Predicting a Powerlifter's Maximum Bench Press Weight

Skylar Liu

Texas A&M University

Introduction:

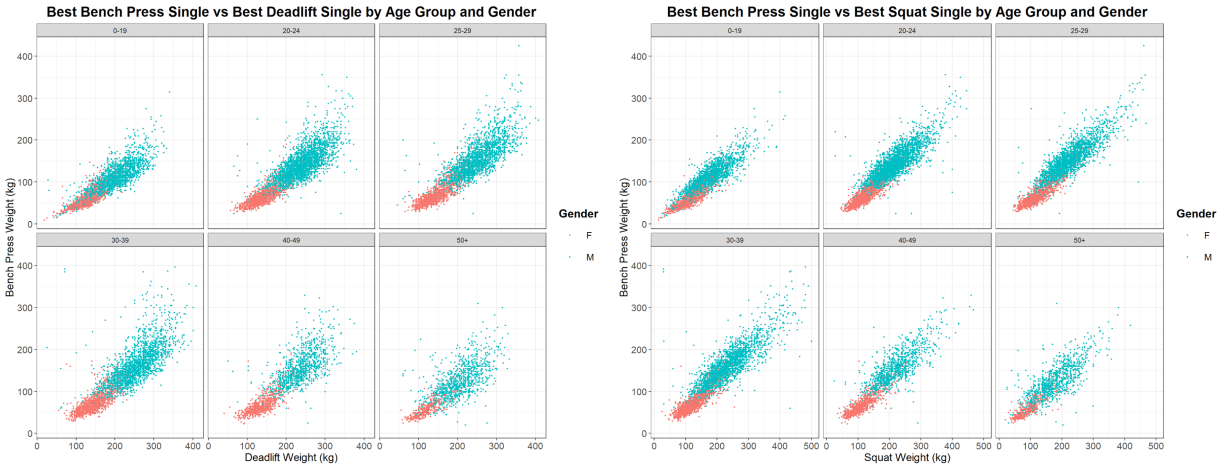
Powerlifting is a rapidly-growing sport that showcases an individual's one-repetition maximum strength on three lifts: squat, bench, and deadlift. As long as the rules are followed, a person must give everything they have to max out that one repetition at a powerlifting meet. The months of training leading up to a competition are crucial, and lay the foundation for a successful meet. Some factors, though, are out of an individual's control, such as age, body weight, and gender, and are accounted for by having separate classes and divisions..

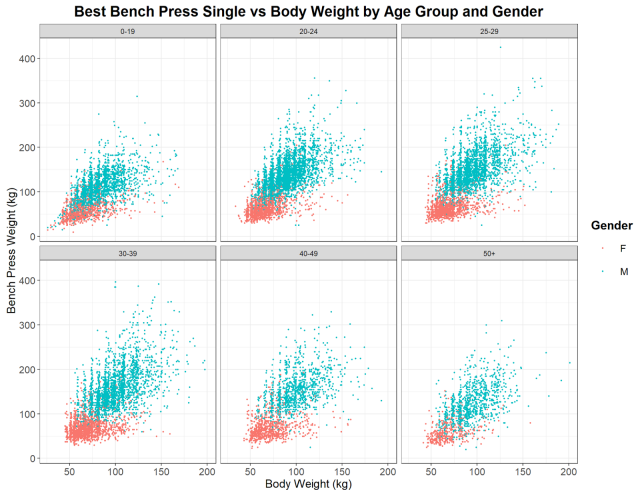
For a competitive powerlifter, it's important to have realistic goals while also being able to scope out future competition. Our client is a top ranked bench press athlete that holds multiple national records; they have reached out to us to analyze a set of data taken from a powerlifting meet to better understand what factors are at play on meet day. Our goal was to identify key factors that influence a person's maximum bench press weight, as well as create max bench press predictions based on these factors.

