**Assignment 2**

*For all of the below, include a link to your code.*

Link to code:
https://github.com/katjadellalibera/CS112/blob/master/Regression%20and%20Bootstrapping/Regression%20and%20Boostrapping%20Code.R

1. Write your own original code that produces a dataset that conforms to the classic univariate regression model. Your data set should have 999 observations and a Normal error term. The slope of the coefficient on your regressor should be positive. Now include a single outlier, such that when you fit a regression to your 1000 data points, the slope of your regression line is negative. Your answer to this question should consist of:
   (a) Your original data-generating equation

$$variable2 = 50 + \frac{variable1}{10} + \text{random noise}$$

   (b) Regression results for the original 999 (copy/paste the "summary" output)

Call:

lm(formula = variable2 ~ variable1, data = data)

Residuals:

| Min | 1Q | Median | 3Q | Max |
|---|---|---|---|---|
| -2.03306 | -0.93114 | 0.00867 | 0.91905 | 2.09153 |

Coefficients:

| | Estimate | Std. Error | t value | Pr(>\|t\|) | |
|---|---|---|---|---|---|
| (Intercept) | **50.048228** | 0.074385 | 672.82 | <2e-16 | *** |
| variable1 | **0.096092** | 0.002552 | 37.65 | <2e-16 | *** |

---

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.139 on 997 degrees of freedom

Multiple R-squared:  0.5871,  Adjusted R-squared:  0.5867

F-statistic:  1418 on 1 and 997 DF,  p-value: < 2.2e-16

(c) Regression results with the outlier included (copy/paste “summary” output)

Call:

lm(formula = variable2 ~ variable1, data = data_with_outlier)

Residuals:

| Min | 1Q | Median | 3Q | Max |
|---|---|---|---|---|
| -198.804 | -1.347 | 0.137 | 1.778 | 5.056 |

Coefficients:

| | Estimate | Std. Error | t value | Pr(>\|t\|) | |
|---|---|---|---|---|---|
| (Intercept) | **53.01489** | 0.42251 | 125.476 | <2e-16 | *** |
| variable1 | **-0.02807** | 0.01431 | -1.961 | 0.0502 | . |

---

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.635 on 998 degrees of freedom

Multiple R-squared:  0.003838,  Adjusted R-squared:  0.00284

F-statistic: 3.845 on 1 and 998 DF,  p-value: 0.05016

(d) A properly-labeled data visualization that shows the regression line based on the original 999 points, and another differentiated regression line (on the same axes) based on 1000 points.

