MINERVA®

Regression and Bootstrapping

CS112 - Fall 2018

Taha Bouhoun

Problem 1:

(a) Data generating equation:

```
set.seed(123)
age = floor(runif(99, 18, 50))
revenue = c()
for (i in 1:99){
  revenue[i] = 2000 * age[i] + rnorm(1, 0, 10000)
}
data = data.frame(age, revenue)
```

(b) Summary of the linear model for the 99 observations:

```
> summary(model1)
Call:
lm(formula = revenue ~ age, data = data)
Residuals:
     Min       1Q   Median       3Q      Max
-23983.5  -6465.2   -918.5   5719.2  20907.2
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4903.46    3388.75   1.447    0.151
age          1880.24      97.68  19.249   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 8857 on 97 degrees of freedom
Multiple R-squared:  0.7925,   Adjusted R-squared:  0.7904
F-statistic: 370.5 on 1 and 97 DF,  p-value: < 2.2e-16
```

(c) Summary of the linear model for the 100 observations:

```
> summary(model2)
Call:
lm(formula = revenue ~ age, data = new_data)
Residuals:
    Min      1Q  Median      3Q     Max
-839456  -37739    8994   52687  101081
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 189596.1    33656.8   5.633 1.69e-07 ***
age          -3890.4      943.5  -4.124 7.82e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 100700 on 98 degrees of freedom
Multiple R-squared:  0.1479,   Adjusted R-squared:  0.1392
F-statistic:    17 on 1 and 98 DF,  p-value: 7.822e-05
```

(d) Data visualization scatterplot:

