Yuhao Chen

CS112 Assignment 2

Regression and Bootstrapping

Fall 2018

Link to The Code:

https://drive.google.com/open?id=17xi4CBW85kS1DzvMigDur622h3c39zrO

Question 1

1. Your original data-generating equation

x=10*runif(99)

y=10+60*x+rnorm(99)

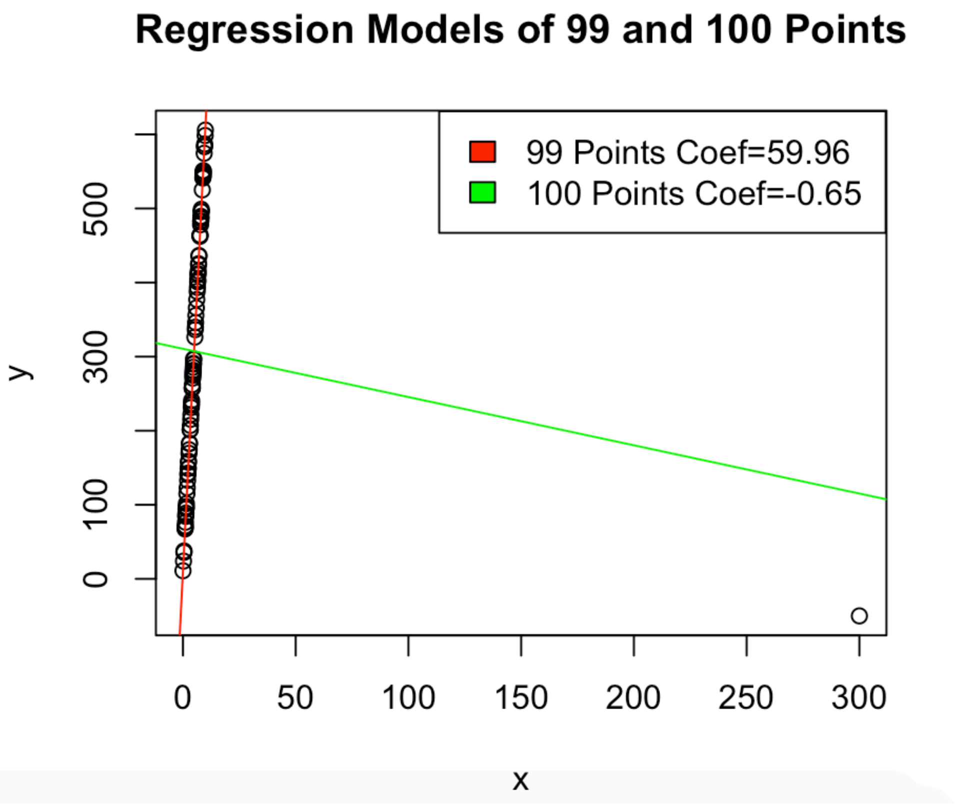
2. Regression results for the original 99 (copy/paste the "summary" output)

```
> summary(lm1)

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.3889 -0.6502 -0.0875  0.5821  2.0903 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 10.28959    0.17920   57.42   <2e-16 ***
x           59.95985    0.03121 1921.07   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.885 on 97 degrees of freedom
Multiple R-squared:      1,	Adjusted R-squared:      1 
F-statistic: 3.691e+06 on 1 and 97 DF,  p-value: < 2.2e-16
```

3. Regression results with the outlier included (copy/paste "summary" output)

```
> summary(lm2)

Call:
lm(formula = y2 ~ x2)

Residuals:
    Min      1Q  Median      3Q     Max 
-299.75 -153.85  -19.95  157.90  301.58 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 310.7262    18.0631  17.202   <2e-16 ***
x2           -0.6518     0.5915  -1.102    0.273    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 174.4 on 98 degrees of freedom
Multiple R-squared:  0.01224,	Adjusted R-squared:  0.002159 
F-statistic: 1.214 on 1 and 98 DF,  p-value: 0.2732
```

4. A properly-labeled data visualization that shows a single scatterplot, the regression line based on the original 99 points, and another differentiated regression line based on 100 points.



5. 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.

Answer: When there are outliers appearing, the model can be largely influenced. Therefore, when doing extrapolation with a trained model, we need to analyze the effectivity of the model first, focusing on whether or not to consider the outliers. If the outliers are ridiculous and appear due to the wrong observation or type-in, we need to ignore them; if the outlier are in the acceptable range, we cannot ignore them.

