MINERVA SCHOOLS AT K.G.I.
Assignment 2
Quang Tran
CS112 Spring 2019

TABLE OF CONTENT

Link to code: https://gist.github.com/quangntran/bae9a75ae0f65df37143359614751cd2

Question 1.

a) Data generating equation:

The true underlying relationship between y and x is $y = 2x + 3$. Errors are drawn from a normal distribution $N(0, 10)$. Hence, the command in R is:

```
y <- 2*x + 3 + rnorm(n, mean = 0, sd = 10)
```

The outlier's x coordinate is 13 standard deviations above the mean of all x values, and its $y-$coordinate is 10 standard deviations below the mean of all the y values. The outlier is therefore at $(4.27546, -99.89386)$.

b) Regression results for the original 999

```
Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max 
-30.002  -6.745  -0.200   6.906  38.005 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   2.6519     0.6415   4.134 3.86e-05 ***
x             2.4868     1.1258   2.209   0.0274 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.36 on 997 degrees of freedom
Multiple R-squared:  0.00487, Adjusted R-squared:  0.003872 
F-statistic:  4.88 on 1 and 997 DF,  p-value: 0.0274
```

c) Regression results with the outlier included

```
Call:

lm(formula = new_y ~ new_x)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-96.705  -6.831  -0.085   6.941  39.947

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.6741     0.6395   7.309 5.52e-13 ***
new_x        -1.8390     1.0926  -1.683   0.0926 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 10.87 on 998 degrees of freedom
Multiple R-squared:  0.002831,  Adjusted R-squared:  0.001832
F-statistic: 2.833 on 1 and 998 DF,  p-value: 0.09264
```

d) Plot

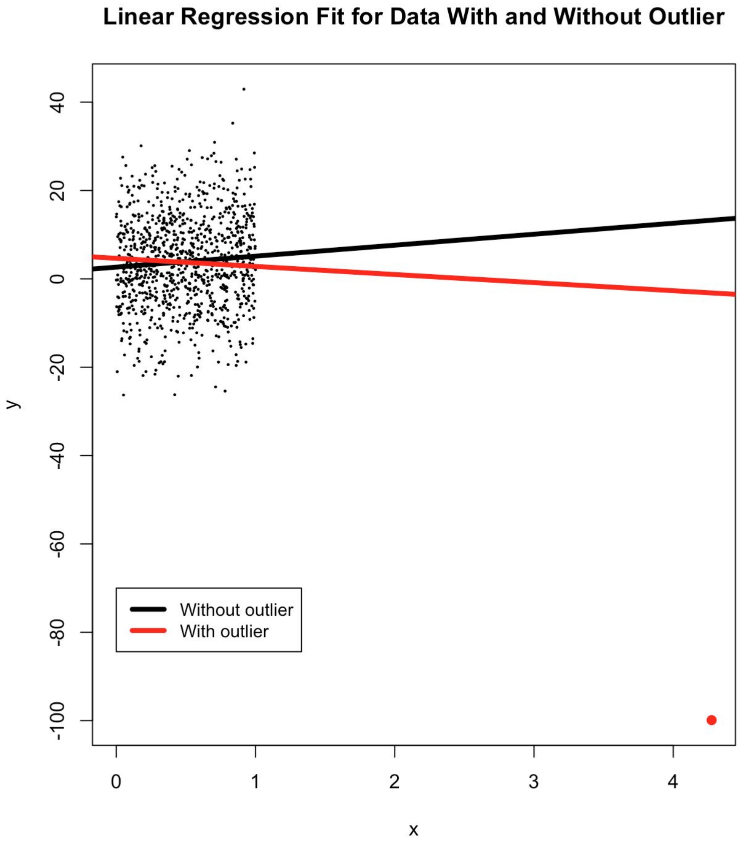

Figure 1. Linear regression fits when including and excluding an outlier (the red dot). When the outlier is omitted, the best fit line (the black line) has a positive slope (2.486); the presence of the outlier pulls the regression line down to a negative slope (the red line with slope -1.839). Using the regression line when the outlier is excluded (the black line) to predict data points that are outside the range of the corresponding data points ([0,1]) may fail to capture a possibly unseen, important relationship between out-of-range data point's x and y values and thus leading to bad predictions, while using the red line biases our prediction even for data points that have x in the range $[0, 1]$.