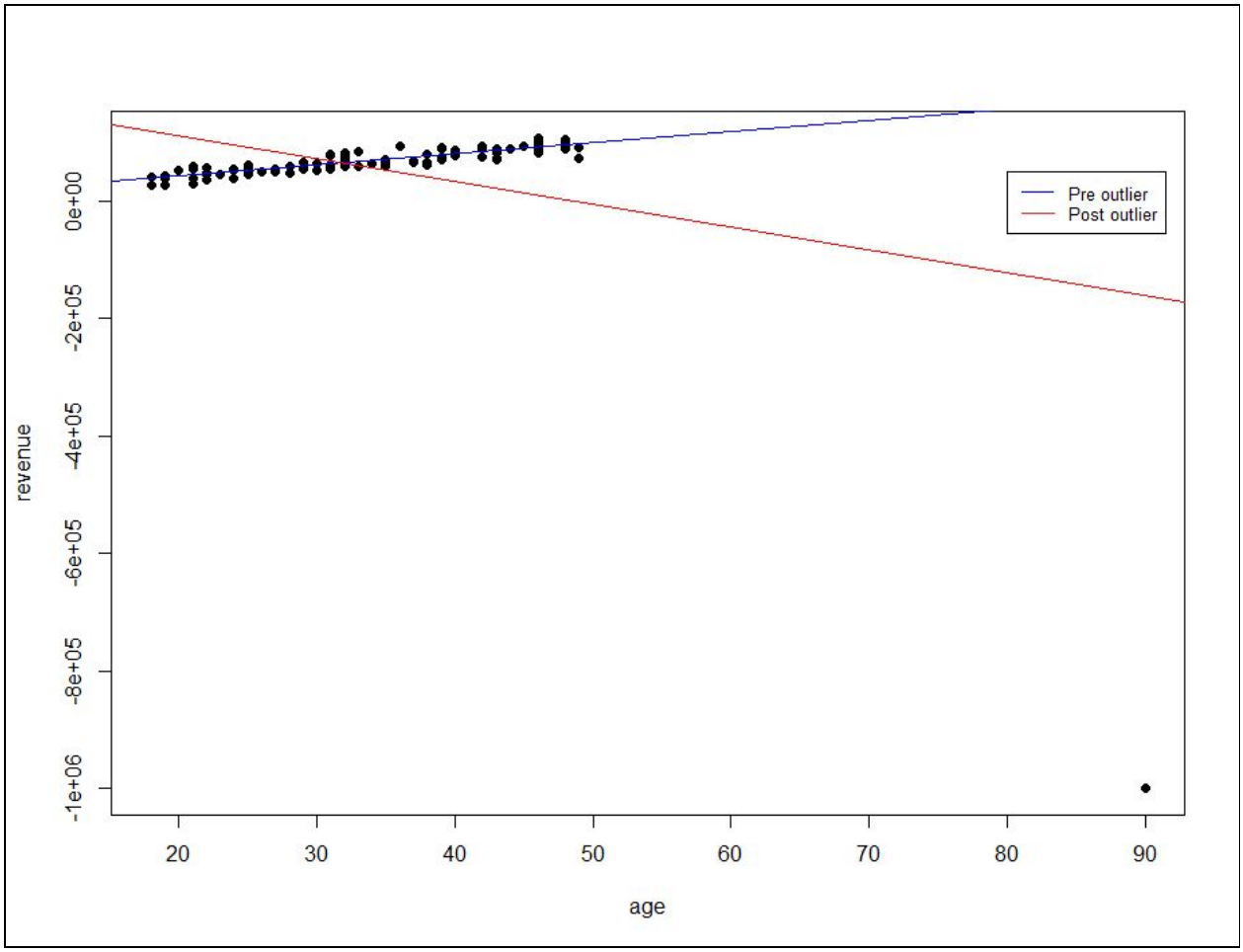

Fig 1. Scatterplot shows the relationship between revenue and age from randomly generalised dataset, along with the linear regression lines that illustrates the effect of outliers and the danger of extrapolation.

(e) Dangers of extrapolation:

The scatterplot shows the relationship between age and revenue before and after including an outlier, extrapolation based on the first model might lead to misleading inferences about the relationships between the input and outcome.

Problem 2:

(a) Tables:

median_edu	Median_re74	Median_re75
10	0	0

	17	18	19	20	21	22	23	24	25	26	27	28	29
2.5%	-6911.563	-6722.514	-6884.75	-6809.972	-6443.01	-6822.462	-6847.979	-6735.374	-6689.413	-6618.319	-6892.604	-6784.561	-6855.225
97.5%	15069.589	15082.222	15080.98	15299.231	14969.03	14788.252	15150.908	14938.170	15237.156	14947.423	14957.836	15099.271	14859.422

	30	31	32	33	34	35	36	37	38	39	40	41	42
2.5%	-6906.612	-6936.75	-6802.007	-7093.354	-6752.024	-6721.694	-6651.749	-6976.758	-6934.259	-7069.494	-6651.929	-6910.07	-7100.277
97.5%	15112.895	15143.59	14918.713	14979.522	15017.666	15335.071	15104.563	15078.429	14978.655	14976.395	15430.274	15185.92	15144.626

	43	44	45	46	47	48	49	50	51	52	53	54	55
2.5%	-7138.255	-6882.408	-6940.351	-6962.39	-7065.058	-6707.802	-6916.989	-7182.703	-6980.219	-7118.776	-7157.175	-7151.466	-7083.325
97.5%	15250.973	15320.828	15119.971	15258.03	15247.956	15414.844	15401.712	15382.047	15412.831	15738.509	15495.898	15353.662	15550.614

Quantile90_edu	Quantile90_re74	Quantile90_re75
12	7628.052	4492.998

	17	18	19	20	21	22	23	24	25	26	27	28	29
2.5%	-5156.297	-5121.328	-5161.663	-5021.174	-4983.256	-4841.839	-5012.852	-5129.074	-5290.042	-4858.19	-4953.229	-4922.525	-4741.881
97.5%	17373.637	17063.267	17256.149	17297.205	17131.358	17018.531	17263.335	16991.745	17085.164	17376.02	17015.320	17411.844	17285.171

	30	31	32	33	34	35	36	37	38	39	40	41	42
2.5%	-5015.843	-5233.041	-5184.754	-4829.112	-4970.119	-4937.815	-5156.839	-5168.701	-5462.694	-5616.707	-5200.30	-5819.117	-5429.048
97.5%	17364.675	16999.562	17053.555	17044.392	17132.292	17524.557	17630.633	17309.427	17442.479	17490.178	17193.87	17426.721	17802.758

