Descriptive Analysis of Powerlifting Competition Data and Predicting Max Deadlift of Competitors

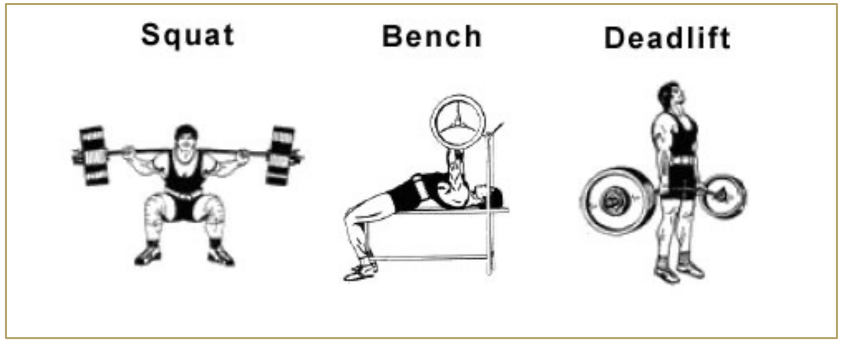
April 2023

Kevin Cho

Introduction

Background

- Powerlifting is a sport where competitors attempt their 1 rep max in 3 lifts:
  - Squat
  - Bench
  - Deadlift
- Powerlifting competitions are held worldwide, and most results can be found online at openpowerlifting.org



Research Questions

- Can a competitor's gender, age, bodyweight, equipment, and max squat/bench be used to predict their max deadlift with linear regression?
- Can variable selection or non-linear methods improve prediction?
- What can be learned by interpreting the coefficients of a linear regression model?
- Which variables are most important for predicting max deadlift?

Data Source and Cleaning

- A public-domain of powerlifting data is maintained by openpowerlifting.org
- The data from 1964-2019 is available on Kaggle
  - Contains data from over 22,000 competitions and 412,000 lifters
  - The original dataset contains 1,048,575 rows x 37 columns
- The following processing steps were performed to clean up and reduce the dataset:
  - Filtered for only competitors who performed all 3 lifts (squat, bench, deadlift) at the same event
  - Filtered for only “USAPL” federation to ensure population is lifting under similar conditions/rules
  - Removed duplicates that can occur when competitors enter multiple divisions
  - Selected columns for gender, equipment, age, bodyweight, max squat, max bench, and max deadlift
  - Removed any rows with missing data
- The processed dataset spans 1997-2019 and is 82,183 rows x 7 columns
- Cook’s distance calculation indicated there are no outliers or influential points

Explanation of Variables

Variable	Type	Description
Sex	Binary	0=Female, 1=Male
Equipment	Binary	0=Raw (belt, knee sleeves, wrist wraps), 1=Single-ply (additional supportive gear)
Age	Continuous	Age in years of competitor
BodyweightKg	Continuous	Bodyweight in kg of competitor
Best3SquatKg	Continuous	Max squat in kg of competitor
Best3BenchKg	Continuous	Max bench in kg of competitor
Best3DeadliftKg	**Continuous**	**Max squat in kg of competitor (Response Variable)**

