Assignment 2

Link to code:

bla-bla.bal.co

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)

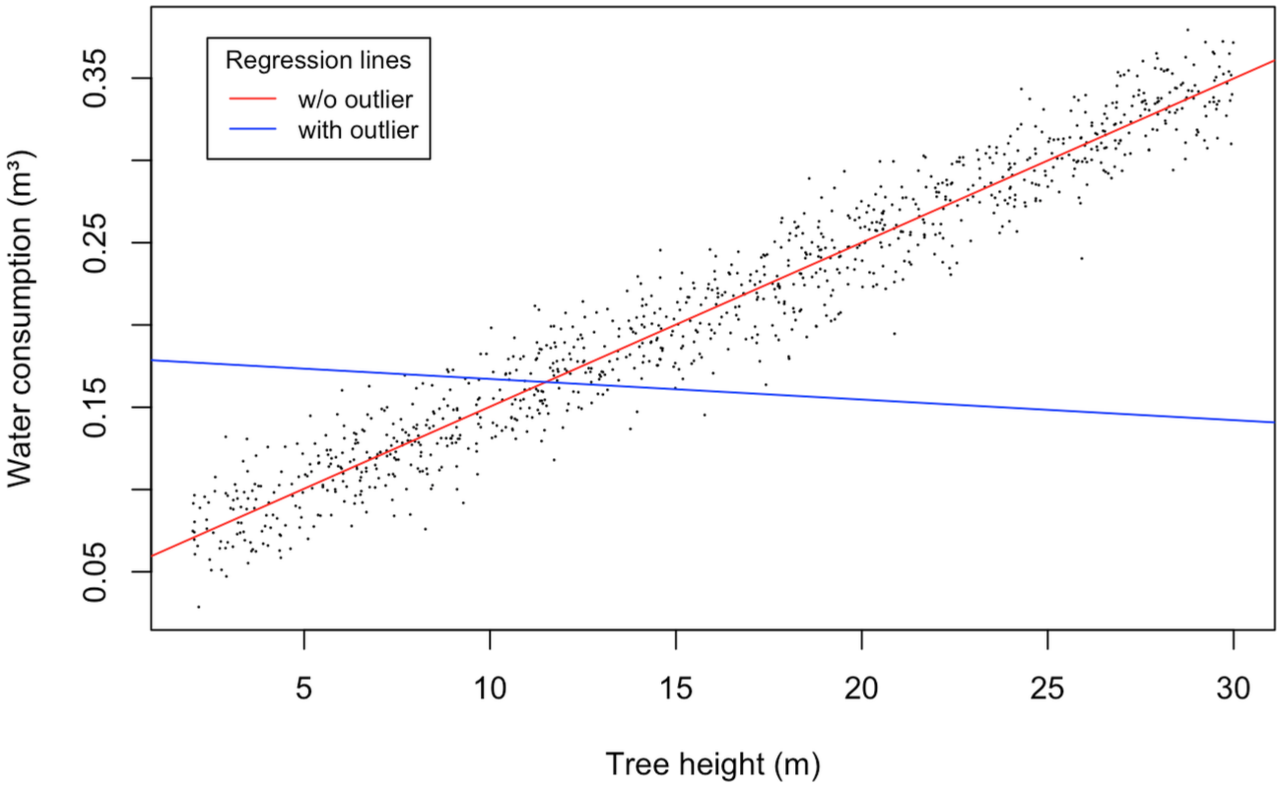

Fig. 1: Visualization of water consumption dependency on height in trees and trend lines showing the overall trend of 999 first points and the trend when the outlier point is included.