Methodology:

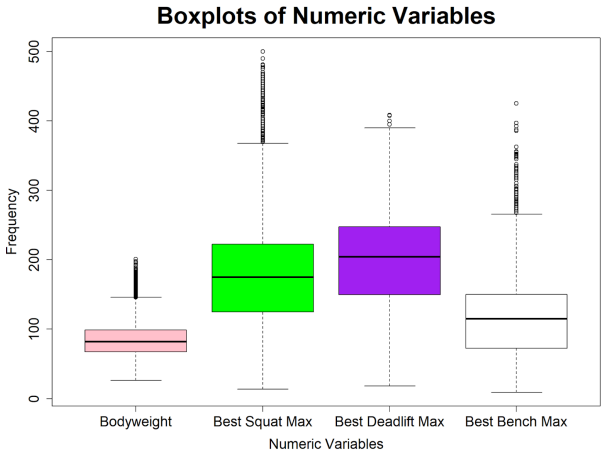
The first step to determining influential factors and estimating one's max bench press weight was to understand and analyze the data. Our data was collected from a national powerlifting meet and included the variables competitor ID, name, gender, equipment, age, body weight, best squat, best deadlift, and best bench. We decided not to use the equipment variable, as it refers to equipment used during a powerlifting meet and isn't applicable to most people. Upon

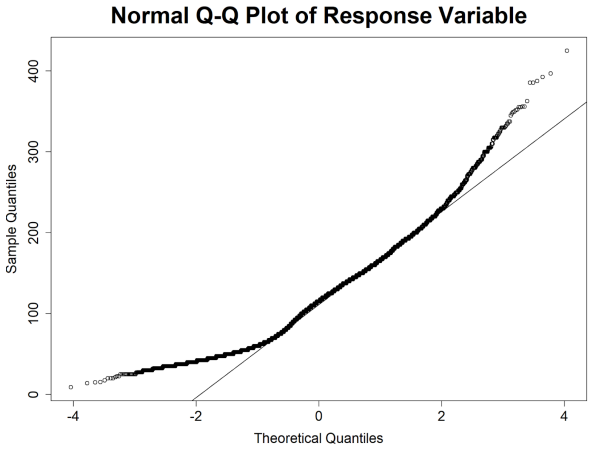
first glance of the data, we noticed some variables contained random negative values when all of the values should be positive, so those negative signs were removed. We were provided with 30,000 data records which were split into a training set of 18,900 records that were used to analyze the data and create the model, and a test set of 11,100 records that were used to test the accuracy of the model. Both sets were found to be missing a few values for age so those records were removed completely since we were still left with plenty of data. We then added one variable that created age groups from the “age” variable which contained a relatively equal number of records (ages 0-19, 20-24, 25-29, 30-39, 40-49, and 50+); the “age” variable was removed and “age group” was used in its place. We proceeded to use R studio to perform analysis and visualize the data for this study. Initial graphs showing the relationship between the provided variables and bench press max weight are shown below:





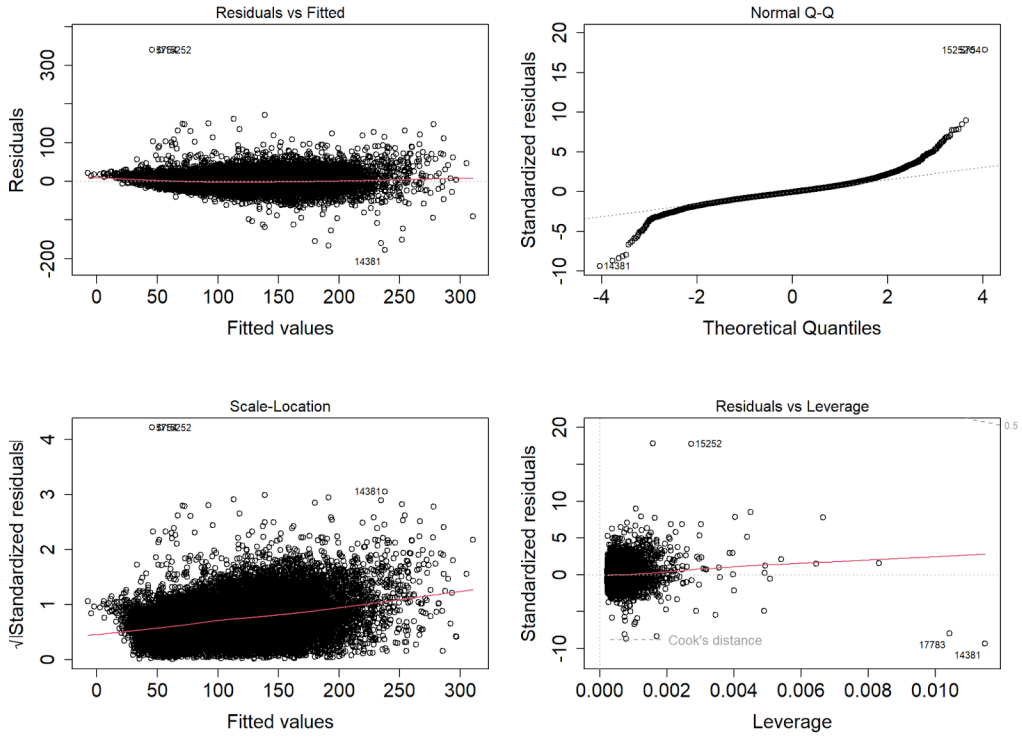
These scatter plots suggested that we have linear relationships between the input variables and max bench press weight. The next thing we looked at is the normality of our data (how much our data is representative of the population) and how each variable was dispersed:





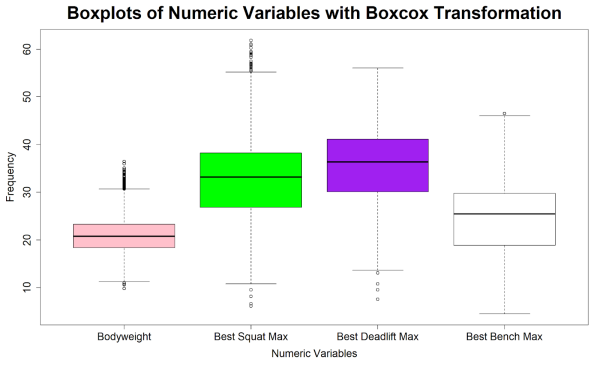
The boxplots show the frequency distribution of our numeric variables and suggest that these variables are skewed right (many data points that are larger than the average) and contain outliers (data points that are much larger than the average and should be looked at further). The Normal Q-Q plot shows how normally distributed our response variable (max bench press

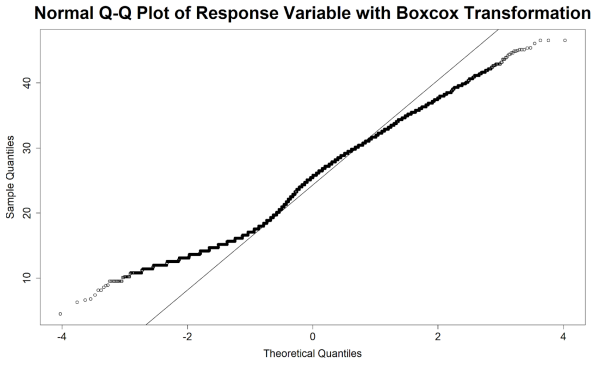
weight) was, and suggests that a transformation of the data was necessary since the line curves on both sides instead of being a completely straight line. We then fit a Gaussian linear model on max bench press weight using body weight, age group, gender, best squat weight, and best deadlift weight. Additional visualizations of the fitted model are provided below:

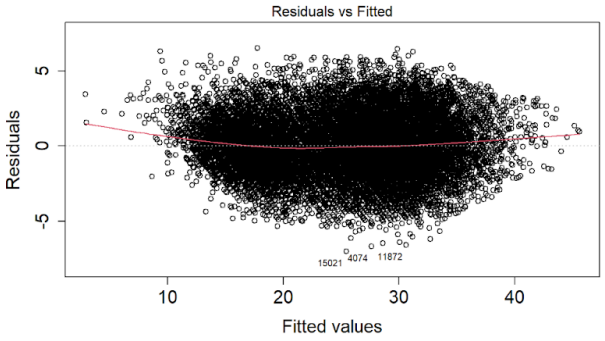


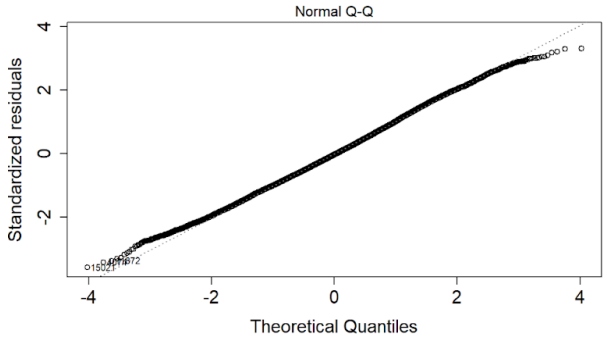
These graphs again show that a transformation of the data was necessary, since the two residual plots on the left were not completely uniform from left to right and were randomly dispersed. We also saw more evidence of outliers on the right two graphs, with the outlying points in the upper right hand and bottom left hand corner of the Normal Q-Q plot, and spread out points in the Leverage graph. We first used Cook's Distance to detect and remove outliers that had a bad influence on our model, as these records most likely belong to individuals that prefer one lift over the other; if we had access to the researchers who collected this data we

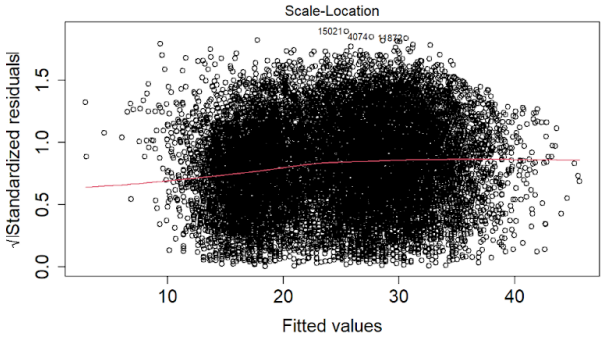
could get a better understanding of what to do with these outliers. Since we retained about 95% of our data without these outliers, we ultimately decided to remove them for this study. Next, we used a Box-Cox transformation on all of the numeric variables since they were all skewed to the right. Our Box-Cox lambda came out to 0.5856, which is similar to taking the square root of each variable. After these transformations and removal of outliers, we obtained the following improved distributions:

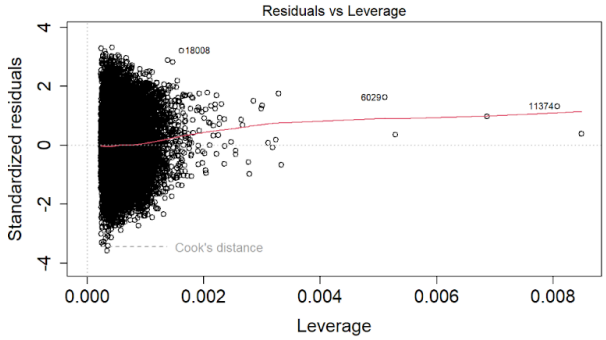






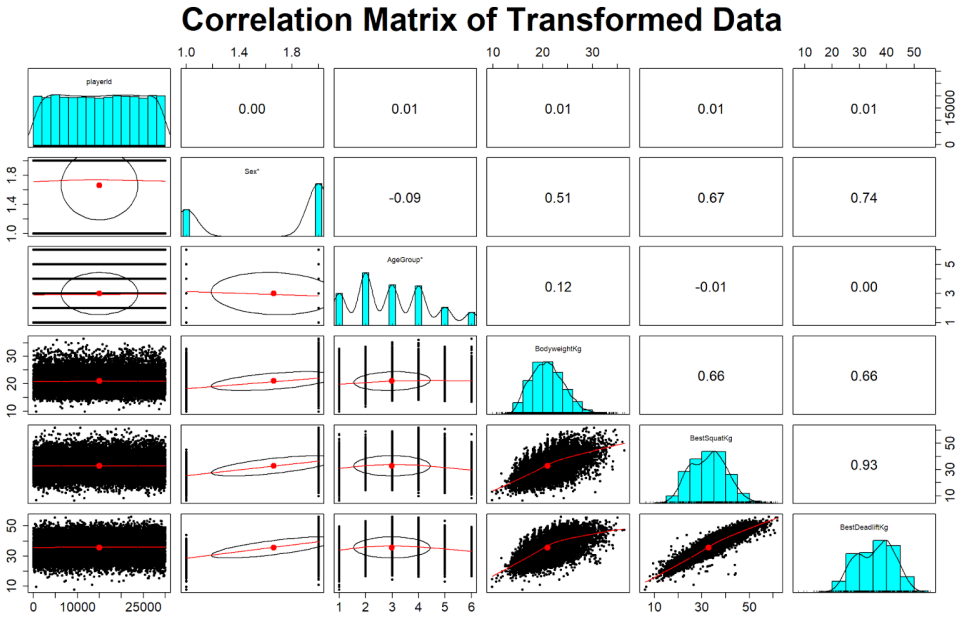






These boxplots are considerably more evenly distributed, and the Q-Q plots follow straighter lines. The residuals are more uniform and randomly distributed, suggesting that our data was now a good representation of the general population. There still appeared to be outliers but since these weren't identified as having a bad influence on our model, they were left in our data set.

The last step in preparing our data was variable selection. We wanted to make sure that we included all important variables but did not include variables that had strong correlations (strong relationships between input variables that could make it hard to detect which variable is actually influencing max bench press weight). Since our data contained both quantitative and categorical variables, we chose to use LASSO for variable selection. LASSO (Least Absolute Shrinkage and Selection Operator) shrinks the data to a central point and results in the not-as-important variables having a coefficient at or near zero. This analysis provided a small coefficient for max deadlift weight, suggesting that we remove this variable. We also created a correlation matrix to visualize the correlations between each variable to aid in variable selection:



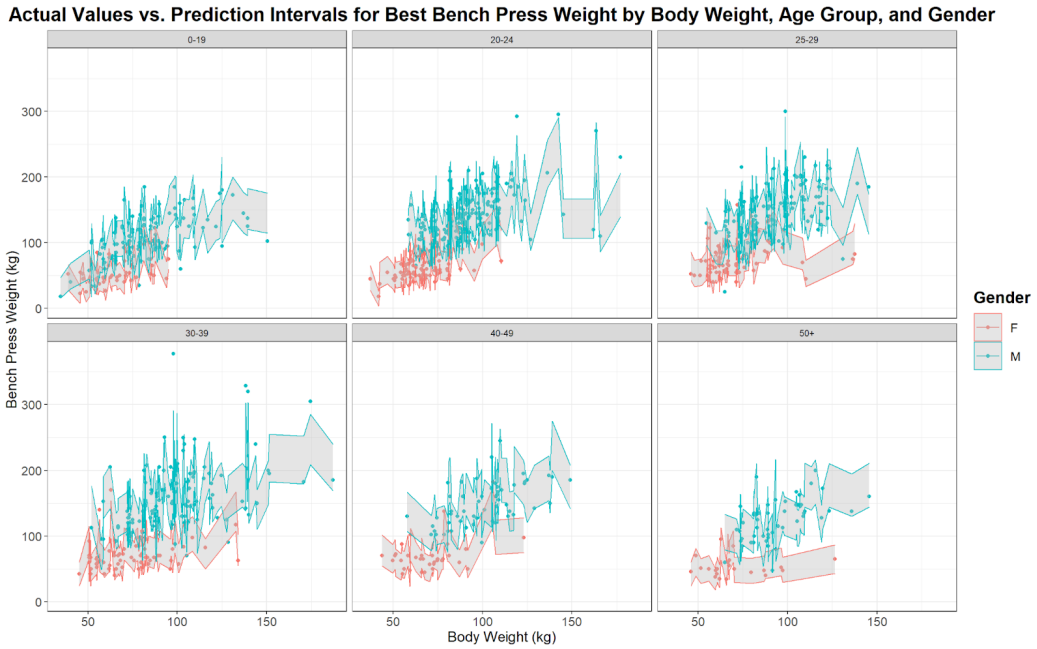
The middle diagonal of this matrix provides histograms of each variable, which show that our data was relatively even and normally distributed. The graphs on the bottom left triangle of the matrix are scatter plots that show the relationship between each variable. The upper right triangle of this matrix shows the correlation between each variable, with a value close to +1 or -1 indicating strong correlation (strong relationship). Best deadlift had a high correlation with 3 other variables, with the highest being a 0.93 correlation with best squat. This suggests that if a person works out, he or she will generally be strong overall, so it may not matter which lift we look at to predict max bench press weight. Since best deadlift had a higher correlation and a lower coefficient in the LASSO analysis than best squat, we removed this variable from our model.

