R Regression Study of Medical Reviews from WebMD

Joshua Kotzker

The original data harvest* collected responses on drugs for medical conditions which the programmer referred to as “common conditions”. My version of the cleaned file, including minor adjustments for the purposes of this study, has been included in the uploaded zip file.

The variables of the dataset are:

Condition- (condition user received treatment for)

Drug- (name of drug used to treat specified condition)

EaseOfUse- (score 1-5 reflecting individual’s opinion on the experience of the administration process of the drug)

Effectiveness- (score 1-5 reflecting individual’s opinion on the effectiveness of the drug)

Satisfaction- (score 1-5 reflecting individual’s opinion on the drug overall)

Indication- (is there FDA approval for the treatment of the specified condition the individual took the drug for existing on the drug’s label)

Price- ($USD estimated price of drug)

Form- (how the drug is administered)

Type- (is the drug only accessible with a prescription, is it available over the counter without a prescription or both)

Reviews- (# submitted reviews for a given drug)

EOUAvg- (average individuals reported ease of use for all drugs reported for a given condition)

EffAvg- (average individuals reported effectiveness for all drugs reported for a given condition)

SatAvg- (average individuals reported satisfaction for all drugs reported for a given condition)

TotalUserScore- (sum of specified drug’s EOU, effectiveness and satisfaction user scores)

The entire linear model can be represented as:

$$
\begin{aligned}
Y = {} & \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8 + \beta_9 x_9 + \beta_{10} x_1 x_2 + \beta_{11} x_1 x_3 + \beta_{12} x_1 x_4 + \beta_{13} x_1 x_5 + \beta_{14} x_1 x_6 + \beta_{15} x_1 x_7 \\
& + \beta_{16} x_1 x_8 + \beta_{17} x_1 x_9 + \beta_{18} x_2 x_3 + \beta_{19} x_2 x_4 + \beta_{20} x_2 x_5 + \beta_{21} x_2 x_6 + \beta_{22} x_2 x_7 + \beta_{23} x_2 x_8 + \beta_{24} x_2 x_9 + \beta_{25} x_3 x_8 + \beta_{26} x_3 x_9 + \beta_{27} x_4 x_8 + \beta_{28} x_4 x_9 \\
& + \beta_{29} x_5 x_8 + \beta_{30} x_5 x_9 + \beta_{31} x_6 x_8 + \beta_{32} x_6 x_9 + \beta_{33} x_7 x_8 + \beta_{34} x_7 x_9 + \beta_{35} x_1 x_2 x_3 + \beta_{36} x_1 x_2 x_4 + \beta_{37} x_1 x_2 x_3 + \beta_{38} x_1 x_2 x_4 + \beta_{39} x_1 x_2 x_5 + \beta_{40} x_1 x_2 x_6 \\
& + \beta_{41} x_1 x_2 x_7 + \beta_{42} x_1 x_2 x_8 + \beta_{43} x_1 x_2 x_9 + \beta_{44} x_1 x_3 x_8 + \beta_{45} x_1 x_3 x_9 + \beta_{46} x_1 x_4 x_8 + \beta_{47} x_1 x_4 x_9 + \beta_{48} x_1 x_5 x_8 + \beta_{49} x_1 x_5 x_9 + \beta_{50} x_1 x_6 x_8 + \beta_{51} x_1 x_6 x_9 \\
& + \beta_{52} x_1 x_7 x_8 + \beta_{53} x_1 x_7 x_9 + \beta_{54} x_1 x_3 x_8 + \beta_{55} x_1 x_3 x_9 + \beta_{56} x_1 x_4 x_8 + \beta_{57} x_1 x_4 x_9 + \beta_{58} x_1 x_5 x_8 + \beta_{59} x_1 x_5 x_9 + \beta_{60} x_1 x_6 x_8 + \beta_{61} x_1 x_6 x_9 + \beta_{62} x_1 x_7 x_8 \\
& + \beta_{63} x_1 x_7 x_9 + \beta_{64} x_2 x_3 x_8 + \beta_{65} x_2 x_3 x_9 + \beta_{66} x_2 x_4 x_8 + \beta_{67} x_2 x_3 x_9 + \beta_{68} x_2 x_4 x_8 + \beta_{69} x_2 x_4 x_9 + \beta_{70} x_2 x_5 x_8 + \beta_{71} x_2 x_5 x_9 + \beta_{72} x_2 x_6 x_8 + \beta_{73} x_2 x_6 x_9 \\
& + \beta_{74} x_2 x_7 x_8 + \beta_{75} x_2 x_8 x_9 + \beta_{76} x_1 x_2 x_3 x_8 + \beta_{77} x_1 x_2 x_3 x_9 + \beta_{78} x_1 x_2 x_4 x_8 + \beta_{79} x_1 x_2 x_4 x_9 + \beta_{80} x_1 x_2 x_5 x_8 + \beta_{81} x_1 x_2 x_5 x_9 + \beta_{82} x_1 x_2 x_6 x_8 + \beta_{83} x_1 x_2 x_6 x_9 \\
& + \beta_{84} x_1 x_2 x_7 x_8 + \beta_{85} x_1 x_2 x_7 x_9 + \varepsilon
\end{aligned}
$$

