Assignment 2

Link to code:

https://github.com/Munchic/cs112/blob/master/Assignment%20II.R

Problem 1

(a) $w = 0.01h + \sigma$, where w represents the water consumption of a tree in cubic meters, h represents its height in meters and σ represents Gaussian noise with the mean of 0.05 and the standard deviation of 0.02

(b) Summary of regression results for the original 999 data points

```
Residuals:
      Min        1Q    Median        3Q       Max
-0.068699 -0.012448  0.000687  0.013234  0.053058
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 5.065e-02  1.384e-03   36.61   <2e-16 ***
Height      9.971e-03  7.761e-05  128.48   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 0.01943 on 997 degrees of freedom
Multiple R-squared:  0.943,        Adjusted R-squared:  0.943
F-statistic: 1.651e+04 on 1 and 997 DF,  p-value: < 2.2e-16
```

(c) Summary of regression results for the original 999 data points and 1 extreme outlier $(h = 30, w = -50)$

```
Residuals:
    Min      1Q  Median      3Q     Max
-50.142  -0.029   0.047   0.127   0.236
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.179771   0.113228   1.588    0.113
Height      -0.001254   0.006345  -0.198    0.843
Residual standard error: 1.591 on 998 degrees of freedom
Multiple R-squared:  3.913e-05,        Adjusted R-squared:  -0.0009628
F-statistic: 0.03905 on 1 and 998 DF,  p-value: 0.8434
```

(d)

