Parametric Bootstrapping

CJ Carani, Dylan Hunt and Matt Kaznikov

02-21-2024

Background:

W.A. Nelson from the Encyclopedia of Ecology states, "Bootstrapping methods are a numerical approach to generating confidence intervals that use either resampled data or simulated data to estimate the sampling distribution of the maximum likelihood parameter estimates". This definition is the backbone reasoning on why and when a statistician should incorporate bootstrapping. The overall idea is that bootstrapping can greatly improve your findings and conclusions of a model's confidence intervals, p-values, F-statistics and hypothesis testing. Bootstrapping generally is also used when there is a lack of observations or there is data missing within the dataset. Bootstrapping solves this issue by resampling the data based off of the current observations that we as statisticians already have to generate plausible values that will allow us to create more concrete assumptions about our dataset. The simulated data will provide us with better reasoning on how well our models perfrom and which model to use by creating a comparison with the use of the new data. Non-parametric bootstrapping uses a technique called sampling with replacement. This procedure randomly selects observations, while putting those observations back into the real dataset.This will create a new sample that will have the same size as the real dataset. These steps are then repeated thousands of time to generate many different datasets for your model to use. The generated datasets use the real dataset to create new observations for our simulated datasets. Parametric bootstrapping uses a model's estimates to generate new datasets. These are the steps to use parametric bootstrapping: 1) Fit a model to the original dataset and record that model's parameters using their estimated values. 2) Next using the recorded estimated parameter values simulate many datasets. 3) For each newly created dataset, record each of the estimated parameters. 4) Then calculate statistics that are important for your conclusions. (p-values, F-statistic, t-statistics, etc.) 5) Create confidence intervals based off of the simulated values. There are multiple reasons for why we would create simulated data using parametric bootstrapping. The major reasons could be if we do not have enough data from our original dataset and need more to create conclusive evidence for a specific idea or hypothesis. Another is if we have doubts about our model assumptions. These doubts could be about normal distribution or skewness within the data.

Methods:

Linear Model Comparison:

Our goal was to view how parametric bootstrapping can help us compare linear models. We crated a full model with four variables and wanted to compare that to a less complicated reduced model with two variables. One way to compare is by comparing the AIC's of the two models and the p-value for the full model using the anova command in RStudio. In addition to being able to run that we could use parametric bootstrapping to generate data based on the reduced model, and then use that data to compare the models. Bootstrapping provides the opportunity to negate any outlier data points, by generating a set number of data points randomly. This allows our estimates to be more exact and allows us to better view which model truly fits the data better.We will do this by generating the f-statistics using parametric bootstrapping and using that to see if our initial measured f-statistic was able to be generated due to chance. This will then allow us to see which model is better for the purposed of the Airbnb data.

The first step for comparing LME models was creating the full model comprising of the variables bedrooms, room_type, overall_satisfaction, and reviews. This model serves as our more complex model for the act of comparing a reduced model to a more complicated model using parametric bootstrapping.Then, we created a reduced model comprising of the variables bedrooms and room_type. This model is predicted to not account for as much variability in the model, but will help us determine if our full model is too complicated. After creating the models, we chose to bootstrap each model individually to view how the confidence intervals changed. We then plotted those confidence intervals for our intercept(b0) on a histogram plot. After doing this we bootstrapped the data to fit our reduced model and used the bootstrapped data to compare the models. This allowed us to get a distribution of our f statistic giving us the chance to see if our measured f-statistic was due to chance or if we had evidence the more complex model was better.

Model without Large Dataset and Comparison:

For the elephant dataset there were two models that were created. Because of the lack of explanatory variables both models used age as explanatory variables. However, the second, full, model used the variable age2 which was the age of the elephants squared. This was done to see if there was a quadratic relationship between age of the elephants and their number of matings. Bootstrapping is essential for this test because there are doubts about the size of the dataset and if the model statistics that are calculated can be trusted because of how small the dataset is. The first bootstrapping procedure used was to calculate how significant our explanatory variable age is to multiple models that use randomly simulated datasets. The confidence interval was also calculated for the variable age. The next bootstrapping procedure used was to determine if the second, full model was significant enough against the reduced model. The process first calculated the the estimates for the reduced model then we use the estimates to create the simulated data many times. During each simulation we track the estimates of each model and their F-statistic. This is important because we will compare the first F-statistic created from the original dataset to the 10,000 randomly simulated F-statistics.

