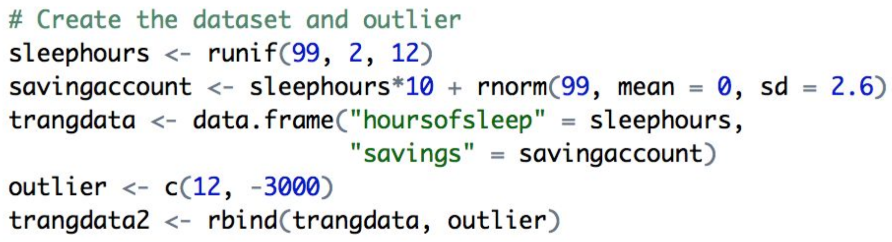
ASSIGNMENT 2

KEY ANSWERS

I, Question 1

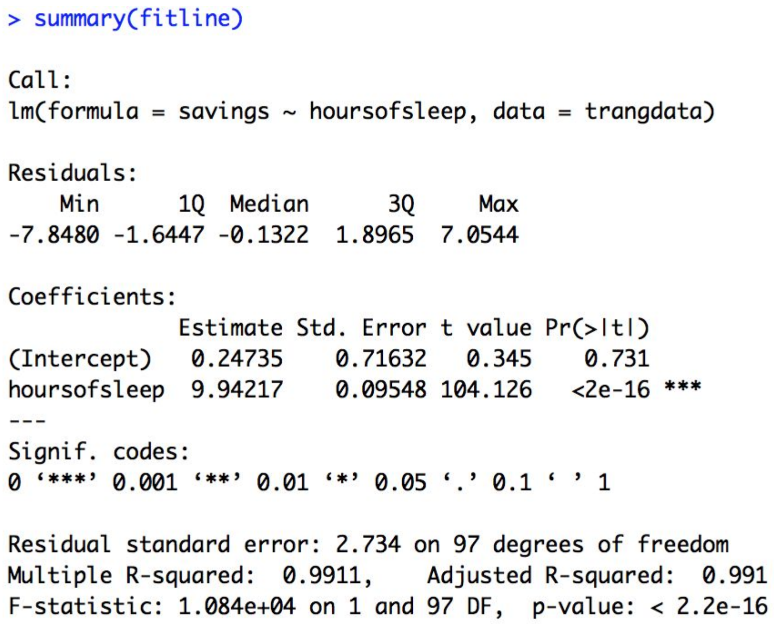
1, Data-generating equation



```
# Create the dataset and outlier
sleephours <- runif(99, 2, 12)
savingaccount <- sleephours*10 + rnorm(99, mean = 0, sd = 2.6)
trangdata <- data.frame("hoursofsleep" = sleephours,
                        "savings" = savingaccount)
outlier <- c(12, -3000)
trangdata2 <- rbind(trangdata, outlier)
```

Figure 1. The code for generating a dataset.

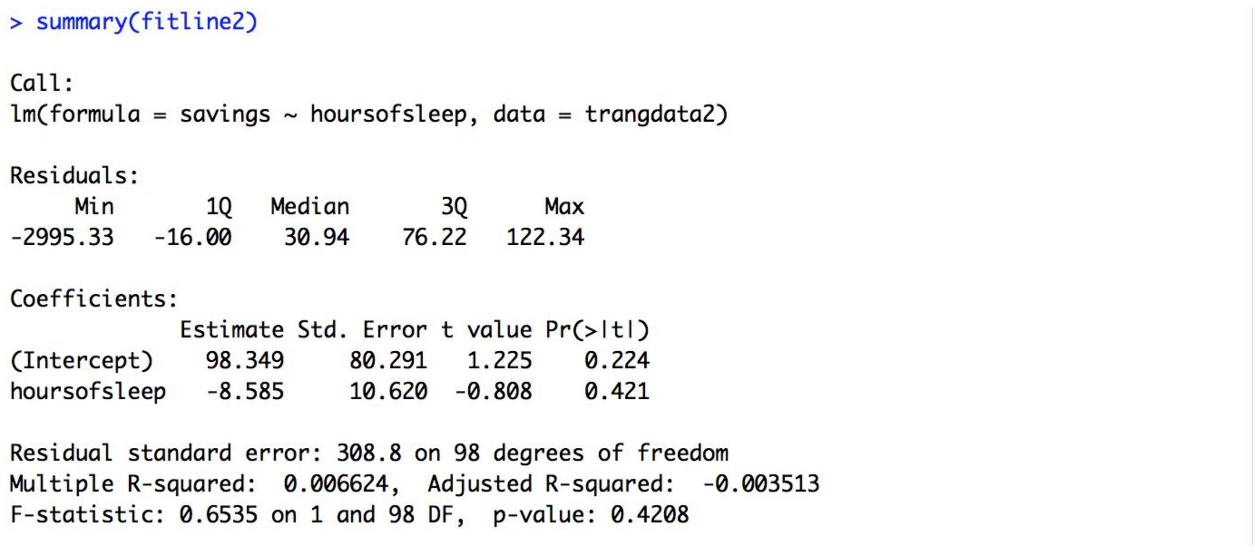
2, Regression results for the original 99



```
> summary(fitline)

Call:
lm(formula = savings ~ hoursofsleep, data = trangdata)

Residuals:
    Min      1Q  Median      3Q     Max 
-7.8480 -1.6447 -0.1322  1.8965  7.0544 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)   0.24735    0.71632   0.345    0.731    
hoursofsleep  9.94217    0.09548 104.126   <2e-16 ***
---
Signif. codes:  
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.734 on 97 degrees of freedom
Multiple R-squared:  0.9911,	Adjusted R-squared:  0.991 
F-statistic: 1.084e+04 on 1 and 97 DF,  p-value: < 2.2e-16
```