Y = total user score

x1 = Price

x2 = Indication, x2 = 0 for On Label, x2 = 1 for Off Label

x3..x7 = Form, where:

Form:	Capsule	Drink	Tablet	Cream	Injection	Other
x3=	1	0	0	0	0	0
x4=	0	1	0	0	0	0
x5=	0	0	1	0	0	0
x6=	0	0	0	1	0	0
x7=	0	0	0	0	1	0

x8..x9 = Type, where:

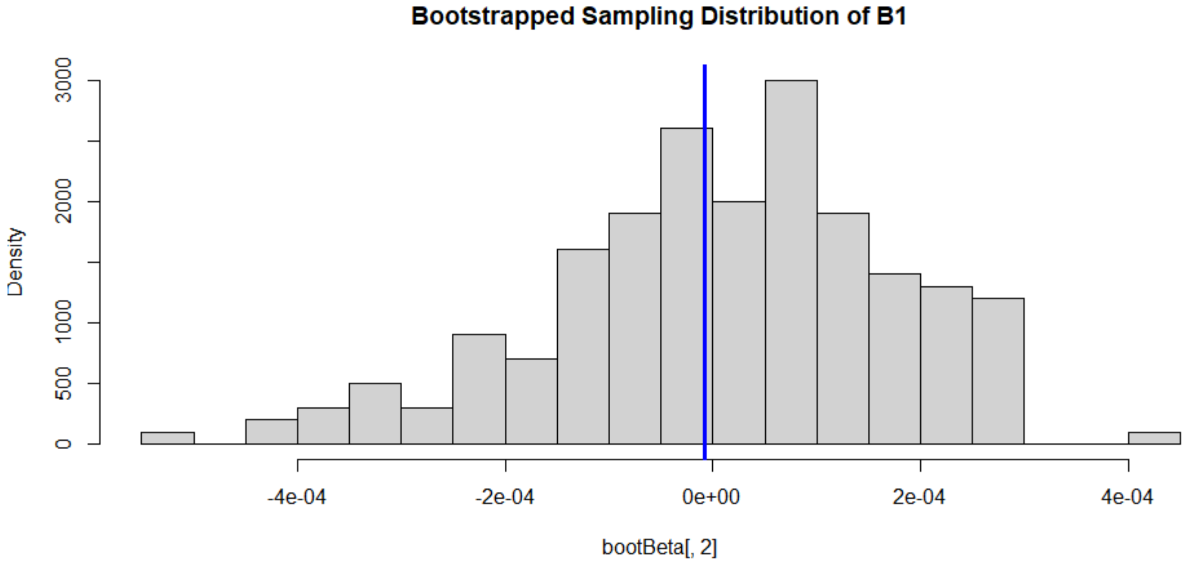
Type	RX	OTC	RX/OTC
x8=	1	0	0
x9=	0	1	0

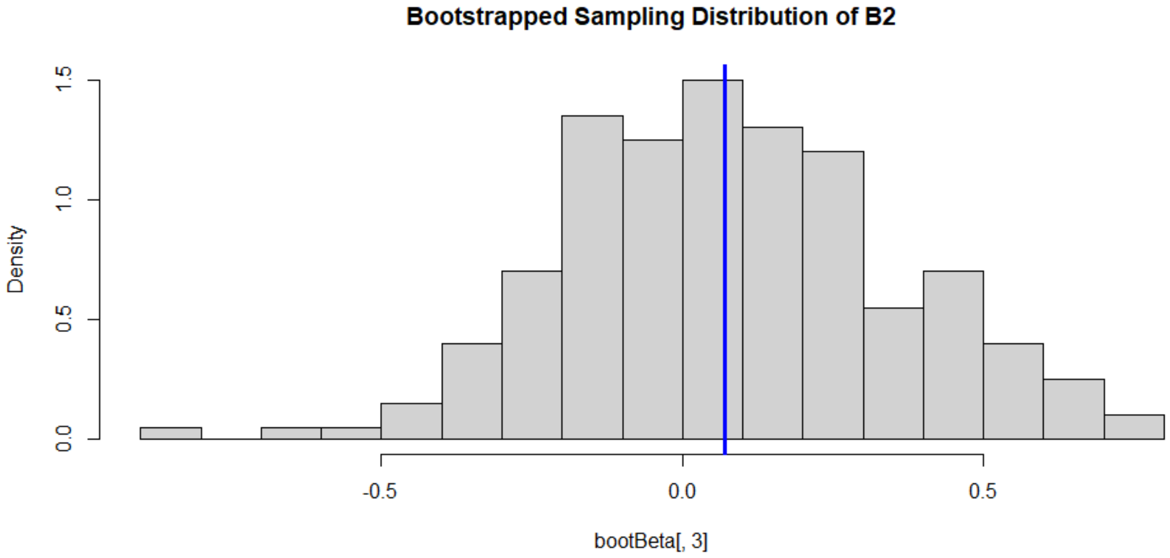
A linear model was chosen due to the initial observations of the data set checking its assumptions. Based on observations from the prior report, the interactions seemed to hold minimal effect on the responses, justifying the shortening of the model. Further justification of the model's shortening is in this project's true nature of being an exercise in bootstrapping. Such a large model would make this process unreasonably convoluted. The model used is:

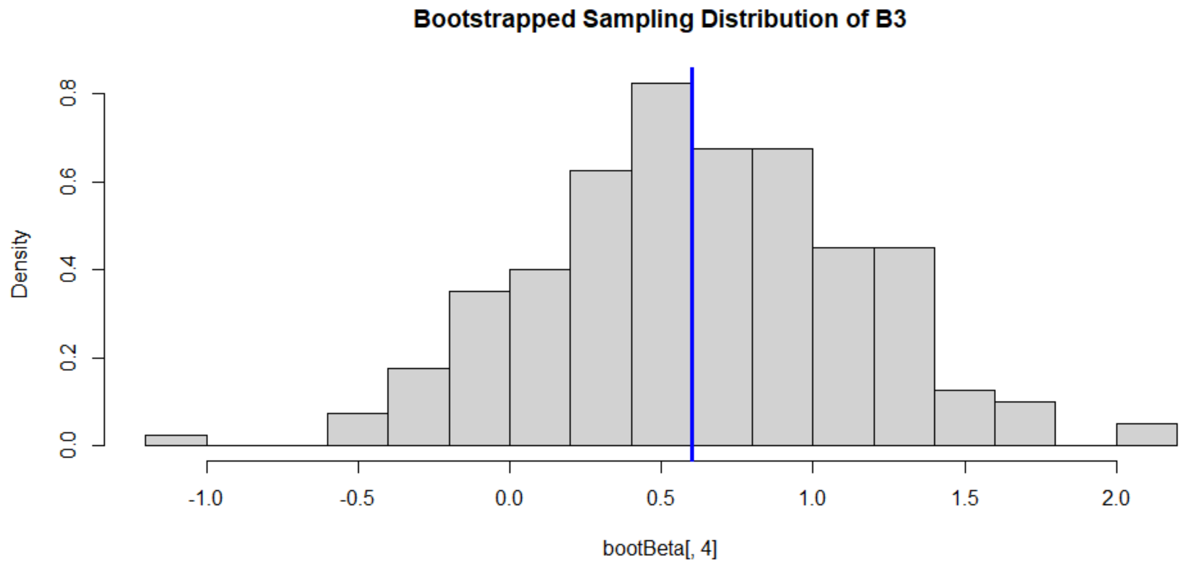
$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8 + \beta_9 x_9 + \varepsilon$$

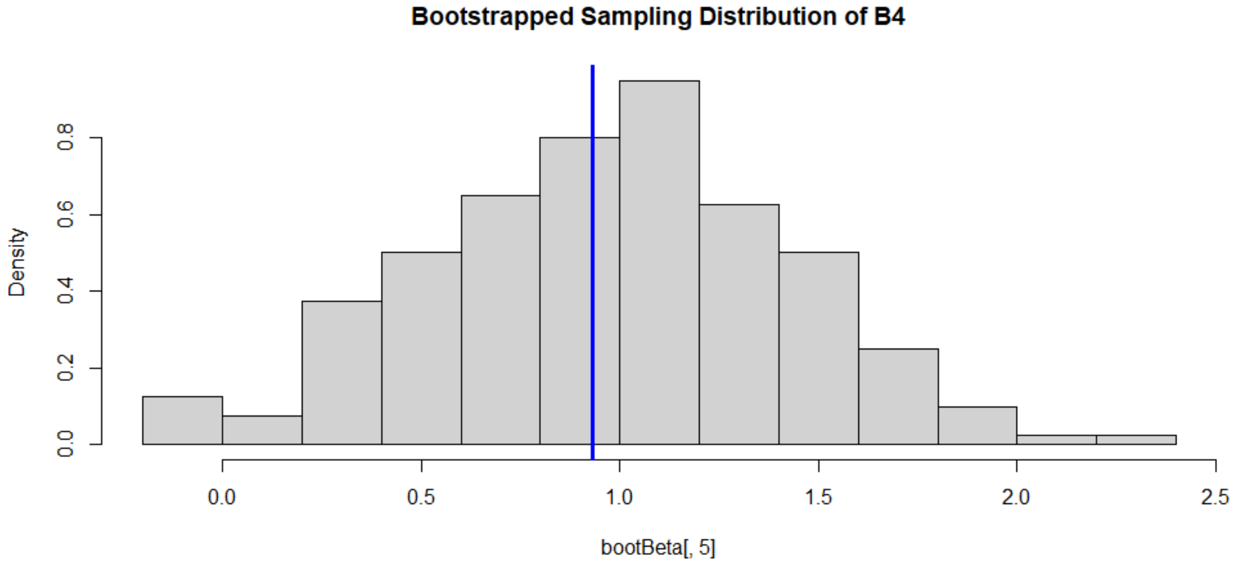
This project aims to explore the possible population effects of each xi on the total user scores for each drug type. To do so, estimates of linear coefficients for each factor need to be made. The beta values of each factor from the original parametric inferences will then be compared to the newly obtained estimated beta values from boot strapped data sets.