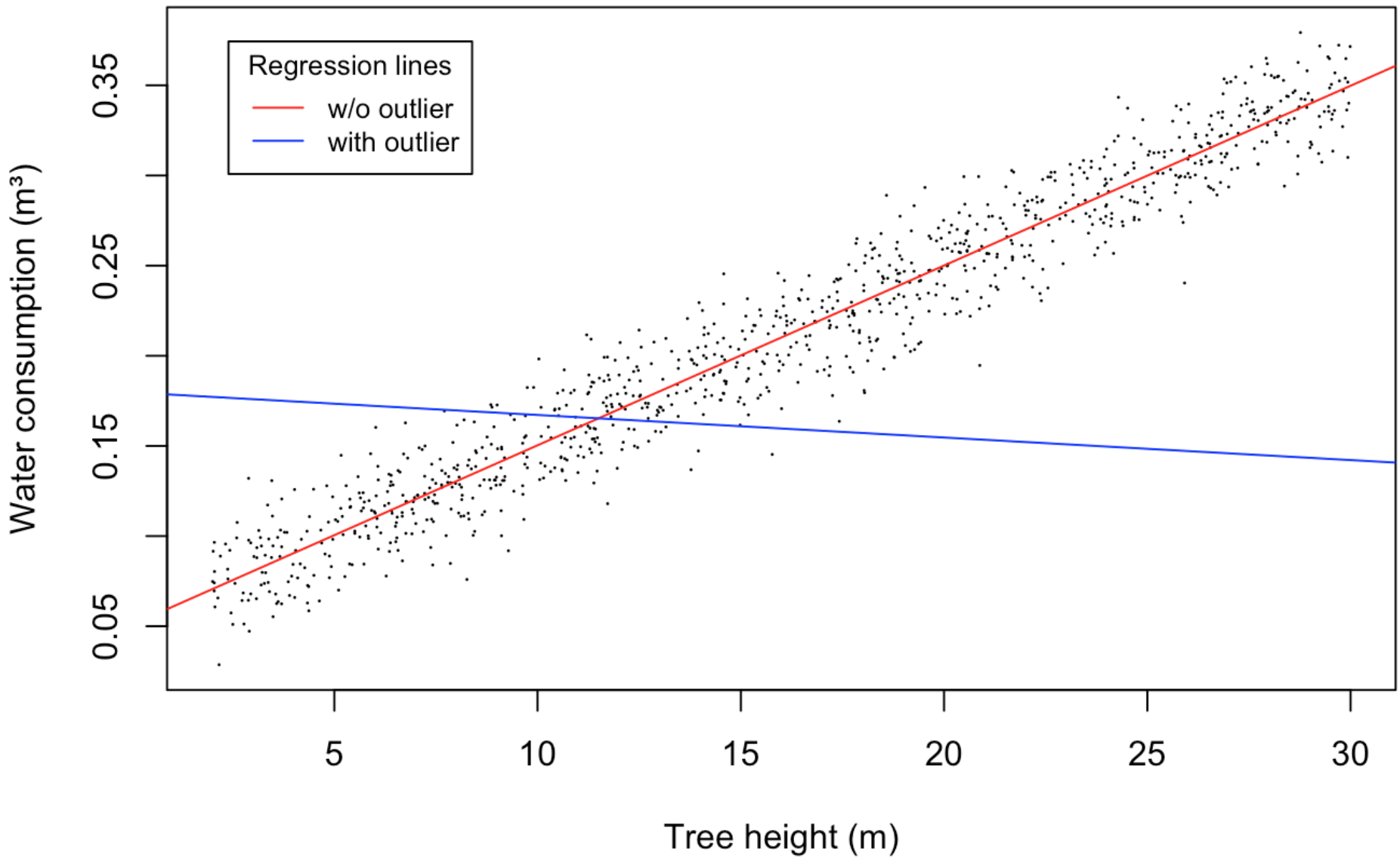

Fig. 1: Visualization of water consumption dependency on height in trees and trend lines showing the overall trend of 999 first points and the trend when the outlier point is included.

(e) As we can see in the experiment above, the general trend for trees should be that the taller a tree is, the more water it consumes. However, in a situation when we have an extreme outlier (e.g., faulty measuring device), we should observe the data carefully before trying to fit a model on the data and extrapolate. We can see that an outlier gave us a completely wrong trend line that would make our extrapolation on unobserved data erroneous.

Problem 2

(a) Confidence intervals for estimated values of **re78** for each age in lalonde dataset. Parameters **educ**, **re74**, **re75** are kept at the medians of their values.

```
        [,17]    [,18]    [,19]    [,20]    [,21]    [,22]    [,23]    [,24]
2.5%  3082.328 3173.319 3228.725 3257.613 3264.533 3242.890 3176.686 3113.182
97.5% 5838.362 5563.847 5323.567 5141.452 4993.536 4898.602 4838.824 4814.426
        [,25]    [,26]    [,27]    [,28]    [,29]   [,30]    [,31]    [,32]    [,33]
2.5%  3025.734 2937.644 2856.244 2797.047 2738.911 2699.09 2665.437 2637.316 2617.836
97.5% 4812.189 4838.055 4871.558 4903.095 4942.411 4987.73 5036.956 5092.854 5142.032
        [,34]    [,35]    [,36]    [,37]    [,38]    [,39]    [,40]    [,41]    [,42]
2.5%  2613.795 2610.696 2603.983 2611.223 2601.347 2586.920 2575.260 2518.351 2462.316
97.5% 5207.872 5292.458 5383.636 5492.588 5618.517 5770.432 5953.045 6172.444 6405.926
        [,43]    [,44]    [,45]    [,46]    [,47]    [,48]    [,49]    [,50]
2.5%  2400.295 2305.976 2201.347 2079.352 1962.892 1834.193 1688.871 1510.424
97.5% 6681.165 6994.627 7354.106 7736.020 8144.641 8598.798 9070.278 9578.134
         [,51]     [,52]      [,53]      [,54]      [,55]
2.5%   1343.059  1158.799   957.1523   749.8935   517.2429
97.5% 10093.756 10645.271 11205.9977 11801.3374 12408.4394
```