Question 2

(a)

	educ	re74	re75
Median	10	0	0

75% percentile	11	824.389	1220.84

Figure 2. Medians and 75% percentiles for educ, re74, and re75 in the lalonde dataset.

		17	18	19	20	21	22	23	24	25	26	27
median	2.5%	3096.255	3355.292	3587.199	3778.009	3923.996	4037.588	4107.917	4144.742	4168.835	4176.423	4187.013
	97.5%	5732.865	5646.636	5594.299	5582.602	5582.099	5635.521	5703.574	5799.354	5901.533	6007.046	6114.378
75%	2.5%	3543.029	3824.434	4061.186	4249.845	4409.602	4535.945	4622.213	4684.277	4726.917	4745.869	4766.214
	97.5%	6303.489	6218.436	6150.759	6114.228	6103.677	6122.254	6183.742	6252.471	6336.857	6434.005	6521.684

		28	29	30	31	32	33	34	35	36	37	38
median	2.5%	4209.278	4221.400	4236.694	4250.158	4262.267	4281.037	4287.693	4310.323	4308.406	4298.788	4271.058
	97.5%	6205.711	6305.523	6396.116	6488.121	6558.996	6614.187	6676.499	6746.048	6821.649	6902.740	6987.685
75%	2.5%	4786.585	4803.091	4812.326	4825.044	4846.487	4847.607	4858.162	4847.300	4836.826	4817.304	4775.190
	97.5%	6616.940	6698.236	6780.672	6859.684	6942.357	7010.858	7088.127	7172.086	7253.772	7351.728	7456.787

		39	40	41	42	43	44	45	46	47	48	49
median	2.5%	4238.088	4191.206	4101.393	3985.722	3841.803	3670.781	3473.765	3251.191	2996.463	2713.139	2390.208
	97.5%	7082.275	7203.035	7351.373	7517.878	7713.357	7921.094	8142.001	8376.424	8647.148	8934.724	9213.871
75%	2.5%	4720.235	4634.109	4534.711	4390.492	4236.659	4042.360	3838.969	3597.858	3313.154	3026.254	2740.712
	97.5%	7568.189	7702.133	7843.295	8015.378	8217.076	8419.576	8636.803	8867.455	9144.744	9442.298	9758.047

		50	51	52	53	54	55
median	2.5%	2049.932	1704.250	1342.562	948.0219	511.2469	76.26293
	97.5%	9534.597	9854.883	10183.529	10547.6095	10948.3364	11345.85502
75%	2.5%	2386.128	2020.783	1667.08	1255.118	818.7243	364.1253
	97.5%	10099.573	10428.067	10775.78	11140.404	11519.9104	11939.9633

Figure 3. 95% confidence intervals for expected value of re78 for age groups ranging from 17 to 55 when educ, re74, and re75 are kept at 1) their medians and 2) their 75% percentiles.

		17	18	19	20	21	22	23	24	25	26	27
median	2.5%	-8826.607	-8137.612	-8560.551	-8294.848	-8094.796	-8109.50	-8362.617	-7974.677	-8111.624	-7648.68	-7625.524
	97.5%	17867.413	17938.978	17635.437	17705.140	17561.595	17939.43	18247.717	17856.327	18334.079	18005.43	18198.166
75%	2.5%	-7997.358	-8050.994	-7979.001	-8125.233	-7851.496	-7655.106	-7610.782	-7265.714	-7323.255	-7442.622	-7321.008
	97.5%	17642.124	18092.969	18318.963	18116.569	18129.823	18506.748	18378.340	18563.269	18531.698	18415.637	18593.050

		28	29	30	31	32	33	34	35	36	37	38
median	2.5%	-7880.23	-7740.098	-7715.504	-7639.468	-7464.947	-7258.553	-7404.345	-7681.604	-7526.585	-6963.419	-7281.132
	97.5%	18380.71	18009.773	18150.272	18188.548	18585.642	18490.253	18544.994	18906.480	18805.058	18588.561	18377.998
75%	2.5%	-7580.553	-7580.762	-7129.693	-7300.275	-6970.496	-6868.584	-6416.037	-7388.383	-7038.806	-6956.719	-6958.159
	97.5%	18578.024	19044.380	19034.625	18798.866	18973.826	18902.684	19300.659	18973.036	18871.230	19159.782	19061.417