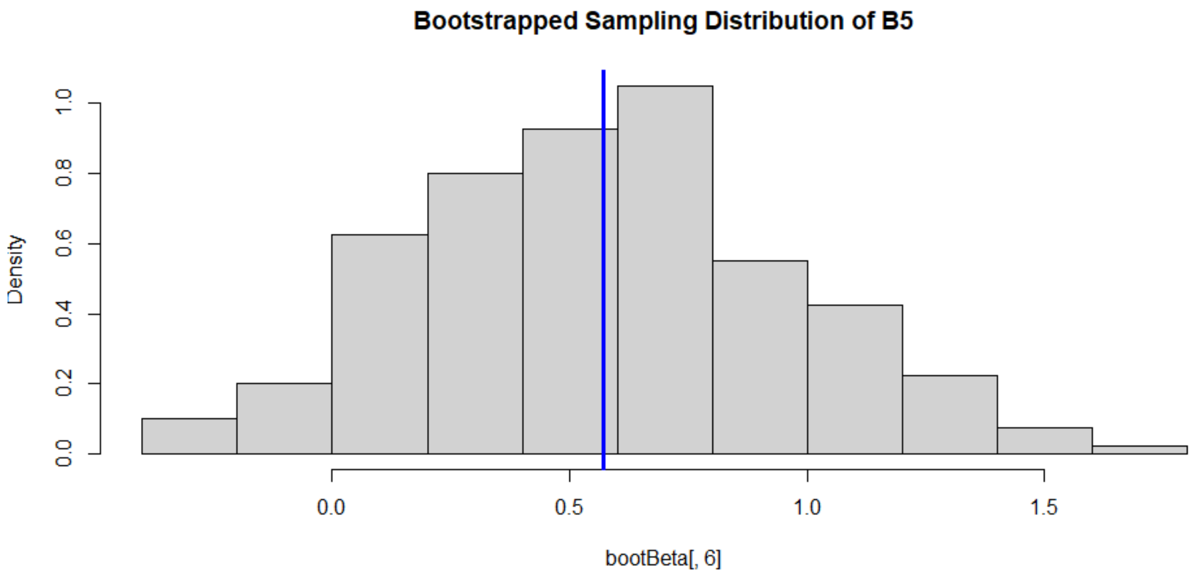
Bootstrapping in this project was aimed at creating samples for estimating linear coefficients, or beta values. Due to the homoscedastic nature of the sample, the initial bootstrapping was accomplished by sampling both the independent variables and the dependent variables' values. To bootstrap under this guideline, the residuals of each observation were gathered from the dataset and set as estimates of the population residuals. Using these residuals and the original x values, artificial beta values are constructed for each of the N datasets constructed from bootstrapping.   The number N bootstrapped datasets is determined by selecting a number deemed to be the largest but computationally as efficient as necessary. For the purposes of this project 100 bootstrapped samples were constructed for each of the 25 studies. The matrix formula Y = XBhat - e represents the recreation of Y data for the bootstrap after the residuals are sampled.   Below are graphs for each of the beta values bootstrapped sampling distributions.

