Prediction of the 1st Percentile of Concentration of Drug E in Blood that is Sufficient to Cause Cardiac Arrest Based on Systolic Blood Pressure Using Resampling Methods

Chris Chen, Alvin Chen, Tim Li, Zihan Li

June 2022

1 Introduction

1.1 Context

Cardiac arrest, defined by Johns Hopkins Medicine [1], is when the heart stops beating suddenly. The lack of blood flow to the brain and other organs make cardiac arrest extremely lethal–over 90% patients die within just a few minutes. By American Heart Association [2], on average, about 14% percent of all death cases in one year in United States results from cardiac arrest. Even if a patient survives, it is difficult for him or her to return to a healthy physical condition: among 200,000 in-hospital cardiac arrests (IHCAs) occur each year in the USA, about 17.6% survive hospital discharge, and only 13.6% of them have favorable neurological outcomes. An analysis of 70 studies involving 142,740 out-of-hospital cardiac arrests (OHCAs) show even lower numbers: about 7.6% survive hospital discharge and less than 10% have favorable neurological outcomes [2].

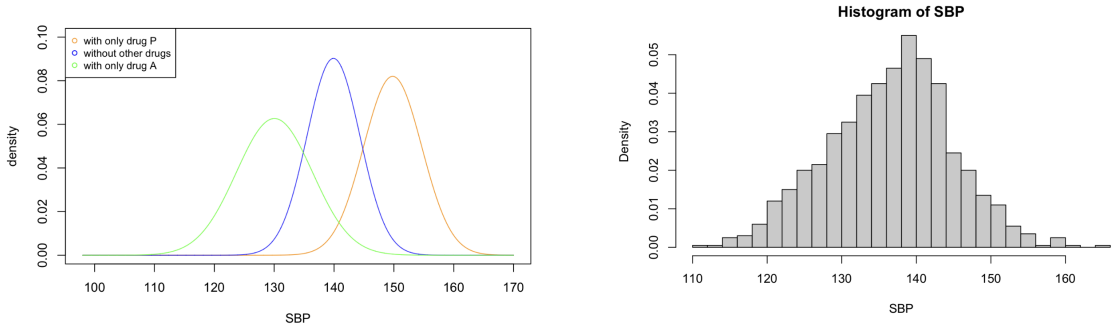
1.2 Significance

A recently published research on a common antibiotic Drug E reported that a high concentration of Drug E in human blood carries a non-negligible risk of causing cardiac arrest. The report also mentioned that the concentration of Drug E in blood that will be sufficient to cause cardiac arrest varies across individuals. One factor associated with this variation was systolic blood pressure, but the details were only briefly mentioned and no scientific statements were given. Knowing that cardiac arrest is a terrifying cause of death with over 90% lethality rate and this rate is even higher among older people who are frequent takers of Drug E, we wish to investigate the relationship between 1st percentile of concentration of Drug E in blood that is sufficient to cause cardiac arrest (C) and systolic blood pressure (SBP). In particular, our group seeks to come up with a model that can help us predict C based on SBP as accurate and precise as possible and hence reduce the probability of older people getting cardiac arrest due to careless usage of drug E as much as possible.

2 Data

2.1 Generation Process

The data generated contains two variables of first percentile of concentration of drug E detected in cardiac arrest patient's bloodstream and the Systolic Blood Pressure measured at 7am each day. We applied a mixture of 3 different multivariate normal distribution when generating the data and we generated 1000 observations. The 3 different multivariate normal distribution reflects the three subpopulations in our population. In particular, the



(a) Subpopulations of SBP

(b) Distribution of SBP

Figure 1: Subpopulations and distribution of SBP

three subpopulations reflects three different scenarios regarding the intake of drug E. The first subpopulation consists of patients who only took drug E. This subpopulation takes up around 50% of the overall population with a sample size of 494. The second subpopulation consists of patients who took both drug E and drug A, which is a known drug that can weaken the effects from drug E. This subpopulation takes up around 40% of the overall population with a sample size of 401. The third subpopulation consists of patients who took both drug E and drug P, which is a known drug that can amplify the effects from drug E. This subpopulation takes up the remaining 10% of the overall population with a sample size of 105. Due to the difference between subpopulations, the parameters of these subpopulations are also different.