Question 2

1. A table with the relevant point estimates (e.g., the bounds of the prediction intervals of y for the different ages, and the medians of the other predictors)

```
       [,14] [,15] [,16]    [,17]     [,18]     [,19]     [,20]     [,21]
2.5%      NA    NA    NA -6662.36 -6839.577 -6735.086 -6807.791 -6577.317
97.5%     NA    NA    NA 15196.25 15098.285 14872.080 15068.488 14937.930
          [,22]     [,23]    [,24]     [,25]     [,26]     [,27]     [,28]
2.5%  -7005.02 -6631.218 -6628.67 -6712.292 -6764.641 -6651.488 -6677.396
97.5% 14987.50 15121.140 15321.73 15120.787 15095.833 15045.038 14917.211
          [,29]     [,30]    [,31]     [,32]     [,33]     [,34]     [,35]
2.5%  -6562.701 -6717.553 -6689.57 -6590.607 -6574.434 -6663.529 -7084.381
97.5% 15015.273 15274.418 15211.22 15031.679 15310.179 15019.754 14883.526
          [,36]     [,37]     [,38]     [,39]     [,40]     [,41]     [,42]
2.5%  -6731.937 -6695.371 -6793.294 -6717.843 -6764.099 -6959.599 -6639.081
97.5% 15129.263 15220.951 15052.206 15330.769 15153.220 15233.153 15177.797
          [,43]    [,44]     [,45]    [,46]     [,47]     [,48]     [,49]
2.5%  -6870.343 -6724.43 -7043.224 -6962.92 -7083.499 -7051.442 -6816.198
97.5% 15020.593 15475.12 15509.989 15423.25 15115.893 15208.657 15252.912
          [,50]     [,51]     [,52]     [,53]     [,54]     [,55]
2.5%  -6936.933 -6916.489 -6856.463 -7176.667 -7013.379 -7255.463
97.5% 15293.982 15414.855 15264.109 15364.337 15661.866 15537.266
```

Table of Confidence Intervals for Predicted Revenues in 1978(Fixed Median)

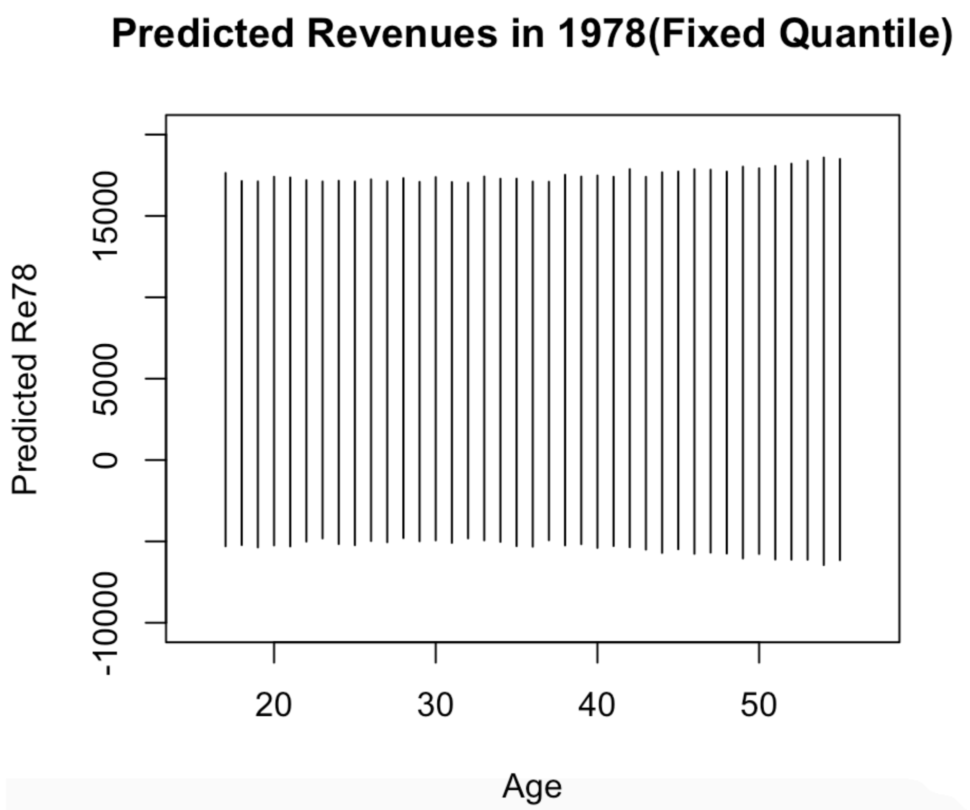
```
       [,14] [,15] [,16]     [,17]     [,18]     [,19]     [,20]    [,21]
2.5%      NA    NA    NA -5217.841 -4994.808 -5359.806 -5240.083 -4771.26
97.5%     NA    NA    NA 17314.702 17276.037 17342.326 17161.841 17211.84
          [,22]     [,23]     [,24]     [,25]     [,26]    [,27]     [,28]
2.5%  -5310.052 -4992.704 -5049.981 -4881.293 -5122.958 -4848.93 -5147.146
97.5% 17363.814 17322.154 17231.646 17192.657 17134.799 17321.13 17153.340
         [,29]     [,30]     [,31]     [,32]     [,33]     [,34]     [,35]
2.5%  -4793.92 -5061.749 -4852.703 -5226.298 -5246.929 -5272.722 -5152.235
97.5% 17416.65 17269.258 17538.646 17552.330 17348.530 17437.959 17301.378
         [,36]     [,37]     [,38]     [,39]     [,40]     [,41]     [,42]
2.5%  -5396.73 -5398.916 -5226.554 -5199.616 -5275.308 -5218.445 -5217.166
97.5% 17169.64 17394.870 17423.655 17644.714 17498.856 17343.481 17272.252
          [,43]     [,44]     [,45]     [,46]     [,47]     [,48]     [,49]
2.5%  -5323.718 -5497.535 -5482.606 -5621.285 -5861.223 -5828.517 -5897.163
97.5% 17549.730 17423.249 17721.412 17901.274 18080.399 17964.429 18252.224
          [,50]     [,51]     [,52]     [,53]     [,54]    [,55]
2.5%  -6208.732 -6240.545 -5983.171 -5930.477 -6452.278 -5994.13
97.5% 17980.199 18012.895 17992.053 18319.592 18577.951 18319.67
```

