Deadlifting: What Matters?

James Nguyen

MATH180

Professor Harish Bhat

November 6, 2020

Abstract

This paper investigates what factors are significant when predicting the performance of a powerlifter's deadlift score. A linear model is made as a baseline so when a higher order model is present, it will have something to compare to. In addition, the model is then tested to see whether how accurate the models are at predicting gender. The models do prove to be able to predict a powerlifter's deadlift score as well as predicting gender given some predictors. Body weight and weight class are shown to be very impactful in determining a powerlifter's deadlift score.

Introduction

Powerlifting is a demanding sport with an emphasis in physical strength in various parts of the body. Within powerlifting, there are three types of scores measured at powerlifting meets; deadlift, benching, and squatting. Powerlifting is all about who can lift the most for their weight class, but which factors are most impactful in predicting an athlete's powerlifting score, more specifically deadlifting? In addition, are all the factors independent of each other or is there some interaction between them? Lastly, is there a linear relationship between the factors or is there a higher-order relationship that can more accurately describe how they interact?

The predictors that are columns in the data for this study are gender (male/female), equipment (wrap type), age (years), bodyweight (kilograms), and weight class (kilograms). In addition to the column predictors, there are speculations of predictors that deal with interaction(s) between predictors, but this will become more defined in the testing methodology. The responses are the athletes' highest score in deadlifting measured in kilograms.

All the data used in this study is found on Open Powerlifting, a public database with scores and stats of athletes who participated in International Powerlifting Association sanctioned meets. The actual archive holds millions of athletes' scores, but some rows are missing values in columns. This is fine since there are a bunch of extra columns that are not being investigated for this study. However, any athlete missing any of the three responses or other important data such as age, weight, etc. will be removed from the sample. Lastly, this study will only be looking at the "Open" class of lifters because this is the most popular age range. To clarify, the requirements for being part of the data; has at least one score for the three type of lifting, the age is within the "Open" class age range, and the row has no empty entry. After removing all data that is not fit for the set, there are a little over 200k athletes remaining. Males and Females are turned into numeric values, 2 and 1 respectively. Equipment is also given numeric values; 5 for wraps, 4 for unlimited, 3 for single-ply, 2 for raw, 1 for multi-ply.

Methodology

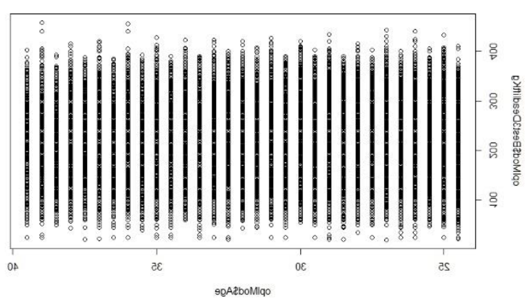
Before starting, the table will still need to be shortened again. This study is only using the top 10000 values because there is too much variance between lifters past this amount. Although it is not measured, it is observed with a graph (See Below). In addition, using a sample size as low as 1000 also does not have enough data to produce a strong enough graph.

After the table is truncated, five of the predictors in the column will be measured for its significance using linear regression. This model's residual error and average error after 10-fold cross validation will be measured. The 10-fold cross validation is using set random seeds and is repeated ten times. Using this, there is a starting point for estimating the weight a powerlifter can lift given the five predictors. From here on, the idea will be to find a model that can show significant evidence that predictors can be more closely represented by a better model. The better model is achieved by guessing and checking different model types until the 10-fold cross validation error no longer goes down. This includes using other types of regression such as polynomial or logarithmic and looking for interaction between the predictors. There is no definitive ending to this, but there will be a model with lower error. After achieving this model, the residual errors can be compared to see whether one is overfitting or if both seem to be fine.

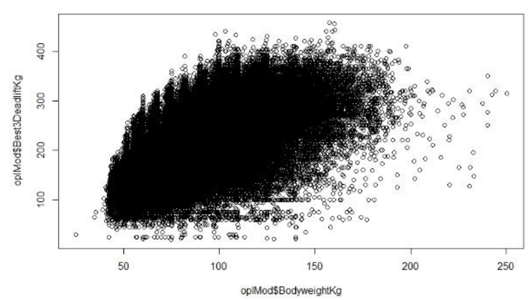
Once the linear and better model are achieved, the same models will be used to predict the gender of the lifter. The training set and test set consists of a 80-20 split randomly across a constant seed. The accuracy between the two models will be compared here, but they are independent of the regression analysis done in the previous part. It is just interesting to see how accurate the models would be when predicting relationships between themselves.