(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]    [,25]

2.5%  3003.907 3072.344 3141.402 3200.350 3258.908 3312.202 33
47.960 3374.216 3376.847

97.5% 5251.365 5186.650 5124.042 5063.327 5001.240 4952.641 49
22.982 4910.126 4899.580

         [,26]    [,27]    [,28]    [,29]    [,30]    [,31]
 [,32]    [,33]    [,34]

2.5%  3381.025 3361.860 3332.874 3287.686 3242.068 3179.495 31
14.422 3050.719 2975.868

97.5% 4906.203 4934.039 4971.773 5014.844 5066.333 5126.525 51
88.836 5263.041 5342.034

         [,35]    [,36]    [,37]    [,38]    [,39]    [,40]
 [,41]    [,42]   [,43]

2.5%  2884.688 2809.413 2722.771 2641.299 2549.555 2460.189 23
68.959 2281.396 2199.71

97.5% 5425.575 5513.414 5598.713 5693.679 5786.343 5882.740 59
79.129 6073.570 6165.46

         [,44]    [,45]    [,46]    [,47]    [,48]   [,49]
 [,50]    [,51]    [,52]

2.5%  2107.292 2002.388 1905.511 1816.258 1738.090 1632.60 153
7.705 1448.420 1351.126

97.5% 6264.356 6359.778 6457.942 6557.034 6653.939 6758.23 685
2.798 6951.372 7052.891

         [,53]    [,54]    [,55]

2.5%  1251.483 1154.657 1048.169

97.5% 7154.437 7256.587 7354.871
```

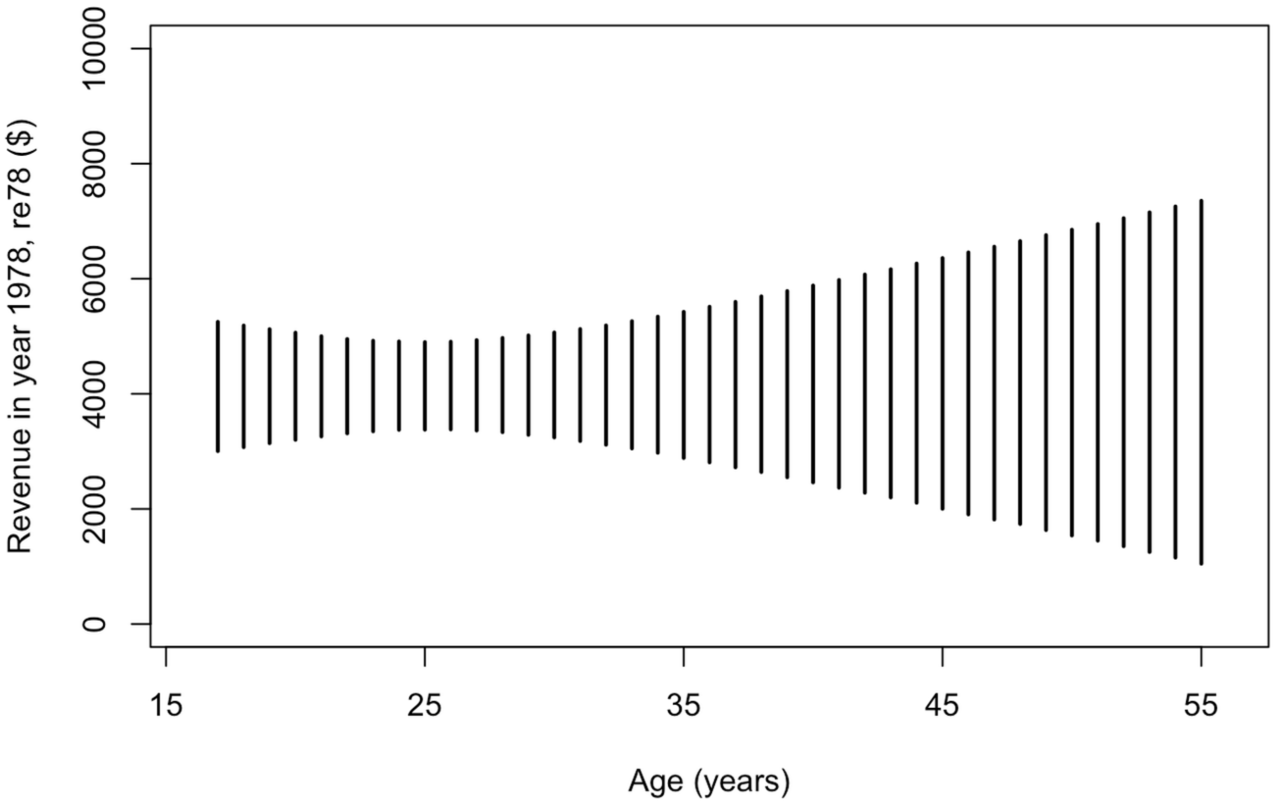

Fig. 2-1: Visualization of the confidence intervals described in the table above.

(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]    [,25]
2.5%  3080.690 3167.794 3230.617 3295.097 3360.114 3407.818 34
48.633 3468.002 3487.899
97.5% 5505.375 5442.311 5374.537 5323.657 5278.784 5239.247 52
12.887 5199.928 5187.756
         [,26]    [,27]    [,28]    [,29]    [,30]    [,31]
 [,32]    [,33]    [,34]
2.5%  3496.949 3497.798 3488.305 3463.054 3417.025 3366.244 33
17.954 3259.684 3199.943
97.5% 5197.853 5211.494 5240.792 5277.715 5332.389 5401.501 54
77.347 5564.976 5649.577
         [,35]    [,36]    [,37]    [,38]    [,39]    [,40]
 [,41]    [,42]    [,43]
```

```
2.5%  3133.169 3058.260 2974.717 2895.218 2817.154 2738.093 26
61.846 2580.342 2478.873

97.5% 5729.347 5830.032 5929.290 6021.261 6131.030 6233.358 63
37.795 6432.795 6532.663

         [,44]    [,45]    [,46]    [,47]    [,48]    [,49]    
 [,50]    [,51]    [,52]

2.5%  2396.239 2312.464 2224.007 2131.464 2048.620 1963.859 18
67.848 1780.429 1693.672

97.5% 6638.450 6741.677 6852.011 6959.400 7072.548 7190.438 72
98.444 7407.179 7508.895

         [,53]    [,54]    [,55]

2.5%  1599.236 1508.547 1415.648

97.5% 7618.737 7726.110 7839.324
```