Figure 2. The summary of the regression line for the original 99 points in the dataset.

3, Regression results with the outlier included



```
> summary(fitline2)

Call:
lm(formula = savings ~ hoursofsleep, data = trangdata2)

Residuals:
     Min       1Q   Median       3Q      Max 
-2995.33   -16.00    30.94    76.22   122.34 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    98.349     80.291   1.225    0.224
hoursofsleep   -8.585     10.620  -0.808    0.421

Residual standard error: 308.8 on 98 degrees of freedom
Multiple R-squared:  0.006624,  Adjusted R-squared:  -0.003513 
F-statistic: 0.6535 on 1 and 98 DF,  p-value: 0.4208
```

Figure 3. The summary of the regression line for the all the points (including the outlier) in the dataset.

4, Data visualization

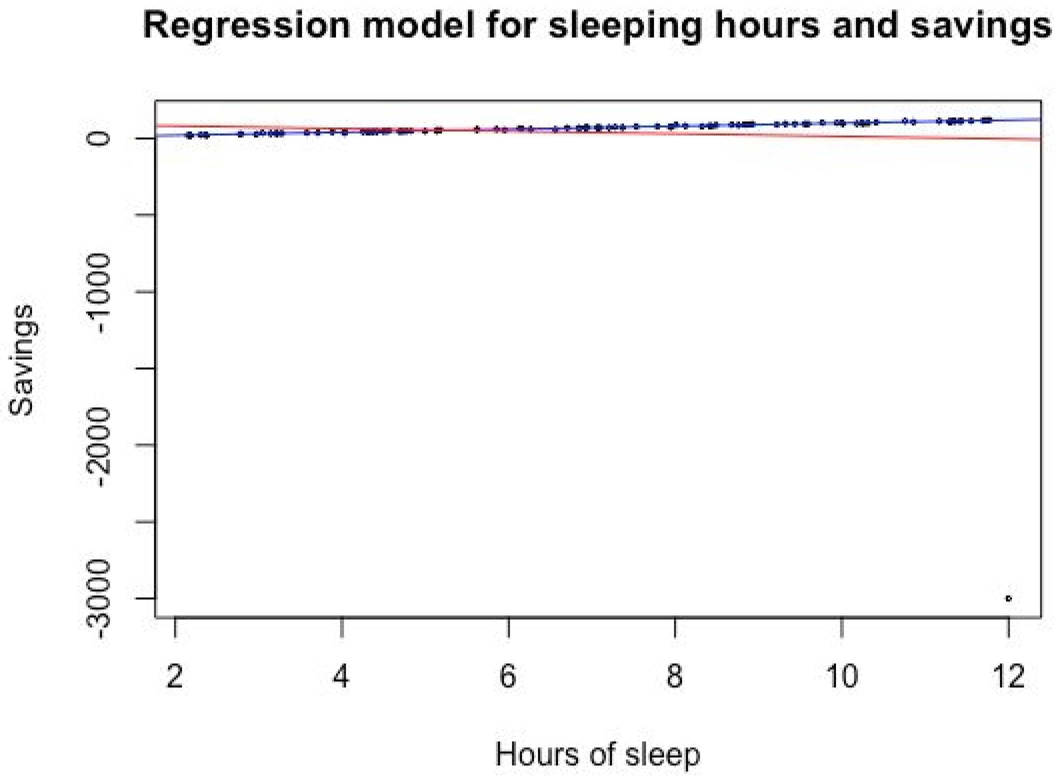


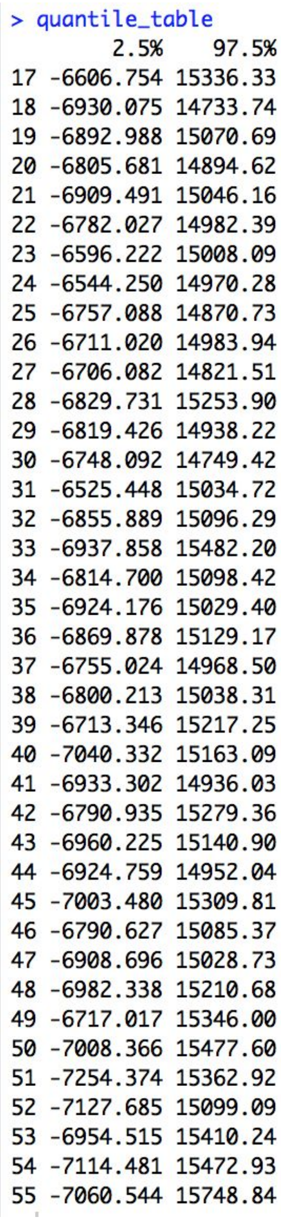
Figure 4. The plot for the points in the dataset and two regression lines.

5, Caption

The graph shows the relationship between the hours of sleep and savings for 100 people. As we can see, if we don't take into account the outlier, the trend is upward, and if we do, it is downward. Therefore, we should not rely on extrapolation, since we can misinterpret the overall trend just because of one extreme outlier.

II, Question 2

1, When variables are held at their medians



```
> quantile_table
        2.5%    97.5%
17 -6606.754 15336.33
18 -6930.075 14733.74
19 -6892.988 15070.69
20 -6805.681 14894.62
21 -6909.491 15046.16
22 -6782.027 14982.39
23 -6596.222 15008.09
24 -6544.250 14970.28
25 -6757.088 14870.73
26 -6711.020 14983.94
27 -6706.082 14821.51
28 -6829.731 15253.90
29 -6819.426 14938.22
30 -6748.092 14749.42
31 -6525.448 15034.72
32 -6855.889 15096.29
33 -6937.858 15482.20
34 -6814.700 15098.42
35 -6924.176 15029.40
36 -6869.878 15129.17
37 -6755.024 14968.50
38 -6800.213 15038.31
39 -6713.346 15217.25
40 -7040.332 15163.09
41 -6933.302 14936.03
42 -6790.935 15279.36
43 -6960.225 15140.90
44 -6924.759 14952.04
45 -7003.480 15309.81
46 -6790.627 15085.37
47 -6908.696 15028.73
48 -6982.338 15210.68
49 -6717.017 15346.00
50 -7008.366 15477.60
51 -7254.374 15362.92
52 -7127.685 15099.09
53 -6954.515 15410.24
54 -7114.481 15472.93
55 -7060.544 15748.84
```

Figure 5. A table of the 95% confidence intervals for re78 of all ages, when other variables are held constant at their medians.