	43	44	45	46	47	48	49	50	51	52	53	54	55
2.5%	-5228.241	-5587.351	-5581.18	-5900.957	-5627.625	-5862.95	-5707.377	-6002.44	-6190.48	-6070.047	-6231.509	-6494.395	-6250.459
97.5%	17620.020	17723.600	17763.62	17894.557	17819.230	18131.24	18264.013	18362.10	18588.54	18226.266	18294.993	18552.857	18545.040

(b) Scatterplots of the simulated confidence intervals:

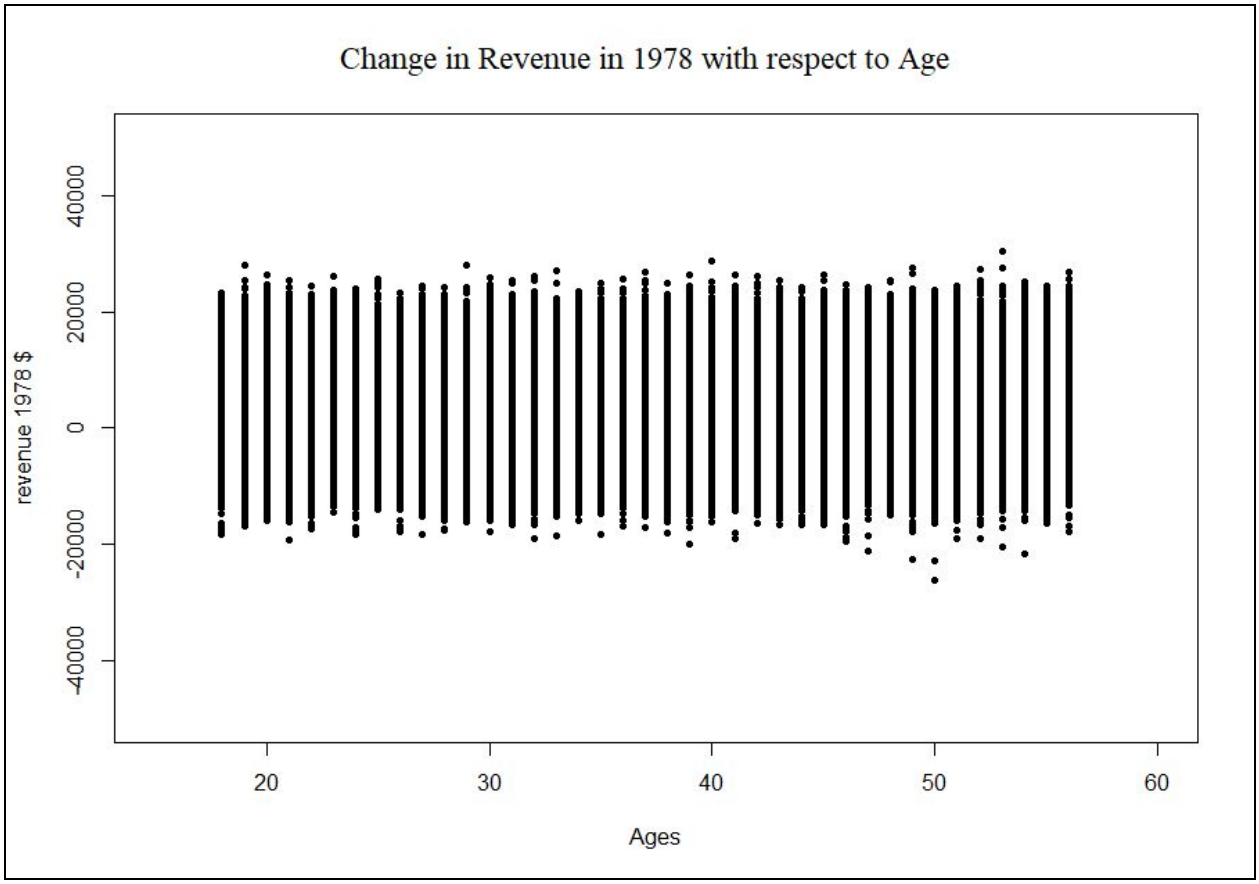

Fig 2. Scatterplot highlights simulation of the confidence intervals of revenues in 1978 and age (holding the other predictors at their median) from the control group of Lalonde dataset.