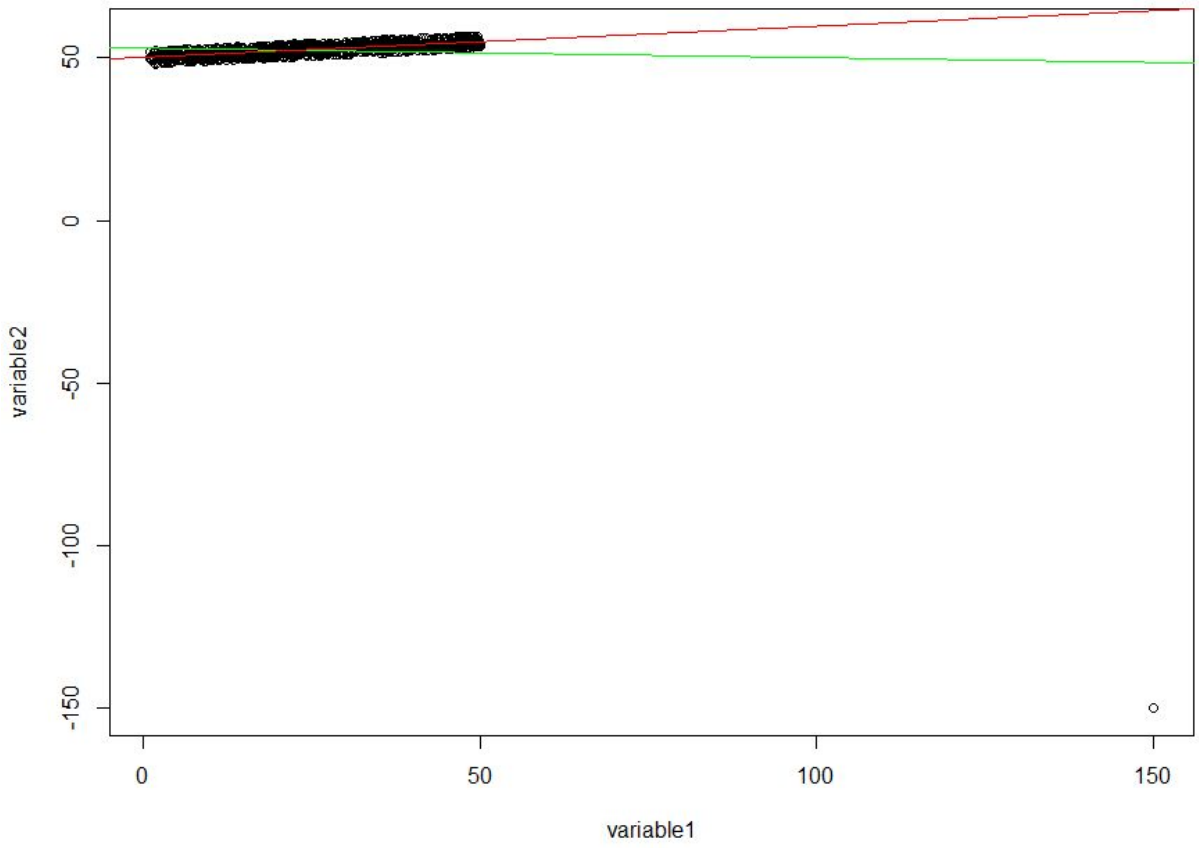


***Figure 1:*** *variable 2 as a function of variable 1 with an outlier in the bottom right corner. The red line is the linear fit to the data without consideration of the outlier which has a positive slope. The green line is a linear fit of the same data this time including the outlier in the regression.*

(e) 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.

Extrapolation is creating predictions based on the expectation that the current trend will persist. The danger of this goes both ways, the one illustrated in the figure above is that a single outlier can have a large influence on the predictions made, as illustrated by the change in coefficients for variable 1 from 0.096 to -0.028, which is a problem if the point is erroneous. On the other hand, we also shouldn't extrapolate from the data in the original section, because the trend displayed did not hold up for at least one data point and instead try to collect more data especially between the points displayed here and adjust our model as needed.[1]

[1] #regression: throughout the assignment, I show my understanding of regression, its' dangers and potential in analyzing data

2. *NOTE: FOR THIS PROBLEM (AND THIS PROBLEM ONLY), USE ONLY THE CONTROL GROUP. DO NOT USE ANY UNITS FOR WHICH TREATMENT == 1.* Using the Lalonde data set and a linear model that predicts re78 as a linear additive function of age, educ, re74, re75, educ*re74, educ*re75, age*re74, age*re75, age*age, and re74*re75, estimate:
    - the 95% interval of expected values for re78, for every unit (i.e., each age 17-55, spanning the age range in the data set), using simulation (i.e., 10000 simulated predictions for every row from 10000 sets of coefficients). You should not incorporate simulated sigmas, and you should hold educ, re74, and re75 at their medians. Even include ages that are not covered by the data (e.g., 47, 49, etc.).

Medians of variables:

- educ:10
- re74: 0
- re75 : 0

| age | Lower-bound (95% confidence interval) | median | Upper-bound (95% confidence interval) |
|---|---|---|---|
| 17 | 3082 | 4466 | 5835 |
| 18 | 3168 | 4371 | 5570 |
| 19 | 3214 | 4283 | 5333 |
| 20 | 3235 | 4197 | 5142 |
| 21 | 3240 | 4126 | 4995 |
| 22 | 3217 | 4059 | 4892 |
| 23 | 3159 | 4001 | 4833 |
| 24 | 3073 | 3950 | 4808 |
| 25 | 2989 | 3909 | 4805 |
| 26 | 2918 | 3877 | 4828 |
| 27 | 2847 | 3853 | 4852 |
| 28 | 2768 | 3834 | 4870 |
| 29 | 2712 | 3826 | 4911 |
| 30 | 2665 | 3825 | 4961 |
| 31 | 2626 | 3834 | 5004 |
| 32 | 2598 | 3848 | 5047 |
| 33 | 2589 | 3870 | 5102 |
| 34 | 2577 | 3896 | 5180 |
| 35 | 2583 | 3932 | 5255 |