In addition, the multivariate normal distribution we used for data generation has a parameter of covariance of -0.5 between the independent variable of SBP and dependent variable of C. This covariance allow us to generate a negative relationship between SBP and C since this relationship is well-known in pharmaceutical industry.

2.2 Independent Variable

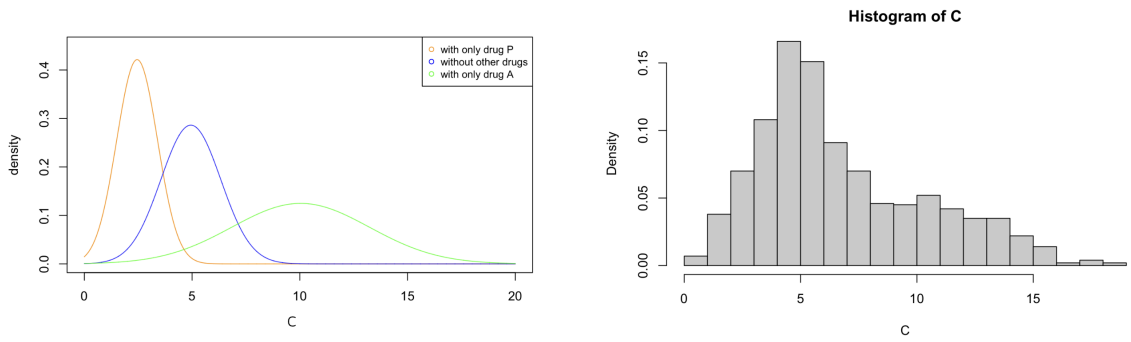
The independent variable in our data is each patient's systolic blood pressure measured at 7am each day (SBP) in the unit of mmHg. The data of SBP is a mixture from the three subpopulations mentioned in previous section. As shown in Figure 1, these subpopulations share very different characteristics: the subpopulation reflecting drug intake of only drug E has a mean of 140 mmHg with a variance of 20, the subpopulation reflecting drug intake of drug E and drug A has a mean of 130 mmHg with a variance of 40, the subpopulation reflecting drug intake of drug E and drug P has a mean of 150 mmHg with a variance of 20.

On the histogram of distribution of the mixture of SBP as shown in Figure 1, we observe a normal shaped distribution centered around 140 mmHg.

2.3 Dependent Variable

The dependent variable in our data is the first percentile of concentration of drug E detected in cardiac arrest patient's bloodstream in the unit of μmol/mL. Like data of SBP mentioned previously, the data of C is a mixture from the three subpopulations as well. As shown in Figure 2, these subpopulations share very different characteristics: the subpopulation reflecting drug intake of only drug E has a mean of 5 μmol/mL with a variance of 2, the subpopulation reflecting drug intake of drug E and drug A has a mean of 10 μmol/mL with a variance of 10, the subpopulation reflecting drug intake of drug E and drug P has a mean of 2.5 μmol/mL with a variance of 1.

On the histogram of distribution of the mixure of C as shown in the right of Figure 2, we notice that the distribution is heavily skewed to the right.



(a) Subpopulations of C

(b) Distribution of C

Figure 2: Subpopulations and distribution of C

3 Statistical Analysis

The ultimate goal of our project is to construct a model that can aid us in making predictions of C based on information on the patient's SBP and prescriptions. Both a parametric approach and a nonparametric approach will be discussed; the accuracy of the sample estimates will be assessed via bootstrapping and the models constructed will be compared and evaluated based on their performance on test data.

3.1 Parametric Model

We use C as the dependent variable and SBP as independent variable to perform a linear regression analysis. That is, we propose a model of the form

$$C_i = \beta_0 + \beta_1 \cdot SBP_i + \epsilon_i, \epsilon_i \overset{\text{i.i.d.}}{\sim} N(0, 10), i = 1, ..., n \tag{1}$$

Analysis are all conducted with the statistical computing software R 4.1.2 and the IDE Rstudio on macOS Big Sur Version 11.6. The default significant level is $\alpha = 0.05$ if not suggested otherwise.

