Abstract

Knowing the best way to predict how much someone can lift is useful information, especially for powerlifters and their coaches. This study looked at data from an open archive of all the female powerlifters in the world and how much they bench pressed based on how much they deadlifted, squatted, their age, bodyweight, and the equipment they used. We compared three different linear models: a multiple linear model using all of the explanatory variables, a multiple linear model using only how much each powerlifter deadlifted and squatted as explanatory variables, and a simple linear model using only how much they squat as the explanatory variable. Based on hypothesis testing, we found that the simple linear model using only squatting weight as the explanatory variable was the most useful model at predicting how much a female powerlifter can lift.

The intent of this project is to find a model that can best predict how much female powerlifters can bench press based on how much they can lift in other events, their age and weight, and what equipment they use. The data used in this project comes from openpowerlifting.org, an open archive where statistics on powerlifters and the competitions they participated in are uploaded onto a permanent and accessible website. Information such as where the meet was located, when it took place, how much they lifted, what age group they participated in, and so on are types of data points that can be included. Researchers from around the globe use data from the website to conduct research in a variety of fields. Since the data is uploaded consistently, the data in this project comes from the October 2020 version.

Data

The candidate explanatory variables are BodyweightKg, Deadlift, Squat, Age, and Equipment. BodyweightKg, Deadlift, Squat, and Age are explanatory numerical variables while Equipment is a categorical explanatory variable.I selected these variables because I believe that they can help predict how much someone can Bench Press due to the fact that Deadlift and Squat are also compound movements and Equipment type has been shown to increase lifts.

Due to the international nature of powerlifting, all of the numerical variables- besides Age- in this analysis are in kilograms (BodyweightKg, Squat, Bench, Deadlift).

For the equipment variable, the different levels are multi-ply, raw, single-ply, straps, unlimited, and wraps. To give context, powerlifters can either compete in either raw or equipped contests. Raw lifting means that lifters use little to no special equipment. Wraps, straps, unlimited, single-ply, and multi-ply fall under the category of equipped lifting. Wraps means using Velcro wraps around the wrists to provide stability and prevent slipping, which could cause injury. Straps do not protect the wrist, but rather increase the amount of weight one can grip. Multi-ply and single-ply refer to the type of bodysuit that powerlifters use. Powerlifting suits are made of high tensile strength fabric. When an athlete goes into the start position for a lift, the suit stretches and creates elastic energy, which helps the powerlift lift more. Single-Ply means the suit has one layer of high tensile strength fabric, whereas Multi-Ply means two layers or more. More layers mean more elasticity and more strength. It's estimated that equipped lifting allows lifters

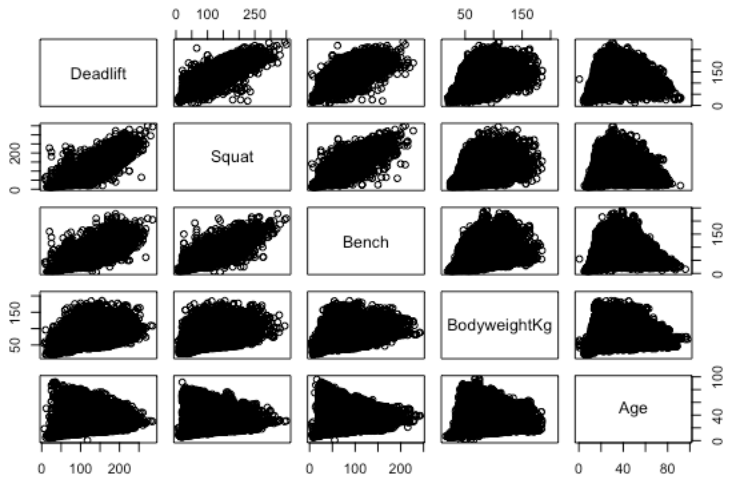
to lift approximately 115% of their max (Siem, 2016). This ability to allow lifters to lift more via equipment type is why the Equipment variable has been included in the project.