(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]    [,2
2]     [,23]     [,24]

2.5%  -6795.655 -6915.214 -6886.698 -6972.526 -6789.421 -6564.
40 -6746.938 -6879.621

97.5% 15026.503 14945.284 14875.525 14992.310 14845.333 14965.
07 15211.281 15202.563

          [,25]     [,26]     [,27]     [,28]     [,29]    [,3
0]     [,31]     [,32]

2.5%  -6655.543 -6582.921 -6763.697 -6814.285 -6593.833 -6995.
21 -6792.403 -6482.636

97.5% 15268.466 15094.670 14926.598 15006.214 15012.030 15121.
89 15170.123 14991.999

          [,33]     [,34]     [,35]    [,36]    [,37]     [,3
8]     [,39]     [,40]     [,41]

2.5%  -6660.117 -6846.906 -6922.325 -6807.94 -7004.88 -6947.66
6 -6913.793 -6779.311 -6822.862
```

```
97.5% 14940.592 14939.204 14998.268 15132.58 14943.07 15273.28
1 15131.700 15066.810 14847.824
         [,42]     [,43]    [,44]     [,45]     [,46]     [,4
7]     [,48]     [,49]     [,50]
2.5%  -6737.61 -6725.481 -6876.79 -7000.327 -6674.509 -6899.02
7 -6909.321 -6738.276 -6893.289
97.5% 15168.26 15368.092 15204.23 15054.544 15013.654 15256.99
7 15228.000 15292.945 15441.206
          [,51]     [,52]    [,53]     [,54]     [,55]
2.5%  -7128.928 -6921.078 -7199.89 -7184.698 -6987.804
97.5% 15307.595 15372.343 15617.25 15438.537 15586.922
```

(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]     [,2
2]     [,23]    [,24]
2.5%  -6611.257 -6416.205 -6363.339 -6272.143 -6413.38 -6600.6
87 -6451.852 -6605.88
97.5% 15181.713 15186.449 15108.780 15149.649 15442.44 15140.1
53 15132.334 14926.01
          [,25]     [,26]     [,27]     [,28]     [,29]     [,
30]     [,31]     [,32]
2.5%  -6572.564 -6625.892 -6497.097 -6400.379 -6547.107 -6671.
589 -6445.356 -6357.819
97.5% 15417.343 15336.237 15145.882 15075.704 14949.972 15020.
754 15457.132 15316.567
         [,33]     [,34]     [,35]     [,36]     [,37]     [,3
8]     [,39]     [,40]
2.5%  -6520.41 -6393.692 -6782.923 -6477.596 -6562.636 -6446.6
08 -6540.733 -6574.787
97.5% 15119.08 15315.609 15351.740 15522.360 15425.293 15212.8
58 15208.076 15440.712
```

```
          [,41]     [,42]     [,43]     [,44]     [,45]     [,
46]    [,47]     [,48]
2.5%  -6624.206 -6574.095 -6796.896 -6670.936 -6665.127 -6583.
075 -6649.03 -6682.051
97.5% 15479.063 15621.804 15547.980 15299.077 15582.879 15910.
814 15386.75 15696.081
          [,49]     [,50]     [,51]     [,52]     [,53]    [,5
4]     [,55]
2.5%  -6653.799 -6620.547 -6454.553 -6826.836 -6951.541 -6865.
65 -6950.507
97.5% 15826.058 15709.670 16010.326 15814.054 15996.906 15859.
63 15824.097
```

