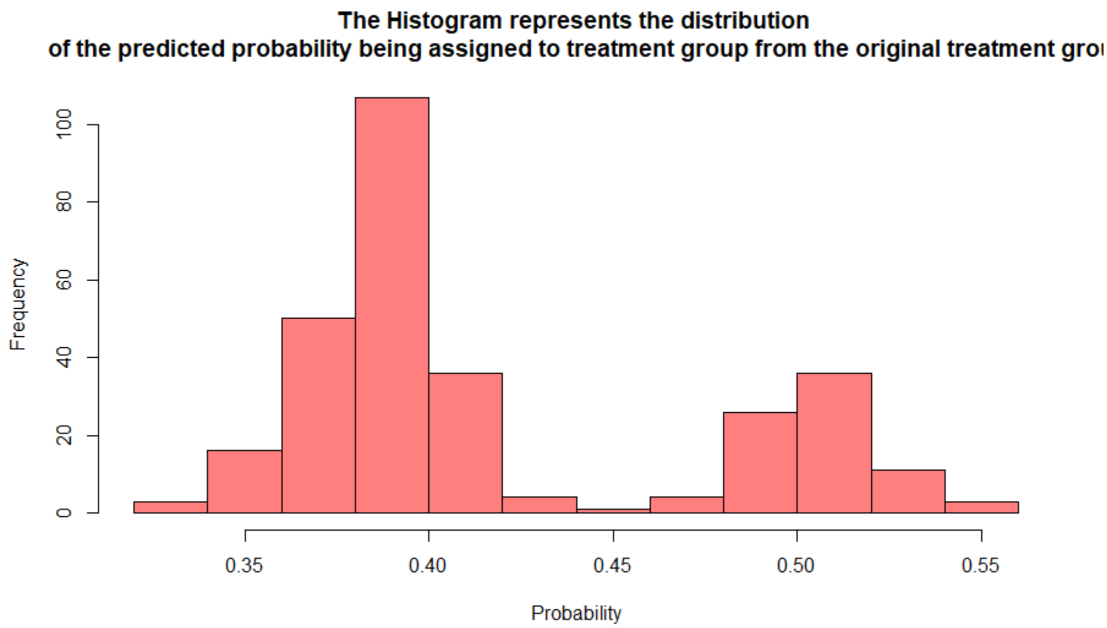

Figure 5.2: The histogram represents the distribution of the estimated probability of being assigned to the treatment group from the data initially from the treatment group.

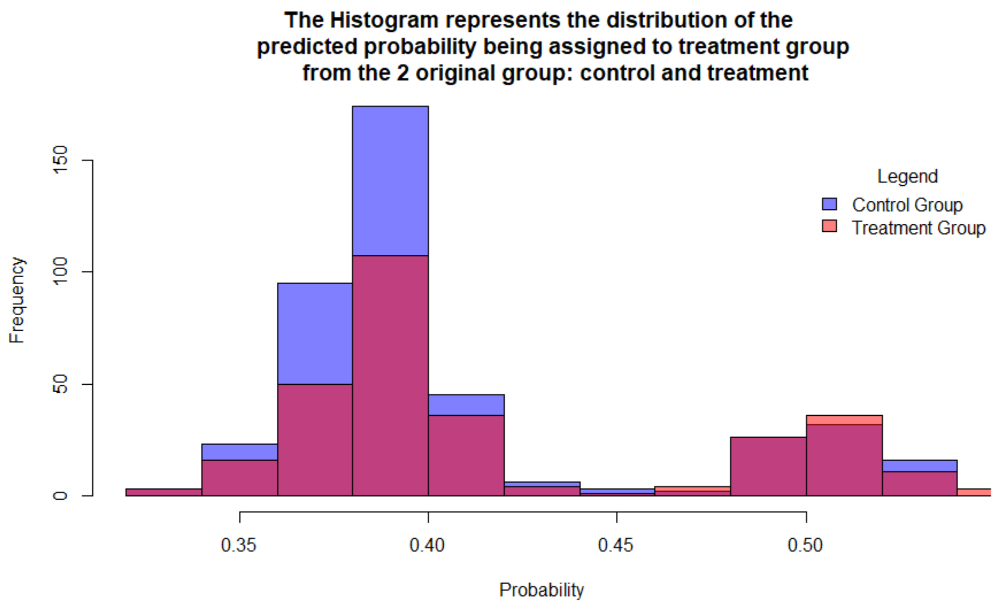

Figure 5.3: The overlapping histogram represents the distribution of the estimated probability of being assigned to the treatment group from the data initially from the control group (blue) and the treatment group (red).