| | | | |
|---|---|---|---|
| 36 | 2577 | 3975 | 5342 |
| 37 | 2584 | 4025 | 5442 |
| 38 | 2587 | 4087 | 5579 |
| 39 | 2565 | 4147 | 5732 |
| 40 | 2541 | 4224 | 5916 |
| 41 | 2500 | 4308 | 6129 |
| 42 | 2460 | 4401 | 6368 |
| 43 | 2403 | 4503 | 6620 |
| 44 | 2327 | 4603 | 6904 |
| 45 | 2245 | 4722 | 7225 |
| 46 | 2145 | 4853 | 7573 |
| 47 | 2056 | 4987 | 7951 |
| 48 | 1933 | 5123 | 8362 |
| 49 | 1777 | 5273 | 8800 |
| 50 | 1600 | 5425 | 9251 |
| 51 | 1434 | 5591 | 9762 |
| 52 | 1229 | 5757 | 10284 |
| 53 | 1030 | 5932 | 10824 |
| 54 | 807 | 6122 | 11378 |
| 55 | 577 | 6315 | 11973 |

- the 95% interval of expected values for re78, for every unit, using simulation (i.e., 10000 simulated predictions for every row from 10000 sets of coefficients). You should not incorporate simulated sigmas, and you should hold educ, re74, and re75 at their 75% quantiles.

75% quantiles for variables:

- educ: 11
- re74: 139
- re75: 650

| age | Lower-bound (95% confidence interval) | median | Upper-bound (95% confidence interval) |
|---|---|---|---|
| 17 | 3206 | 4694 | 6194 |
| 18 | 3294 | 4605 | 5926 |
| 19 | 3370 | 4523 | 5680 |

| | | | |
|---|---|---|---|
| 20 | 3419 | 4448 | 5494 |
| 21 | 3447 | 4381 | 5342 |
| 22 | 3435 | 4326 | 5224 |
| 23 | 3393 | 4276 | 5139 |
| 24 | 3345 | 4231 | 5111 |
| 25 | 3277 | 4194 | 5108 |
| 26 | 3202 | 4163 | 5124 |
| 27 | 3137 | 4143 | 5148 |
| 28 | 3069 | 4134 | 5190 |
| 29 | 3028 | 4132 | 5251 |
| 30 | 2980 | 4137 | 5307 |
| 31 | 2944 | 4153 | 5375 |
| 32 | 2903 | 4171 | 5438 |
| 33 | 2875 | 4198 | 5510 |
| 34 | 2868 | 4230 | 5583 |
| 35 | 2862 | 4273 | 5669 |
| 36 | 2852 | 4320 | 5773 |
| 37 | 2828 | 4377 | 5904 |
| 38 | 2811 | 4444 | 6076 |
| 39 | 2768 | 4520 | 6243 |
| 40 | 2747 | 4604 | 6436 |
| 41 | 2726 | 4695 | 6666 |
| 42 | 2694 | 4791 | 6905 |
| 43 | 2641 | 4905 | 7172 |
| 44 | 2567 | 5017 | 7481 |
| 45 | 2480 | 5140 | 7824 |
| 46 | 2374 | 5272 | 8177 |
| 47 | 2267 | 5415 | 8580 |
| 48 | 2157 | 5570 | 8981 |
| 49 | 2018 | 5718 | 9445 |
| 50 | 1862 | 5872 | 9939 |
| 51 | 1688 | 6032 | 10462 |
| 52 | 1534 | 6219 | 10992 |
| 53 | 1355 | 6409 | 11550 |

| | | | |
|---|---|---|---|
| 54 | 1123 | 6599 | 12150 |
| 55 | 909 | 6811 | 12781 |

- the 95% prediction interval for re78, for every unit (i.e., each age, spanning the age range in the data set), using simulation (i.e., 10000 simulated predictions for every row from 10000 sets of coefficients). You will need to incorporate simulated sigmas, and you should hold educ, re74, and re75 at their medians.

