California State University - Northridge

Powerlifting

Christian Vega

Nicholas Zermeno

Sammana Kabir

MATH 444 Statistical Modeling

Section 11901-SP2022

Professor Adriano Zambom

May 12th 2022

Abstract:

This paper explores the study of the prediction of finding the best features/predictors that plays an important role in achieving maximum deadlift for athletes partaking in powerlifting competition. The project is implemented in R on the dataset found from Kaggle which includes all the features related to deadlifting. After preprocessing the dataset which included handling missing values, multicollinear and irrelevant data, a model building algorithm was applied to find the optimal model for achieving the maximum deadlift. In the end, the accuracy of the model was measured with adjusted R-squared as the indicator.

Introduction:

The purpose of this project is to predict the predictors that facilitate achieving maximum deadlift. The dataset chosen for this project shows the results of various powerlifting competitions from OpenPowerlifting database most recently updated as of April 2019. The competitors in this dataset competed to lift the most weight for their class in three separate categories: Squat, Bench Press, and Deadlift. The objective of this project was to gauge prediction criteria for the best deadlift. The initial motivation to use this data included an interest in being able to predict how well competitors would perform on their best deadlift. In addition to this, other motivations include that our group has a profound interest in fitness that we felt would give us a slight advantage in intuitive insight that we would be able to employ in our analysis.

Methodology:

The first issue that needed to be addressed with the data was gauging exactly what parts contained missing values. We addressed this issue by running a function in R to determine the ratio of missing data for each attribute as a percentage of overall data points. Since we are evaluating best deadlift as our response variable, we decided it would be best to exclude all the data points that have the particular value of deadlift missing because an analysis of how best deadlift would be predicted has no use for data that has no value in that category. In order to make the data easier to work with we assigned values of zero to all data points that contained no value in order to make the data easier to work with. In addition to this, we saw it as most suitable to omit the use of calculated scores in any candidate models. These scores include Glossbrenner, wilks, McColloch as well as IPF points. The rationale for excluding this data from final models is because they are all directly related to the real data of the lifters, Glossbrenner score is determined directly by bodyweight, wilks is determined by the actual lifts, and McColloch is determined by weight lifted as well as age. Since the goal of our model is to predict deadlift based on the raw data of each of the athletes alone, it would be problematic to include other data that is calculated directly from that raw data due to the obvious multicollinearity that would result. In addition we also condensed the date category into a decade column, this was to simplify the model and extrapolate the date in an easier to analyze manner. We also ended up dropping the name and division because the levels were over fifty thousand each, which would not have been practical for the scope of our analysis. In addition, we also consolidated the country name column to better organize the country data. We also checked for multicollinearity among our candidate predictors. Bodyweight and weight class was run with KramerV test and chi squared test at which point it was determined it would be best to remove federation. For

age-class and age there was more missing data in age, which caused us to take the median of age-class at the point imputed into the age column. After this was done we removed the age-class column because it's easier to work with continuous data rather than categorical. After this we got rid of MeetState, MeetCountry and MeetPlace because it was not feasible to run models given the constraints of R and our computers as the number of levels in these variables made computation performance slow down drastically. After this we added the data frame into the multicollinearity function that was created earlier to check for covariance. In addition, we also had to remove the best deadlift that was below zero. The reasoning behind removing negative as well as zero max deadlift was that a negative or zero value indicates that the athlete failed at every attempt or was not interested in deadlifts, this type of data would not be useful in this analysis because we are trying to determine a way to predict a max deadlift and reasonable criteria would be to insure every data point we use has a valid attempt. After this we ran MICE on body weight because this was the only missing values we had left

Analysis:

For our analysis we ran full and null models and ran SignifReg forward and backward selection using adjusted R-squared as the selection criteria and found that both searches yielded the same model. We used the test and train split to find the PRESS score at 76035400, given the massive amount of data in this data set we found this to be an ok score.The split we used was 70 percent train and 30 percent test. After this we ran the model chosen to get the summary.

m3 <- lm(formula = MaxDeadlift ~ Sex + MaxSquat + Event + BodyweightKg + Decade + Equipment + Tested + Age, data = dt)

summary(m3)