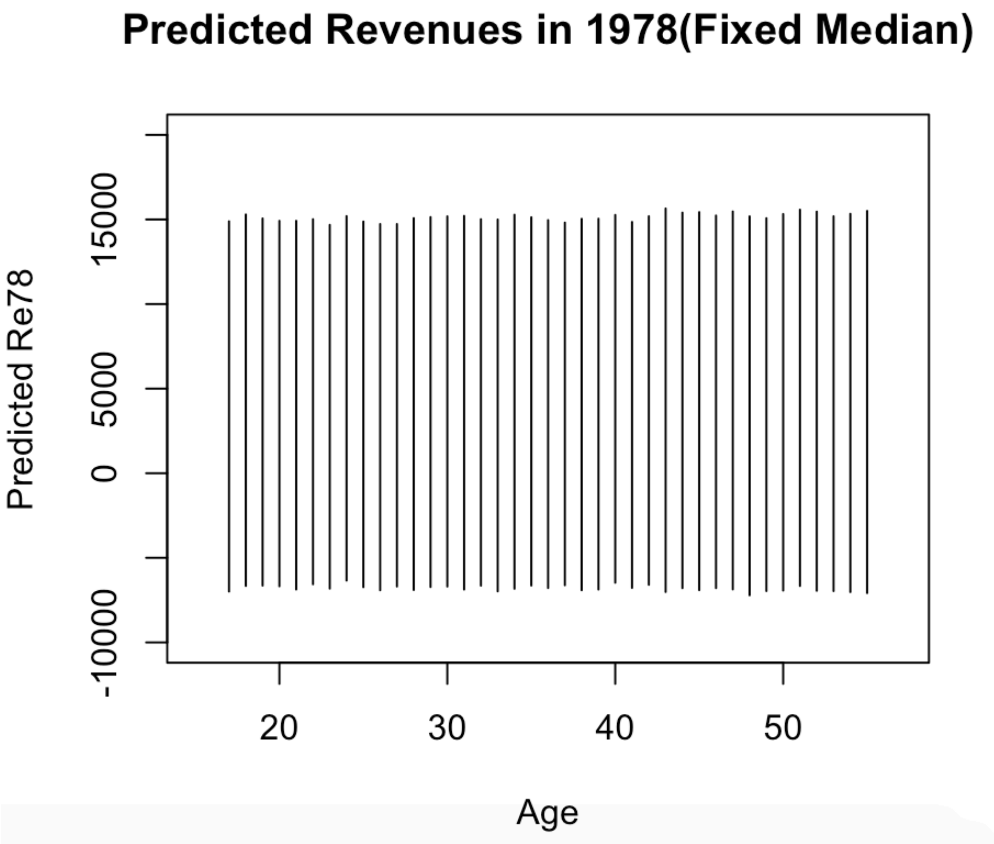
Table of Confidence Intervals for Predicted Revenues in 1978(Fixed Quantile)