3.1.1 One Linear Model per Subgroup

Mentioned previously, the report did not specify "how C varies across individuals". Hence our first guess was that C varies across individuals based on SBP and their prescriptions. Therefore, one linear regression of C on SBP was constructed for each subpopulation. The intercepts, slopes, and the slope-corresponding p-values are shown in the table below.

Subpopulation	Intercept	Slope	Slope p-value
1	7.42	-0.0322	0.141
2	6.88	-0.0138	0.303
3	11.0	-0.00776	0.739

Surprisingly, there does not exist any significant linear relationship between C and SBP. All slope p-values are greater than 0.05, and the magnitude of the slopes are tiny (all smaller than 0.05). Empirical bootstrap slope p-values were calculated for these three subpopulations and the results yielded were 0.491, 0.499, 0.498, correspondingly. We could not find any significant linear relationship between C and SBP within each subgroup.

3.1.2 One Linear Model for the Entire Population

However, hinted by our previous biomedical knowledge, we believe some type of linear relationship should exist since high level of SBP is associated with cardiomyopathy, which

is in turn strongly related to susceptibility to cardiac arrest; at the same time, higher C increases the patient's susceptibility to cardiac arrest. Therefore, we expect to see a negative linear relationship between C and SBP, logic being it takes lower C for higher SBP individuals to get cardiac arrest. A new linear regression of C on SBP is performed, but this time on all the data, i.e. no subpopulation constraints. A significant linear relationship is found and the result is provided below:

Intercept	Intercept p-value	Slope	Slope p-value
43.7	<2e-16	-0.270	<2e-16

3.2 Residual Analysis

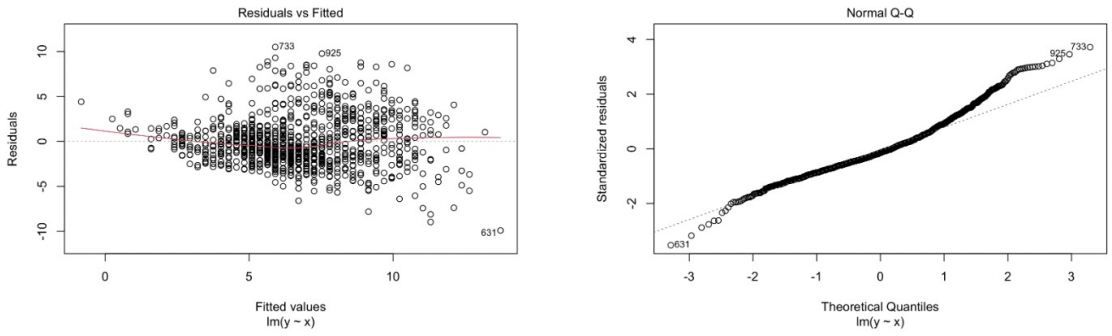


Figure 3: Residual plot (left) and Normal QQ-plot (right) of the linear model

A natural step after any linear regression is to conduct residual analysis. Mentioned previously, we assumed the errors to be i.i.d. normal with mean 0 and variance 10 and now we take steps to check if these assumptions are satisfied. Figure 3 shows the residual plot and the normal qq-plot. As shown, the residuals appears to have a cone shape from 0 to 6, and no pattern after 6. This indicates that the identical assumption might not hold. Namely, heteroskedasticity might exist. A variance test is performed to the residuals that are smaller than 6 and residuals that are greater than or equal to 6. A F-statistic of 0.381 and a corresponding p-value of $8.34 \cdot 10^{-8}$ are yielded. This means heteroskedasticity indeed exists and certain steps need to be taken while bootstrapping to deal with this violation. The qq-plot shows that our distribution has a very long right tail comparing to a normal distribution, so the normality assumption doesn't hold, either. A Durbin-Watson test is then performed on the residuals and the p-value is 0.152, which means the independence assumption is satisfied. Cook's distance is calculated for each observation; and we can spot a few points of high leverage comparing to others in Figure 4. Steps also have to be taken to deal with this problem.