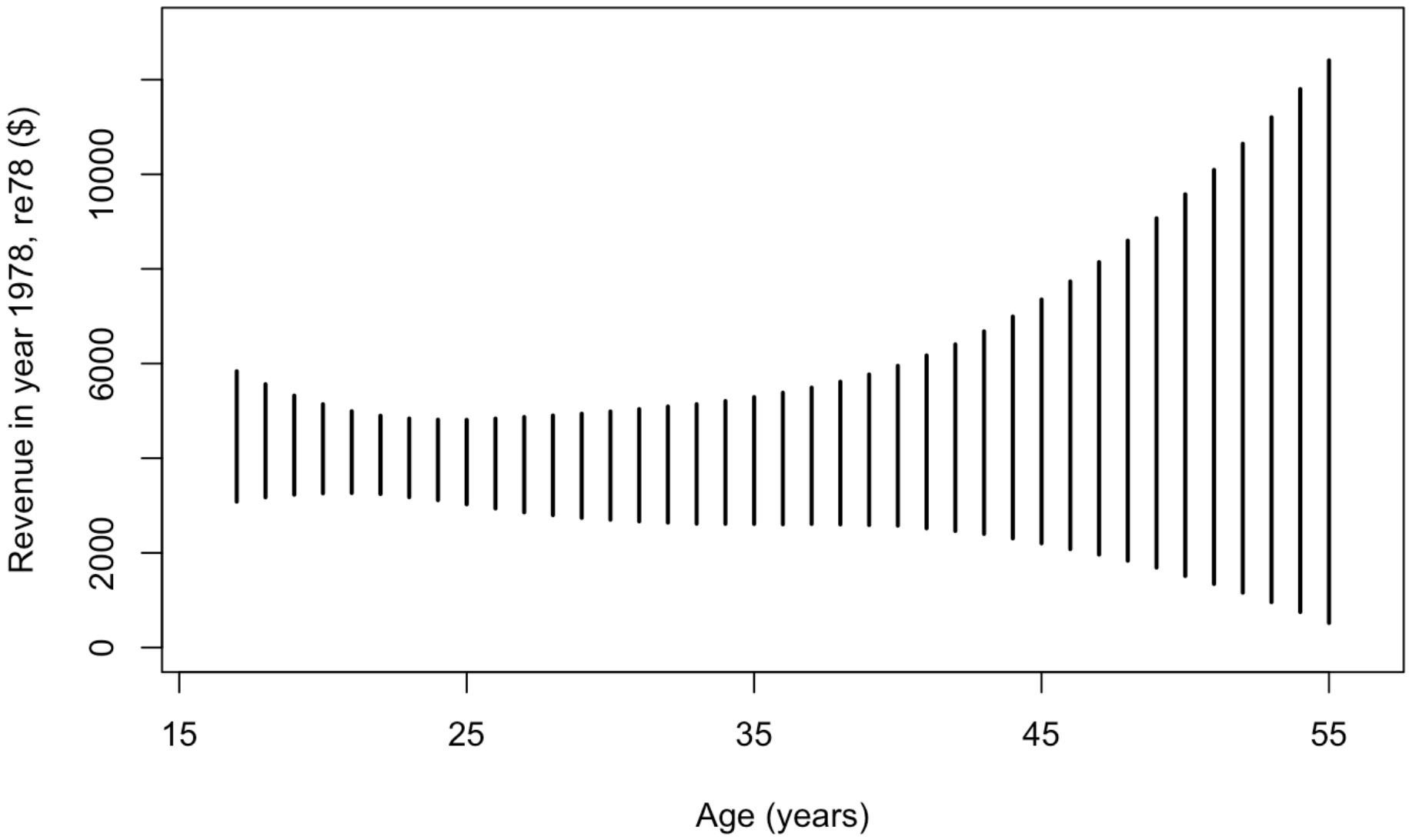

Fig. 2-1: Visualization of the confidence intervals described in the table above.

(b) Confidence intervals for estimated values of **re78** for each age in lalonde dataset. Parameters **educ**, **re74**, **re75** are kept at the 75% quantiles of their values.

```
         [,17]    [,18]    [,19]    [,20]    [,21]   [,22]    [,23]    [,24]
2.5%  3178.442 3268.677 3343.584 3387.170 3413.339 3412.52 3384.094 3334.698
97.5% 6213.251 5923.140 5682.316 5487.325 5342.345 5230.93 5163.839 5136.348
         [,25]    [,26]    [,27]    [,28]   [,29]    [,30]    [,31]    [,32]    [,33]
2.5%  3284.145 3225.671 3157.376 3096.317 3048.38 3001.187 2984.376 2956.541 2939.452
97.5% 5130.539 5134.956 5159.066 5198.471 5236.65 5278.270 5319.906 5380.320 5450.909
         [,34]    [,35]    [,36]    [,37]    [,38]    [,39]    [,40]    [,41]    [,42]
2.5%  2919.157 2908.606 2902.906 2895.011 2883.867 2863.617 2833.133 2778.442 2706.540
97.5% 5526.486 5620.027 5730.962 5851.140 6001.278 6178.718 6394.980 6628.314 6882.336
         [,43]    [,44]    [,45]    [,46]    [,47]    [,48]    [,49]     [,50]
2.5%  2654.892 2572.982 2479.414 2356.629 2226.363 2082.962 1945.089  1781.686
97.5% 7177.407 7493.528 7859.194 8250.722 8654.429 9081.426 9538.157 10068.149
         [,51]     [,52]     [,53]      [,54]      [,55]
2.5%   1591.59  1395.714  1219.536   998.2171   785.0984
97.5% 10610.18 11157.635 11738.446 12361.0853 12994.3558
```