The correlation matrix between numerical variables is as follows:

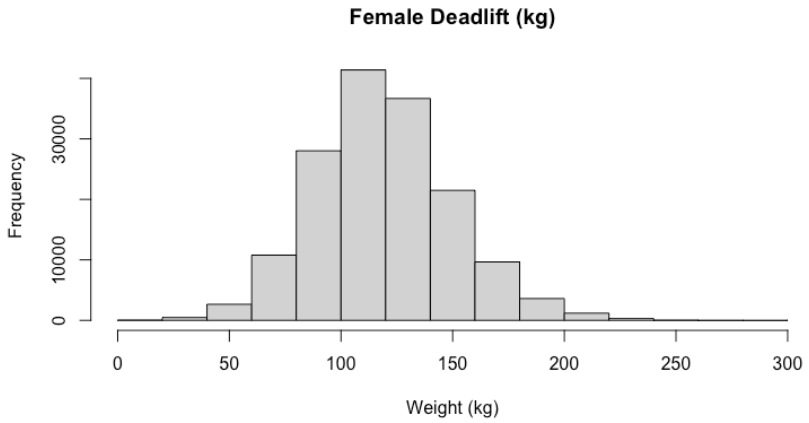
```
> cor(temp_powerlift,use="complete.obs")
                Deadlift       Squat      Bench BodyweightKg         Age
Deadlift     1.00000000  0.86264993 0.78309370    0.3512406  0.04576868
Squat        0.86264993  1.00000000 0.84314555    0.3442522 -0.05995648
Bench        0.78309370  0.84314555 1.00000000    0.3169408  0.05847171
BodyweightKg 0.35124056  0.34425222 0.31694076    1.0000000  0.10576005
Age          0.04576868 -0.05995648 0.05847171    0.1057600  1.00000000
```

Looking at the correlation matrix, there are not any correlations greater than 0.9 between any of the numerical variables. For the most part, there is a moderately strong relationship between the variables while there is a moderately weak relationship between BodyweightKg and the other variables.There is actually a negative correlation between Age and squatting weight and very weak correlations between Age and any other variable.

The scatterplot matrix of numerical variables is as follows:

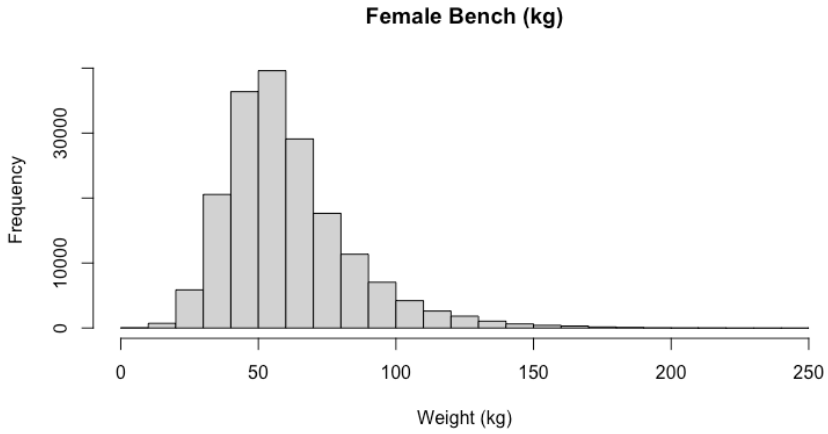


Looking at the scatterplot matrix, there looks to be strong, positive, linear relationships between all of the three compound movements (Deadlift, Squat, and Bench). The other scatterplots don't look to be nearly as strong or as linear, especially for Age as the explanatory variables against the other numerical variables.



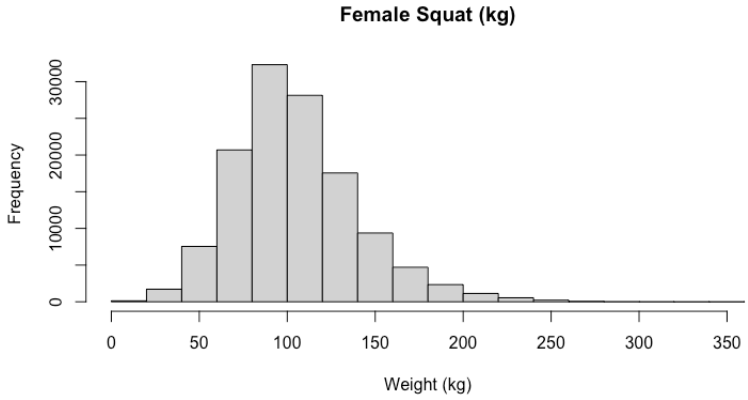

```
> summary(powerlift$Deadlift)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
    6.8   100.0   120.0   121.1   140.0   285.0  355202
> sd(powerlift$Deadlift, na.rm = TRUE)
[1] 30.64094
```