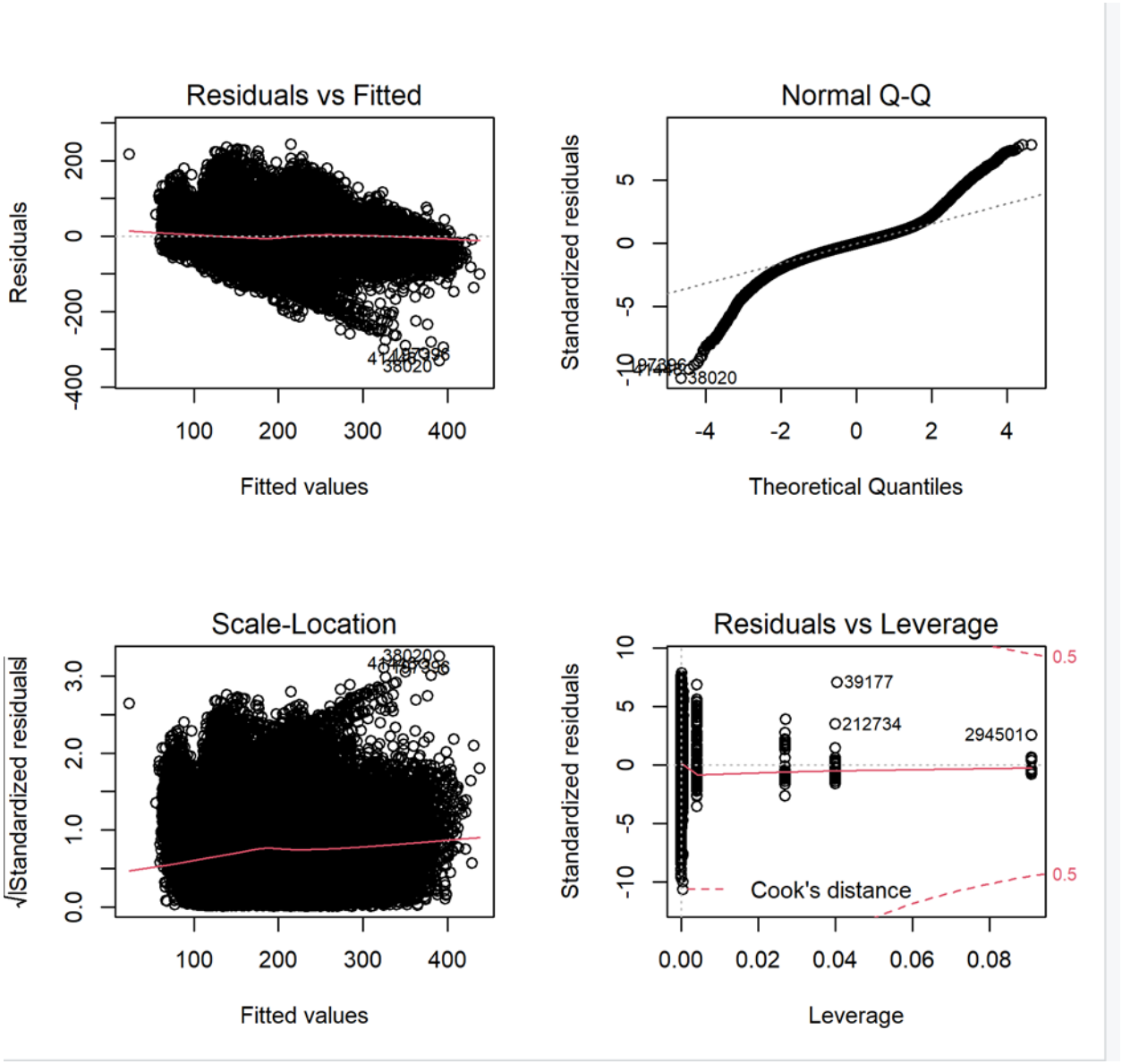
```
MaxSquat             < 2e-16 ***
EventD               < 2e-16 ***
EventSBD             < 2e-16 ***
EventSD              < 2e-16 ***
BodyweightKg         < 2e-16 ***
Decade1970-1980        0.368
Decade1980-1990     1.24e-07 ***
Decade1990-2000     2.19e-13 ***
Decade2000-2010      < 2e-16 ***
Decade2010-2020      < 2e-16 ***
EquipmentRaw         < 2e-16 ***
EquipmentSingle-ply  < 2e-16 ***
EquipmentStraps      < 2e-16 ***
EquipmentWraps       < 2e-16 ***
TestedYes            < 2e-16 ***
Age                  < 2e-16 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 30.2 on 277681 degrees of freedom
Multiple R-squared:  0.7635,    Adjusted R-squared:  0.7634
F-statistic: 5.272e+04 on 17 and 277681 DF,  p-value: < 2.2e-16

> 
```

Given that the F statistic was 5.272e+04 and the p value was 0.7634, we had sufficient evidence to reject the null hypothesis. This indicated to us that the variability of max deadlift was explained by the selected predictors of sex (pre-factored as dummy variable), maxsquat, event, bodyweight in kilograms, decade, equipment, tested for performance enhancement (pre-factored

as dummy variable), and age to a degree of 76.34 percent. Once this test was done we moved next to examine the residual diagnostic plots shown in the figure below.