		39	40	41	42	43	44	45	46	47	48	49
median	2.5%	-7244.797	-7429.747	-7049.068	-7308.62	-6956.779	-7255.828	-7459.452	-7762.523	-7828.432	-7565.573	-7907.243
	97.5%	18682.172	18748.747	18744.520	18976.00	18947.395	19021.265	18771.351	18960.829	19191.724	19108.523	18742.772
75%	2.5%	-7115.555	-6853.477	-6835.342	-6917.815	-6799.321	-6732.563	-7112.753	-6955.413	-6757.736	-7044.35	-7179.913
	97.5%	19022.770	19080.032	19099.920	19127.689	19374.352	18891.603	19451.520	19527.490	19690.465	19332.83	19682.734

		50	51	52	53	54	55
median	2.5%	-7440.355	-7548.939	-7890.491	-8087.044	-8304.093	-8323.85
	97.5%	19423.133	19261.523	19546.649	19422.253	19651.171	19735.48
75%	2.5%	-7444.565	-7390.775	-7536.117	-7557.513	-7578.784	-8281.057
	97.5%	19901.022	19891.809	19968.690	19908.740	20025.507	20238.983

Figure 4. 95% prediction intervals of re78 for age groups ranging from 17 to 55 when educ, re74, and re75 are kept at 1) their medians and 2) their 75% percentiles.

(b)