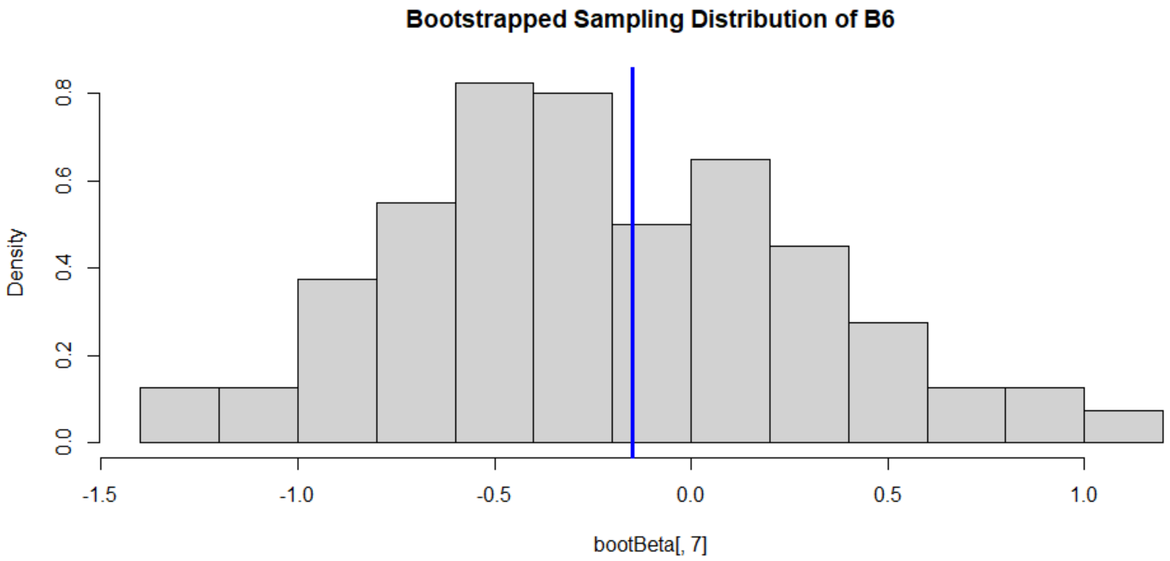

Bootstrapped Sampling Distribution of B1
Density
0
500
1000
2000
3000
-4e-04
-2e-04
0e+00
2e-04
4e-04
bootBeta[, 2]

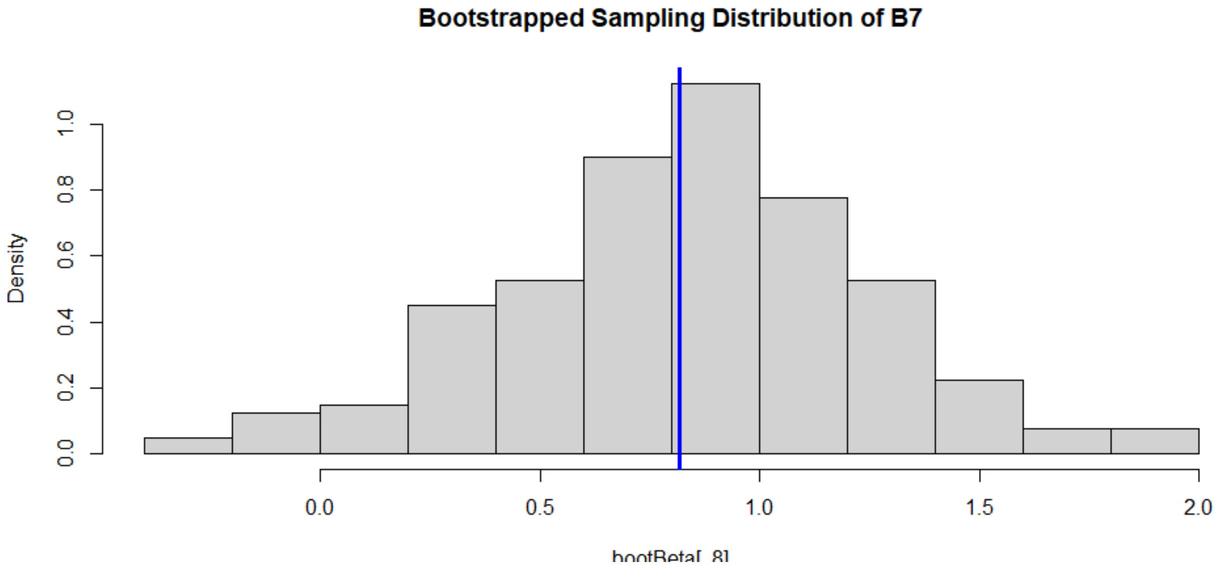

Bootstrapped Sampling Distribution of B2
Density
0.0
0.5
1.0
1.5
-0.5
0.0
0.5
bootBeta[, 3]