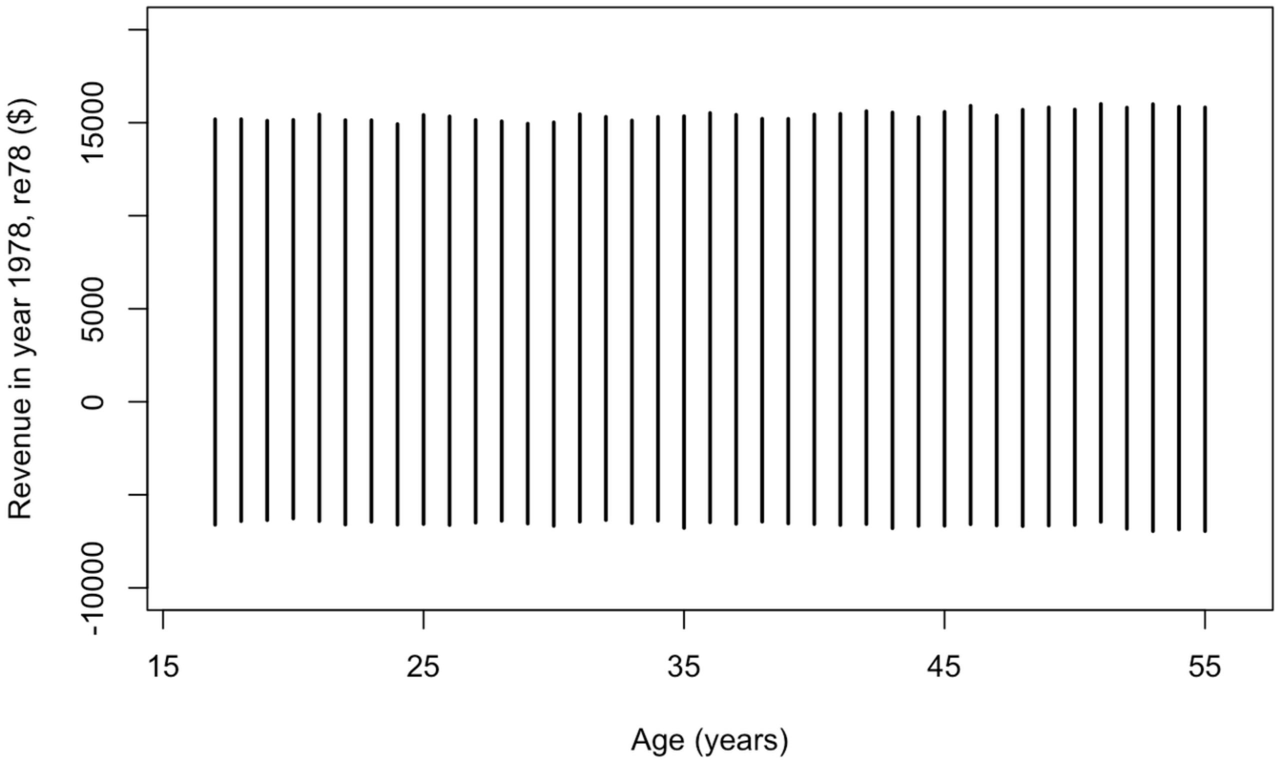

Fig. 2-2: Visualization of the confidence intervals of predicted values described in the table above.

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.