| age | Lower-bound (95% interval) | median | Upper-bound (95% interval) |
|---|---|---|---|
| 17 | -6366 | 4436 | 15582 |
| 18 | -6202 | 4398 | 15129 |
| 19 | -6673 | 4288 | 15174 |
| 20 | -6636 | 4202 | 14983 |
| 21 | -6976 | 4132 | 14935 |
| 22 | -6668 | 4147 | 14814 |
| 23 | -6676 | 4021 | 15063 |
| 24 | -6847 | 4014 | 14686 |
| 25 | -6935 | 3796 | 14725 |
| 26 | -6890 | 3806 | 14793 |
| 27 | -6948 | 3832 | 14554 |
| 28 | -6945 | 3865 | 14741 |
| 29 | -7227 | 3790 | 14544 |
| 30 | -7191 | 3968 | 14618 |
| 31 | -6930 | 3826 | 14544 |
| 32 | -7183 | 3884 | 14503 |
| 33 | -7166 | 3746 | 14722 |
| 34 | -7154 | 3820 | 14866 |
| 35 | -6999 | 3751 | 14667 |
| 36 | -6857 | 3975 | 14847 |
| 37 | -6772 | 4011 | 14887 |
| 38 | -7071 | 4026 | 15278 |
| 39 | -6605 | 4099 | 15285 |
| 40 | -6801 | 4080 | 15190 |

| | | | |
|---|---|---|---|
| 41 | -6440 | 4413 | 15261 |
| 42 | -6482 | 4428 | 15443 |
| 43 | -6344 | 4418 | 15640 |
| 44 | -6475 | 4703 | 15736 |
| 45 | -6297 | 4925 | 15904 |
| 46 | -6399 | 4821 | 15818 |
| 47 | -6287 | 4952 | 16211 |
| 48 | -6083 | 5217 | 16275 |
| 49 | -5975 | 5237 | 16510 |
| 50 | -6059 | 5486 | 16898 |
| 51 | -6102 | 5582 | 17137 |
| 52 | -6106 | 5682 | 17614 |
| 53 | -5917 | 5812 | 17975 |
| 54 | -5823 | 6186 | 18452 |
| 55 | -5844 | 6412 | 18585 |

- the 95% prediction interval for re78, for every unit, using simulation (i.e., 10000 simulated predictions for every row from 10000 sets of coefficients). You will need to incorporate simulated sigmas, and you should hold educ, re74, and re75 at their 75% quantiles.

| age | Lower-bound (95% confidence interval) | median | Upper-bound (95% confidence interval) |
|---|---|---|---|
| 17 | -6187 | 4717 | 15721 |
| 18 | -6235 | 4560 | 15615 |
| 19 | -6218 | 4591 | 15257 |
| 20 | -6327 | 4492 | 15203 |
| 21 | -6532 | 4329 | 15063 |
| 22 | -6676 | 4423 | 15158 |
| 23 | -6422 | 4341 | 15308 |
| 24 | -6812 | 4141 | 15162 |
| 25 | -6484 | 4170 | 15089 |
| 26 | -6821 | 4177 | 15015 |
| 27 | -6698 | 4222 | 14965 |

| | | | |
|---|---|---|---|
| 28 | -6598 | 4173 | 14920 |
| 29 | -6754 | 4090 | 14902 |
| 30 | -6535 | 4185 | 15015 |
| 31 | -6663 | 4218 | 15371 |
| 32 | -6891 | 4088 | 14892 |
| 33 | -6729 | 4280 | 15185 |
| 34 | -6415 | 4294 | 14869 |
| 35 | -6773 | 4234 | 15096 |
| 36 | -6697 | 4303 | 15272 |
| 37 | -6535 | 4275 | 15262 |
| 38 | -6484 | 4392 | 15447 |
| 39 | -6233 | 4555 | 15288 |
| 40 | -6220 | 4538 | 15555 |
| 41 | -6137 | 4635 | 15413 |
| 42 | -6270 | 4704 | 15560 |
| 43 | -6058 | 5006 | 15744 |
| 44 | -6226 | 5041 | 15875 |
| 45 | -6085 | 5254 | 16366 |
| 46 | -5718 | 5334 | 16578 |
| 47 | -5789 | 5366 | 16656 |
| 48 | -5708 | 5634 | 17018 |
| 49 | -5582 | 5614 | 17166 |
| 50 | -5736 | 5858 | 17397 |
| 51 | -6050 | 6052 | 17920 |
| 52 | -5438 | 6176 | 18033 |
| 53 | -5646 | 6457 | 18661 |
| 54 | -5689 | 6586 | 18808 |
| 55 | -5554 | 6874 | 18999 |

Your answer to this question should consist of the following:

(a) 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)

(b) 1 figure for the 2 interval analyses with expected values, and 1 figure for the 2 interval analyses with predicted values. 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 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.).