3.3 Assessment of Sample Estimates Accuracy and Model Performance

Cross validation is used to assess model performance and resampling methods, in particular bootstraps, are implemented to assign measures of accuracy to the sample estimates. We use three different bootstrap methods to assess the accuracy of the sample estimates.

An empirical bootstrap is done firstly–a method that directly bootstraps on the original pairs of data. This is the desired solution when conditions are ideal, i.e. the assumptions we made about the model are all satisfied. However, as we've explained in the residual analysis section, almost all assumptions fail: high leverage points linger, heteroskedasticity exists, and the normality assumption does not hold. A residual bootstrap is then introduced to deal with high leverage points by bootstrapping on residuals instead of on original data

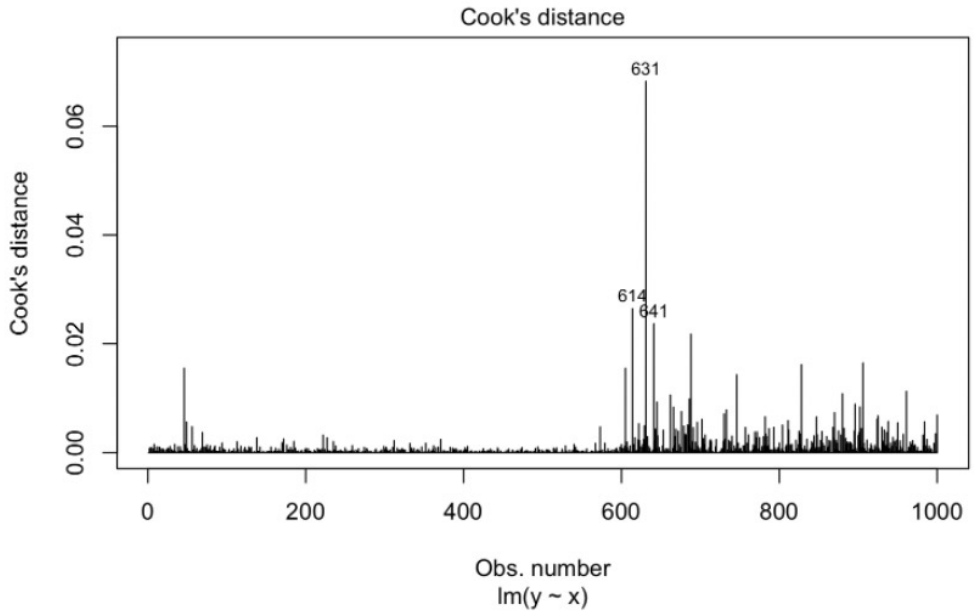


Figure 4: Cook's distance of observations

pairs. A wild bootstrap is done to counter heteroskedasticity by randomly multiplying the residuals by a random variable v_i with mean 0 and standard deviation 1. Idea behind wild bootstrap is to "homogenize" the variances by resampling the response variables based on the residual values while leaving the regressor at sample value.

The estimates of coefficients, the biases, variances, and MSEs using all three bootstrap schemes will all be provided in the Results section. Combining with a 10-fold CV, we select the best set (the set that gives the lowest test MSE) of the coefficients for each method and additionally record the test MSEs. They will be provided in the Results section.

3.4 Nonparametric Model

Instead of assuming the data comes from a linear relationship, the nonparametric method derives a model based on the distribution of existing data without assuming any parameters. In particular, we use a local method called kernel density estimation to predict C. This approach is justified given the sample size $n = 1000$. The prediction at each SBP is calculated as following:

$$\hat{p}_n(SBP_i) = \frac{1}{nh}\sum_{i'=1}^{n} K(\frac{SBP_{i'} - SBP_i}{h}) \tag{2}$$

Careful choices of the kernel function and bandwidth are crucial to the success of the prediction. In this study, we select the kernel function to be the normal distribution in favor of its tail smoothing and we use a 10-fold cross-validation to select the best bandwidth among 1 to 10 since the smallest unit in SBP is 1. Best is defined in terms of smallest test MSE returned. The test MSE from the model with the best bandwidth is recorded and will be provided in the Results Section.