Data & Analysis

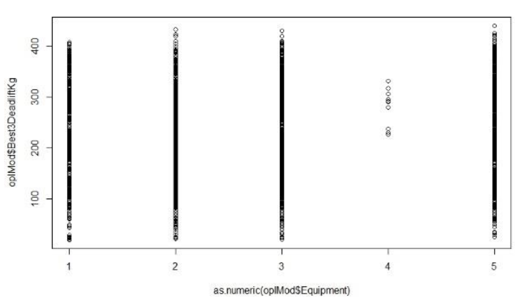
Pre-Model Predictions:



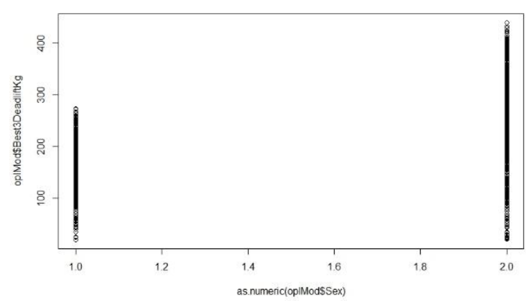
Graph 1.1: Age vs. Deadlift (All Data)



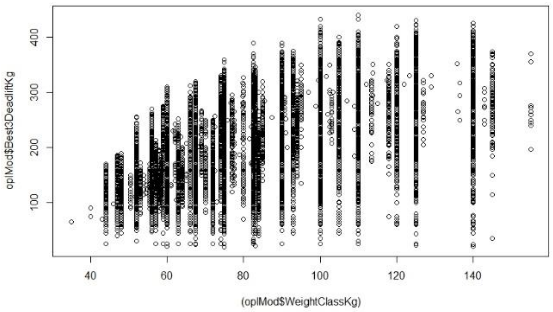
Graph 1.2: Bodyweight vs. Deadlift (All Data)



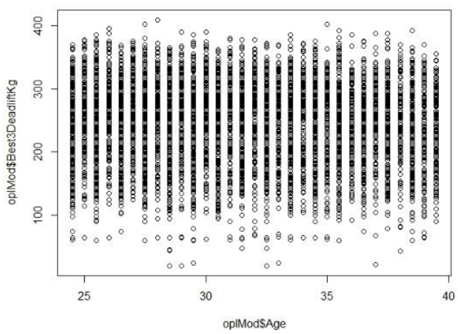
Graph 1.3: Equipment vs Deadlift (All Data)



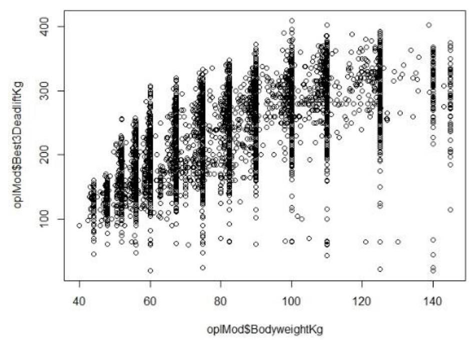
Graph 1.4: Gender vs Deadlift (All Data)



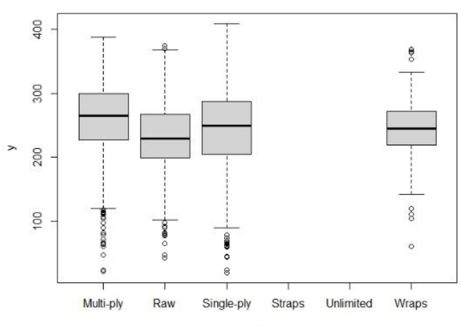
Graph 1.5: Weight Class vs Deadlift (All Data)



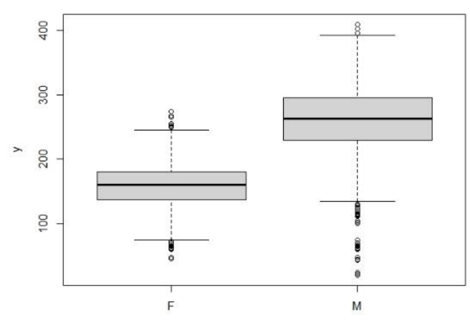
Graph: 2.1 Age vs. Deadlift (Short Data)



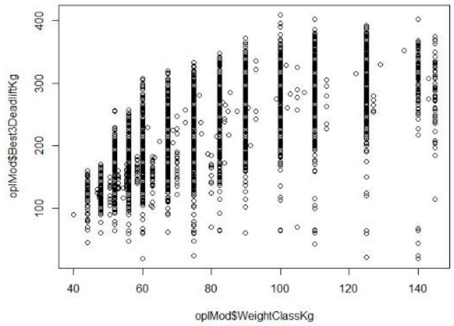
Graph 2.2: Bodyweight vs. Deadlift (Short Data)



Graph 2.3: Equipment vs Deadlift (Short Data)



Graph 2.4: Gender vs Deadlift (Short Data)



Graph 2.5: Weight Class vs Deadlift (Short Data)

By truncating some of the data, there are less outliers and there is a more defined path. This is probably due to many scores where the lifter is not as experienced and so they drag down the shape that is already present. In addition, their inexperience also varies so there is more variation than if everyone was equally inexperienced. Very noticeable graphs are 1.2 and 2.2, 1.4 and 2.4, and 1.5 and 2.5. Due to the variance, truncating the data brought better results.