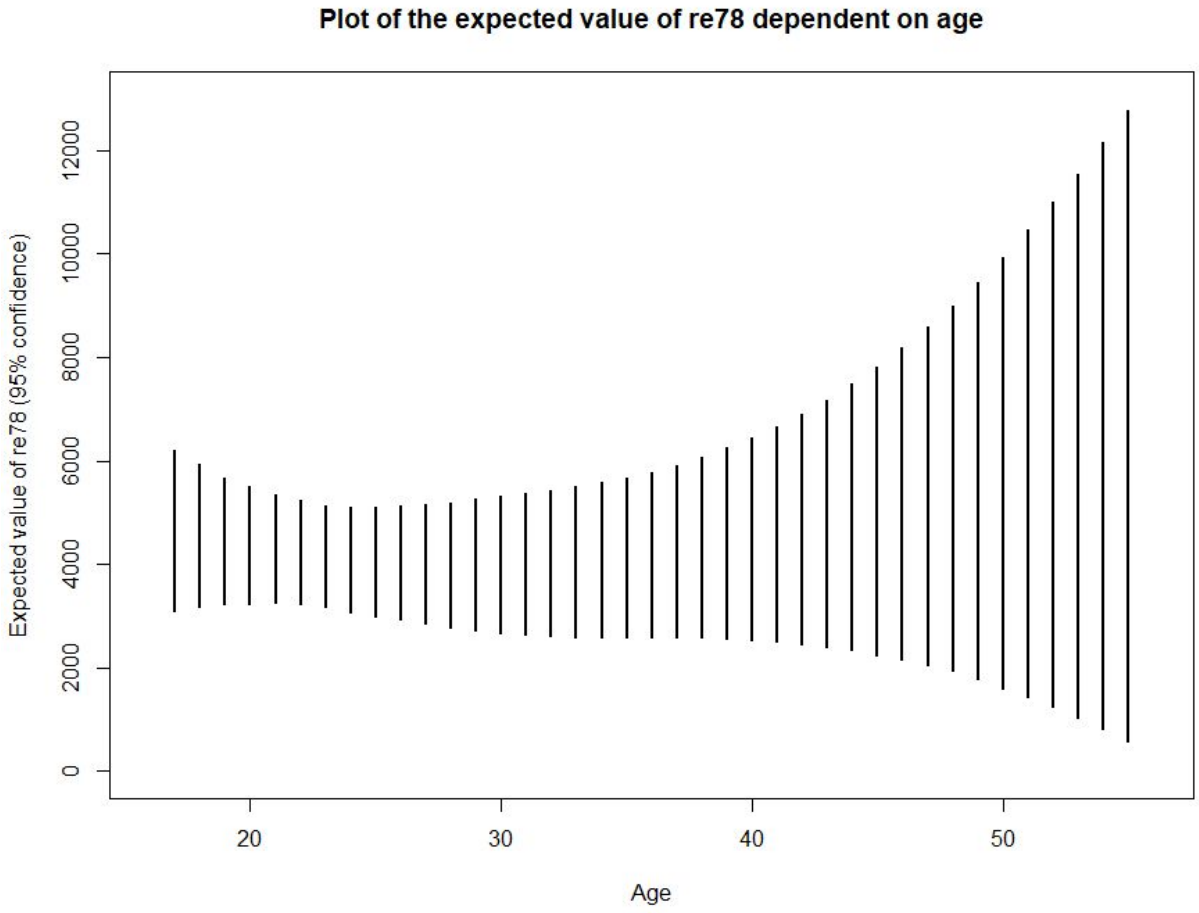


***Figure 2:*** *The 95% confidence interval of the expected values for re78 from the linear model described above. It appears to be fairly confident for younger respondents and very uncertain for older respondents*