(c) Confidence intervals for predicted values of **re78** for each age in lalonde dataset. Parameters **educ**, **re74**, **re75** are kept at the median of their values.

```
          [,17]     [,18]     [,19]     [,20]     [,21]     [,22]     [,23]     [,24]
2.5%  -6584.463 -6401.241 -6715.459 -6730.401 -6521.605 -6533.894 -7084.047 -7105.373
97.5% 15546.529 15153.406 15480.657 15107.591 15062.492 15070.724 15026.674 14978.602
          [,25]     [,26]     [,27]     [,28]     [,29]     [,30]     [,31]     [,32]
2.5%  -6923.564 -6926.516 -7082.141 -7119.689 -7427.565 -7085.777 -7066.803 -6798.886
97.5% 14586.843 14782.418 14679.167 14963.294 14798.700 14445.794 14675.625 14751.441
```

```
          [,33]     [,34]     [,35]     [,36]     [,37]    [,38]     [,39]     [,40]
2.5%  -6905.701 -7094.634 -7026.992 -7143.753 -6995.483 -6979.50 -6705.096 -6814.899
97.5% 14990.215 14830.796 15125.200 14778.777 15043.552 15028.99 14851.195 15105.474
          [,41]     [,42]     [,43]     [,44]     [,45]     [,46]     [,47]     [,48]
2.5%  -6671.82 -6455.179 -6764.099 -6673.873 -6542.782 -5942.453 -6298.096 -6273.239
97.5% 15497.34 15376.468 15609.800 15713.353 15969.642 16053.679 16220.890 16346.487
          [,49]     [,50]     [,51]     [,52]     [,53]     [,54]     [,55]
2.5%  -6021.98 -6153.612 -6182.086 -5782.019 -6009.703 -6040.258 -5971.774
97.5% 16904.14 16824.798 17345.575 17748.115 17621.846 18184.867 18796.903
```

(d) Confidence intervals for predicted values of **re78** for each age in lalonde dataset. Parameters **educ**, **re74**, **re75** are kept at the 75% quantiles of their values.

```
          [,17]     [,18]     [,19]     [,20]     [,21]    [,22]     [,23]     [,24]
2.5%  -6310.024 -6289.663 -6607.508 -6488.818 -6532.356 -6930.18 -6649.508 -6701.028
97.5% 15099.216 15336.291 15395.419 15424.879 15248.927 15353.77 15152.169 15087.898
          [,25]     [,26]     [,27]     [,28]     [,29]     [,30]     [,31]     [,32]
2.5%  -6864.134 -6696.118 -6896.632 -6624.876 -6732.413 -6842.644 -6670.164 -6870.014
97.5% 15130.341 15072.223 15237.399 14869.229 15095.770 14846.473 15086.881 15382.504
          [,33]     [,34]     [,35]    [,36]     [,37]    [,38]     [,39]     [,40]
2.5%  -6804.742 -6627.224 -6565.683 -6735.09 -6668.053 -6677.62 -6417.439 -6366.435
97.5% 15355.483 14799.682 15095.312 15303.61 15385.594 15093.76 15380.521 15393.578
          [,41]     [,42]     [,43]    [,44]     [,45]    [,46]     [,47]     [,48]
2.5%  -6281.389 -6255.979 -6237.162 -6060.56 -6072.505 -6015.75 -5935.556 -5626.214
97.5% 15593.739 15707.562 16071.514 16107.43 16160.083 16511.52 16489.703 16941.496
          [,49]    [,50]     [,51]     [,52]     [,53]     [,54]     [,55]
2.5%  -5701.38 -5617.27 -5734.906 -5417.691 -5911.211 -5926.523 -5791.233
97.5% 17329.00 17810.55 17639.415 18157.143 18566.154 18643.752 19361.417
```