One could reasonably argue that a kernel density estimation should be done separately on each of the subpopulations. However, to be consistent with parametric approaches (so that the comparison of model performance can be meaningful), we choose to abandon this part.

3.5 Assessment of Model Performance and Prediction Uncertainty

To assess model performance and prediction uncertainty, we resort to the three different bootstrap methods (empirical, residual, and wild bootstrap) again. Namely, a 99.8% normal confidence interval is constructed for the predictions using each of the three bootstrap methods. Details will be provided in the Results Section.

4 Results

4.1 Parametric Model

		Coefficient	Bias	Variance	MSE
classical t-test	$\widehat{\beta}_0$	43.6765			
	$\widehat{\beta}_1$	-0.2698			
paired bootstrap	$\widehat{\beta}_0$	43.42197	-0.2533588	2.915679	2.979869
	$\widehat{\beta}_1$	-0.2677195	0.002046262	0.0001482694	0.0001524566
residual bootstrap	$\widehat{\beta}_0$	43.71814	0.03484374	2.606416	2.60763
	$\widehat{\beta}_1$	-0.2698818	-0.0001256981	0.0001380611	0.0001380769
wild bootstrap	$\widehat{\beta}_0$	43.69086	0.01939072	2.911443	2.911819
	$\widehat{\beta}_1$	-0.2698906	-0.000090681	0.0001479145	0.0001479303

Figure 5: parametric regression result with coefficients, bias, variance, and MSE

Figure 5 demonstrates the sample estimates of linear model coefficients and their corresponding accuracy, measured in terms of bias, variance, and MSE using the three bootstrap resampling methods. From the classical t-test, the estimate of intercept is 43.7 and the estimate of slope is -0.270. Both estimates have significant p-values $<$ 2.2e-16. Accuracy of these sample estimates are then calculated by the three bootstrap resampling methods. In general, the estimated intercepts and slopes from three bootstrap methods are approximately the same, and they differ only by a little to the estimates given by the classical t-test. However, each bootstrap method gives a different set of bias, variance, and MSE for each of the coefficient estimates.

As explained in the Statistical Analysis section, almost all the assumptions we made on the distribution of error are invalid. Consequently, one should expect empirical bootstrap to perform the worst among all three bootstrap methods. It is indeed what we observe: bias, variance, and hence MSE are all the largest.

Residual bootstrap is designed to counter the increase in variance of bootstrap estimates caused by high leverage points. Therefore, unsurprisingly, residual bootstrap gives the lowest variance among all three bootstrap methods, and the bias and MSE given are considerably smaller than those given by the empirical bootstrap.

Wild bootstrap algorithm assumes that the "true" residual distribution is symmetric and centered at 0 and hence can offer advantage over empirical bootstrap and residual

bootstrap on bias, especially for smaller samples. Therefore, expectedly, we can see that wild bootstrap gives the lowest bias, and the variance and MSE are smaller than those given by the empirical bootstrap.

The test MSEs given by a 10-fold cross-validation is presented in the following table.

Resampling Methods	Test MSEs
Empirical bootstrap	4.965
Residual bootstrap	4.959
Wild bootstrap	5.380

Residual bootstrap triumphs the other two methods with a test MSE of 4.959, while wild bootstrap performs gives the highest test MSE: 5.380.

4.2 Nonparametric Model

The best bandwidth selected by cross-validation is 7. Figure 6 shows the test MSE vs smoothing bandwidth. An U-shape curve is observed, implying that too small a bandwidth could lead to underfitting yet too large a bandwidth could cause overfitting. The test MSE given by this bandwidth is 7.858, considerably larger than all the test MSEs we obtained via parametric approaches.

Moreover, three bootstrap methods were again implemented to quantify the uncertainty of prediction. Figure 7 shows the predictions given by kernel estimation with 99.8% normal confidence intervals constructed by the three bootstrap methods.