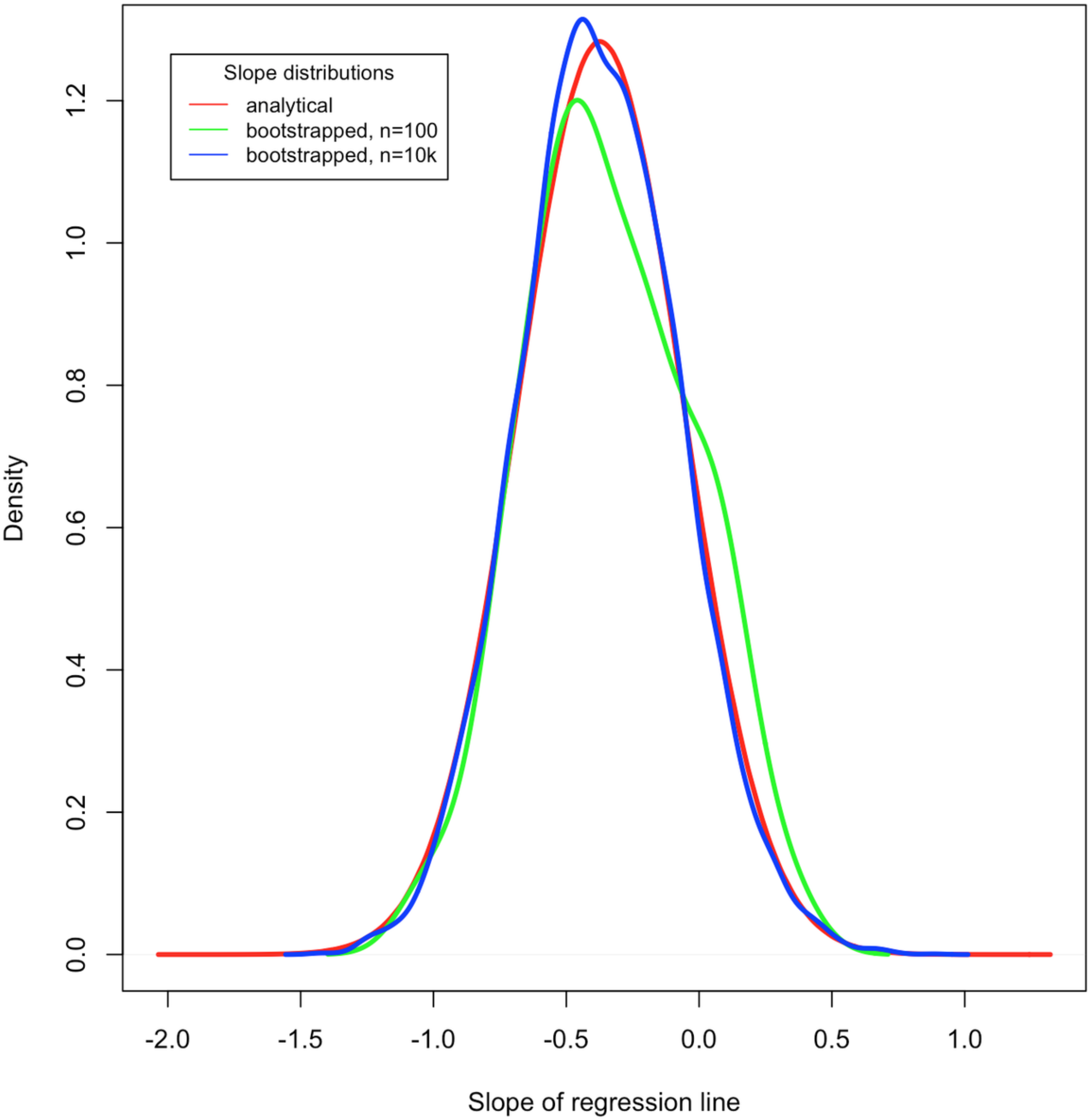

*Fig. 3: Comparison of analytically derived and bootstrapped distributions of slopes in the dependency of plant **weight** on **treatment1**.*

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)