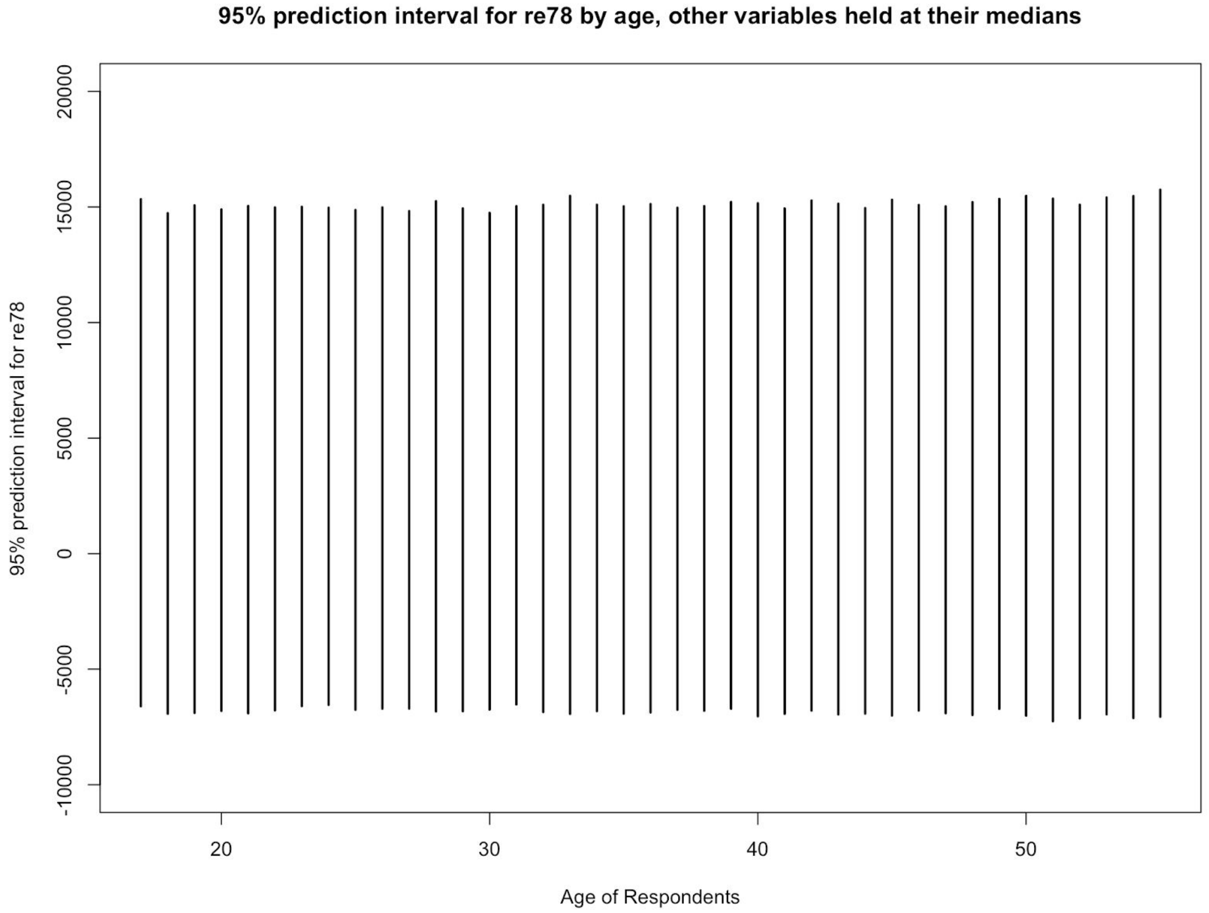
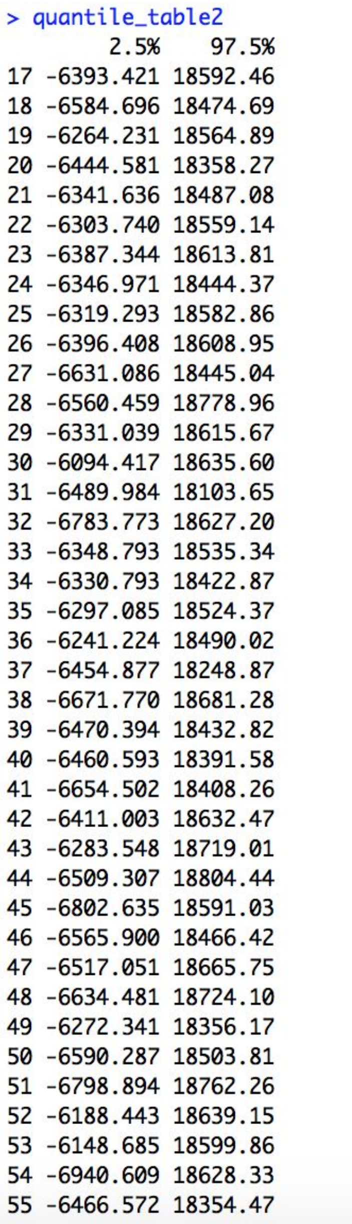


Figure 6. The graph illustrates 95% prediction intervals for re78 by age when other variables are held at their medians.

2, When variables are held at their 90% quantiles



```
> quantile_table2
        2.5%    97.5%
17 -6393.421 18592.46
18 -6584.696 18474.69
19 -6264.231 18564.89
20 -6444.581 18358.27
21 -6341.636 18487.08
22 -6303.740 18559.14
23 -6387.344 18613.81
24 -6346.971 18444.37
25 -6319.293 18582.86
26 -6396.408 18608.95
27 -6631.086 18445.04
28 -6560.459 18778.96
29 -6331.039 18615.67
30 -6094.417 18635.60
31 -6489.984 18103.65
32 -6783.773 18627.20
33 -6348.793 18535.34
34 -6330.793 18422.87
35 -6297.085 18524.37
36 -6241.224 18490.02
37 -6454.877 18248.87
38 -6671.770 18681.28
39 -6470.394 18432.82
40 -6460.593 18391.58
41 -6654.502 18408.26
42 -6411.003 18632.47
43 -6283.548 18719.01
44 -6509.307 18804.44
45 -6802.635 18591.03
46 -6565.900 18466.42
47 -6517.051 18665.75
48 -6634.481 18724.10
49 -6272.341 18356.17
50 -6590.287 18503.81
51 -6798.894 18762.26
52 -6188.443 18639.15
53 -6148.685 18599.86
54 -6940.609 18628.33
55 -6466.572 18354.47
```

Figure 7. A table of the 95% confidence intervals for re78 of all ages, when other variables are held constant at their 90% quantiles.