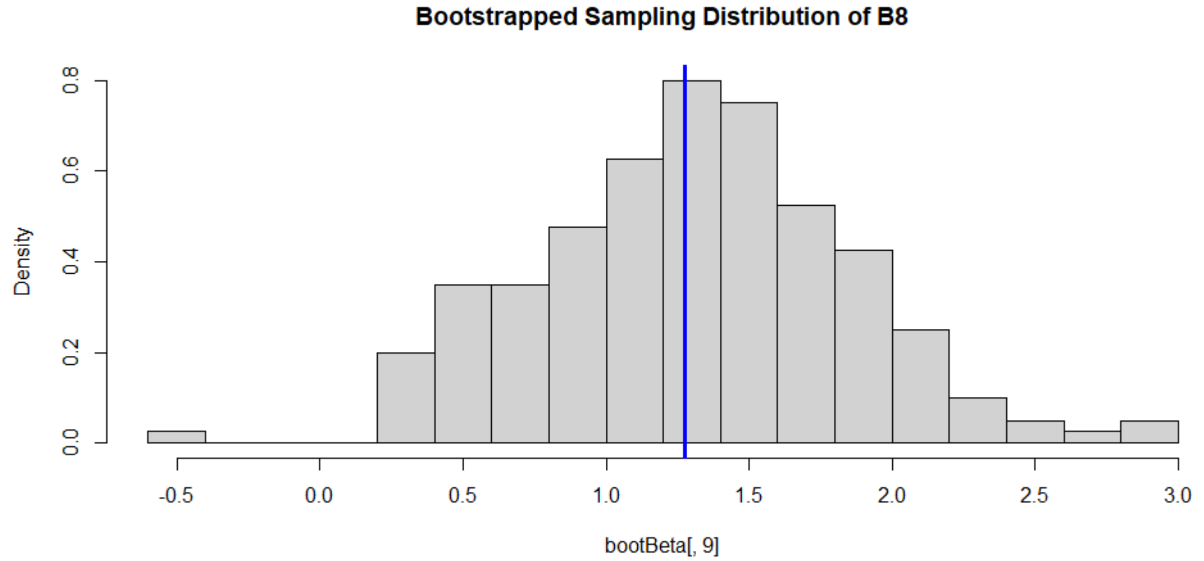

Bootstrapped Sampling Distribution of B3
Density
0.0
0.2
0.4
0.6
0.8
-1.0
-0.5
0.0
0.5
1.0
1.5
2.0
bootBeta[, 4]


Bootstrapped Sampling Distribution of B4
Density
0.0
0.2
0.4
0.6
0.8
0.0
0.5
1.0
1.5
2.0
2.5
bootBeta[, 5]


Bootstrapped Sampling Distribution of B5
Density
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.5
1.0
1.5
bootBeta[, 6]


Bootstrapped Sampling Distribution of B6
Density
0.0
0.2
0.4
0.6
0.8
-1.5
-1.0
-0.5
0.0
0.5
1.0
bootBeta[, 7]