Model Comparison for Multilevel GLM:

The referees dataset allowed us to see how parametric bootstrapping can help us with a multilevel generalized linear model. The bootstrapping taking place in this part of the report is similar to the bootstrapping above, but we are using a more complicated model structure. Additionally, instead of testing for the randomly simulated f-statistic, we are randomly simulating 10,000 chi squared values consistent with the model to see which model is better. This is a continuation of all the work we have done with the other types of models, but we are able to study how bootstrapping can help with generalized linear models.

As done previously, a reduced and a full model was created. The reduced included one explanatory variable and a random slope. This random slope allows us to account for the difference associated with each basketball game. This term was chosen because each basketball game has many factors that can effect the fouls in a game, most notably different referees. A more complex model was then created that used all the previous variables and random slope, but added a random effect for foul difference. After creating the models we ran a bootstrap simulation 10,000 times that created data that fit the models and used this to estimate 10,000 chi squared values. These values were then graphed in a histogram so that we could view if our measured value occurred due to chance. This is important for viewing if our advanced model is better than our reduced model.

Data:

Linear Model Comparison:

When we compared the Full Model to the Reduced Model we are using the Airbnb data which we have used previously in class. This dataset has just over 1500 observations and 14 variables. In the first model which was labeled as "reduced" we used a linear model hence "lm". In this model, the dependent variable also known as the Response Variable is Price. We are trying to predict the Price of an Airbnb based off of the independent variables commonly referred to as the Explanatory Variables which in this case are bedrooms and room types. For the second model we still used the Airbnb data. The second model was labeled as "full" and we also used a linear model hence the “lm” again. In this model the Response Variable was also Price but the difference from the reduced model was that now we used bedrooms, room type, overall satisfaction score, and reviews as the Explanatory Variables. The difference between the two models is that the Full Model is more complex because of the two extra explanatory variables.

Model without Large Dataset and Comparison:

The dataset used for demonstrate an example of a model without large data is the Elephants dataset created by Poole (1989). This dataest was used to figure out factors that contribute to the amount of matings that an elephant does in a given year. The only explanatory factor is AGE of the elephant while the response variable is MATINGS. The only data modifications that were made was the addition of another column that accounted for the AGE variable being squared which is labeled age2. The biggest concern about this data is the lack of observations within the dataset. There are only 41 observations which will not allow for concrete conclusions about the models that will be created. This is a perfect example for parametric bootstrapping.

Model Comparison for Multilevel GLM:

The dataset used to for multilevel generalized linear model parametric bootstrapping was the Referees dataset. This data compiles fouls from 340 college basketball games from 2009-2010. In total 4,972 fouls occurred giving us a large starting data with many occurrences. The Referees dataset’s goal was to find out the likelihood for given teams to be called for a foul in different situations. To do this various variables like home team, visiting team, number of fouls, date, foul difference, game, home team fouls, and visitor fouls were used.

All of these variables created two different levels of data. It is important to distinguish these levels to help interpret the results of our bootstrapping. The first level is the one with more occurrences and measured the characteristics of each individual fouls, while the second level measured the characteristics of each individual game. As you can imagine there are usually many fouls in a game, so there are many fouls with the same game characteristics in the dataset. Variables like home team, visiting team, date, and game are specific to each game and thus are second level variables. While variables like foul differential are specific to a given foul in the game and thus are level one variables. For our models we will be concerned with the variables home team fouls, foul difference, and game.

There are no major concerns with this data, but the goal is to use parametric bootstrapping to see if a more complicated model with a random effect term is better than a reduced model without one.

Results:

Linear Model Comparison:

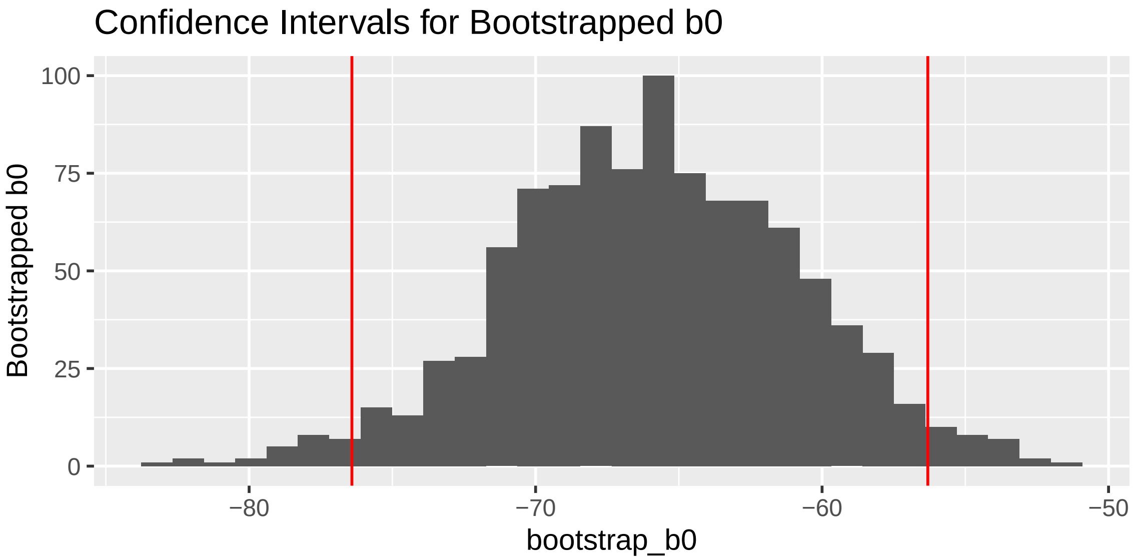
Model 1: Reduced Model

```
##
## Call:
## lm(formula = price ~ bedrooms + room_type, data = Airbnb)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -213.86  -27.47   -8.47   16.53  930.50
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)             69.496      4.982  13.948  < 2e-16 ***
## bedrooms                47.342      2.506  18.892  < 2e-16 ***
## room_typePrivate room  -53.372      3.998 -13.350  < 2e-16 ***
## room_typeShared room   -82.138     10.130  -8.108 1.03e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 68.36 on 1557 degrees of freedom
## Multiple R-squared:  0.4027, Adjusted R-squared:  0.4015
## F-statistic: 349.9 on 3 and 1557 DF,  p-value: < 2.2e-16
```

Model 1: Bootstrapping

```
##   bootstrap_b0     bootstrap_b1    bootstrap_b2        bootstrap_b3
##  Min.   :-82.88   Min.   :39.75   Min.   :-13.33566   Min.   :-33.4869
##  1st Qu.:-69.22   1st Qu.:45.68   1st Qu.: -2.72452   1st Qu.: -7.1049
##  Median :-65.92   Median :47.12   Median : -0.01831   Median : -0.4522
##  Mean   :-65.91   Mean   :47.21   Mean   : -0.04644   Mean   : -0.3882
##  3rd Qu.:-62.40   3rd Qu.:48.97   3rd Qu.:  2.68950   3rd Qu.:  6.5390
##  Max.   :-51.12   Max.   :55.89   Max.   : 12.15880   Max.   : 26.9619
##  bootstrap_sigma
##  Min.   :64.14
##  1st Qu.:67.46
##  Median :68.24
##  Mean   :68.29
##  3rd Qu.:69.12
##  Max.   :72.22
```

Model 1: Confidence Interval for Bootstrapped b0



```
##      2.5%     97.5%
## -76.41179 -56.30857
```

