DATA 605 - Discussion 11

Omer Ozeren

Table of Contents


```
Using R, build a  regression model for data that interests you.  Conduct
residual analysis.
Was the linear model appropriate? Why or why not?
```

For this discussion, I will look at Kaggle's Powerlifting Database dataset.

- It's a dataset containing competitor results in powerlifting from the OpenPowerlifting Database and do residual analysis.

I will build a simple linear regression model of body weights vs best bench press for seniors to see if a linear relation exists between them.

- Dataset can be found here: https://www.kaggle.com/open-powerlifting/powerlifting-database
- Get the data and examine a preview

```
library(data.table)

## Warning: package 'data.table' was built under R version 3.5.3

powerlift <-
read.csv('C:\\Users\\OMERO\\Documents\\GitHub\\DATA605\\openpowerlifting.csv'
)
head(powerlift, n=5)

##              Name Sex Event Equipment Age AgeClass Division BodyweightKg
## 1    Abbie Murphy   F   SBD     Wraps  29    24-34     F-OR         59.8
## 2     Abbie Tuong   F   SBD     Wraps  29    24-34     F-OR         58.5
## 3  Ainslee Hooper   F     B       Raw  40    40-44     F-OR         55.4
## 4 Amy Moldenhauer   F   SBD     Wraps  23    20-23     F-OR         60.0
## 5    Andrea Rowan   F   SBD     Wraps  45    45-49     F-OR        104.0
##   WeightClassKg Squat1Kg Squat2Kg Squat3Kg Squat4Kg Best3SquatKg Bench1Kg
## 1            60       80     92.5      105       NA          105     45.0
## 2            60      100    110.0      120       NA          120     55.0
## 3            56       NA       NA       NA       NA           NA     27.5
## 4            60     -105   -105.0      105       NA          105     67.5
```

```
## 5           110      120    130.0      140       NA          140     70.0
##   Bench2Kg Bench3Kg Bench4Kg Best3BenchKg Deadlift1Kg Deadlift2Kg
## 1     50.0     55.0       NA         55.0       110.0         120
## 2     62.5     67.5       NA         67.5       130.0         140
## 3     32.5    -35.0       NA         32.5          NA          NA
## 4     72.5    -75.0       NA         72.5       132.5        -140
## 5     75.0     80.0       NA         80.0       150.0         160
##   Deadlift3Kg Deadlift4Kg Best3DeadliftKg TotalKg Place  Wilks McCulloch
## 1         130          NA           130.0   290.0     4 324.16    324.16
## 2         145          NA           145.0   332.5     2 378.07    378.07
## 3          NA          NA              NA    32.5     1  38.56     38.56
## 4        -140          NA           132.5   310.0     3 345.61    345.61
## 5         170          NA           170.0   390.0     3 321.25    338.91
##   Glossbrenner IPFPoints Tested Country Federation       Date MeetCountry
## 1       286.42    511.15                   GPC-AUS 2018-10-27   Australia
## 2       334.16    595.65                   GPC-AUS 2018-10-27   Australia
## 3        34.12    313.97                   GPC-AUS 2018-10-27   Australia
## 4       305.37    547.04                   GPC-AUS 2018-10-27   Australia
## 5       274.56    550.08                   GPC-AUS 2018-10-27   Australia
##   MeetState      MeetName
## 1       VIC Melbourne Cup
## 2       VIC Melbourne Cup
## 3       VIC Melbourne Cup
## 4       VIC Melbourne Cup
## 5       VIC Melbourne Cup
```

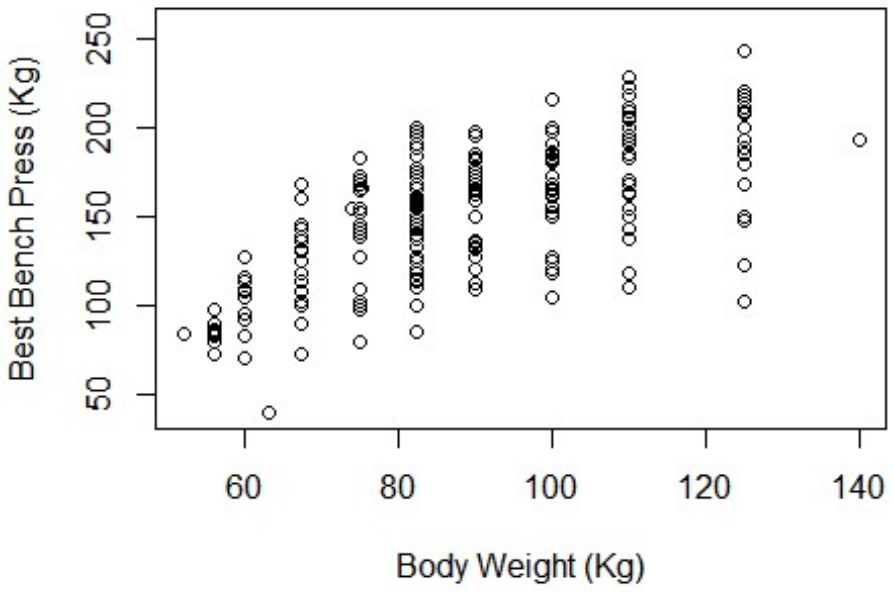
```r
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.5.3
```

```r
powerlift_senior <- powerlift %>% filter(Best3BenchKg > 0 & Division ==
'Senior') %>%
  select(BodyweightKg, Best3BenchKg)
```