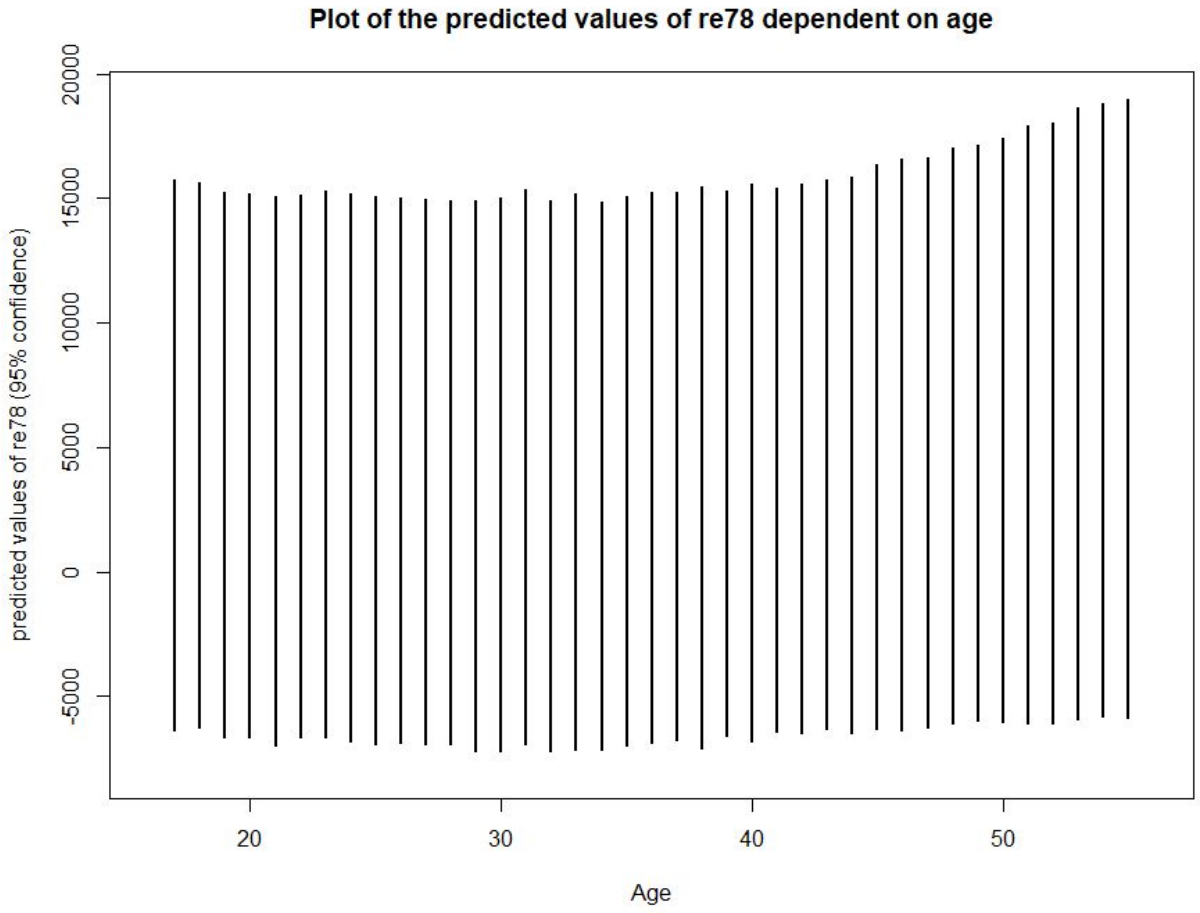


***Figure 3:*** *When considering sigmas, the confidence decreases significantly and the scale increases. The model is not good at predicting the expected earnings depending on age*

E.g.: https://gist.github.com/diamonaj/75fef6eb48639c2c36f73c58d54bac2f

3. Obtain the PlantGrowth dataset in R.

   Specify a regression model in which the dependent variable is *weight* and the independent variable is an indicator of treatment1 (set the value = 1) or control (set the value = 0). This means you will discard observations associated with treatment2.

   Then, bootstrap the 95% confidence intervals for the value of the coefficient for treatment. Then, obtain the analytical confidence interval for the coefficient value using the standard error that pops out of a regression (or equivalently, in R, you can use the *confint* function). Compare the two confidence intervals--one obtained via simulation, the other via the formula.

**NOTE: Make sure that you don't use a 'canned' bootstrap function -- please code the bootstrap routine manually.**

Your answer to this question should consist of the following:

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

| | Bootstrap intercept | Bootstrap indicator | Regression intercept | Regression indicator |
|---|---|---|---|---|
| 2.5% | 4.687 | -0.945 | 4.569 | -1.025 |
| 97.5% | 5.385 | 0.234 | 5.495 | 0.283 |