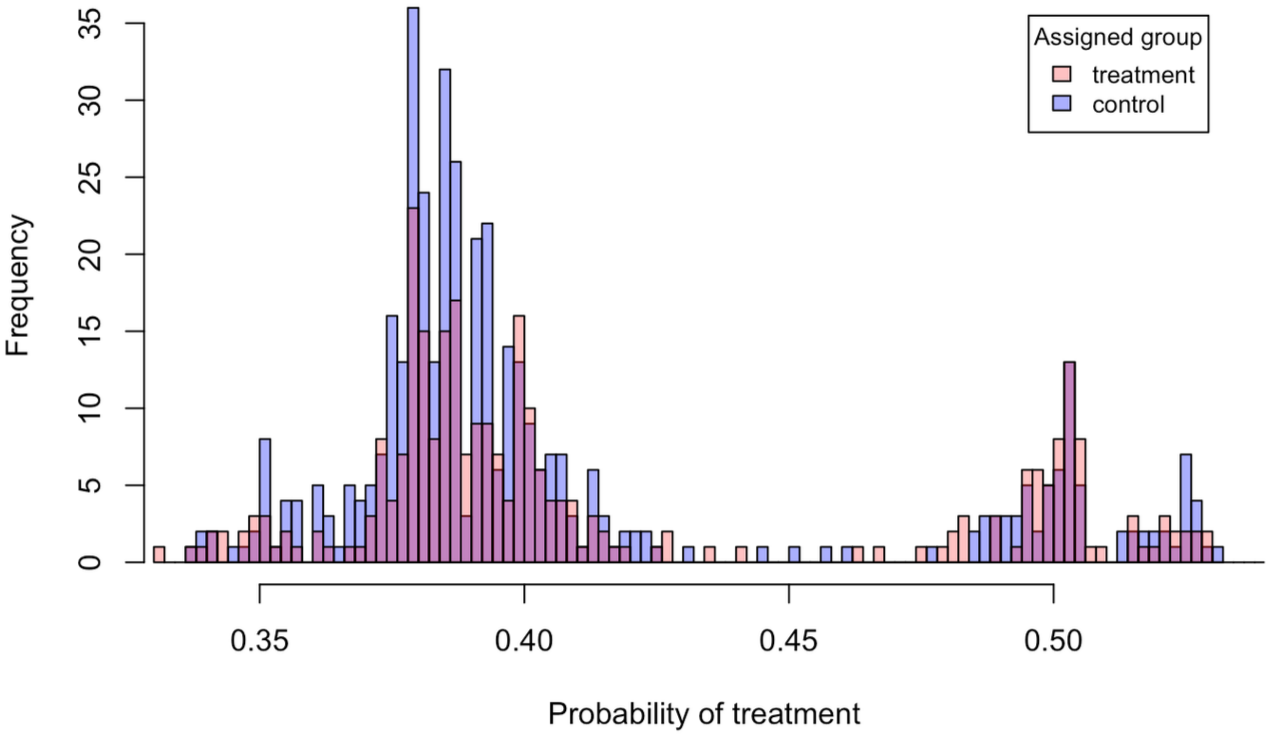

Fig. 4: Comparison of estimated probabilities of treatment assignment for control and treatment groups.

(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).