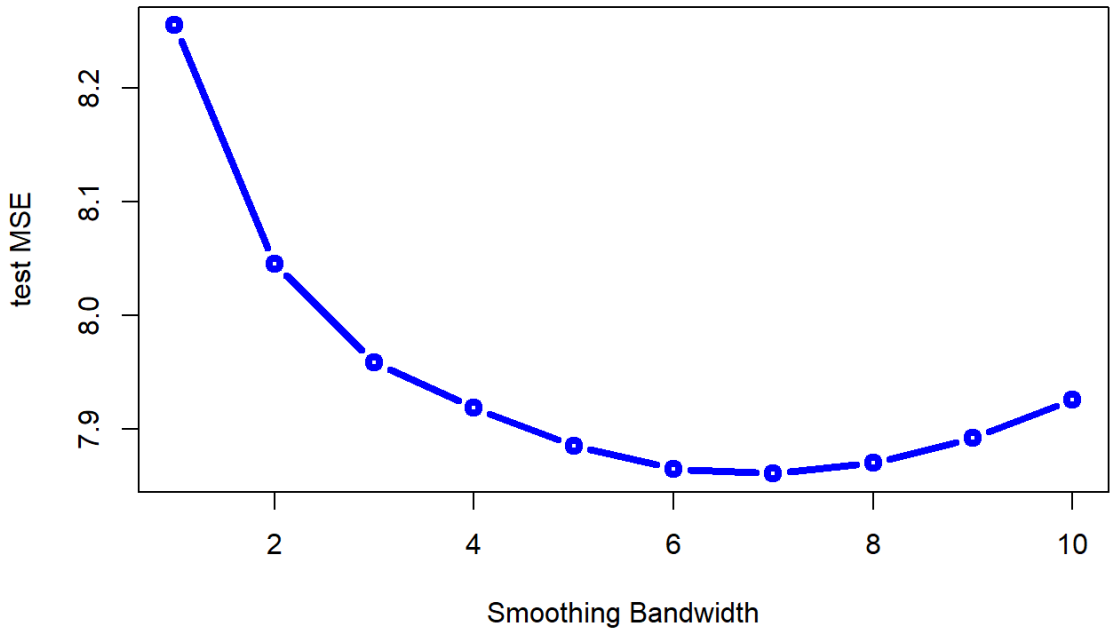
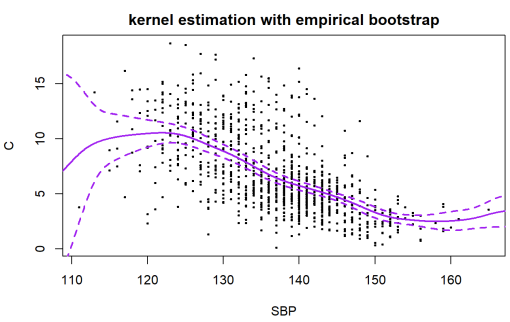


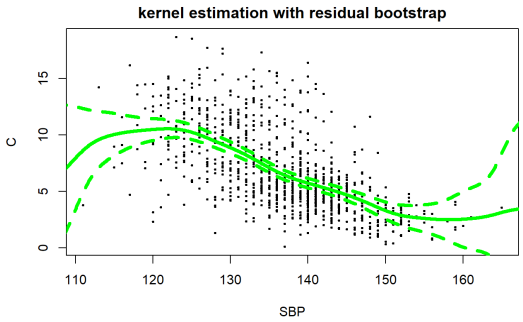
Figure 6: test MSE value for each bandwidth in kernel density estimation

5 Conclusion

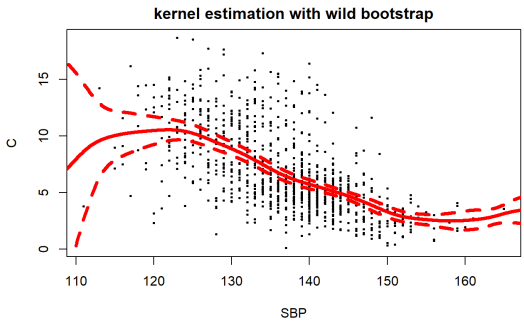
To determine the best model, a careful comparison on all available models is required. First of all, we decide on the approach. A significant linear relationship is found between SBP and C; but one might come up with this question: why a significant linear relationship only exists in the entire population but not in each subpopulation? A possible answer is that the response is discrete, i.e., a given range of SBP is only associated with one C. This