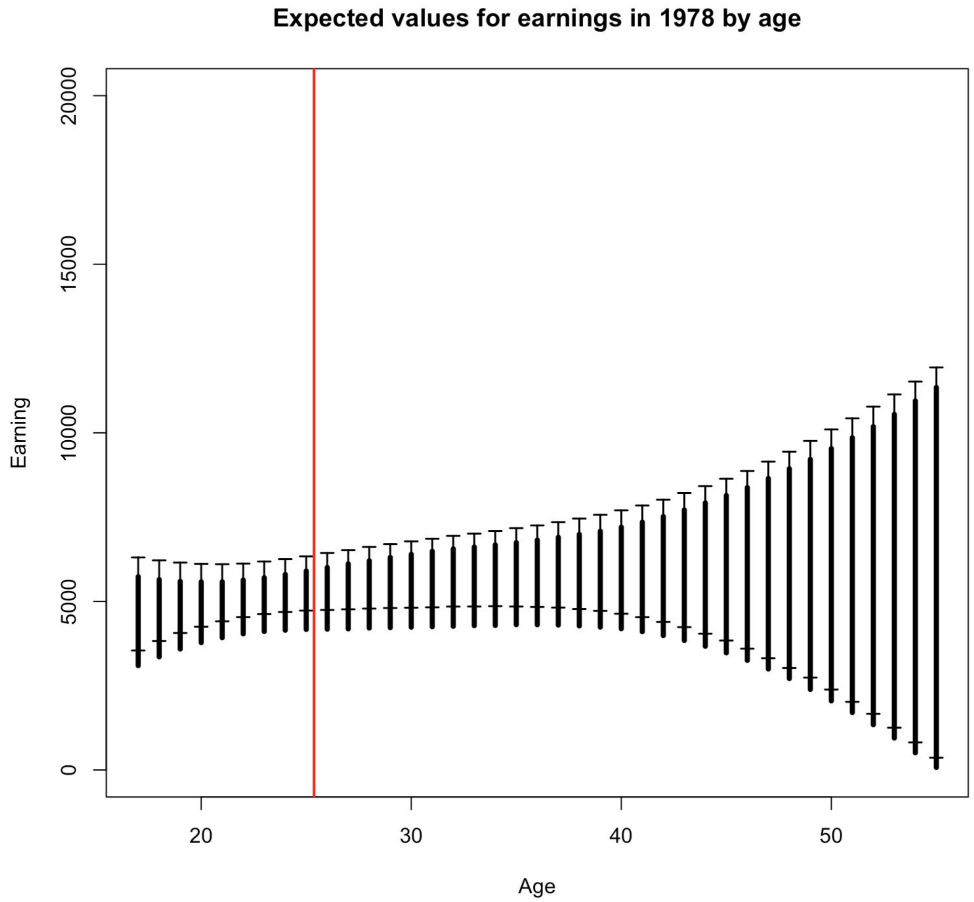

Figure 5. Expected values for earnings in 1978 by age. The bolded intervals without bar heads are for expected re78 when other predictors are kept at their median. The inverals with bar heads are for those when other predictors are kept at their 75% quantile. The vertical red line indicates the mean age (25.3) from the data set. The intervals of both types similar in range for each age, but intervals of the median group are consistently higher (looking like they are up shifted) than those of the 75% quantile group. The intervals are larger for ages that are further from the mean age. We are more certain our estimates with data point closer to the mean.

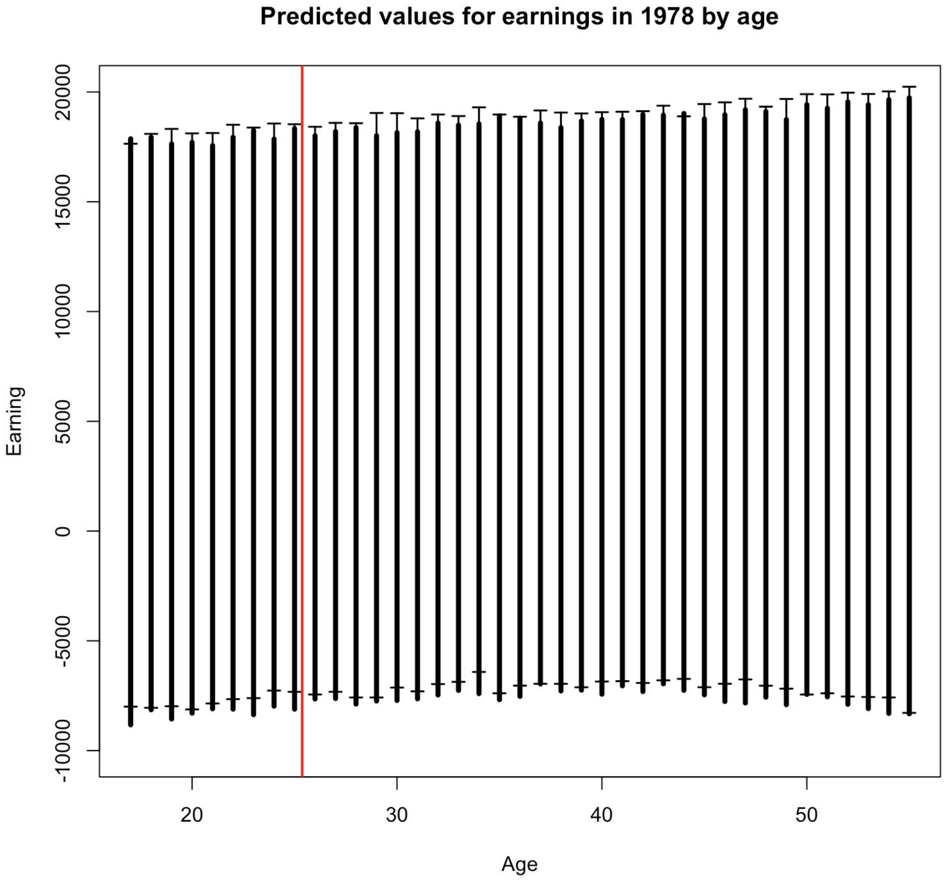

Figure 6. Predicted values for earnings in 1978 by age. The bolded intervals without bar heads are for when other predictors are kept at their median. The inverals with bar heads are for when other predictors are kept at their 75% quantiles. The vertical red line indicates the mean age (25.3) from the data set. The prediction intervals are larger than the confidence intervals for expected values as the former incorporate both fundamental and estimation uncertainties, while the latter neglect the fundamental uncertainty. The prediction intervals also include negative values. There is not a clear and obvious trend of increasing intervals as going further from the mean age as observed in confidence intervals for expected values.

Question 3

	Original Data	Bootstrap
2.5%	-1.025	-0.947
97.5%	0.283	0.232

Figure 7. The 95% confidence intervals for the value of the coefficient for treatment using the PlantGrowth data set and using 10,000 bootstrapped samples.