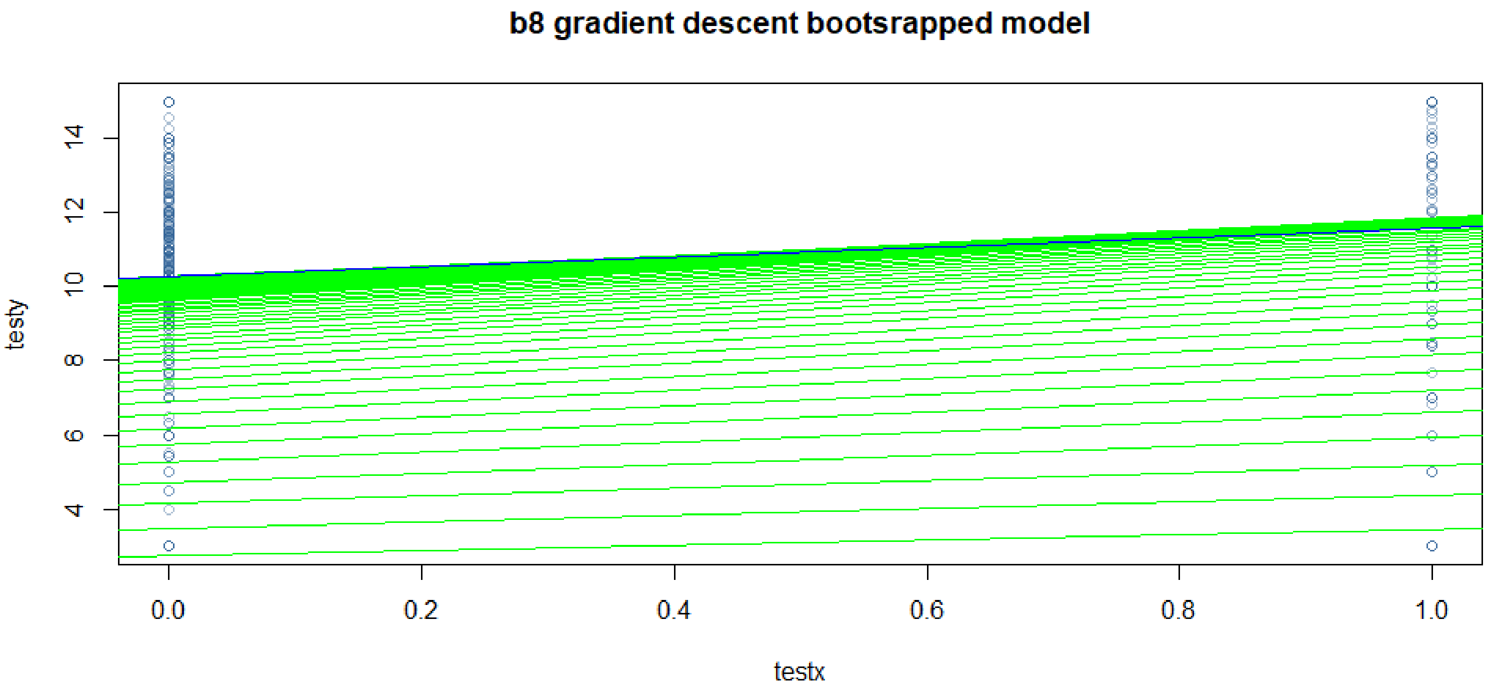


Earlier studies on this dataset established the beta value 8, the coefficient for OTC, as the only which is relevant towards the fitting of the model. Thus, the formula for inference was reduced to:

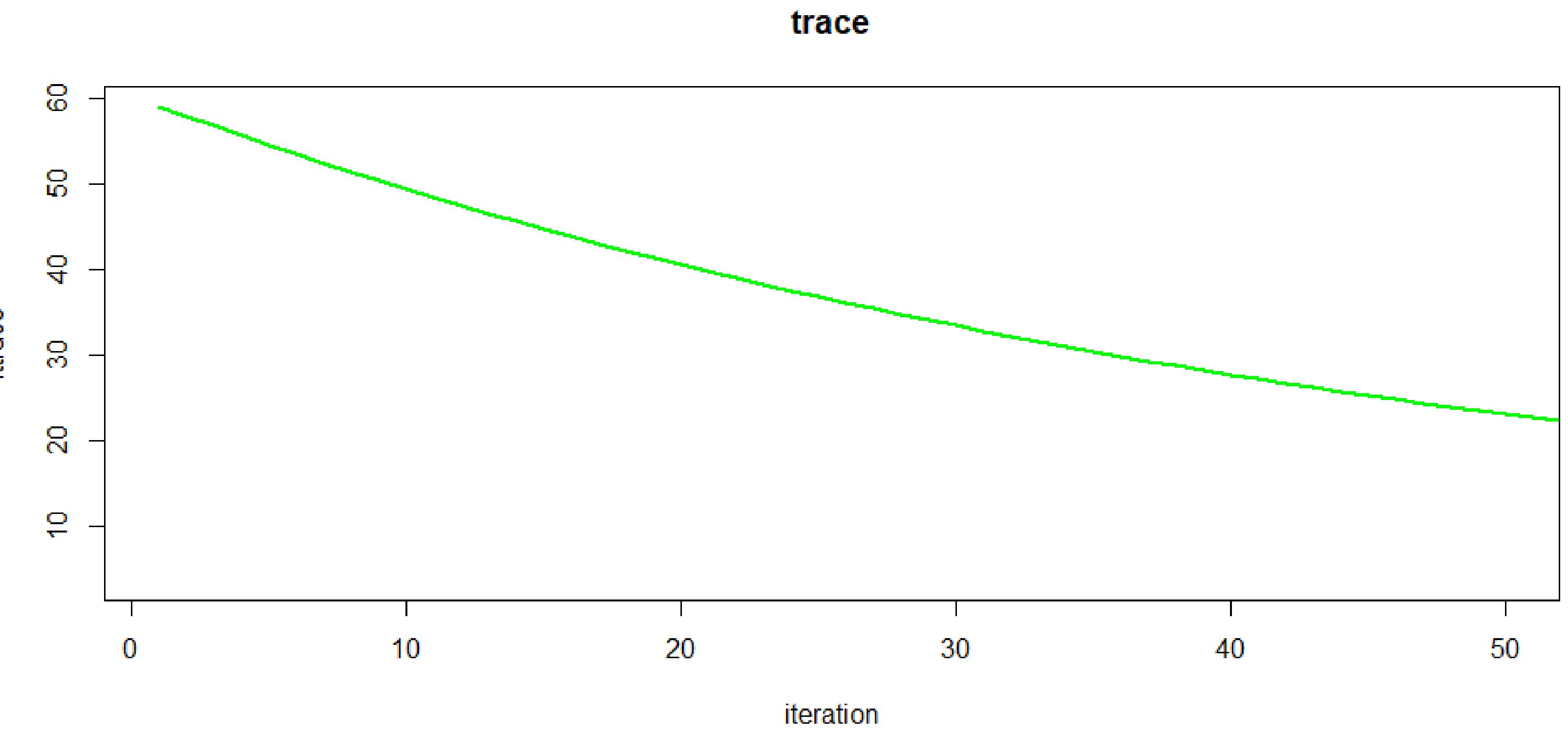
$$Y = \beta_0 + \beta_8 x_8 + \epsilon$$