(a) Empirical Bootstrap on KDE with 99.8% CI



(b) Residual Bootstrap on KDE with 99.8% CI



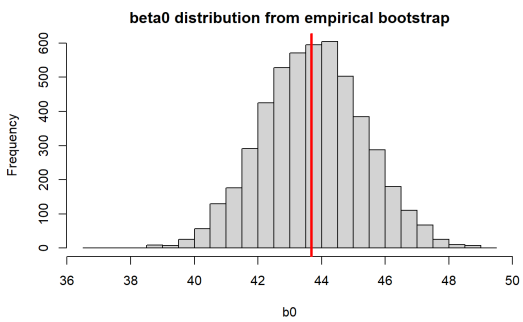
(c) Wild Bootstrap on KDE with 99.8% CI

Figure 7: Bootstrap resampling on kernel density estimation

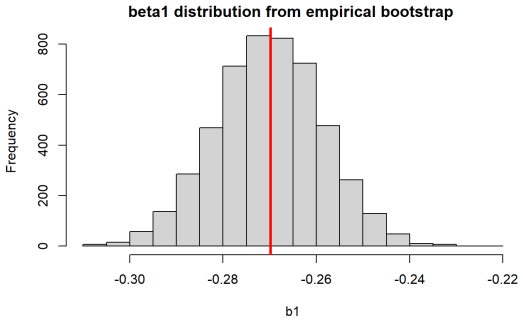
implies that maybe a nonparametric approach is superior to parametric ones. Nevertheless, an alternative explanation is that we simply do not have enough data to observe such a linear relationship, especially when the relationship is weak. Additionally, the test MSEs of parametric models are much lower than that of the nonparametric model, indicating the model fit of parametric models are much better. Therefore, we prefer a parametric model to a nonparametric model.

Secondly, we decide on the resampling scheme. All three bootstrap methods have their strengths and weaknesses, hence it's crucial for us to evaluate what is demanded the most under this study's context. We want the predictions to be as accurate and precise as possible, so naturally the MSE should is the metric we want to compare. Figure 8 shows that all bootstrapped coefficient estimates are approximately normally distributed, so that the MSEs of these estimates are accurate, meaningful, and comparable. Combining MSEs in Figure 5 with the theoretical background of each bootstrap method, we can readily filter out the empirical bootstrap method.

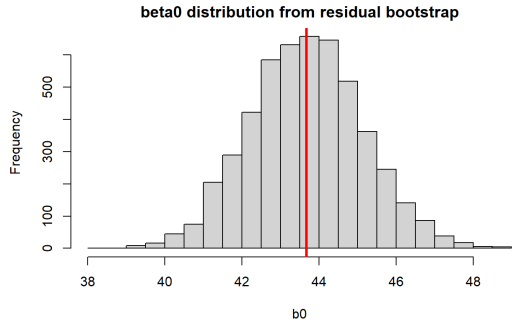
However, choosing the better between residual bootstrap and wild bootstrap is a more complicated matter. In terms of metrics alone, residual bootstrap is better–lower MSEs for both coefficient estimates, and lower test MSE. However, as we described in the Statistical Analysis section, residual bootstrap is designed to make variances, and hence MSEs of coefficient estimates small. Keep in mind that smaller MSE is not all that we are after–what we want is closer to reality. Therefore, this choice of resampling scheme boils down to this question: what is a more serious problem, heteroskedasticity or high leverage points? Looking at Figure 4 again, we see that the highest leverage is between 0.06 and 0.07. Empirically speaking, this is not very high. Usually a point with leverage above 0.25 is only considered problematic. But the variance test we done in residual analysis gives a test statistic of $F = 0.381$, i.e., the variance of a group of the residuals is more than twice than that of the rest. This is a more serious problem. Thus, our group prefers wild bootstrap for that it deals