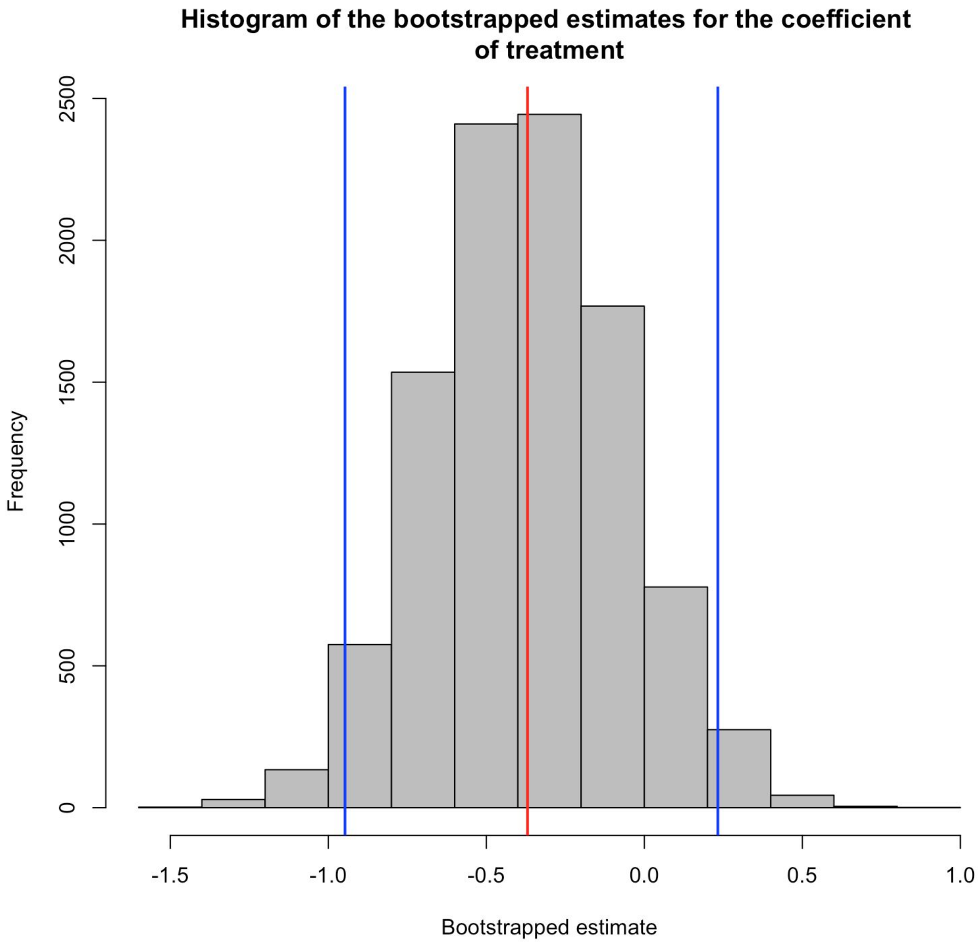

Figure 8. Histogram for the bootstrap estimates of the treatment's coefficient using 10,000 bootstrapped samples. The red line at -0.369 indicates the mean of the estimates. The blue lines represent the lower bound and upper bound of the estimate confidence interval.

Conclusions:

- The distances between the mean and the two bounds are roughly the same (0.57 and 0.60). This is because the distribution of the estimated coefficients are roughly normal. We have also observed that the distribution increasingly resembles a normal one as the number of bootstrapped samples used increased.