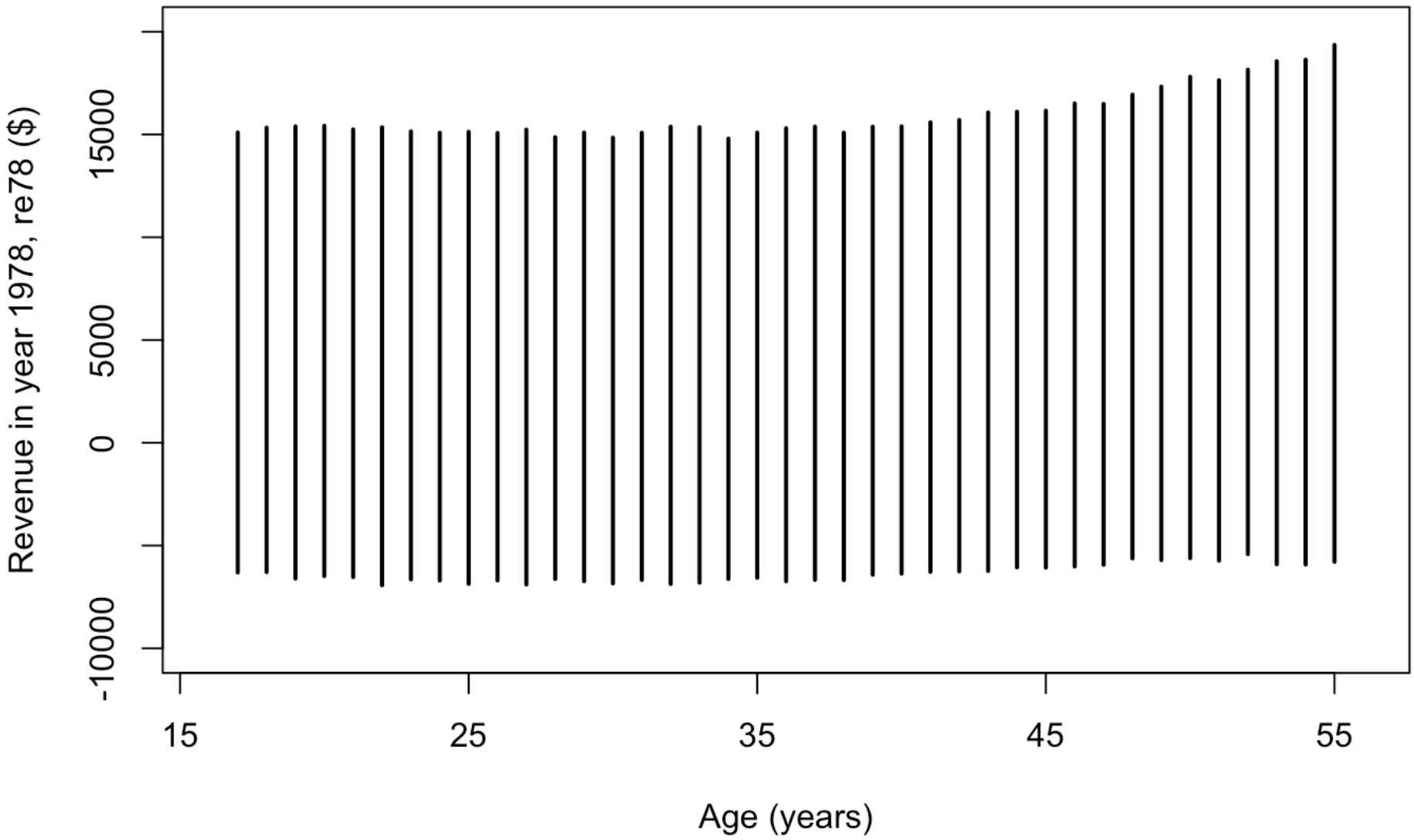

Fig. 2-2: Visualization of the confidence intervals of predicted values described in the table above.

Problem 3

	Intercept	Slope
2.5%	4.56934	-1.02530
97.5%	5.4946602	0.2833003

Table 1-1: Analytically obtained confidence intervals for intercept and slope of regression line of weight dependency on treatment1 in the PlantGrowth dataset.

	Intercept	Slope
2.5%	4.688997	-0.9445979
97.5%	5.388579	0.2270725

Table 1-2: Confidence intervals obtained by bootstrapping the observations 10,000 times for intercept and slope of the same regression line of weight dependency on treatment1 in the PlantGrowth dataset.