Exploratory Data Analysis

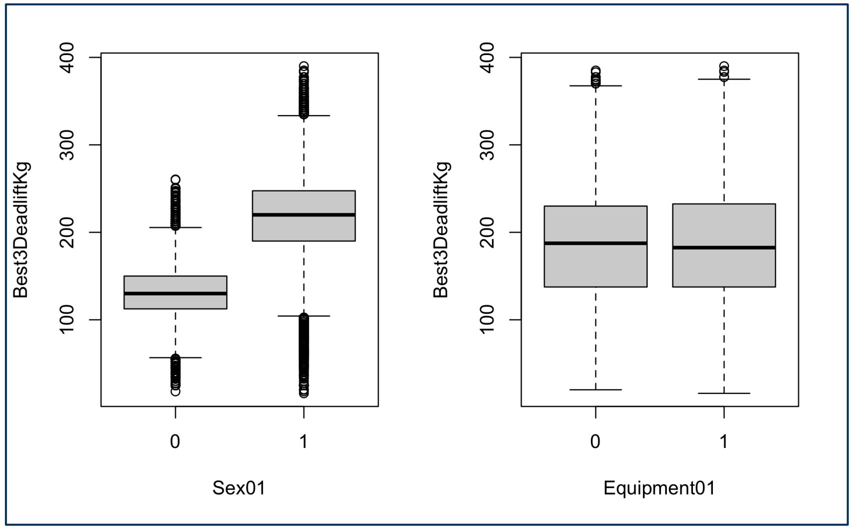


Figure 1 – Boxplots of response vs binary variables

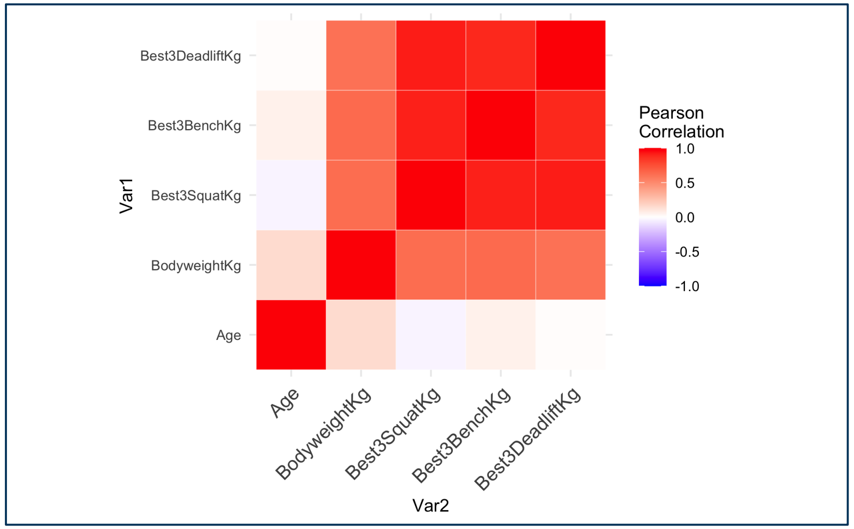


Figure 2 – Heat map of correlation coefficients for continuous variables

- Median max deadlift for males is ~100kg higher than for females
- Median max deadlift for Raw lifters is slightly higher than for Single-ply lifters
- Will explore the conditional relationship of these variables to the response using regression

- Deadlift has high correlation to squat and bench, moderate correlation to bodyweight, and almost no correlation to age
- Expecting squat and bench to be most important variables in regression model
- Squat and bench have high correlation to each other, but VIF calculation indicated multicollinearity should not be an issue