Linear Regression Model: When using the five predictors from the columns into linear regression equations: x_1 = gender, x_2 = equipment, x_3 = body weight, x_4 = weight class, x_5 = age

$$\hat{y}_{Deadlift} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5$$

Coefficient	Value	Standard error	p-value
β_0	39.0838	3.7981	< 2e-16
β_1	67.1977	1.1999	< 2e-16
β_2	0.8849	0.4611	0.055
β_3	-0.1825	0.2850	0.522
β_4	1.4647	0.2829	2.29e-07
β_5	-1.0025	0.0884	< 2e-16

Table 1.1: The coefficients of the linear regression model for deadlifting

Residual standard error	38.65
Multiple R-Squared	0.5833
10-Fold CV	5663.19

Table 1.2: Error values for the linear regression model for deadlifting

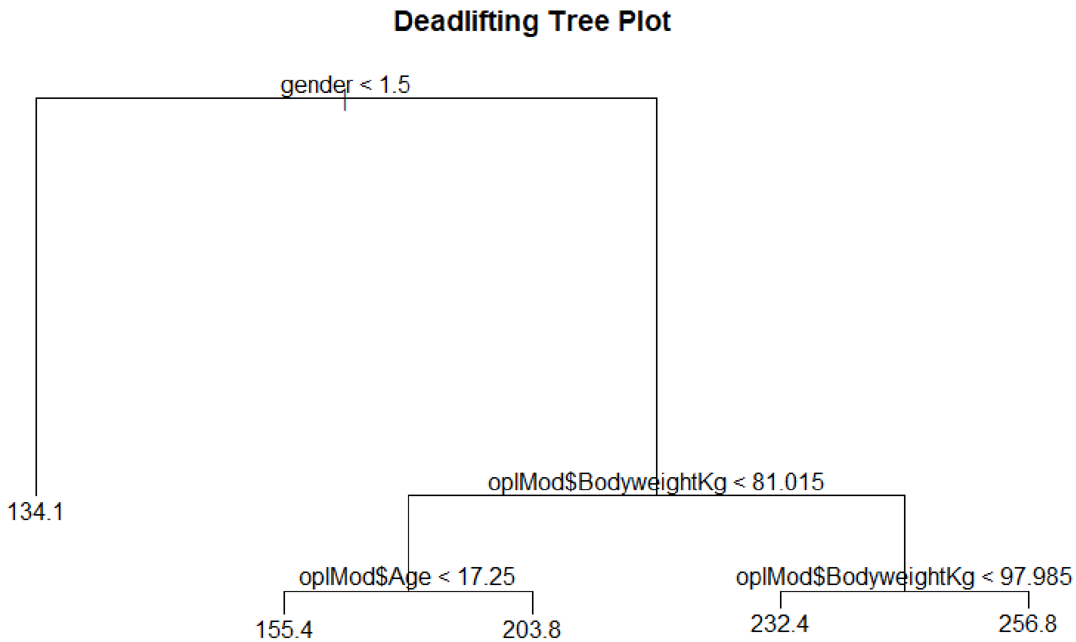


Figure 1.1: The tree plot for a linear model of deadlifting

From the linear regression model, it shows very significant values for everything except equipment and bodyweight. This is probably due to a powerlifter's bodyweight being very close to their weight class and so the weight class ended up being a better predictor than the bodyweight. This is observed more in the better model.

Better Model: A linear model with less predictors than the first. The predictor variables use the same names as the first one: x_3 = body weight, x_4 = weight class x_5 = interaction between bodyweight and weight class

$$\hat{y}_{Deadlift} = \beta_0 + \beta_5 x_5 + \beta_4 x_4$$

Coefficient	Value	Standard error	p-value
β_0	-82.34	5.7700	< 2e-16
β_5	67.1977	0.1313	< 2e-16
β_4	0.8849	0.0007206	< 2e-16

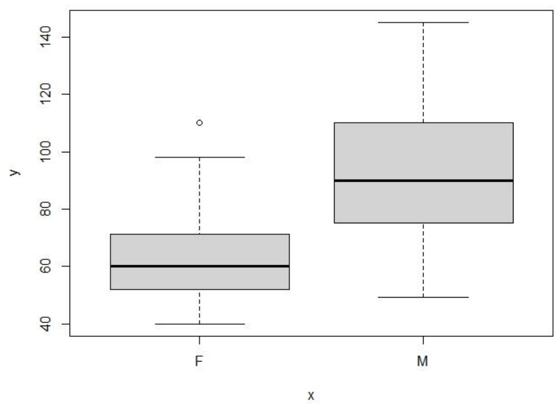
Table 2.1: The coefficients of the better linear regression model for deadlifting