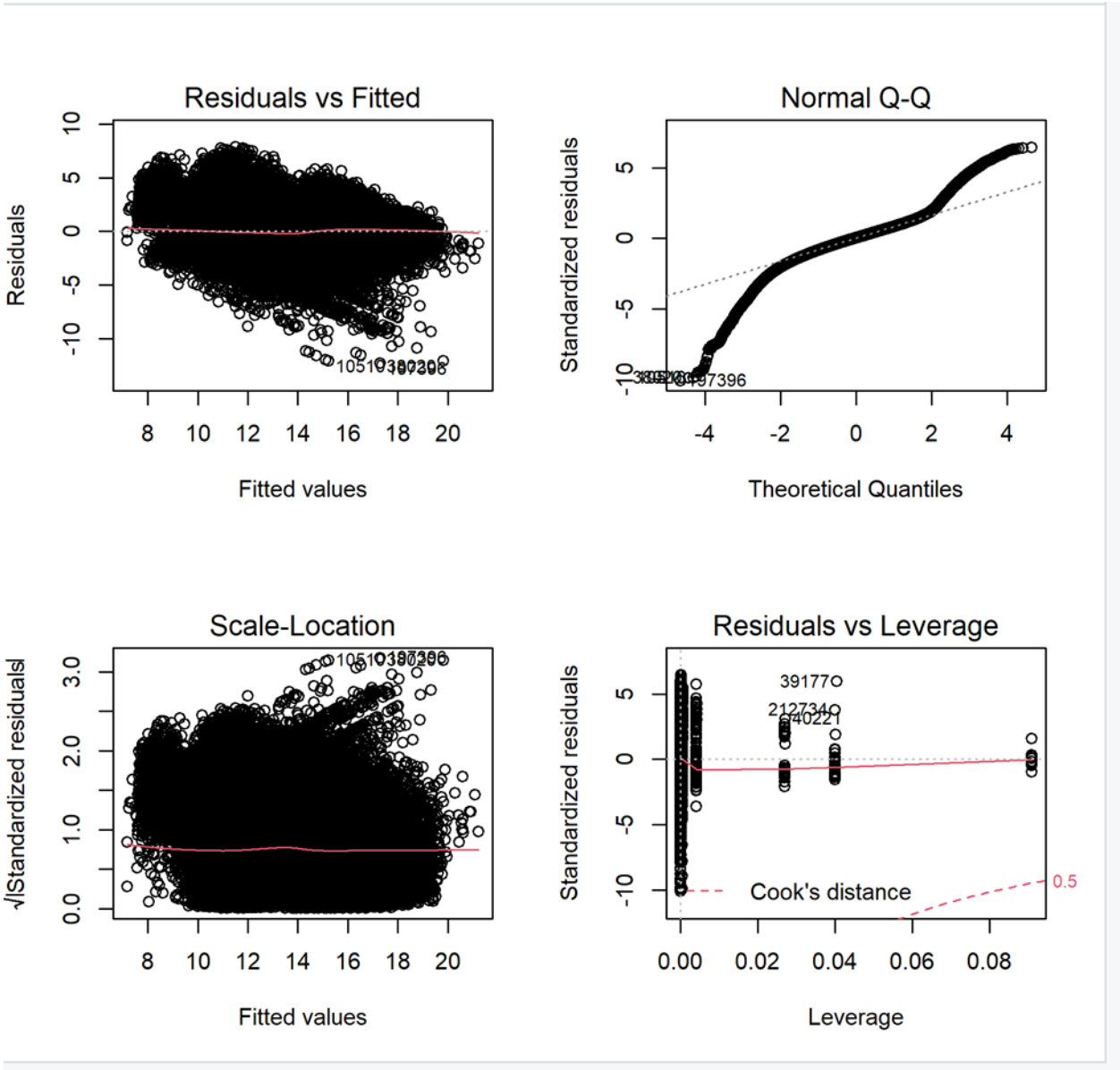


We noticed the residual diagnostic plots for the selected model were less than ideal. The most apparent issue is that we see tails at the ends of the theoretical quantities vs standardized residual plots indicating that the errors are not normally distributed. In addition to this we see slight patterns in the residuals vs fitted plots that challenge the assumption that the errors are independent of the fitted values. We tried a few different transformations such as inverting as well as applying log function to age, in addition to square rooting every attribute that had to do with weight measurements, both lifting and bodyweight.