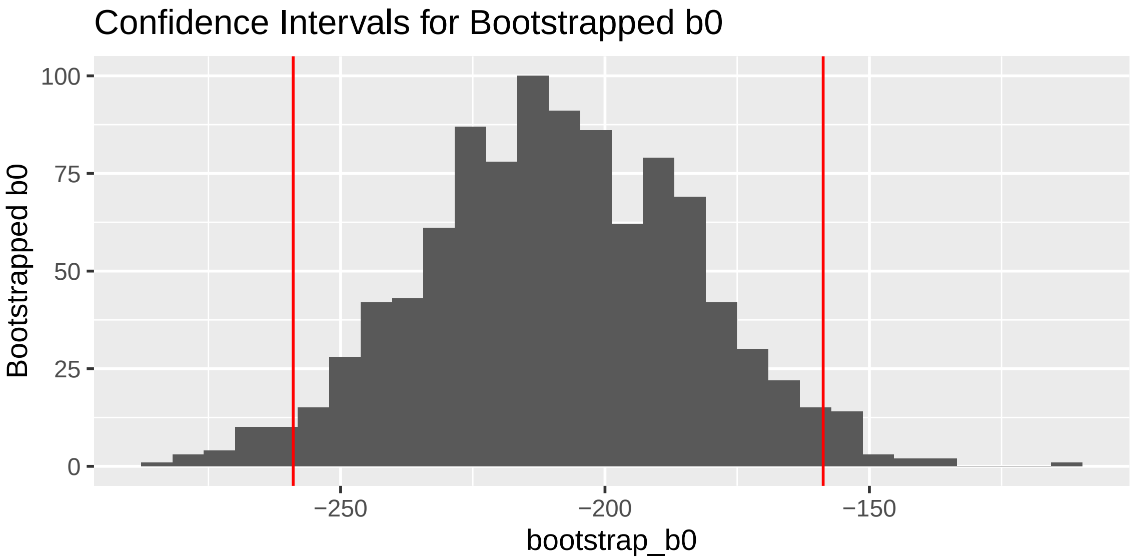
Model 2: Full Model

```
##
## Call:
## lm(formula = price ~ bedrooms + room_type + overall_satisfaction +
##     reviews, data = Airbnb)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -225.14  -28.17   -7.71   17.81  920.90
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           -72.21935   24.62148  -2.933   0.0034 **
## bedrooms               47.76056    2.47712  19.281  < 2e-16 ***
## room_typePrivate room -54.19290    3.95588 -13.699  < 2e-16 ***
## room_typeShared room  -81.64921   10.02502  -8.145 7.73e-16 ***
## overall_satisfaction   30.35449    5.06886   5.988 2.63e-09 ***
## reviews                -0.11371    0.04913  -2.314   0.0208 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 67.55 on 1555 degrees of freedom
## Multiple R-squared:  0.4176, Adjusted R-squared:  0.4157
## F-statistic:   223 on 5 and 1555 DF,  p-value: < 2.2e-16
```

Model 2: Bootstrapping

```
##   bootstrap_b0     bootstrap_b1    bootstrap_b2        bootstrap_b3
##  Min.   :-284.6   Min.   :38.60   Min.   :-12.92615   Min.   :-37.77335
##  1st Qu.:-225.3   1st Qu.:46.06   1st Qu.: -2.60918   1st Qu.: -7.12982
##  Median :-209.5   Median :47.75   Median : -0.08796   Median :  0.03589
##  Mean   :-208.9   Mean   :47.74   Mean   : -0.03818   Mean   : -0.35805
##  3rd Qu.:-190.5   3rd Qu.:49.48   3rd Qu.:  2.79032   3rd Qu.:  6.10441
##  Max.   :-112.6   Max.   :56.44   Max.   : 14.39723   Max.   : 31.56783
##   bootstrap_b4    bootstrap_b5   bootstrap_sigma
##  Min.   :11.59   Min.   :11.59   Min.   :63.61
##  1st Qu.:26.84   1st Qu.:26.84   1st Qu.:66.67
##  Median :30.58   Median :30.58   Median :67.50
##  Mean   :30.53   Mean   :30.53   Mean   :67.54
##  3rd Qu.:34.17   3rd Qu.:34.17   3rd Qu.:68.44
##  Max.   :46.86   Max.   :46.86   Max.   :71.09
```

Model 2: Confidence Intervals for Bootstrapped b0

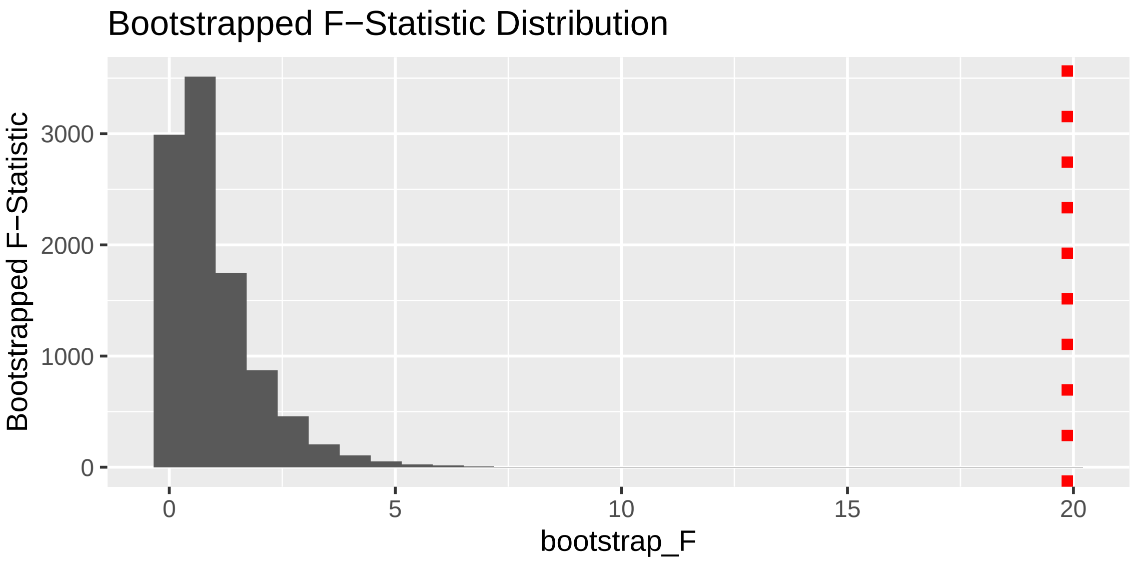


```
##      2.5%     97.5%
## -258.9743 -158.7401
```