The distribution of female deadlifts looks to be normally distributed. There do not seem to be any clear outliers. The median of the distribution is 120 kg while the mean is 121.1 kg. The spread runs from 6.8 kg to 285 kg.




```
> summary(powerlift$Bench)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
    5.0    47.5    57.5    62.4    72.5   242.5  331807
> sd(powerlift$Bench, na.rm = TRUE)
[1] 23.27161
```

The distribution of female bench looks to be skewed right. It is possible that there are outliers, but more data exploration via 1.5 * IQR Rule would be necessary to confirm. The median is 57.5 kg while the mean is 62.4 kg. The spread runs from 5 kg to 242.5 kg.

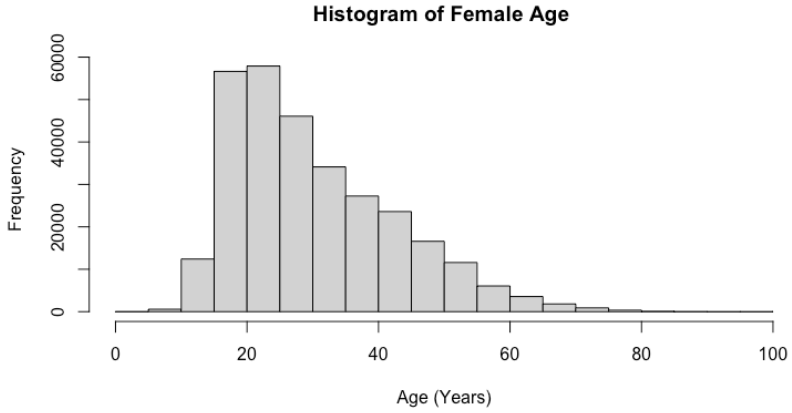



```
> summary(powerlift$Squat)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
    6.8    82.5   102.5   106.6   125.0   350.0  385045
> sd(powerlift$Squat, na.rm = TRUE)
[1] 35.3958
```

The distribution of female squat weight looks to be skewed slightly right. It is difficult to tell if there are any outliers by looking at the histogram, so we could check using the 1.5 * IQR Rule. The median is 102.5 kg while the mean is 106.6 kg. The spread runs from 6.8 kg to 350.0 kg.

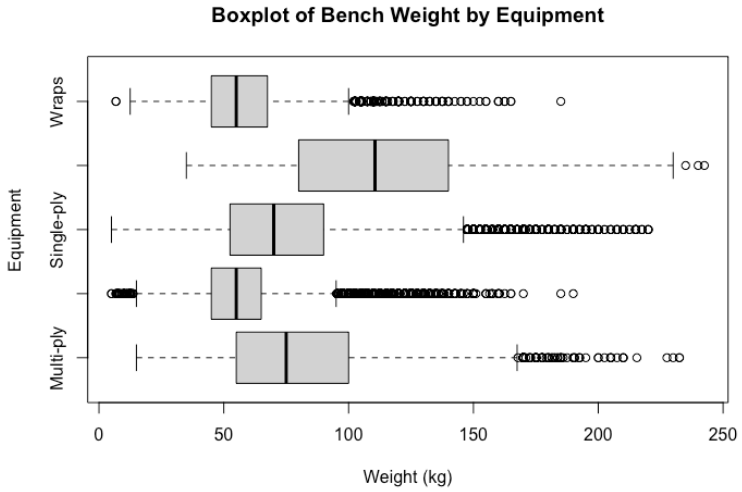

```
> summary(powerlift$BodyweightKg)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  15.88   55.55   64.14   67.53   75.00  205.66    6853
> sd(powerlift$BodyweightKg, na.rm = TRUE)
[1] 17.41229
```

The distribution of female looks to be skewed right. There do not seem to be any clear outliers, but we could use the 1.5 * IQR Rule to check. The median is 64.14 kg while the mean is 67.53 kg. The spread runs from 15.88 kg to 205.66kg.




```
> summary(powerlift$Age)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
   0.50   20.50   27.50   30.41   38.50   98.00  211953
> sd(powerlift$Age, na.rm = TRUE)
[1] 12.44516
```