	Value
Median of Educ	10
Median of Re74	0
Median of Re75	0
90% Quantile of Educ	12
90% Quantile of Re74	7628.052
90% Quantile of Re75	4492.998

Table of Medians and 90% Quantiles of The Other Predictors

2. 2 figures showing the scatterplots (one for the analysis holding predictors at their medians, and other for the analysis holding predictors at their 90% quantiles). The “scatterplots” don’t have to show the original data--all I am interested in are the prediction intervals for each age. Each of these figures should show how the prediction intervals’ change over time (i.e., over the range of ages in the data set). Be sure to label your plot’s features (axis, title, etc.).



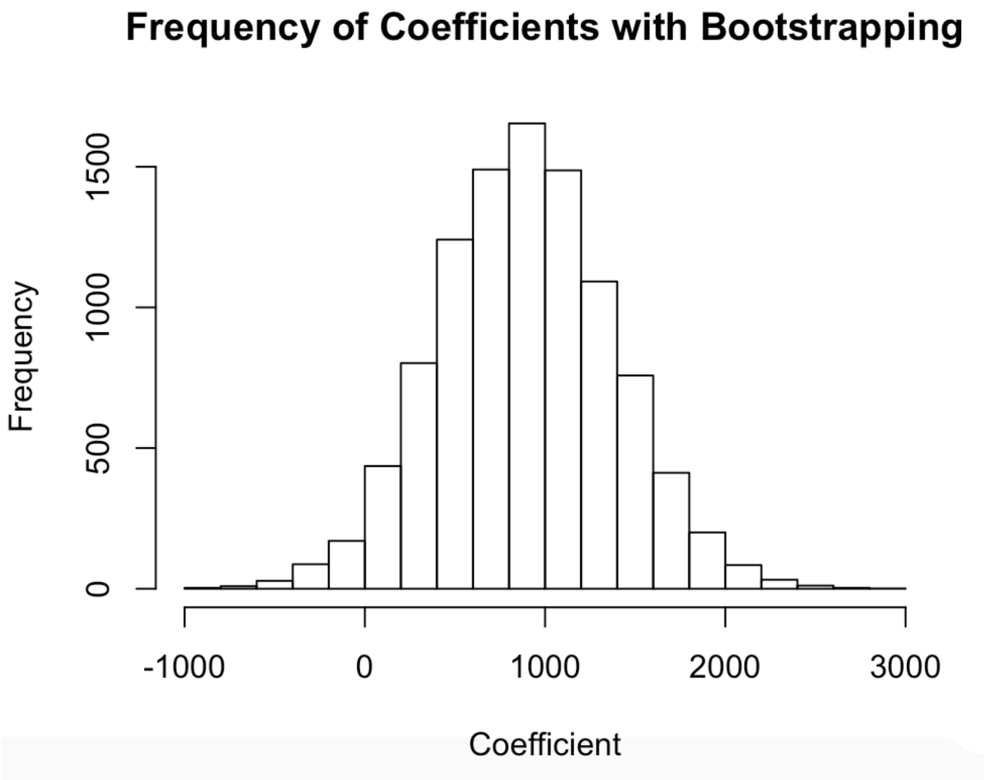


Question 3

1. A table with the relevant results (bounds on the 2 confidence intervals).

Type	2.5%	97.5%
Analytical Confidence Interval of Coeffient	-42.48831	1866.237
Bootstrap Confidence Interval of Coefficient	-40.52635	1813.134

2. 1 histogram (properly labeled) showing your bootstrap-sample results. How you do this one is up to you.



3. No more than 3 sentences summarizing the results and drawing any conclusions you find relevant and interesting.

From the analytical and bootstrapping methods, we get the 95% confidence intervals of coefficients: (-42.5,1866.2) and (-40.5,1813.1), which are quite similar. This activity makes me understand how to generate a large amount of samples with bootstrapping, and how bootstrapping decrease the error and generate normal distribution as the graph presents.

Question 4

Write a function (5 lines max) that takes Ys and predicted Ys as inputs, and outputs R2. Copy/paste an example using the nsw.dta data (from #3 above) that shows it working.