ANOVA Test

```
## Analysis of Variance Table
##
## Model 1: price ~ bedrooms + room_type + overall_satisfaction + reviews
## Model 2: price ~ bedrooms + room_type
##   Res.Df     RSS Df Sum of Sq      F   Pr(>F)
## 1   1555 7094695
## 2   1557 7275958 -2   -181263 19.864 3.03e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Bootstrapping F-Statistic



```
## [1] 0
```

Linear Model Comparison Results:

In the reduced model, the coefficient for bedrooms is 47.342. This means, holding room_type constant, each additional bedroom is associated with an increase in the predicted price by 47.342. All other coefficients have extremely low p-values (< 0.001), indicating that they are statistically significant. In summary, the model suggests that both the number of bedrooms and room type are significant predictors of the price due to the F-statistics p-value being very low. For our Full Model, for each additional bedroom, the predicted price increases by 47.76056 units, holding other variables constant, which is very comparable to the reduced model. Furthermore, for each unit increase in overall satisfaction, the predicted price increases by 30.35449 units, holding other variables constant & for each additional review, the predicted price decreases by 0.11371 units, holding other variables constant. The F-statistic (223) with a very low p-value (< 2.2e-16) suggests that the overall model is statistically significant. We also conducted bootstrapping on both models. After bootstrapping the reduced model, the mean values for the bootstrap coefficients are close to the estimates from our original model, suggesting stability in the estimates. For the full model, the mean values for the bootstrap coefficients are also close to the estimates from the original model, which also suggests stability in the estimates. After doing the Annova test, the table suggests that the full model (Model 1) with additional predictors (overall_satisfaction and reviews) provides a significantly better fit than the reduced model (Model 2). The low p-value (< 0.001) indicates that the improvement in model fit is statistically significant. Therefore, based on this analysis, the full model is preferred over the reduced model. We also created two histograms to visualize the bootstrap distribution and confidence intervals for the intercept (bootstrap_b0) in both the reduced and full models. The red vertical lines indicate the 2.5th and 97.5th percentiles of the bootstrap distribution, forming a 95% confidence interval.The 95% confidence interval for bootstrap_b0 in the reduced model is approximately [-75.81, -56.62]. On the flip side, The 95% confidence interval for bootstrap_b0 in the full model is approximately [-252.94, -164.03]. The confidence interval for bootstrap_b0 in the full model is a little bit wider and shifted to the left compared to the confidence interval in the reduced model which can be explained by the additional predictors of overall satisfaction and reviews. Lastly, we conducted a bootstrap hypothesis test comparing the full model (Mb_Full) and the reduced model (Mb_Red) using the F-statistic. The histogram represents the distribution of F-statistics obtained from bootstrapped samples. The red dotted vertical line indicates the original F-statistic value of 19.864 from our ANOVA test. With a p-value of 0, this suggests that the F-statistic from our original model is extremely unlikely to have occurred by chance under the null hypothesis. The low p-value provides evidence to reject the null hypothesis that there is actually no difference between the reduced and full models. Therefore, based off of our bootstrap analysis, we can conclude that we have statistical evidence in favor of the full model over the reduced model.

Model without Large Dataset and Comparison:

Summary of Model 1 & Model 2

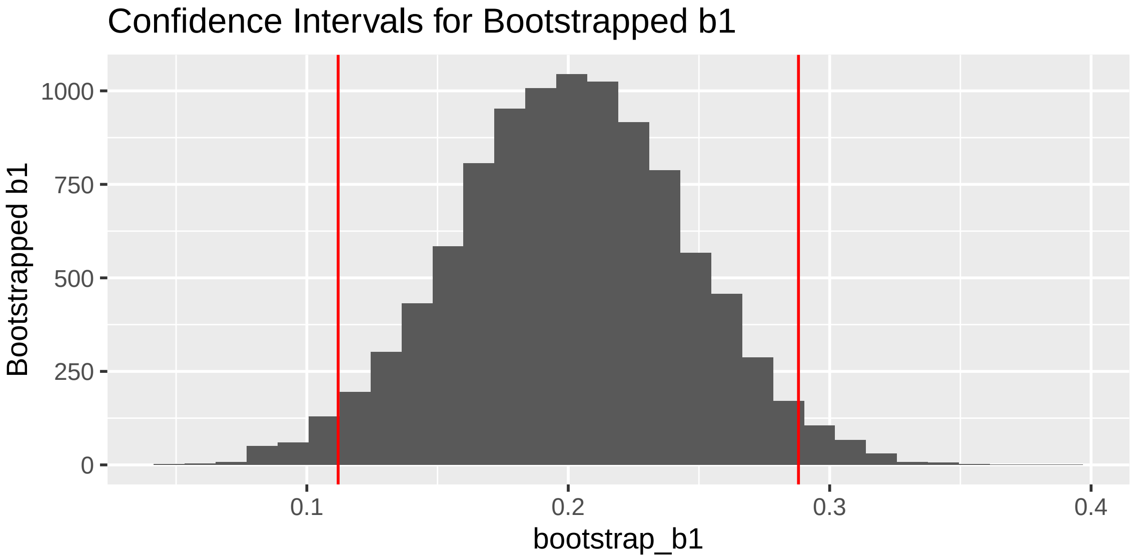
```
##
## Call:
## lm(formula = MATINGS ~ AGE, data = elephants)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.1158 -1.3087 -0.1082  0.8892  4.8842
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4.50589    1.61899  -2.783  0.00826 **
## AGE          0.20050    0.04443   4.513 5.75e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.849 on 39 degrees of freedom
## Multiple R-squared:  0.343,  Adjusted R-squared:  0.3262
## F-statistic: 20.36 on 1 and 39 DF,  p-value: 5.749e-05
```