The distribution looks to be skewed right. There do not seem to be any outliers, but more data exploration should be done to confirm via the 1.5 * IQR Rule. The median age is 27.5 years while the mean is 30.41. The spread runs from 0.5 to 98 years. It is possible that there was an error when inputting the age, because 6-month olds do not participate in powerlifting competitions, as far as we know after doing some light research.



```
> by(powerlift$Bench, powerlift$Equipment, summary)
powerlift$Equipment: Multi-ply
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  15.00   55.00   75.00   80.66  100.00  232.50    7903
-------------------------------------------------------
powerlift$Equipment: Raw
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
   5.00   45.00   55.00   57.16   65.00  190.00   93991
-------------------------------------------------------
powerlift$Equipment: Single-ply
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
   5.00   52.50   70.00   74.56   90.00  220.00  213178
-------------------------------------------------------
powerlift$Equipment: Straps
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
     NA      NA      NA     NaN      NA      NA       3
-------------------------------------------------------
powerlift$Equipment: Unlimited
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
   35.0    80.0   110.6   115.9   138.8   242.5      74
-------------------------------------------------------
powerlift$Equipment: Wraps
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
   6.80   45.00   55.00   57.96   67.50  185.00   16658
> by(powerlift$Bench, powerlift$Equipment, sd, na.rm = TRUE)
powerlift$Equipment: Multi-ply
[1] 33.943
-------------------------------------------------------
powerlift$Equipment: Raw
[1] 17.44028
-------------------------------------------------------
powerlift$Equipment: Single-ply
[1] 29.32397
-------------------------------------------------------
powerlift$Equipment: Straps
[1] NA
-------------------------------------------------------
powerlift$Equipment: Unlimited
[1] 43.34979
-------------------------------------------------------
powerlift$Equipment: Wraps
[1] 18.34745
```

The distribution for all of the levels seem to be skewed right because the means are larger than the medians. For each level, there are high outliers, and there are low outliers for “Raw” and “Wraps”. The median for “Multi-ply”, “Raw”, “Single-ply”, “Straps”, and “Wraps” respectively (in kg) are 75, 55, 70, 110.6, and 55. The mean for “Multi-ply”, “Raw”, “Single-ply”, “Straps”, and “Wraps” respectively (in kg) are 80.66, 57.16, 74.56, 115.9, and 57.96. The spread for “Multi-ply”, “Raw”, “Single-ply”, “Straps”, and “Wraps” respectively (in kg) are 15 to 232.5, 5 to 190, 5 to 220, 35 to 242.5, and 6.8 to 185.

Model Selection and Diagnostics

Model 1: Multiple Linear Regression (all selected explanatory variables)

The output of the multiple linear regression model of Bench against all selected explanatory variavles (Equipment, Age, BodyweightKg, Squat, and Deadlift) is as follows:

```
Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         -0.778746   0.318660  -2.444 0.014535 *
Deadlift             0.149554   0.002404  62.201  < 2e-16 ***
Squat                0.373408   0.002233 167.206  < 2e-16 ***
BodyweightKg         0.030151   0.002315  13.027  < 2e-16 ***
EquipmentRaw        -4.508398   0.262695 -17.162  < 2e-16 ***
EquipmentSingle-ply -0.950993   0.262749  -3.619 0.000295 ***
EquipmentWraps      -6.051784   0.272985 -22.169  < 2e-16 ***
Age                  0.164791   0.003055  53.937  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.78 on 96605 degrees of freedom
  (414921 observations deleted due to missingness)
Multiple R-squared:  0.7367,    Adjusted R-squared:  0.7367
F-statistic: 3.861e+04 on 7 and 96605 DF,  p-value: < 2.2e-16
```