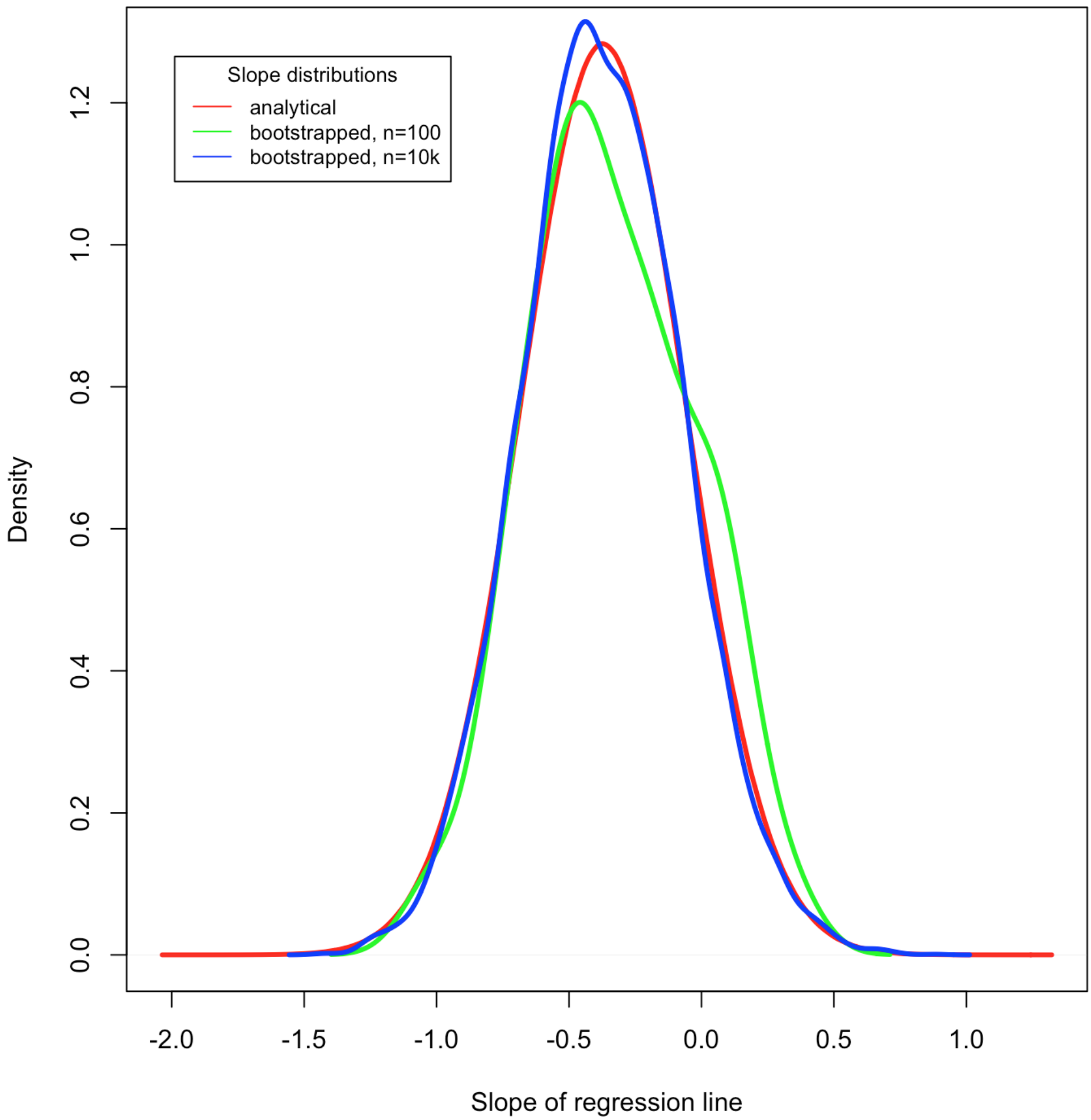

*Fig. 3: Comparison of analytically derived and bootstrapped distributions of slopes in the dependency of plant **weight** on **treatment1**.*

In the above tables and figure, we can see that the bootstrapped values of the linear model parameters are quite close to their analytical estimates. In fact, as we increase the number of resamples, the distribution of bootstrapped estimates gets more similar to the form of the analytically derived one.