The function:

```
r_squared=function(y,predicted_y){
  mean_y=mean(y)
  return (sum((predicted_y-mean_y)**2)/sum((y-mean_y)**2))
}
```

The results we get by using our function to calculate the R-squared of our predicted re78:

```
> r_squared(re78,predict_re78)
[1] 0.004871571
```

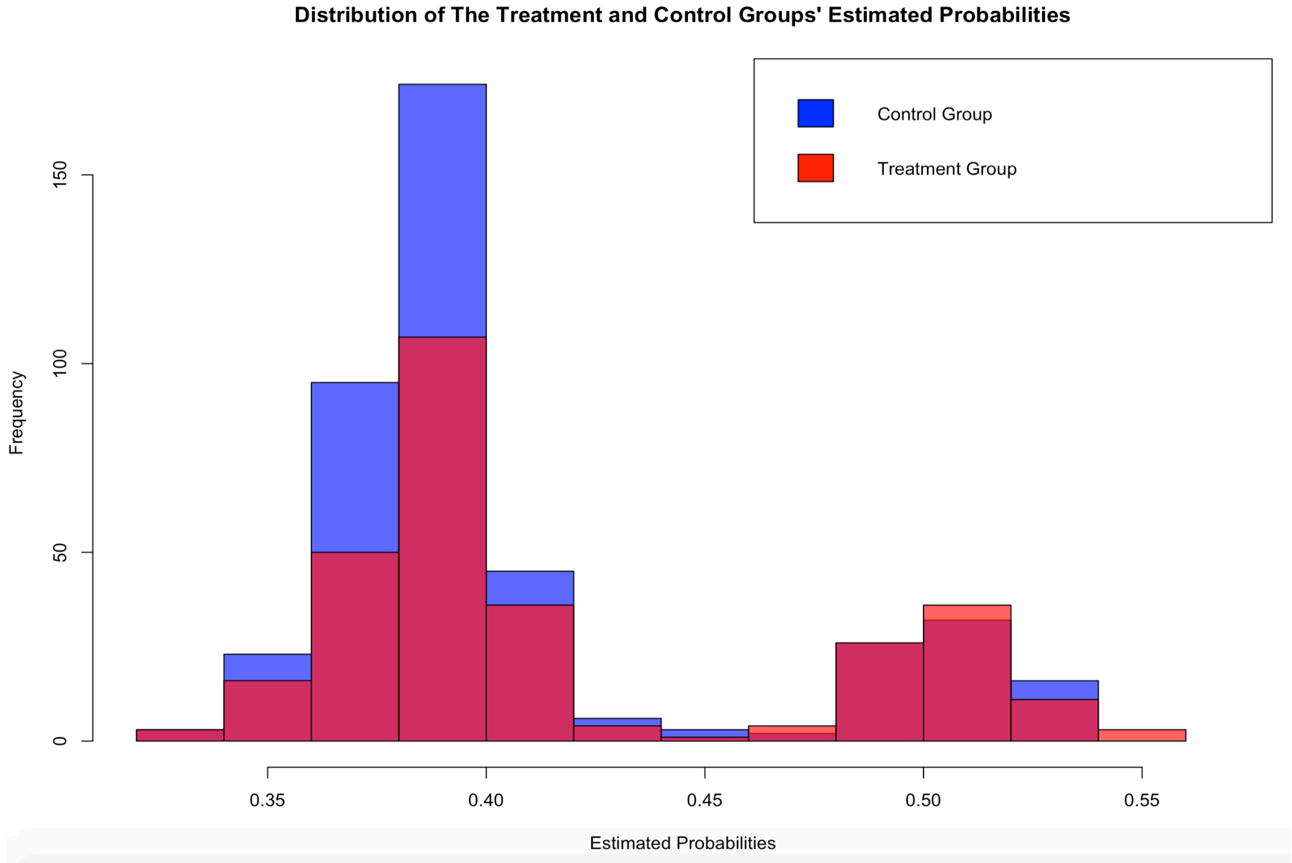
The result of R-squared we get from summary(lm) function:

```
Multiple R-squared:  0.004872
```

We can find that the results are the same.

Question 5

1. Two properly labeled histograms: one in red (showing the distribution of the treatment group's estimated probabilities) and one in blue (showing the distribution of the control group's estimated probabilities). Extra credit for a legend in the plot.



2. No more than 3 sentences summarizing the differences between the two distributions of estimated probabilities, and whether/not your results are surprising and/or intuitive.

Answer: From the graph, we can find that the overall shapes of the histograms for both control and treatment groups are similar(both lean to left), while the histogram for control group leans left more. This observation tells us that although most of the data in both control and treatment groups are predicted to be a data in control group, the data in control group is predicted more likely to be assigned to a data in control group. The annotation suggests that the treatment does influence the performance of the experimental objects, but the influence is not much obvious.