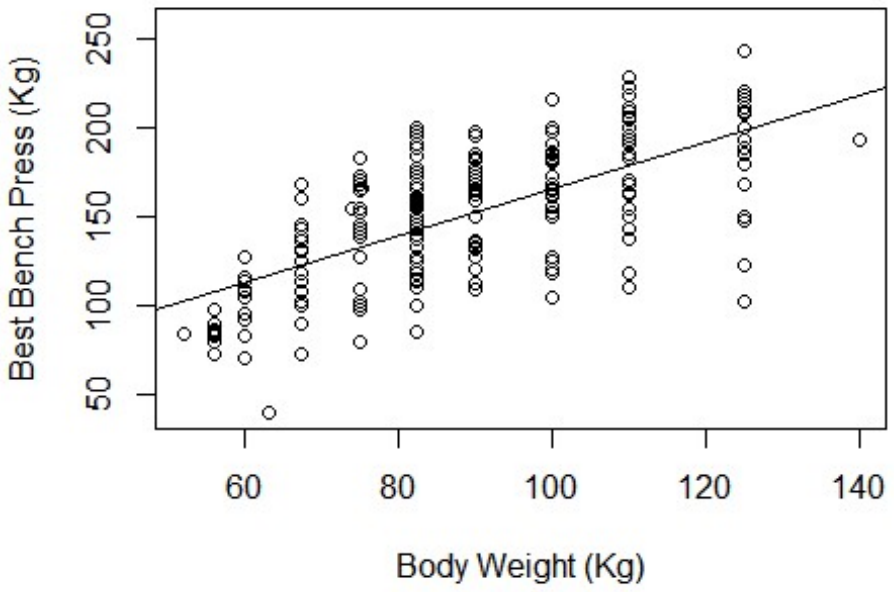
Visualize the data (EDA)

```r
with(powerlift_senior, plot(BodyweightKg, Best3BenchKg, xlab = "Body Weight
(Kg)",
                            ylab = "Best Bench Press (Kg)"))
```



Residual Analysis

```
lm_powerlift_senior <- lm(Best3BenchKg ~ BodyweightKg, data =
powerlift_senior)
with(powerlift_senior, plot(BodyweightKg, Best3BenchKg,xlab = "Body Weight
(Kg)",
                            ylab = "Best Bench Press (Kg)"))
abline(lm_powerlift_senior)
```



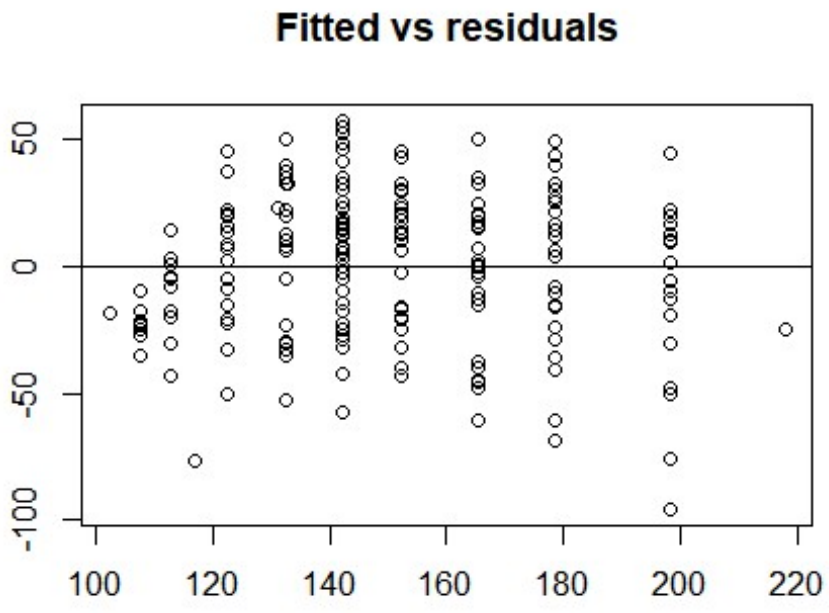

```
summary(lm_powerlift_senior)

##
## Call:
## lm(formula = Best3BenchKg ~ BodyweightKg, data = powerlift_senior)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -95.697 -20.860   4.566  20.359  57.206
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   34.0119     9.2492   3.677 0.000297 ***
## BodyweightKg   1.3135     0.1023  12.838  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 28.77 on 218 degrees of freedom
##   (12 observations deleted due to missingness)
## Multiple R-squared:  0.4305, Adjusted R-squared:  0.4279
## F-statistic: 164.8 on 1 and 218 DF,  p-value: < 2.2e-16
```