Based on the output, the intercept is -0.779. At the 0.05 significance level, an intercept p-value of 0.0145 is statistically significant evidence against the null hypothesis that the intercept is equal to zero. The slope coefficient for Deadlift is 0.150. A p-value of <2e-16 is sufficient evidence to reject the null hypothesis that the slope coefficient is equal to zero. The slope coefficient for Squat is 0.373 and has a p-value of <2e-16, which is sufficient evidence to reject the null hypothesis that the slope coefficient is equal to zero. The slope coefficient for BodyweightKg is 0.0301 and has a p-value of <2e-16, which is sufficient evidence to reject the null hypothesis that the slope coefficient is equal to 0. The slope coefficient for EquipmentRaw is -4.508 and has a p-value of <2e-16 which is sufficient evidence to reject the null hypothesis that the slope coefficient is equal to zero. The slope coefficient for EquipmentSingle-ply is -0.951 and has a p-value of 0.000295 which is sufficient evidence to reject the null hypothesis that the slope coefficient is equal to zero. The slope coefficient for EquipmentWraps is -6.051 and has a p-value of <2e-16 which is sufficient evidence to reject the null hypothesis that the slope coefficient is equal to zero. The slop coefficient for Age is 0.165 and has a p-value of <2e-16 which is sufficient evidence to reject the null hypothesis that the slope coefficient is equal to zero.

With an adjusted R-squared value of 0.7367, 73.67% of the variation in Bench weight can be explained by the multiple linear model. Additionally, since the F-test p-value is <2.2e-16, this indicates that the model is useful and appropriate, and that each explanatory variable offered predictive power to the model.

```
> vif(model1)
                 GVIF Df GVIF^(1/(2*Df))
Deadlift     4.460569  1        2.112006
Squat        5.159611  1        2.271478
BodyweightKg 1.235690  1        1.111616
Equipment    1.356000  3        1.052067
Age          1.067256  1        1.033081
```

There are no GVIF values over 10, which indicates that there is no multicollinearity between variables.

Model 2: Multiple Linear Regression (Deadlift + Squat)

We decided to run another multiple linear regression, this time only including Deadlift and Squat.

The summary output for model 2 is as follows:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.637632   0.139604   4.567 4.94e-06 ***
Squat       0.389200   0.001857 209.640  < 2e-16 ***
Deadlift    0.148373   0.002149  69.057  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.94 on 111385 degrees of freedom
  (400146 observations deleted due to missingness)
Multiple R-squared:  0.722,     Adjusted R-squared:  0.7219
F-statistic: 1.446e+05 on 2 and 111385 DF,  p-value: < 2.2e-16
```

Based on the output, the intercept is 0.6376 and had a p-value of 4.94e-06, which is statistically significant to reject the null hypothesis that the intercept is 0. Deadlift has a slope coefficient of 0.1484 while Squat has a slope coefficient of 0.3892. All of the slope coefficients are statistically significant at 0.05 significance level with a p-value of <2e-16 for each coefficient.

With an adjusted R-squared value of 0.7219, 72.19% of the variation in Bench weight can be explained by the multiple linear model. Additionally, since the F-test p-value is <2.2e-16, this indicates that the model is useful and appropriate, and that each explanatory variable offered predictive power to the model.

```
> vif(model2)
   Squat Deadlift
3.923558 3.923558
```

There are no VIF values over 10 so there is no multicollinearity between variables.

Model 3: Simple Linear Regression (Bench~Squat)

Here we chose to do a simple linear regression between Bench and Squat because the correlation matrix indicated that there was a strong relationship between the two variables more so than with the other explanatory variables.

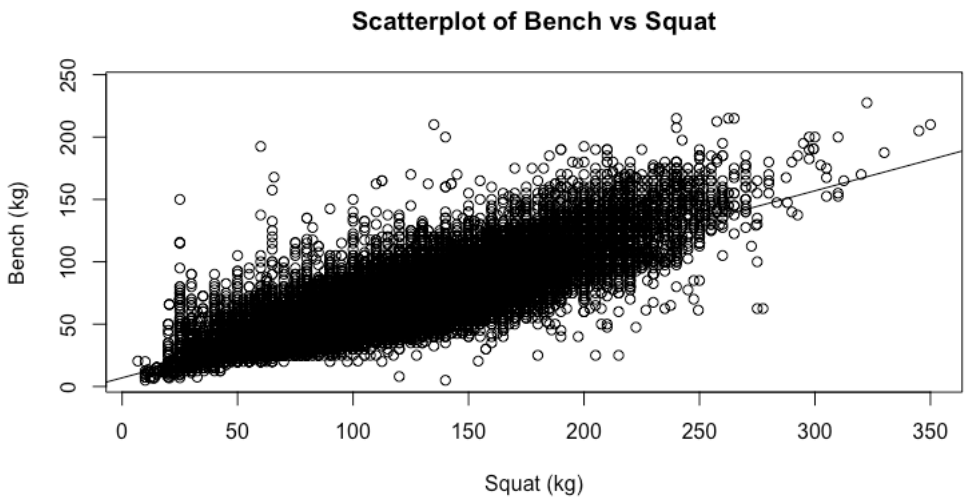
The model summary output is as follows:

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 6.9603675  0.1052337   66.14   <2e-16 ***
Squat       0.5000466  0.0009398  532.07   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.18 on 115278 degrees of freedom
  (396254 observations deleted due to missingness)
Multiple R-squared:  0.7106,    Adjusted R-squared:  0.7106
F-statistic: 2.831e+05 on 1 and 115278 DF,  p-value: < 2.2e-16
```