Once we had our data prepared with the variables narrowed down, we refit the model. We then transformed the test data set that was set aside earlier using the same method used for our training set. Using the refit model, we tested it on the test data set and created prediction intervals that provided a range of possible values to compare our test results to the actual values. Lastly, we back-transformed the test data and prediction intervals to have values with the same scale that we started with so that we could visualize and understand the results.

Results:

We looked at a subset of available variables to use as input for our goal of identifying key factors and predicting a person's maximum bench press weight. With the large amount of data provided for analysis, we created a model that can be used to successfully predict a person's maximum bench press weight based on body weight, age group, gender, and maximum squat weight. A portion of the provided data that was not used during model creation was set aside to test the model's accuracy (our test set). The model yields point estimates as well as prediction

intervals, which demonstrate the uncertainty associated with specific estimates. These intervals contain a range of possible values for a person's predicted bench press max, to which we can be 95% confident that the value will fall within this range. Since we tested the model on part of our available data, we compared these results to the actual max bench press weight values provided. In the following graphs, we took a random sample of 1,000 records from the test data since using all 11,000 values of the test data congested the graph and was impossible to read. The points represent the actual max bench press weights while the lines containing the shaded regions are the prediction intervals.



This graph matrix represents the relationship between body weight and max bench press weight, separated by age group and gender. There was a slight upward trend, but body weight was overall not a consistent factor in determining a person's max bench press weight.

This graph represents the relationship between max squat weight and max bench press weight, separated by age group and gender. There was a significantly more stable relationship between how much a person can squat and their max bench press, suggesting that a person's overall strength is better at indicating their bench press strength than intrinsic factors such as body weight, age, or gender. As for intrinsic factors, gender seemed to be more indicative than age or body weight at predicting bench press strength.

These results were validated using several methods, including RMSE, MAE, and adjusted R^2. RMSE (Root Mean Squared Error) measures the distance between the actual and predicted values, with the goal being a value close to zero. The RMSE for our model came out to 2.5366. We also calculated MAE (Mean Absolute Error), which also measures the distance between the actual and predicted values but isn't as sensitive to outlier values. Since our data contained an abundance of outliers, MAE was a more relevant validation statistic. We would also like for this

value to be close to zero, and we obtained an MAE of 1.8453. The last validation statistic we calculated was adjusted R^2, which explains the extent of which the variables of interest explained max bench press weight and accounts for any correlation between the variables of interest. A perfect adjusted R^2 is 1.0, indicating a perfect relationship; our adjusted R^2 came out to 0.8757. These values indicated that our model was a decently good fit. Our data indicated that there were a lot of outliers, possibly due to a person favoring one lift over the other, so our model could be improved by better accounting for these values.

Conclusion:

For this consulting project, we examined a set of variables to identify influential variables and create a model that predicts a person's maximum bench press weight. Our model accurately provided an estimated value along with prediction intervals based on age group, gender, body weight, and maximum squat weight, with max squat weight and gender being the strongest predictors. A strong correlation existed between max squat weight and max deadlift weight, suggesting that strength carries over regardless of the specific exercise. Our data contained a sizable amount of outliers, so the next step would be to examine the reason for these outliers and better account for them to increase the accuracy of our model. We hope these findings will benefit future competitors in reaching their maximum bench press potential.