The heteroscedastic nature of the residuals for only OTC warranty an additional bootstrap approach. The x values from OTC and y values were sub sampled at size m. The number m for the sample size of each bootstrap should be as close as possible to the original sample size, as larger sample sizes for any level of inference will reduce error. I decided to use sample sizes of 400 to potentially emulate a real life scenario in which a perfect resample is not feasible.

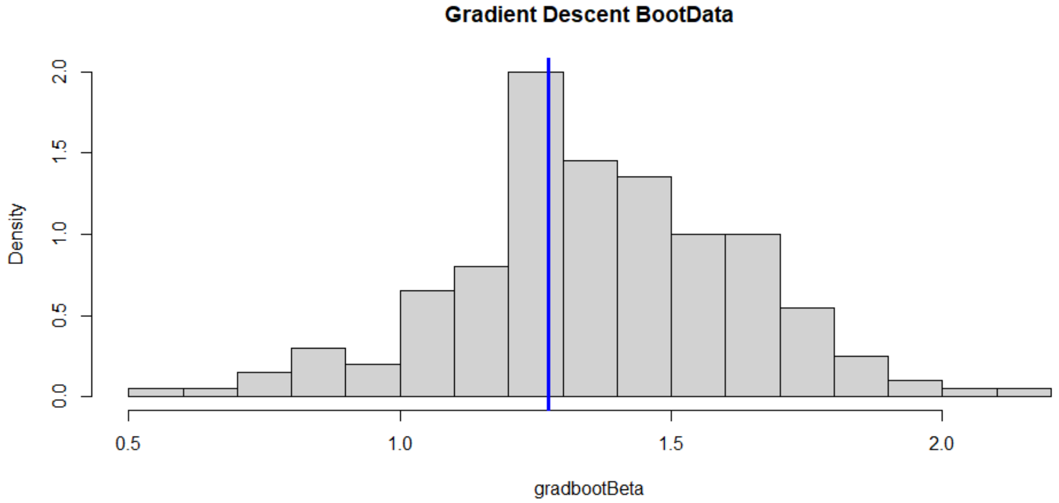
Using gradient descent, convergence to the estimated bootstrapped beta values was established for every iteration of bootstrap. In total, it was determined that 50 total bootstrap iterations deemed a computationally acceptable number of bootstraps, under the typical guidelines of normal distribution sizes. One of the results is pictured below:



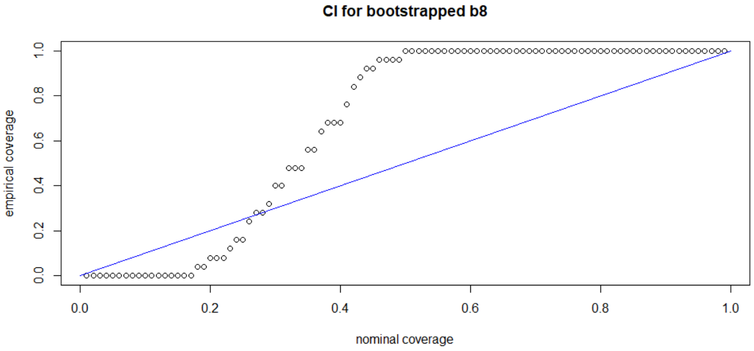
Its iterated trace:



The values converged to are estimated to be similar to the originally calculated beta value, as represented by this histogram:



A confidence interval was then conducted for the coefficient. Coverage for it is shown below:



A larger number of simulations would be necessary to fix the coverage.

This model shows that there is an existence of correlation between a drug being available OTC or not. It has been established that there is a statistically significant interaction between this variable and people's responses. People tend to rate drugs available over the counter moderately higher than drugs that are not. This makes sense, as seeking out a doctor before taking medicine can be annoying and costly.Overall, despite the simplicity of the proposed model, this study seems as it warrants at the very least a continuation of study of this dataset.

The biggest issue present in the model presented is the exclusion of interaction. These interactions likely play a highly significant role in the determination of the data, given the sheer quantity of interactions present. Calculating every one of those interactions by hand would prove to be an arduous task and is likely best suited for a program.

https://www.kaggle.com/datasets/thedevastator/drug-performance-evaluation?select=Drug_clean.csv

https://zenodo.org/record/3571494#.ZElO4XbMJPZ