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

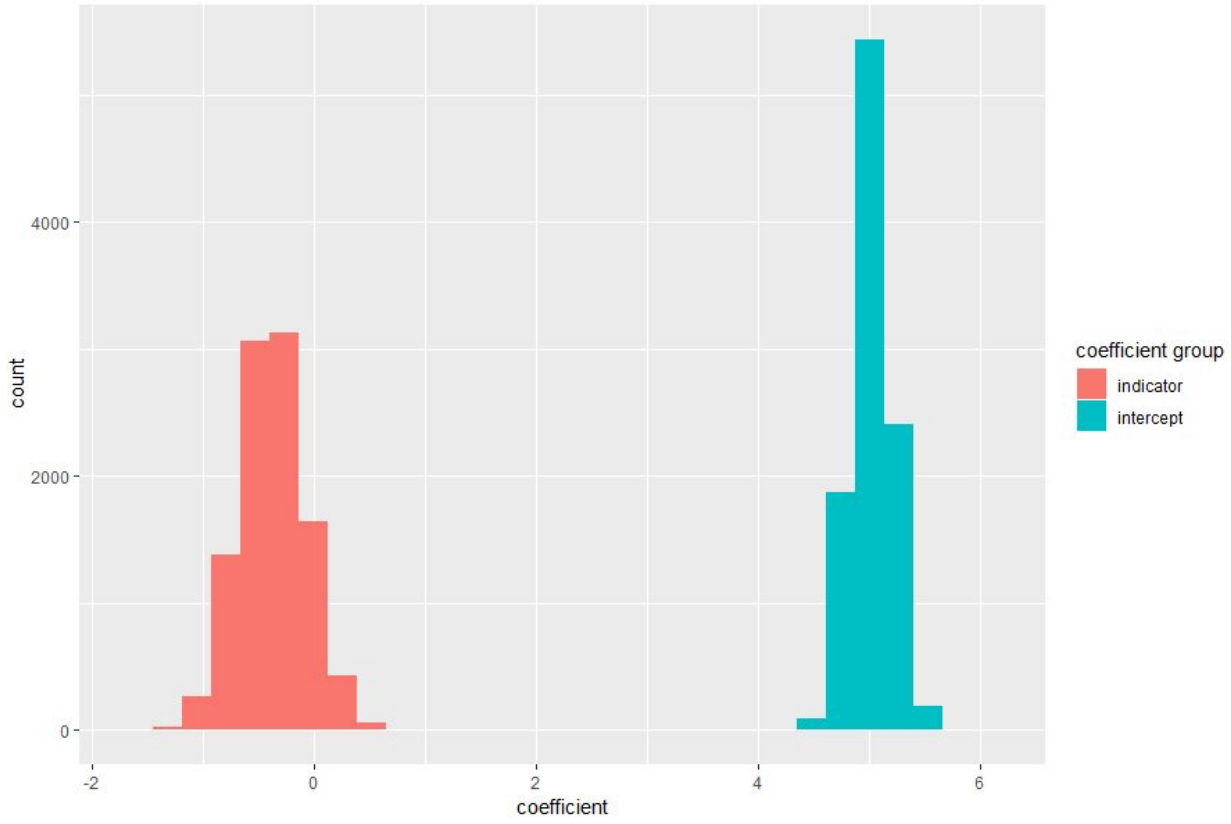


***Figure 4:*** *Histograms of the coefficients obtained through a bootstrap simulation*[2]

[2] #simulation: I understand the value of bootstrapping the small dataset to simulate a larger one.

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

Both the traditional regression result and the bootstrapped result agreed mostly on the coefficients of the intercept and indicator (treatment or non-treatment) as they influence weight. The impact of the intercept is approximately 5 and the treatment has a slight negative effect (-0.375).

4. Write your own function (5 lines max) that takes Ys and predicted Ys as inputs, and outputs $R^2$. Copy/paste an example using the *PlantGrowth* data (from #3 above) that shows it working.

r_function<- function(Y,pred_y) cor(Y,pred_y)^2

Tested on the expected weight from the linear model of plant growth and the actual weight, it returns the $R^2 = 0.073$. This is in agreement with the multiple R-squared from the summary of the linear model

5. Obtain the *nsw.dta* dataset from http://users.nber.org/~rdehejia/data/nswdata2.html. Read the description of this data set provided on the page. If you proceed with this work in R (recommended) use the *foreign* library to open it (so you can use *read.dta*).

   Use this *nsw.dta* dataset to estimate the probability of being assigned to the treatment group (vs. the control group) for every observation in the data set. Your logistic regression model should be a linear additive function of all predictors available to you -- no interaction terms needed. NOTE: re78 is not a predictor because it postdates the treatment. (In other words, it's an outcome.)