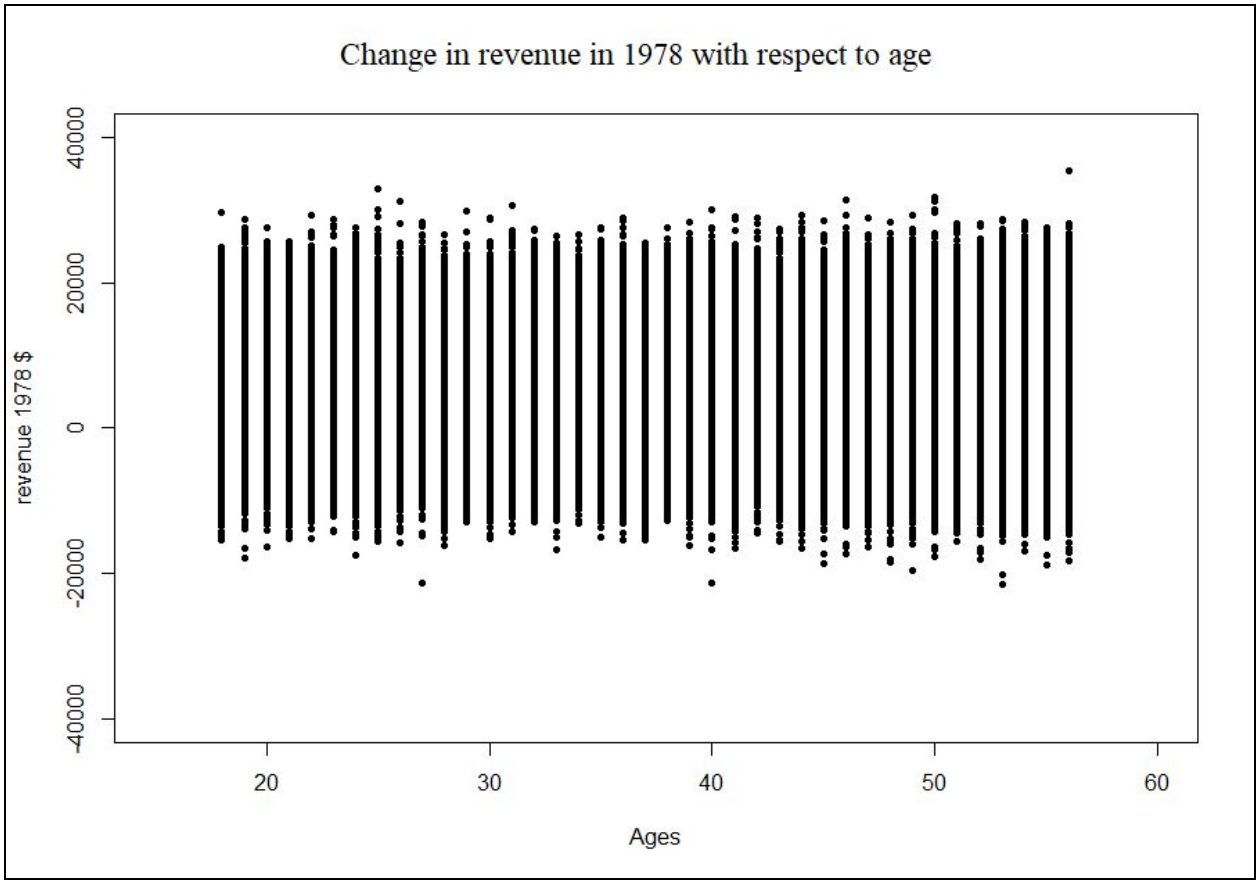

Fig 3. Scatterplot highlights simulation of the confidence intervals of revenues in 1978 and age (holding the other predictors at their 90% quantile) from the control group of Lalonde dataset.