TRANSFORMED MODEL DIAGNOSTICS

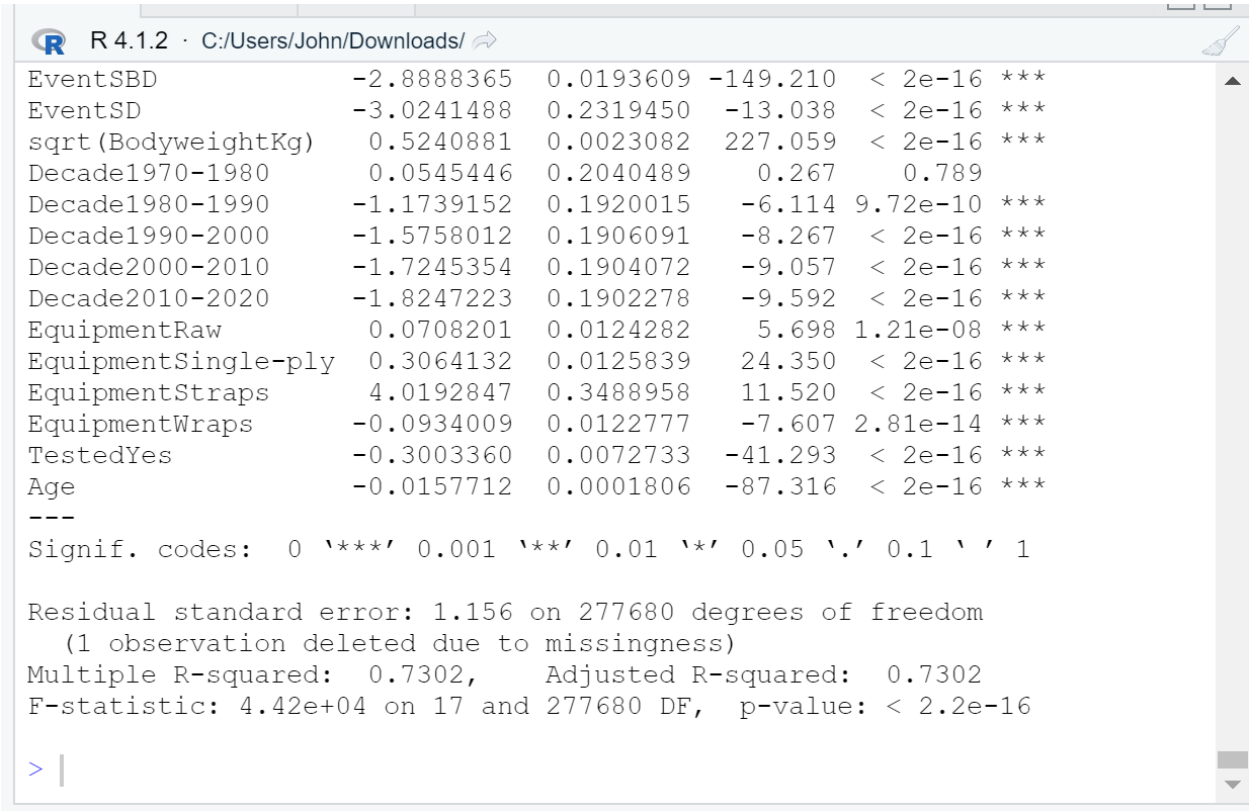
After encountering issues with the residual diagnostic plots of the forward selected model. The best competing model we attempted to apply was the square root transformation to every attribute that involved weight measurements. The diagnostic plots of the transformed model are shown in the figure below:

```
m3 <- lm(formula = sqrt(MaxDeadlift) ~ Sex + sqrt(MaxSquat) + Event + sqrt(BodyweightKg)
+  Decade + Equipment + Tested + Age, data = dt)
```



In addition to the results of the transformation not being much better than the non transformed model with the exception of slightly better residuals vs fitted values, we actually observed a slight drop in the adjusted R squared value from 0.7634 to 0.7302 as shown in the figure below.

TRANSFORMED MODEL SUMMARY



After exhaustive analysis of differing transformations, we came to the conclusion that it would be best to keep the original model selected in the forward search of the SignifReg package.

Obstacles:

One major obstacle we faced with this project is that the dataset is very large, with more than one million data points. This had a significant impact on the computation time it took to get basic analysis tools to render such as diagnostic plots and residual histogram plots. In addition to this, the amount of missing data as well as the somewhat controversial practice of allowing 3 and 4 year olds to compete in a weight lifting competition posed concerns both ethically as well as practically in the scope of this project. The practicality concern had to do with the idea that

analyzing lifters so young could skew the linearity of relation between performance and age due to the fact that 3 and 4 year old lifters certainly would not be expected to have reached puberty. In addition to this, the international community does not recognize competitors under 8 years of age. Ultimately we decided to incorporate the restriction by international standards set and made the cutoff for age in our analysis to be starting from 8 years old.

Conclusion:

What we observed in this project was that in forward and backward selection, on rare occasions the same final model will be selected, this happens to be the case in our findings in this analysis. In addition to this we have found that the tools employed in our methodology and analysis were definitely consistent with our initial insight beforehand that the IPF, GlossBrenner, Wilks, as well as Mcculoch scores would not be suitable as predictors to our analysis due the obvious collinearity they carried with the direct attributes of the athletes. This insight was supported and confirmed by our results.