   Your answer to this question should consist of the following:

   (a) 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.

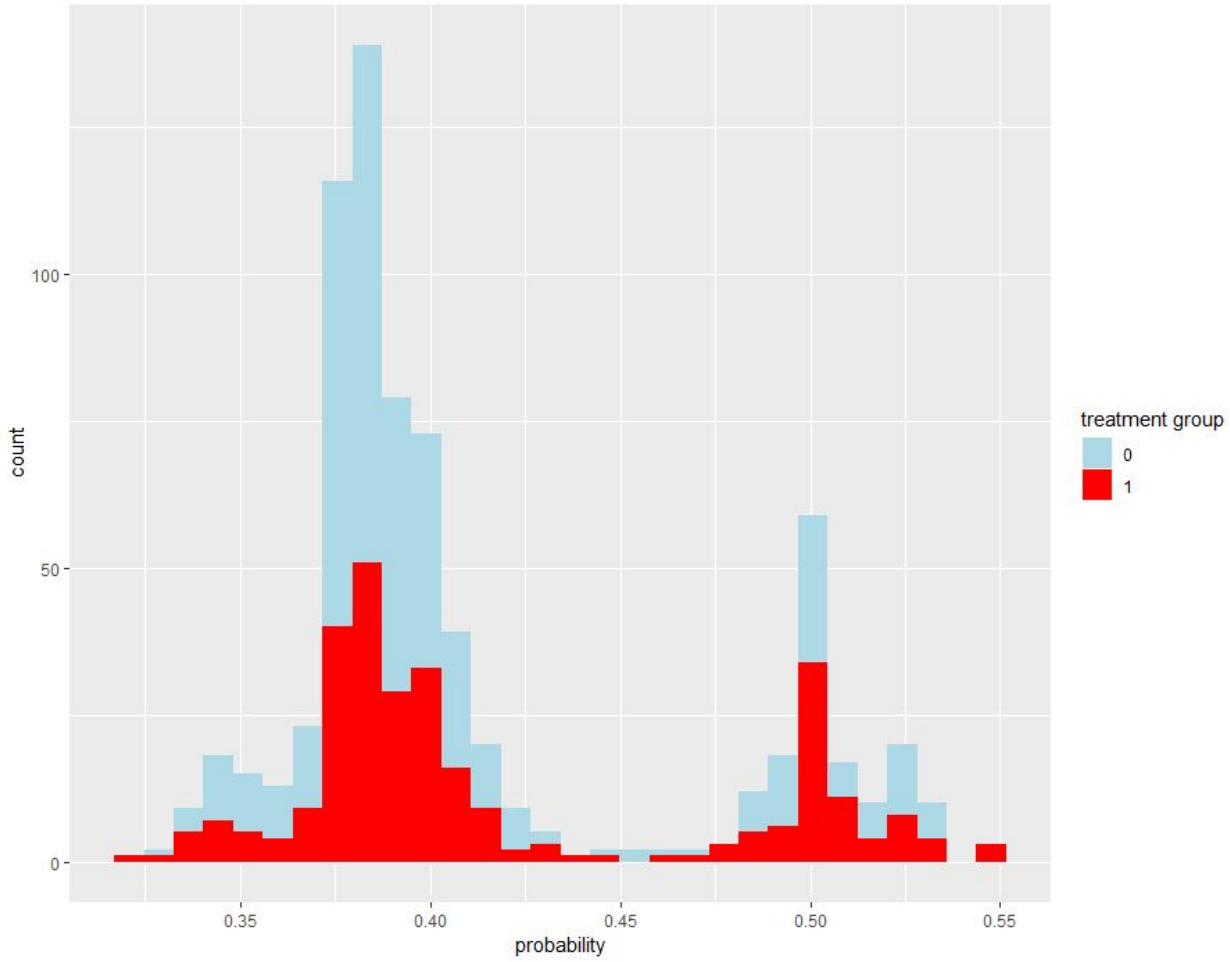


***Figure 5:*** *A plot of the two distributions of probabilities with the treatment group in red and the control group in blue.*[3]

(b) 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.

The distributions were surprisingly similar, with most of the observations being predicted as control group in both the control group and treatment group. This can be explained by the size of the control group in general and suggests that the groups are too similar to build a strong predictive model

[3] #dataviz: I clearly label and describe my plots and make this one even more clear by putting the distributions on the same axis.