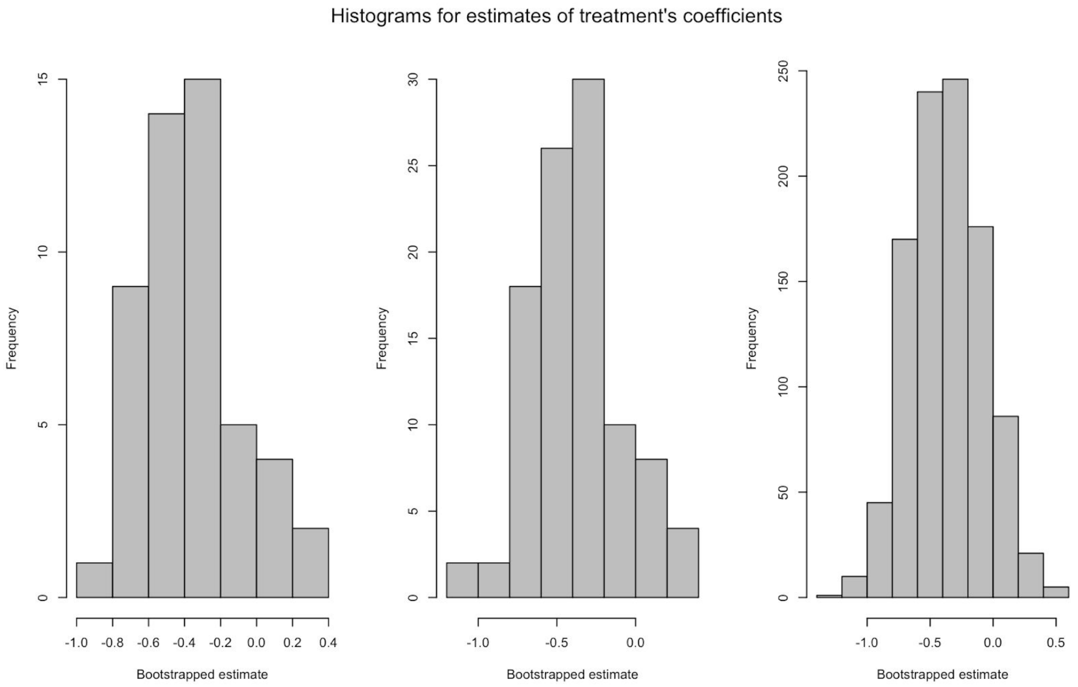

Figure 9. Histogram for the bootstrap estimates of the treatment's coefficient using 10 (left), 100 (center), and 1000 (right) bootstrapped samples.

- The analytical interval is larger than the bootstrap one.
- The average of all the bootstrapped estimates tend toward the coefficient for treatment obtained using the original data, as the number of bootstrapped samples used increases. The length of the confidence intervals tend to stabilize as more bootstraps used.