Problem 3:

(a) Table of the relevant confidence intervals:

	Lower bound (2.5%)	Upper bound (97.5%)
Bootstrapping sample	**-40.8437**	**1852.7283**
Analytical	**-40.52635**	**1813.134**

*the results of the bootstrapping is based on setting the seed at 7

(b) Histogram of the bootstrap-sample results:

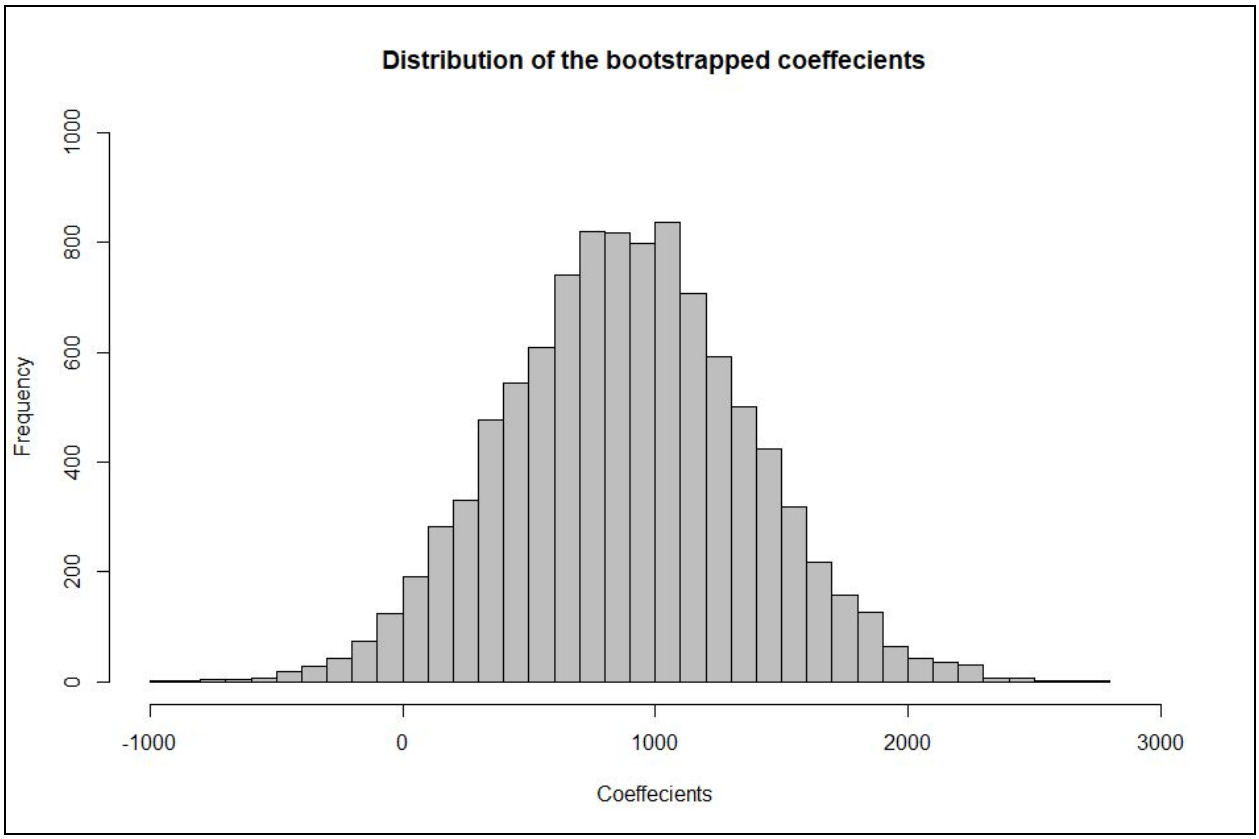

Fig 4. Histogram of the bootstrapping coefficient distribution for the linear model that outputs revenue ini 1978 with treatment status as a predictor.

(c) Summarizing the results:

The results of the bootstrapped confidence intervals are close the analytical results. It's interesting to mention that the distribution of the bootstrapped coefficients are nearly normal and the coefficient (treatment effect) is more likely to be positive based on the confidence interval.