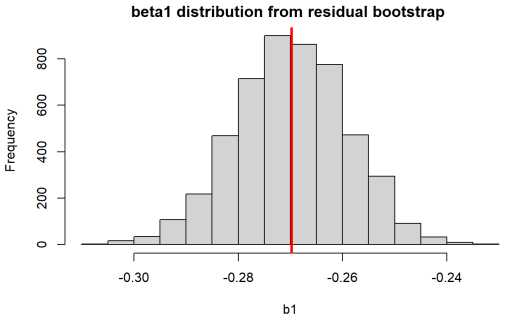
(a) intercept distribution from empirical bootstrap



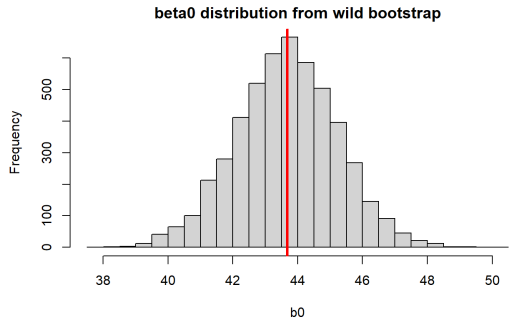
(b) slope distribution from empirical bootstrap



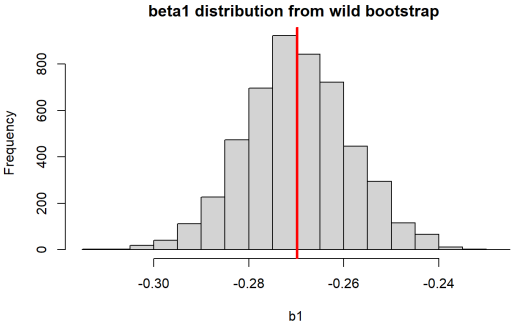
(c) intercept distribution from residual bootstrap



(d) slope distribution from residual bootstrap



(e) intercept distribution from wild bootstrap



(f) slope distribution from wild bootstrap

Figure 8: Distribution of coefficients from three bootstrap methods

with a more serious problem–it closer reflects reality. The final model we construct, is hence

$$C_i = 43.69086 - 0.26989 \cdot SBP_i \tag{3}$$

6 Discussion

This study is strongly goal-oriented. We kept one thing in mind throughout the whole process–how can we achieve the goal we set at the beginning of the study? That is, how can we construct a model that can help us predict C based on SBP as accurate and precise as possible? To construct a model that can predict C based on SBP, we seek help from linear regression and kernel estimation. To assess and improve the accuracy of the coefficient estimates and predictions using limited amount of data (one of the subpopulations only contained 105 observations), we resort to resampling methods. The need of more data and the idea that we do not possess any knowledge or belief on what the parent distribution should be like made nonparametric bootstrap the best resampling method for us.

However, this made our range of resampling methods narrow in some sense. Comparing to other groups, our study exploited all kinds of yet only bootstrap methods. We considered

adding other elements into our study, for example utilizing KS-test or Permutation test to differentiate subpopulations of C, or implementing Monte-Carlo to calculate the mean, median, and standard deviation of each subpopulation. However, due to the length restriction of this assignment, we decide to forgo those parts to keep our work "whole"–we want a complete and strong logic link between each part of our study and we desire to answer the question with the best we can. In the future, more exploratory analysis could be incorporated into such kind of studies when length is not longer a hard requirement.

7 Acknowledgement

We would like to express our special thanks to Professor Thompson for giving us this opportunity to work on such intriguing project, to put the things we learn in class into test, and to do something that is of real-world significance. We would also like to place our gratitude to group 3 and group 13 for their insightful comments and their help on our project.

References

[1] Johnathan Chrispin. Cardiac arrest. *Johns Hopkins Medicine*, 2022.

[2] Soar J. Wenzel V. et al. Nolan, J. Cardiopulmonary resuscitation and management of cardiac arrest. *Nature*, 2012.