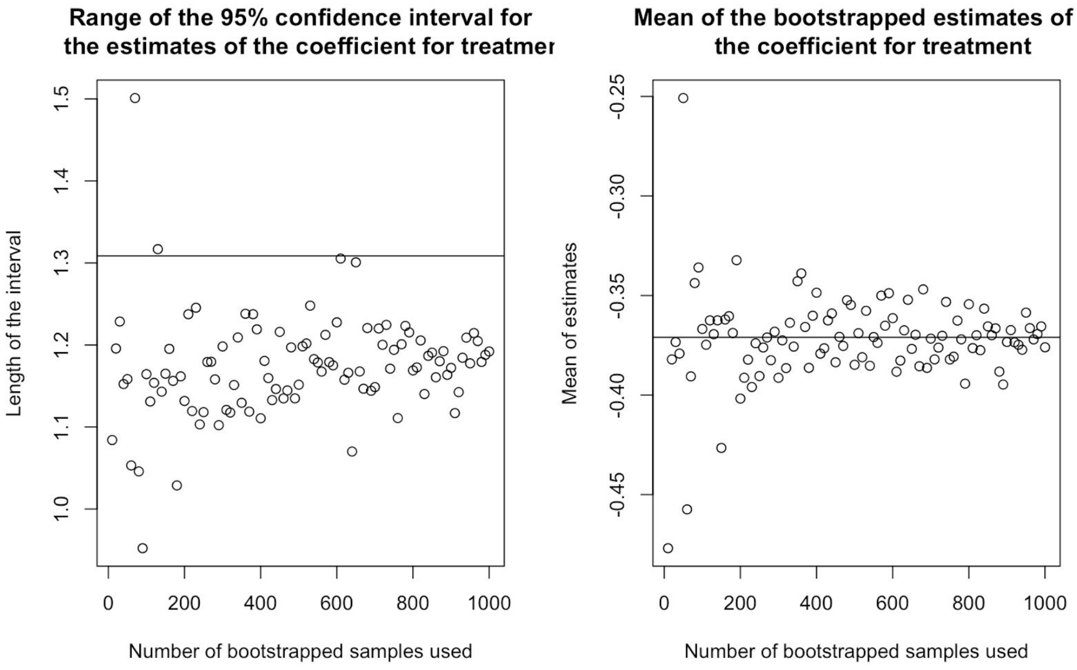

Figure 10. Scatter plots for 1) the range of the 95% confidence intervals of the coefficient (left) and 2) the average of the bootstrap estimates for the coefficient of the treatment (right), for different number of bootstrap samples. In the left panel, the horizontal line indicates the range of the analytical interval (0.283--1.025=1.308). The lengths stabilize as the number of bootstraps increases. In the right panel, the horizontal line indicates the coefficient of the treatment using the linear regression model on the original data. The averages of the bootstrap estimates converge, and they also converge to this line.

Question 4.

```
> lm.fit3 <- lm(weight~group, data=new_df)
> rsquared(new_df$weight, predict.lm(lm.fit3))
[1] 0.0730776
```

The result is the same as the one output by the software:

```
> summary(lm.fit3)$r.squared
[1] 0.0730776
```

Question 5.

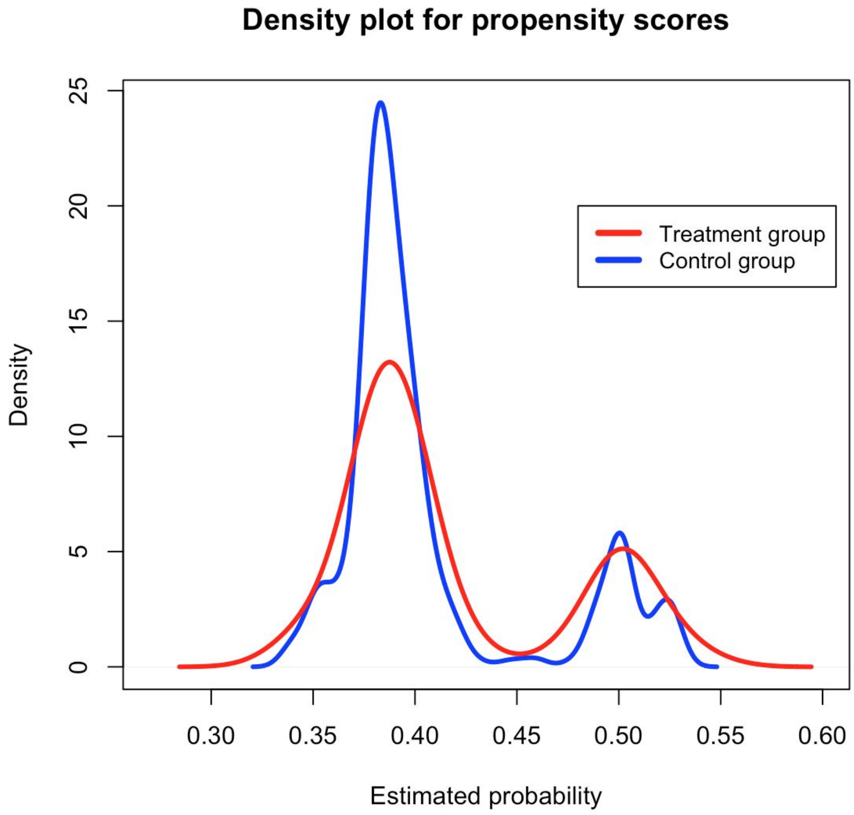

Figure 11. Propensity score density for treatment group and control group.

1. The shapes of the two distributions look very similar. That is, there are two bumps at the very similar positions (~0.36 and 0.5, respectively). This shows that the propensities of the two groups are similar and thus show that the data is randomized well (there is no bias towards any group to receive the treatment). This is good for causal inference, and the analysis on this data set would be close to that of an RCT (or maybe the data set is from an RCT itself.)
2. The higher bump at 0.35 for the control group can be attributed to the fact that there are much more controls than treated units (there are 297 units that are treated and 425 units that are controlled). This may be common in RCT as treatment may be expensive to administer.
3. What surprises me is the stark contrast of the smoothness between the two groups' distributions.