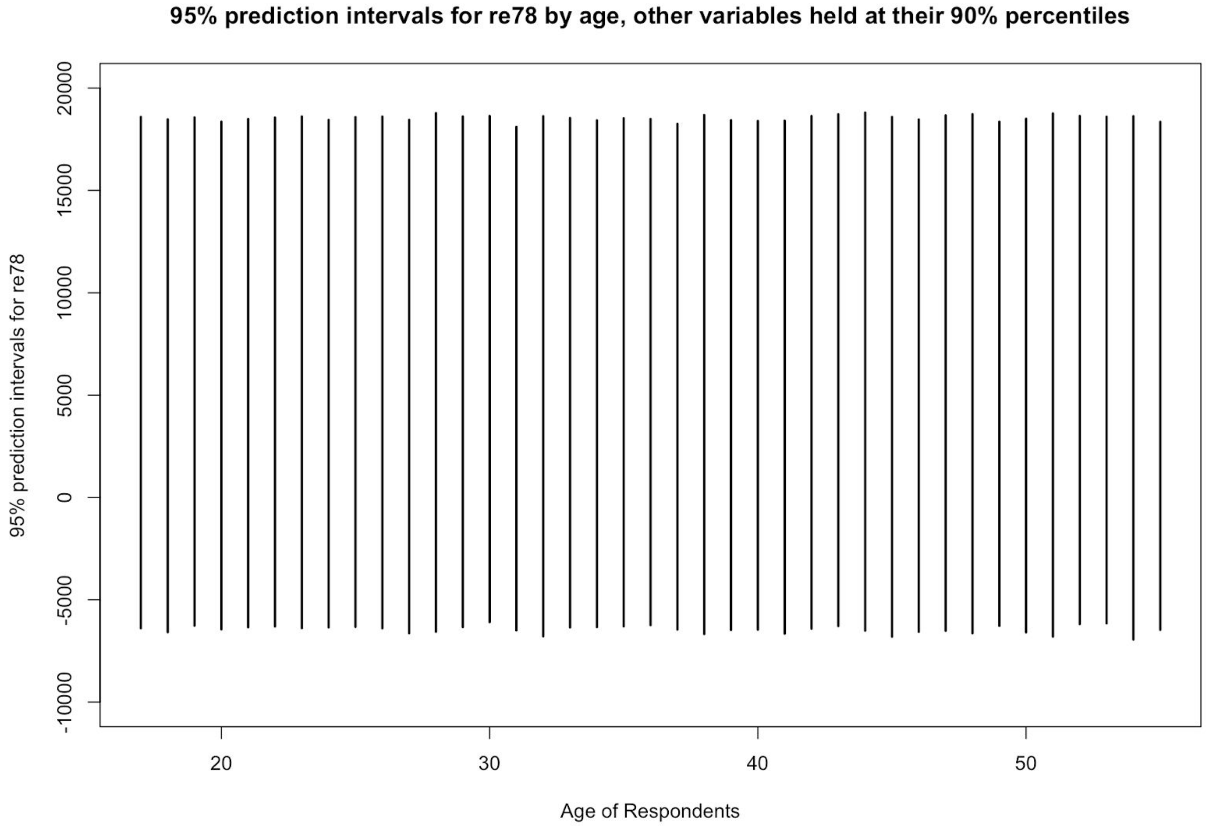
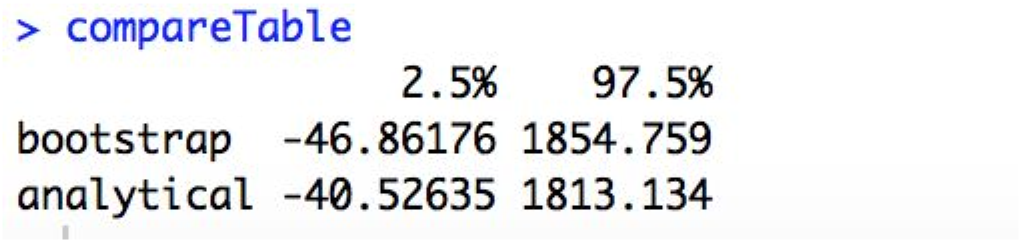


Figure 8. The graph illustrates 95% prediction intervals for re78 by age when other variables are held at their 90% percentiles.

III, Question 3

1, Table of confidence intervals



```
> compareTable
                 2.5%    97.5%
bootstrap  -46.86176 1854.759
analytical -40.52635 1813.134
```

Figure 9. A table that compares the two confidence intervals for re78 obtained in two different ways.