Linear Regression Assumptions

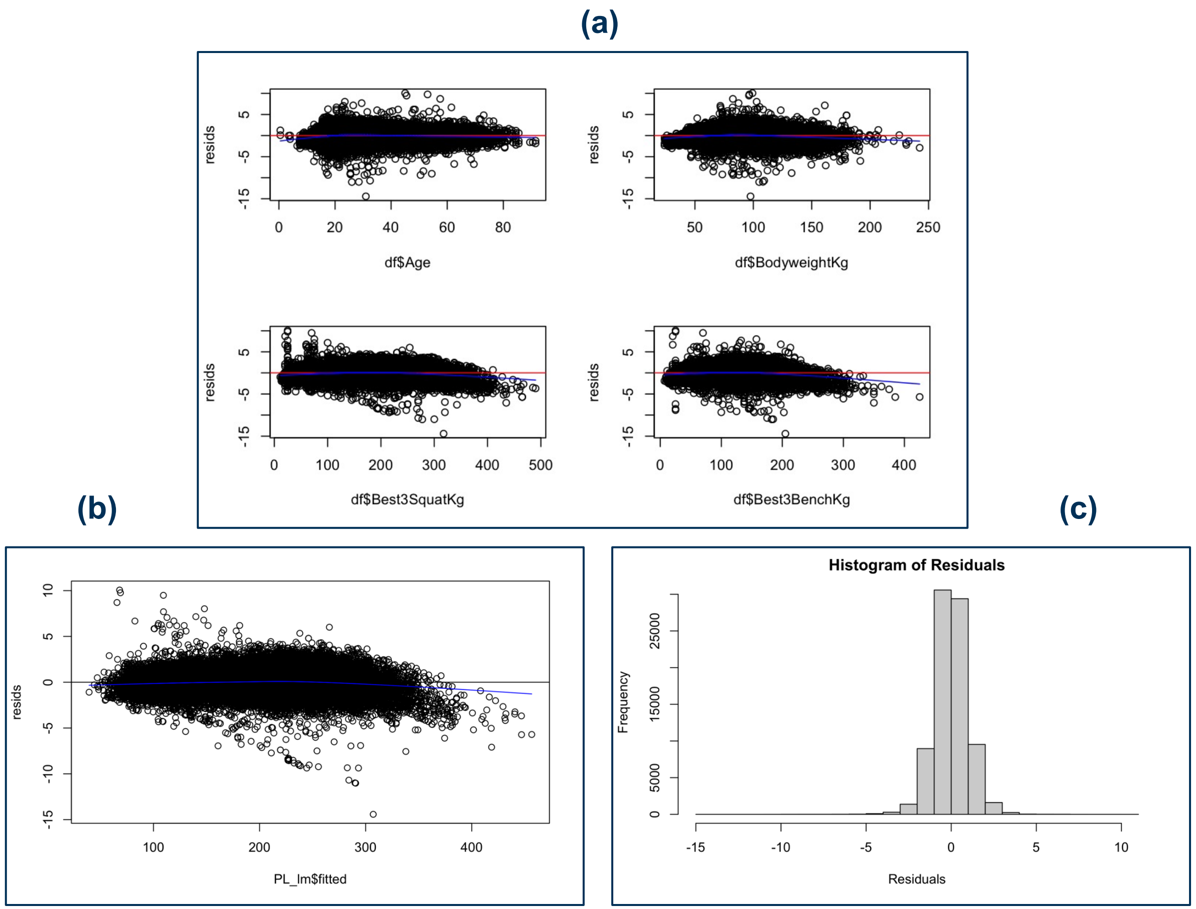


Figure 3 – Plots to check assumptions (a) Standardized Residuals vs Predictors (b) Standardized Residuals vs. Fitted Values (c) Histogram of Standardized Residuals

- Linear Regression model assumptions were checked before proceeding to see if regression is reasonable for this dataset
- Used a linear regression model with all 6 predictors on processed dataset
- Figure 3a checks linearity, 3b checks constant variance, and 3c checks normality
- All assumptions hold reasonably well enough to proceed
- Linearity/constant variance assumptions start to not hold at high values, will be mindful that linear regression starts to overestimate max deadlift for the small population of very strong/heavy lifters

Methodology

- Processed dataset was split to 70% training (57,528 samples) and 30% test (24,655 samples)
- Mean Squared Error (MSE) used as error metric for max deadlift prediction
- All models except Random Forest were further evaluated with 10 runs of Monte Carlo CV

#	Model	Variable Selection?	Description
1	Linear Regression with all 6 predictors	No	Baseline model
2	Linear Regression with 4 best predictors	Yes	Exhaustive search to find best subset with lowest residual sum of squares
3	Linear Regression with stepwise AIC selection	Yes	Combination of forward/backward stepwise selection to minimize AIC
4	LASSO Regression	Yes	Estimates coefficients by accounting for L1-norm penalty; penalty parameter tuned by minimizing Mallow's Cp
5	Ridge Regression	No	Estimates coefficients by accounting for L2-norm penalty; penalty parameter tuned by generalized cross validation
6	Random Forest	No	Non-linear ensemble method to improve prediction

Model Comparison Results

	Model <chr>	Train_MSE <dbl>	Test_MSE <dbl>	CV_MSE <dbl>	CV_variance <dbl>
1	LM w/ all predictors	382.6141	383.7175	384.9885	18.7409
2	LM with k=4 best predictors	383.7912	384.7560	386.0864	19.6394
3	LM with stepwise AIC	382.6141	383.7175	384.9885	18.7409
4	LASSO Regression	382.6141	383.7175	384.9885	18.7409
5	Ridge Regression	382.6141	383.7167	384.9882	18.7405
6	Random Forest	343.2596	350.6215	350.6215*	NA