Based on the output, the intercept value is 6.9604 and has a p-value of <2e-16 which his sufficient evidence to reject the null hypothesis that the intercept is zero. The squat slope coefficient is 0.5000 and has a p-value of <2e-16. The interpretation of the slope means that for every 1 kg increase in squat weight, bench weight will increase by 0.5 kg.

With an adjusted R-squared value of 0.7106, 71.06% of the variation in Bench weight can be explained by the linear relationship between Bench and Squat. Additionally, since the F-test p-value is <2.2e-16, this indicates that the model is useful and appropriate, and that Squat offers predictive power to the simple linear model.



The above scatterplot indicates a strong positive relationship between Bench and Squat. The relationship looks linear, and the regression line obtained from the simple linear model appears to fit the distribution well.

AIC

```
> AIC(model1, model2, model3)
       df      AIC
model1  9 733630.1
model2  4 849023.8
model3  3 883766.4
```

The AIC values indicate that model 1 is the preferred model to use, however since the models were made based on a different number of data points, more emphasis will be placed on adjusted R-squared values, SVAs, and F-tests when selecting the final model to analyze.

Model Selection

The first model (multiple linear regression of Bench against all explanatory variables) has a statistically significant f-test p-value, indicating that the model is useful. All slope coefficient p-values were statistically significant. It also has the lowest AIC value, indicating that it is the more appropriate model. In terms of SVAs, while the qqplot did not indicate normality, having a sample size larger than 40 means we can proceed with procedures. Some of the residual plots were better than others in terms of equal variance and linearity. The strip chart of equipment seemed to have equal variance across equipment types. The Adjusted R-squared value is 0.7367, which is the highest of the three models.

The second model (multiple linear regression of Bench against Squat + Deadlift) also had a statistically significant f-test p-value. All slope coefficient p-values were statistically significant. The AIC value was larger than the value for model 1 and less than the AIC value of model 2. All the SVAs for model 2 were satisfied. The Adjusted R-squared value is 0.7219, which is only around 1% less than the most complex model.

The third model (simple linear regression Bench ~ Squat) had a statistically significant f-test p-value. The slope coefficient had a statistically significant p-value. The third model had the highest AIC. The residual plot for model 3 had equal variance and no clear nonlinear patterns. The normality condition was satisfied as well. The Adjusted R-squared value is 0.7106, which is the lowest of the three models.

While the models are very similar, we will choose model 3 because the coefficients in the linear model were statistically significant, there was a statistically significant f-test p-value, completely satisfied SVAs, and a relatively high Adjusted R-squared value. In this case, the AIC was not completely helpful because the models had a different number of data points, so that test holds the least weight in deciding. If we look at the Adjusted R-squared between all 3 models, having only one predictor (like in model 3) managed to garner a 0.7106. Adding the additional predictors, thereby making a more complicated model, only increased the Adjusted R-squared by 2% at most (as seen in model 1). Additionally, not all the SVAs in model 1 were satisfied while they were all satisfied in models 2 and 3. While the additional variables in the other models add predictive power, the increase is relatively small, so in this case it seems that it would be best to stick with the simpler model to get an as accurate prediction as the more complex models would give. The more complex the model, the more likely that the model is tailored to this particular dataset and generalizability suffers. Additionally, simpler models are easier to understand and is computationally efficient.