Problem 4:

(a) Function that outputs R^2

```
R2 <- function(ys, yhat){
  ym = mean(ys)
  sse = sum((ys - yhat)^2)
  ssto = sum((ys - ym)^2)
  return(1-(sse/ssto))}
```

(b) Testing the function on *nsw.dta*

```
> hats <- predict(model3, nsw)
> R2(nsw$re78, hats)
[1] 0.004871571

> summary(model3)$r.squared
[1] 0.004871571
```

Problem 5:

(a) Histograms of the probabilities of assignment:

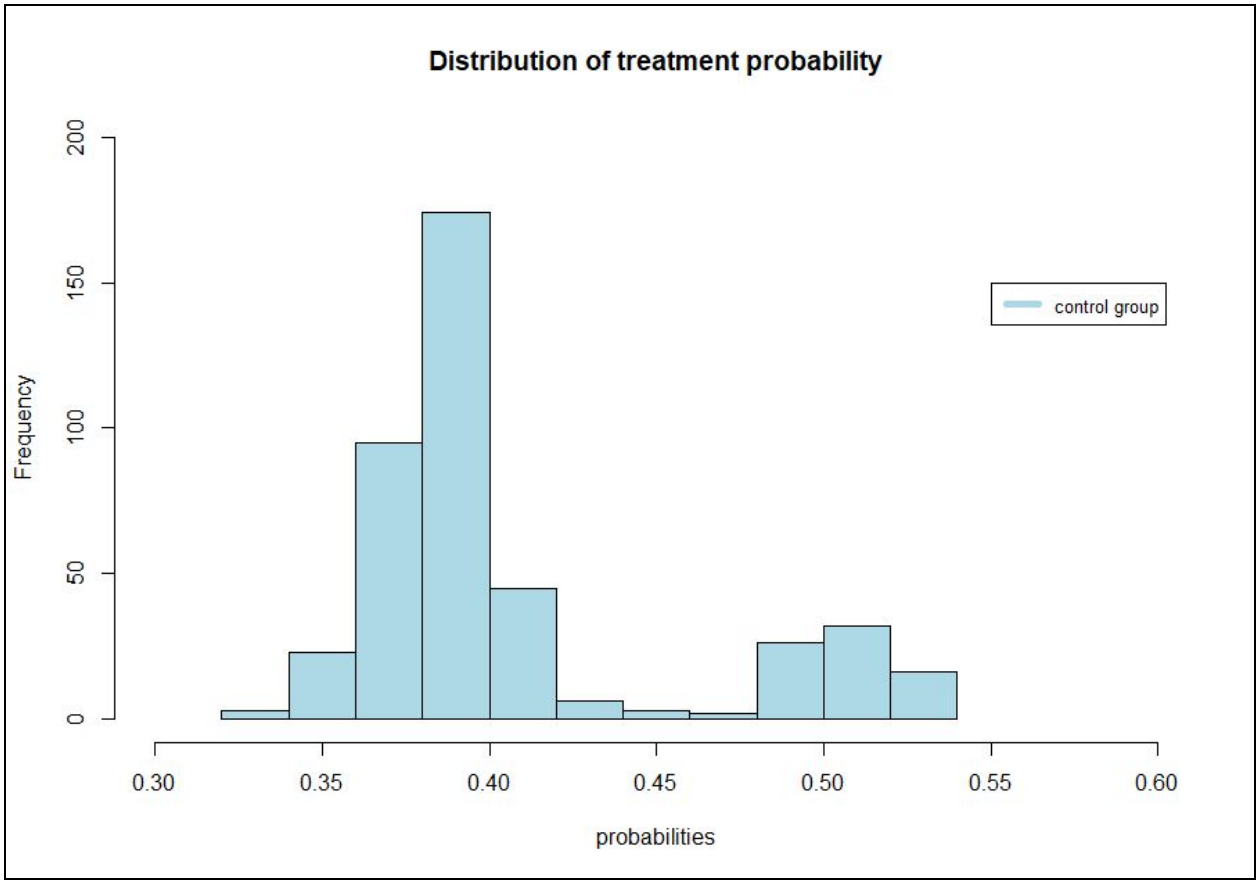

Fig 5. Estimated probabilities for being assigned to treatment for the control group.