```
##
## Call:
## lm(formula = MATINGS ~ AGE + age2, data = elephants)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.0461 -1.2021 -0.1683  0.9962  4.9539
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.526838   9.464669   0.161    0.873
## AGE         -0.131194   0.514562  -0.255    0.800
## age2         0.004414   0.006821   0.647    0.521
##
## Residual standard error: 1.863 on 38 degrees of freedom
## Multiple R-squared:  0.3502, Adjusted R-squared:  0.316
## F-statistic: 10.24 on 2 and 38 DF,  p-value: 0.0002772
```

Model 1: Bootstrapping

```
##   bootstrap_b0      bootstrap_b1     bootstrap_sigma
##  Min.   :-11.446   Min.   :0.04248   Min.   :1.139  
##  1st Qu.: -5.622   1st Qu.:0.17082   1st Qu.:1.694  
##  Median : -4.531   Median :0.20079   Median :1.836  
##  Mean   : -4.519   Mean   :0.20069   Mean   :1.838  
##  3rd Qu.: -3.434   3rd Qu.:0.23078   3rd Qu.:1.978  
##  Max.   :  1.325   Max.   :0.38599   Max.   :2.661  
```

Model 1: Confidence Intervals for Bootstrapped b1

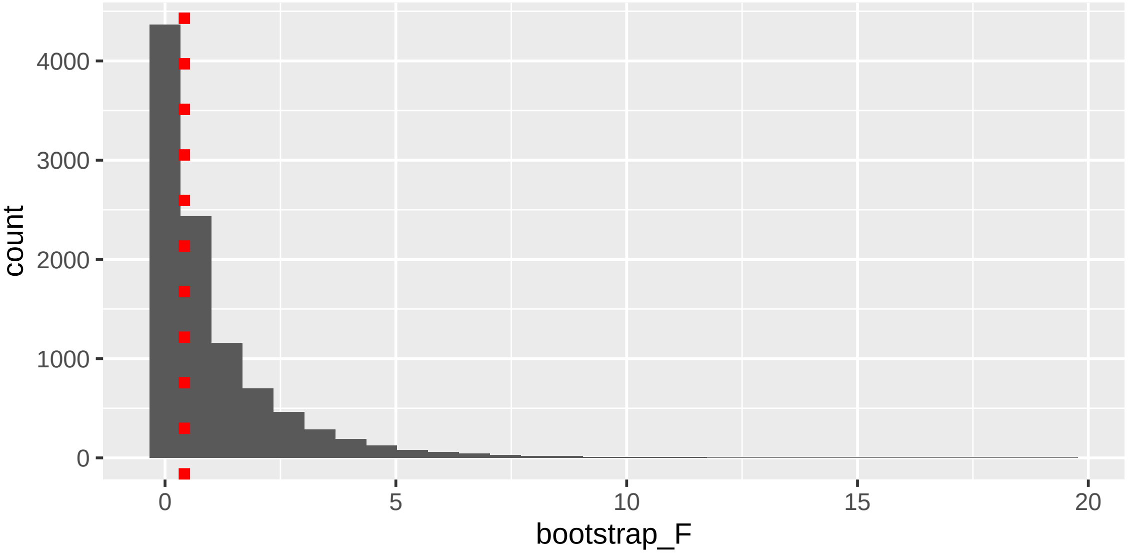


```
##      2.5%     97.5% 
## 0.1119849 0.2880765 
```

ANOVA Test

```
## Analysis of Variance Table
## 
## Model 1: MATINGS ~ AGE
## Model 2: MATINGS ~ AGE + age2
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     39 133.28                           
## 2     38 131.83  1    1.4526 0.4187 0.5215
```

Bootstrapping F-Statistic



###

Model without Large Dataset and Comparison Results:

In E_M1, for every additional year of age for an elephant, on average we expect to see the number of matings increase by 0.2. In E_M2, for every additional year of age for an elephant, we expect on average the number of matings to quadratically increase by a factor of 0.0058. The bootstrap confidence interval in the first histogram shows that the estimated parameter based on the real data is significant when compared against the bootstrapped data. The second histogram created is used to compare the F-statistic of the full model to the bootstrapping data. The F-statistic is very common which does not give us evidence that the full model is better than the reduced model.