Summary of Statistical Findings

The selected model output is as follows:

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 6.9603675  0.1052337   66.14   <2e-16 ***
Squat       0.5000466  0.0009398  532.07   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.18 on 115278 degrees of freedom
  (396254 observations deleted due to missingness)
Multiple R-squared:  0.7106,    Adjusted R-squared:  0.7106
F-statistic: 2.831e+05 on 1 and 115278 DF,  p-value: < 2.2e-16
```

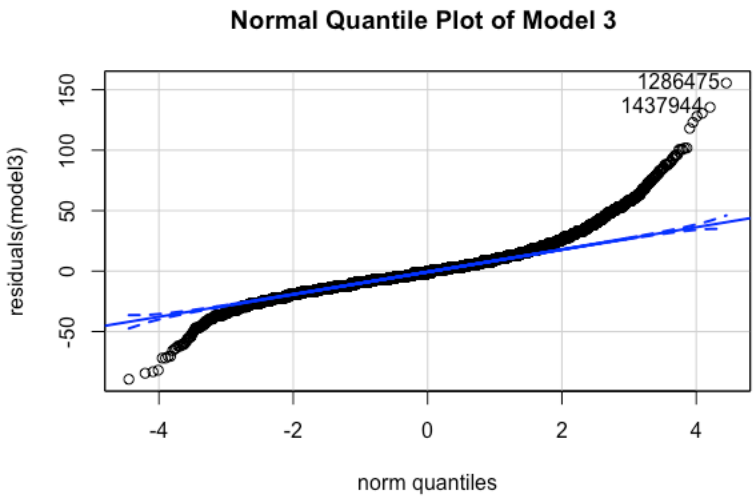
The fitted model equation is as follows:

$$\mu\,[Bench|Squat] \;=\; 6.9603675 \;+\; 0.5000466(Squat)$$

It makes sense that Squat has a positive association with Bench powerlifters do not just train on compound movement, they train in bench, squat, and deadlift events. If someone is able to squat more, then it is intuitive that they would be able to bench more as well.

The magnitude makes sense since bench press is an upper body focused compound movement and arm muscles are smaller and less capable of lifting weight than legs do since some of largest muscles in the body are located in the lower body (*What is the strongest muscle in the human body?*, n.d.). Therefore, powerlifters generally lift less when they bench press. In that way, it makes sense that the increase in bench press is not 1 to 1.

Sampling Variability Assumptions (SVAs):



It looks like many data points lie outside the dashed boundary lines, but since the sample size is greater than 40, we can satisfy the normality condition here.

The variance seems to be equal across the residual plot. There do not seem to be any striking or concerning nonlinear patterns.

Conclusion

The model does a good job of predicting Bench weight. The summary output has statistically significant p-values for all of the slope coefficient in the model. The Adjusted R-squared value is 0.7106, which means that 71.06% of the variability in Bench can be explained by the linear relationship between Bench and Squat, which is a large portion of the data. The SVAs are satisfied, meaning there are no issues with normality or variance that could interfere with statistical procedures. Like we mentioned before when selecting the model, it seemed best to choose the simpler model for reasons such as computational efficiency, ease of understanding, and similarly accurate prediction power compared to more complex models.

Predictions:

```
> predict(model3, new = newData, interval = "confidence")
       fit      lwr      upr
1 86.46777 86.35128 86.58427
> predict(model3, new = newData, interval = "prediction")
       fit     lwr      upr
1 86.46777 64.5538 108.3817
```

CI interpretation: The estimated Bench for all female powerlifters who squat 159 kg is 86.47 kg. 95% CI: 86.35, 86.58. We are 95% confident that the true mean bench press weight for all female powerlifters who squat 159 kg is between 86.35 kg and 86.58 kg.

PI interpretation: The estimated Bench for a female powerlifter who squats 159 kg is 86.47 kg. 95% PI: 64.55, 108.38. We are 95% confident that the bench press weight for a female powerlifter who squats 159 kg is between 64.55 kg and 108.38 kg.

Bibliography

Keys, M. (2012). *Powerlifting a Brief History – Cast Iron Strength*. Castironstrength.com. https://www.castironstrength.com/powerlifting-a-brief-history/

Rankings. (2020). Openpowerlifting.org. https://www.openpowerlifting.org/

Siem, B. (2016, July 8). *Raw vs Equipped Powerlifting*. BarBend; BarBend. https://barbend.com/raw-vs-equipped-powerlifting/

What is the strongest muscle in the human body? (n.d.). Library of Congress, Washington, D.C. 20540 USA. https://www.loc.gov/everyday-mysteries/item/what-is-the-strongest-muscle-in-the-human-body/#:~:text=The%20gluteus%20maximus%20is%20the