2, The histogram

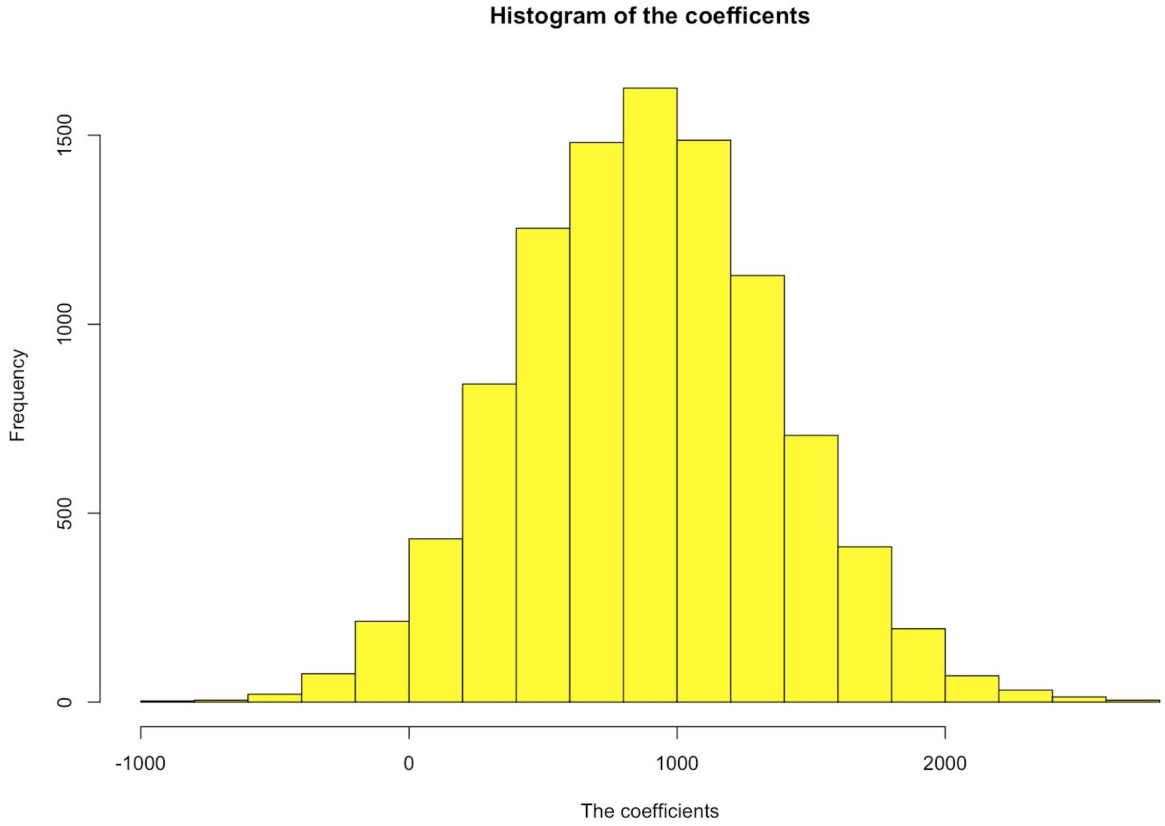
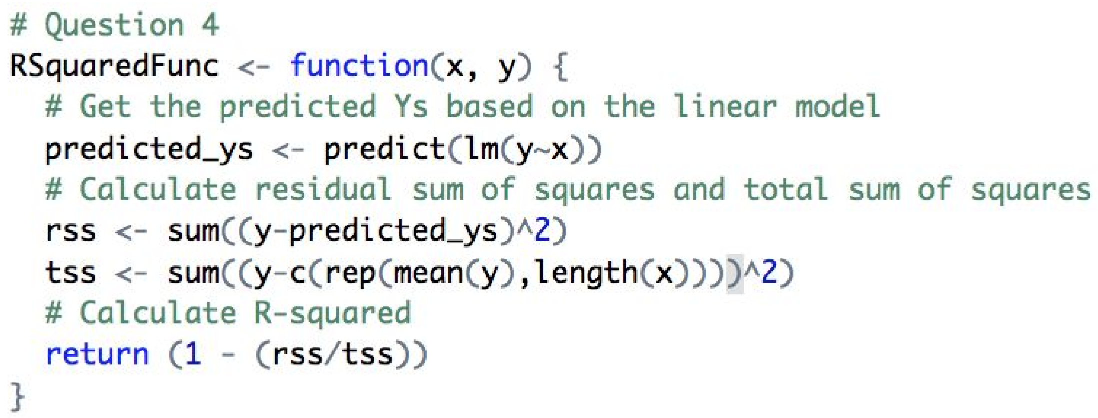


Figure 10. The graph illustrates the frequencies of the coefficients of 10,000 bootstrapped samples.