Problem 4

```
r.squared <- function(y.act, y.pred) {
  y.mean = mean(y.pred)
  SSR = sum((y.act - y.pred)^2)
  SST = sum((y.act - y.mean)^2)
  return(1 - SSR / SST)
}
```

```
w.pred <- predict(plant.weight.lm, plant.growth)
```

```
plant.growth$weight
r.squared(plant.growth$weight, w.pred)
[1] 0.0730776
```

Quite poor coefficient of determination showing that there seems to be little to no effect from treatment 🙂

Problem 5

(a)

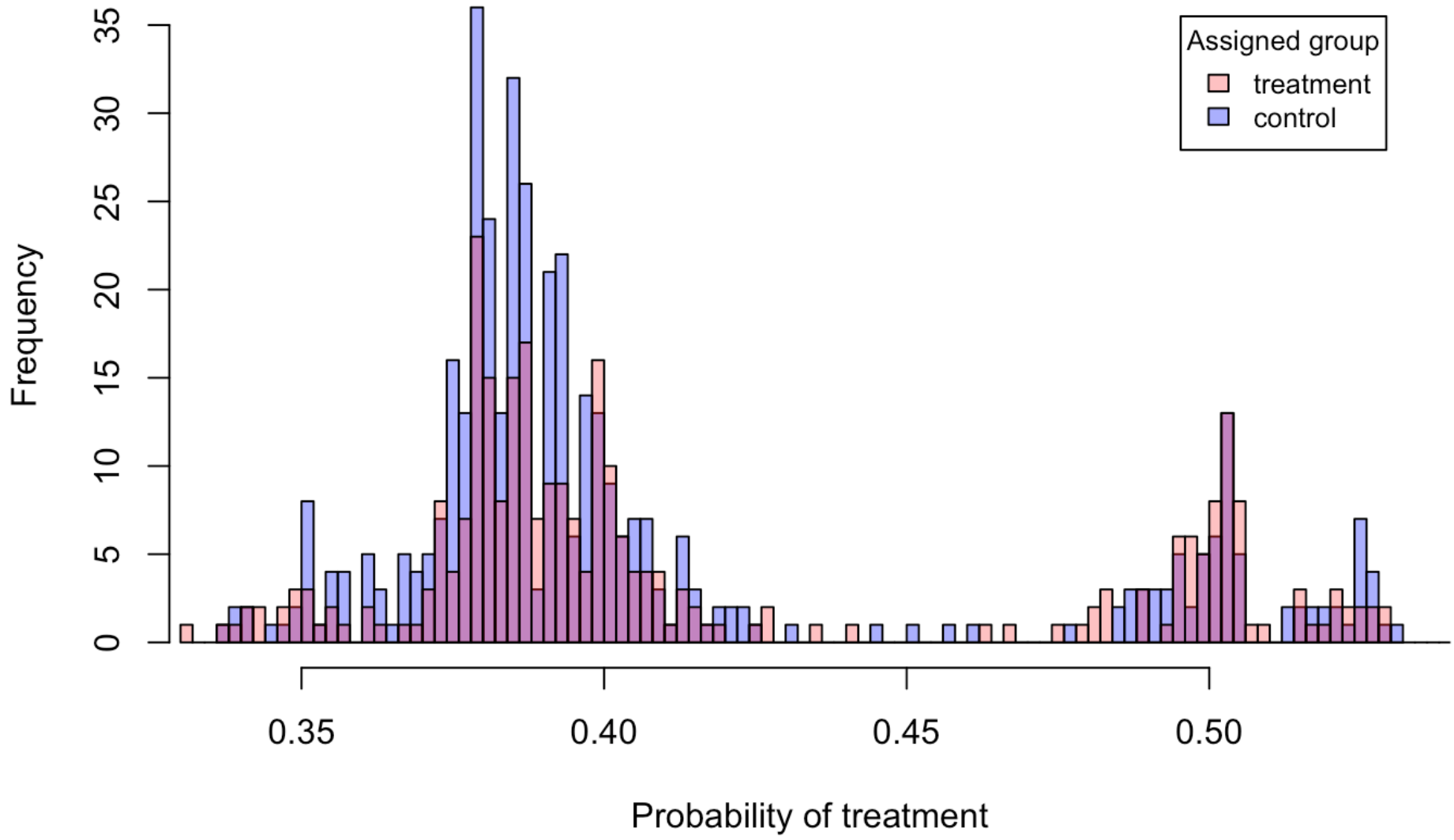

Fig. 4: Comparison of estimated probabilities of treatment assignment for control and treatment groups.

(b) As we can see in the figure above, the distributions of treatment/control assignment are very similar. This implies that this interventional study was an RCT, meaning that for any individual, there is an equal chance of being assigned to treatment group. In this case, the peak frequency for estimated probability of assigning treatment is at about **0.36**, meaning that we we have about 36% of people in treatment group and 64% of people in the control group which approximately corresponds to the count of people in both groups (297 and 425).