*Imputed from Test MSE

Table 1 – Performance metrics of all models

- Model 2 (Best Subset) forces variable selection; the unselected predictors were age and bodyweight
  - The CV MSE is only ~1 unit higher than the baseline model, so performance is very similar
  - However, partial F-tests found that the unselected predictors are still significant with predictive power
- Model 3 (Stepwise) and 4 (LASSO) ended up selecting all predictors so their results are the exact same as Model 1 (baseline)
- Model 5 (Ridge) was tuned to have a small penalty term, therefore its estimated coefficients and performance ended up being effectively the same as the baseline model
- The non-linear nature of Model 6 (Random Forest) was able to improve MSE from baseline by ~35 units
- Random forest does not have coefficients to interpret, so the descriptive analysis on the next slide will be performed on Model 1 (the best regression model)

Descriptive Analysis

```
Call:
lm(formula = Best3DeadliftKg ~ ., data = df)

Residuals:
     Min       1Q   Median       3Q      Max 
-282.161  -11.771   -0.155   11.785  196.898 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  44.777918   0.314838  142.22   <2e-16 ***
Sex01        16.439803   0.208570   78.82   <2e-16 ***
Equipment01 -16.518415   0.187159  -88.26   <2e-16 ***
Age           0.091070   0.006271   14.52   <2e-16 ***
BodyweightKg -0.035686   0.004289   -8.32   <2e-16 ***
Best3SquatKg  0.673516   0.002890  233.07   <2e-16 ***
Best3BenchKg  0.240393   0.004202   57.22   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 19.57 on 82176 degrees of freedom
Multiple R-squared:  0.8878,	Adjusted R-squared:  0.8878 
F-statistic: 1.084e+05 on 6 and 82176 DF,  p-value: < 2.2e-16
```

Figure 4 – Model summary for linear regression using all predictors trained with full processed dataset

- R-squared value indicates 89% of the variability in the response variable is explained by the regression model
  - Reasonably good fit, can proceed with interpreting relationship between predictors and response using coefficients
- Each interpretation is made with the condition that all other variables are held constant
- Although the coefficients for age and bodyweight are statistically significant, they are very small and not of much practical significance
- Squat and bench have the most predictive power
  - With every 1kg of max squat gained, max deadlift is expected to increase 0.67kg
  - With every 1kg of max bench gained, max deadlift is expected to increase 0.24kg
- Males are expected to have a 16.4kg higher max deadlift than females of equivalent profile
- Single-ply lifters are expected to have a 16.5kg lower max deadlift than Raw lifters of equivalent profile

Conclusions (Part 1)

- Best performing regression model was the baseline model with all predictors
  - R-squared of 0.89 suggests reasonable fit
  - Variable selection and Ridge could not improve prediction and reduce MSE
- Random Forest, a non-linear method, could improve prediction
  - Expect it was able to improve prediction for the small population of very heavy/strong lifters where the linear regression model assumptions were starting to not hold
- All 6 predictors found to be statistically significant in linear regression
  - However, age and bodyweight had very small coefficients and are not of practical significance
- Squat and bench have most predictive power for predicting max deadlift
  - Expect this is more so correlation than causation, a lifter’s max squat/bench is an indication of their training level and general strength
  - In the real world, a lifter increases their max in the 3 lifts by training each lift; it is unlikely one can increase their deadlift by only training squat/bench

Conclusions (Part 2)

- Males are expected to have a 16.4kg higher max deadlift than females of equivalent profile
  - This suggests males of females of equivalent bodyweight that can squat/bench the same amount will not deadlift the same amount on average
  - Males could possibly have a mechanical advantage in deadlift due to anatomical differences, would need to be confirmed with additional physiological/biomechanical research
- Single-ply lifters are expected to have a 16.5kg lower max deadlift than Raw lifters of equivalent profiles
  - Extra lifting gear used in Single-ply events could possibly be more advantageous to squat and bench that it is to deadlift
  - I.e., the extra gear may help a generally “weaker” lifter achieve a squat and bench similar to that of a Raw lifter of equivalent bodyweight, gender, age; but the extra gear may not help the “weaker” lifter match the Raw lifter’s deadlift

Future Work Ideas

- Repeat this analysis but use squat or bench as the response variable instead of deadlift
  - Would be interesting to see if the same conclusions are confirmed or if new observations are made
- Extend analysis to a larger population of lifters
  - Explore if a model trained with data from the USAPL federation would have similar error when tested on lifters of another federation or country
  - New models might need to be fitted for lifters of other federations/countries; interpreting their coefficients might uncover new observations