3, Conclusion

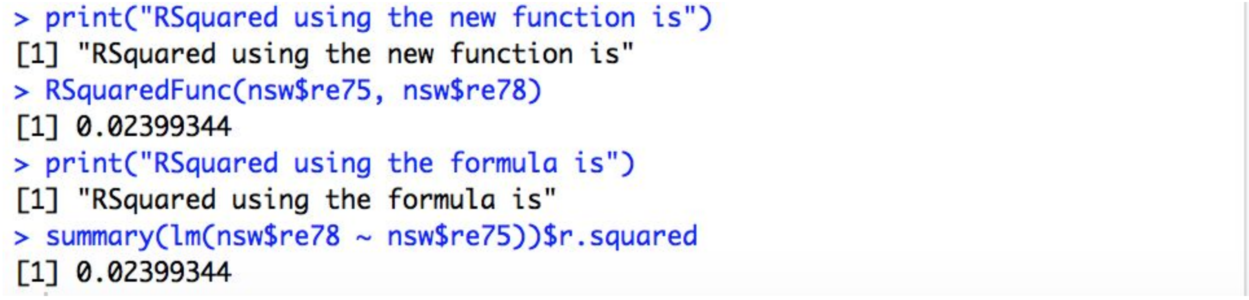
In general, the analytic confidence interval for the coefficient value obtained by using the standard error is similar to that using bootstrapping. And the coefficients we obtained from bootstrapping are pretty normally distributed.

IV, Question 4



```
# Question 4
RSquaredFunc <- function(x, y) {
  # Get the predicted Ys based on the linear model
  predicted_ys <- predict(lm(y~x))
  # Calculate residual sum of squares and total sum of squares
  rss <- sum((y-predicted_ys)^2)
  tss <- sum((y-c(rep(mean(y),length(x))))^2)
  # Calculate R-squared
  return (1 - (rss/tss))
}
```

Figure 11. The code shows the manual function to get R-squared.



```
> print("RSquared using the new function is")
[1] "RSquared using the new function is"
> RSquaredFunc(nsw$re75, nsw$re78)
[1] 0.02399344
> print("RSquared using the formula is")
[1] "RSquared using the formula is"
> summary(lm(nsw$re78 ~ nsw$re75))$r.squared
[1] 0.02399344
```

Figure 12. The R-squared values obtained in two different ways.

V, Question 5

1, The histograms

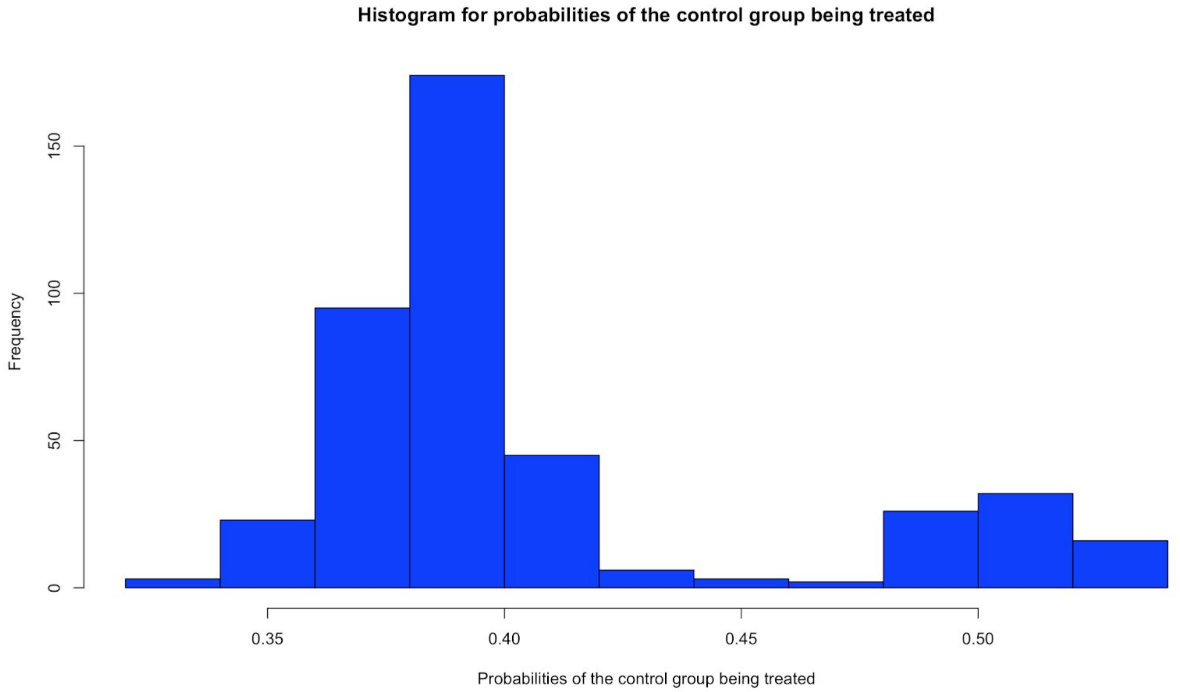


Figure 13. The graph illustrates the frequencies of the probabilities by which the observations in the control group could have been treated.