Model Comparison for Multilevel GLM:

Summary of GLM1 & GLM2:

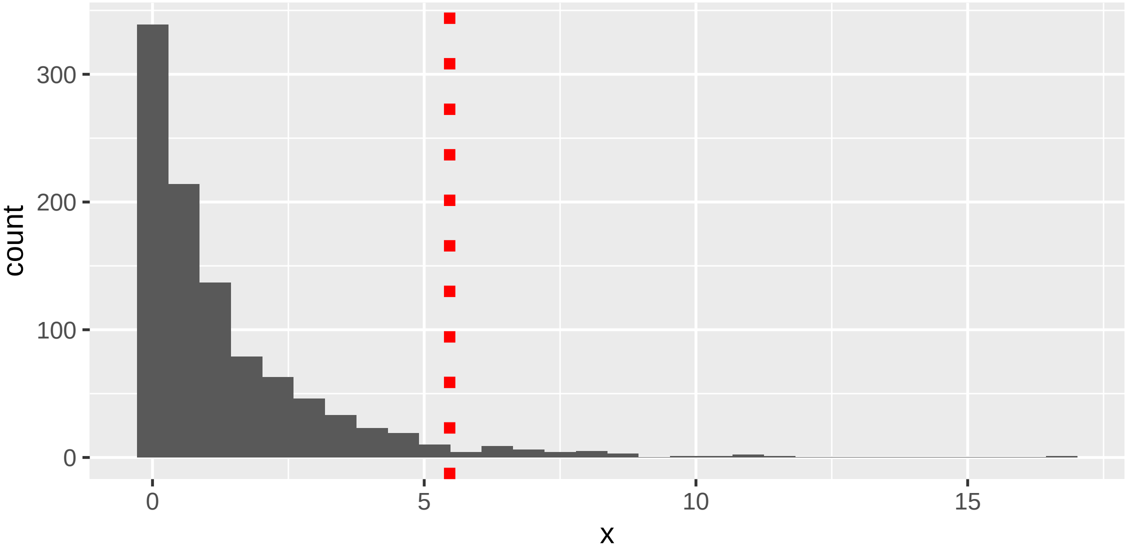
```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: foul.home ~ foul.diff + (1 | game)
##    Data: refdata
##
##      AIC      BIC   logLik deviance df.resid
##   6792.5   6812.1  -3393.3   6786.5     4969
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.6995 -0.9055 -0.6518  0.9655  1.6849
##
## Random effects:
##  Groups Name        Variance Std.Dev.
##  game   (Intercept) 0.273    0.5225
## Number of obs: 4972, groups:  game, 340
##
## Fixed effects:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.18886    0.04434  -4.259 2.05e-05 ***
## foul.diff   -0.26821    0.03895  -6.887 5.71e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##           (Intr)
## foul.diff 0.368
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: foul.home ~ foul.diff + (foul.diff | game)
##    Data: refdata
##
##      AIC      BIC   logLik deviance df.resid
##   6791.1   6823.6  -3390.5   6781.1     4967
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.6399 -0.9087 -0.6349  0.9528  1.7687
##
## Random effects:
##  Groups Name        Variance Std.Dev. Corr
##  game   (Intercept) 0.294141 0.54235
##         foul.diff   0.001235 0.03514  -1.00
## Number of obs: 4972, groups:  game, 340
##
## Fixed effects:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.15684    0.04637  -3.382 0.000719 ***
## foul.diff   -0.28533    0.03835  -7.440    1e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##           (Intr)
## foul.diff 0.192
## optimizer (Nelder_Mead) convergence code: 0 (OK)
## boundary (singular) fit: see help('isSingular')
```

ANOVA Test