From figure 5.1 and 5.2, the shape of the distributions of quite similar and when we overlap the two histograms (figure 5.3), the histogram of the control group is higher, which is because the control group has more data points. We want to further investigate the density plot of these two sets of probability values from 2 groups.

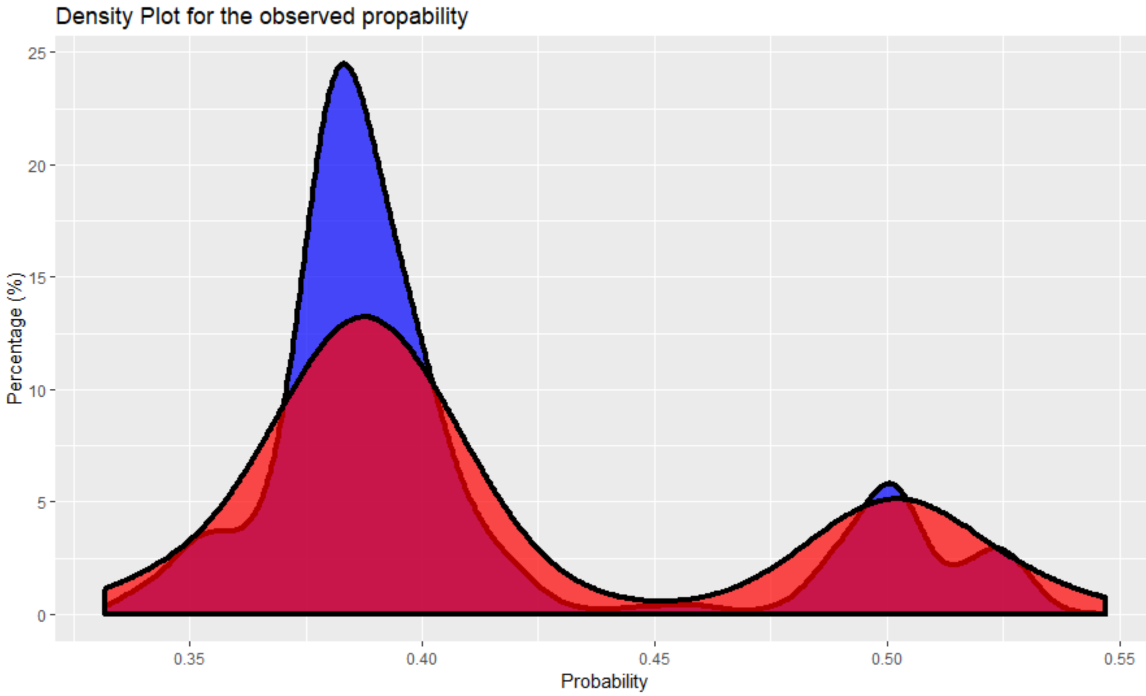

Figure 5.3: The overlapping density represents the distribution of the estimated probability of being assigned to the treatment group from the data initially from the control group (blue) and the treatment group (red).

Call P(1) the probabilities of being assigned to the treatment group.

The mean of P(1) of the control group is 0.4068407, whereas the mean of P(1) of the treatment group is 0.4178206, which are quite similar to each other and equal to the ratio of the treatment group (287) and the total group (712) (approximately 0.4031). This would imply other variables (age, black, etc.) cannot distinguish between whether the person will be assigned to the treatment group or not. One possible explanation is that the treatment was randomly assigned to the participants: the treatment is assigned randomly, independently of any other variables: we cannot predict accurately if a person will be assigned to the treatment based on other given variables (age, educ, etc.).

Link to code: https://drive.google.com/open?id=1k6JWFDWHexH25DfK15wQejEDGLRC-C_J