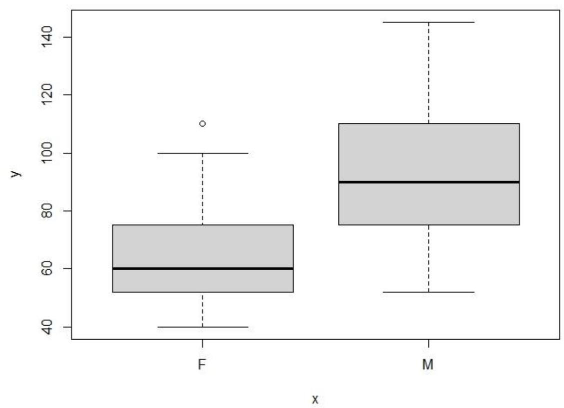
Residual standard error	42.63
Multiple R-Squared	0.4931
10-Fold CV	5337.02

Table 2.2: Error values for the better regression model for deadlifting

*Tree is not available for this model because of the interaction

Although, there were plans to use better regression models than linear, a modified linear model ended up being the best that was made so far. The error over 10-fold cross validation decreased, but not by much (a little over 5%), but the residual standard error went up. This is likely due to the first model overfitting with all the extra predictors when these few could have done the job. If a lot of the predictors in the first model was significant, why did this model end up with less error? This is most likely due to these variables doing a much better job at summarizing the other predictors without overfitting the data.





Graph 3.1 Gender vs Bodyweight

Graph 3.2 Gender vs Weight Class

Despite gender showing very strong significance in the linear regression model, the weight class and bodyweight predictor can sum up which gender much better than itself can. Due to this, the gender predictor is dropped from the better model. Equipment as seen in table 1.1 is not significant and so it was also discarded. Graph 2.2 shows age is really all over the place, not sure how it had high significance though. Lastly, table 1.1 also showed bodyweight to not have significance in the linear model but when it interacts with weight class, that's different. This is probably due to the closer someone's bodyweight is to a weight class, the higher they are predicted to deadlift; which is not shown alone since weight class is more significant than bodyweight.

Classification: With these two models, predicting gender will be measured along the two. Instead of predicting deadlift, it is now predicting gender. Train and test sets is defined and used in both models.

$$\widehat{Gender}_{Linear} = \beta_0 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5$$

$$\widehat{Gender}_{Better} = \beta_0 + \beta_5 x_5 + \beta_4 x_4$$

	Female	Male
Female Prediction	136	89
Male Prediction	188	1587

Table 3.1: Linear model of predicting genders

$$Accuracy_{Linear} = \frac{(Correct\ Male\ predictions) + (Correct\ Female\ Predictions)}{(All\ predictions)} = .8615$$

	Female	Male
Female Prediction	117	77
Male Prediction	207	1599

Table 3.2: Better model of predicting genders

$$Accuracy_{Better} = \frac{(Correct\ Male\ predictions) + (Correct\ Female\ Predictions)}{(All\ predictions)} = 0.8580$$

The accuracy of the linear model when predicting genders is lower slightly in the better model than the linear model. Even the values have very similar figures; female correctly predicted, male correctly predicted, etc. >85% is good enough to say these models are more than likely accurate enough to predict gender.

Conclusion

With the two regression models found, it is possible to predict a powerlifter's deadlift score. With the linear model, it is possible to do so given their weight class, age, and gender. Although this model overfits the data, it still shows it is possible to get within the range. With the better model, it is possible to do with only the body weight and the weight class and it does so without overfitting the data at the cost of higher variance. Using both models it is possible to predict which gender a person is given their age, body weight, weight class, and equipment.

Reflection

The project was difficult when finding a better model because there are so many types of regression; logistic, linear, poisson, etc. In addition, finding out possible two-predictor interactions between the variables. It was a lot of guess and checking work and it is disappointing to see how simple the "best" model was. The other hard thing was learning to format someone else's data into a form that is applicable for the project since OpenPowerlifting is a large dataset. In terms of improvement, it would be helpful to find if it was possible to have three predictors interacting with each other but guessing and checking this much probably is not good. Maybe using machine learning to take care of the guess work may find a better method, but for now, this is the best model.

References

Open Powerlifting. (2020). OpenPowerlifting. https://www.openpowerlifting.org/

Greg Nuckols. (2018, July 5). How Sex, Strength, and Age Affect Strength Gains in Powerlifters. *Stronger by Science*. https://www.strongerbyscience.com/predict-strength-gains/ (Did not cited anything yet, using it as supplementary material right now to make better guesses about the higher-order predictors. This will hopefully save me time when guessing and checking.)