```
## Data: refdata
## Models:
## glm1: foul.home ~ foul.diff + (1 | game)
## glm2: foul.home ~ foul.diff + (foul.diff | game)
##      npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## glm1    3 6792.5 6812.1 -3393.3   6786.5
## glm2    5 6791.1 6823.6 -3390.5   6781.1 5.4682  2    0.06495 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Bootstrapping Chisq



Simulation Based P-Value

```
## [1] 0.037
```

Model Comparison for Multilevel GLM Results:

Our first model gives us the relationship between fouls on the home team and the foul difference with a random slope for game. The second model has thee same relationship, but adds a random effect for the foul difference. The estimates of the two models are quite a bit different due to the added random effect. A bootstrap is then run from the reduced model and both models are fit to the simulated data. Then our histogram displays the distribution of Chi Squared values that is produced from this bootstrapping. We find that our observed Chi squared value in our anova test before the simulation is on the far right end of the distribution. This then helps us calculate our simulated P-value. We get a simulated P-value of 0.037. This P-Value is smaller that the previously observed P-value. This result is consistent with the idea that when boundary constraints are at issue in estimating variance and correlation of random effects, theory based P-value tests are too conservative leading to P-values that are estimated higher than they actually occur. Our simulated P-value supports this hypothesis that our above P-value was likely too large ue to a conservative approach to calculating the P-values in the ANOVA test. The simulated P-value also tells us that we do not have enough evidence to say that our full model is better than our reduced model. Thus, we cannot say that that addition of a random effect for foul differential is proven to be beneficial in modeling the odds a fouls is called on a home team.

Conclusions:

Parametric Bootstrapping is useful for generating simulated data from model estimates to solve problems that occur to dataset size, issues with assumptions, or comparing models. This method of bootstrapping can be used on various different types of models. In this report we use this method on linear regression models, generalized linear models, and models with a lack of data. In all instances parametric bootstrapping is useful for being able to generate additional data to increase the confidence in estimates relevant statistics, like chi squared and f-statistic. One of the more useful ways to use parametric bootstrapping is by using it to generate additional data to test various models to see which one is more useful for explaining a given dataset. We used this in all three of our examples and then used parametric bootstrapping to estimate the f-statistic and chi squared values. Parametric bootstrapping is very useful for generating additional data to be able to make a conclusive argument for a hypothesis or finding.

References:

https://www.sciencedirect.com/topics/earth-and-planetary-sciences/bootstrapping#:~:text=Parametric%20bootstrapping%20
https://stat455-w22.github.io/stat455-w22-notes/multilevel-generalized-linear-models.html#parametric-bootstrapping https://bookdown.org/roback/bookdown-BeyondMLR/ch-GLMM.html#cs:refs