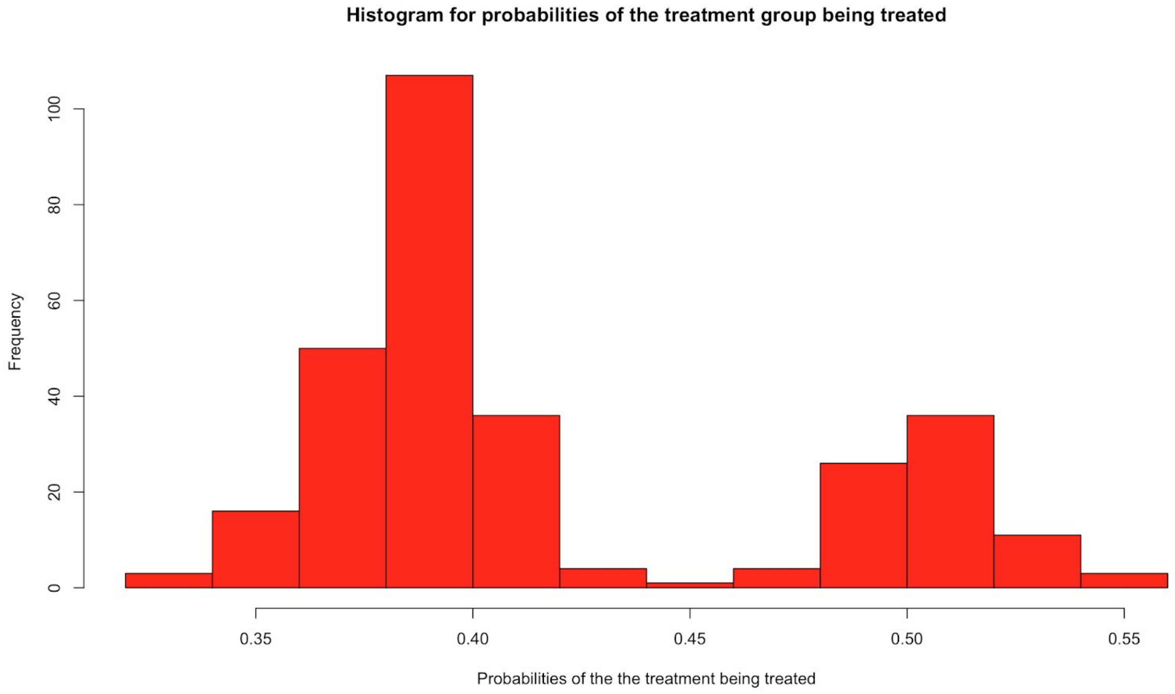
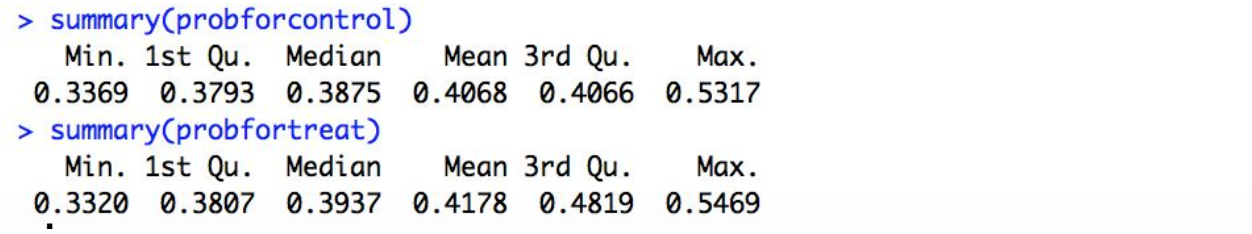


Figure 14. The graph shows the frequencies of the probabilities by which the observations in the treatment group could have been treated.

2, Summary



```
> summary(probforcontrol)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 0.3369  0.3793  0.3875  0.4068  0.4066  0.5317 
> summary(probfortreat)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 0.3320  0.3807  0.3937  0.4178  0.4819  0.5469 
```

Figure 15. Summaries for the two distributions above.

In general, the two distributions of estimated probabilities for the treatment and control groups are pretty similar to each other. This result is intuitively reasonable because we use the same linear model to predict the probabilities and it is very likely that the assignments to either treatment or control for each observation are arbitrary.

APPENDIX

All the codes:

https://gist.github.com/trangnguyenvn1398/e8d985290c57bb2c9e98159698a2a479