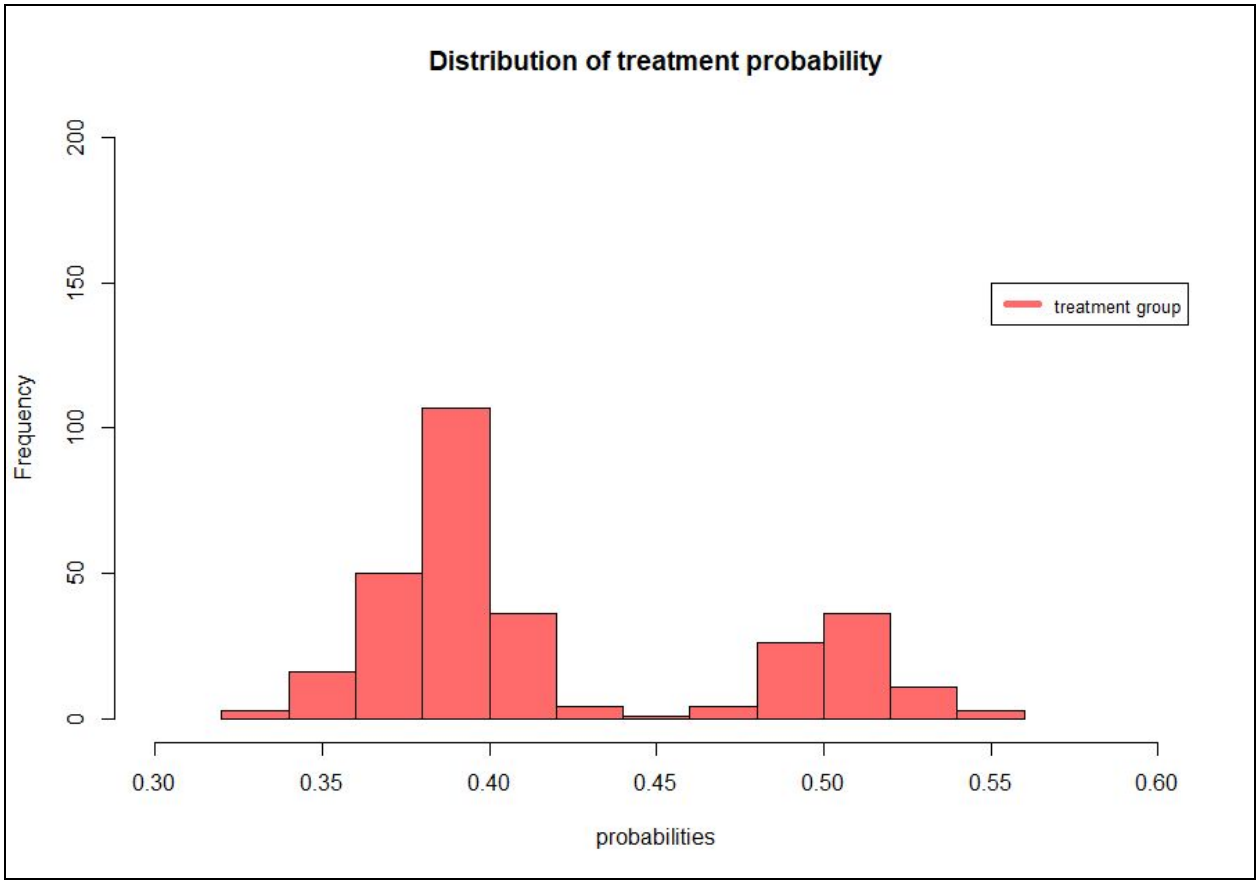

Fig 6. Estimated probabilities for being assigned to treatment for the treatment group.

(b) Comparing between the two histograms:

After fixing the same dimensions for both histograms, we can conclude that:

- Treatment group is smaller than the control group.
- The two histograms have nearly the same distribution (bimodal)
- For both groups, the model predicted that most subjects would be assigned to control rather than treatment.

The results aren't surprising as it proves that the sampling was random and it didn't depend on any of the predictors.

Appendix:

Link to the assignment's code:

https://gist.github.com/Tahahaha7/5819d2a9f0e7f28552dba59903de7638