- Equation of line is

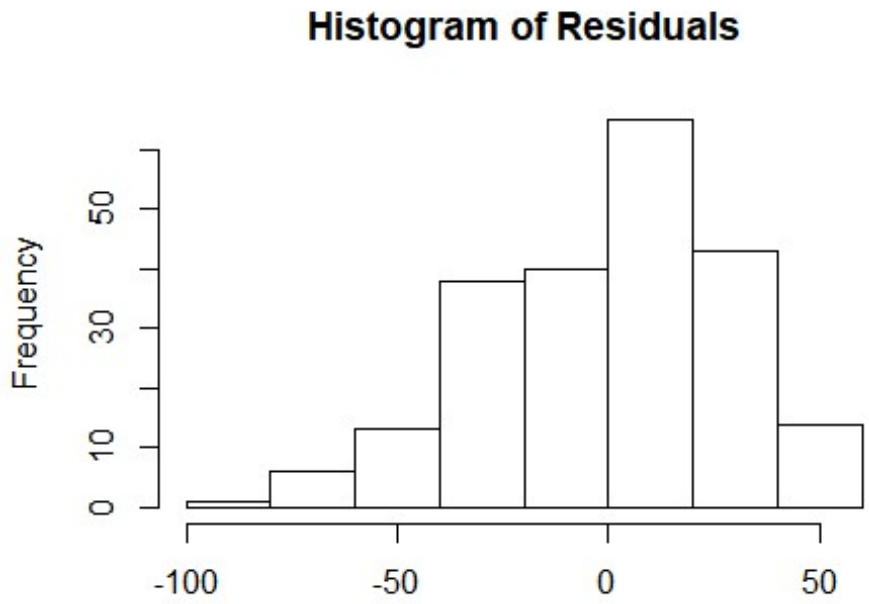
$$\hat{bestbenchpress} = 0.1376 + 1.2804 * \hat{bodyweight}$$

Residual plots

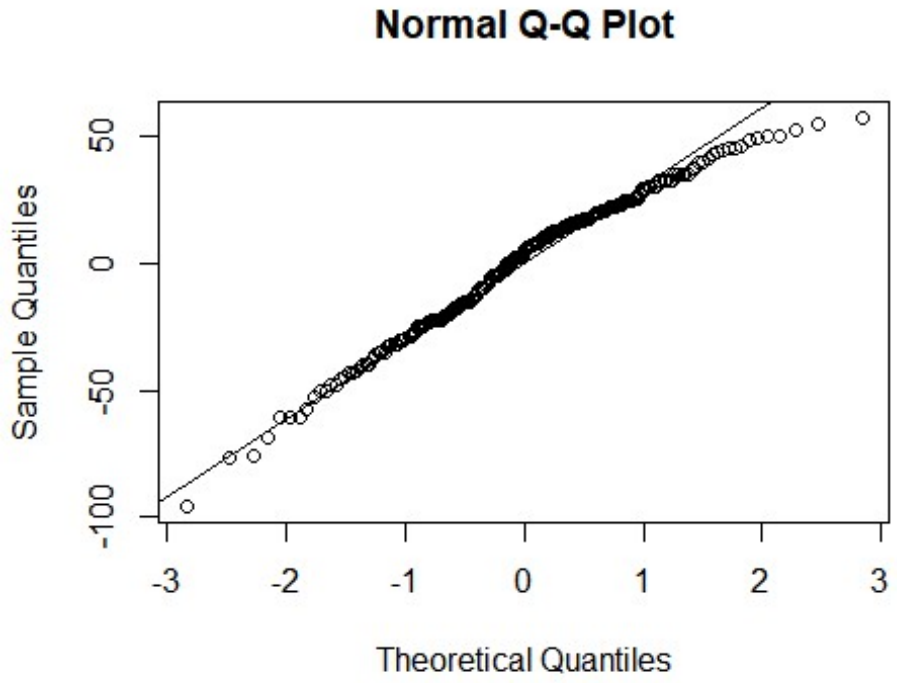
```r
plot(fitted(lm_powerlift_senior), resid(lm_powerlift_senior),
     main = "Fitted vs residuals", xlab = "", ylab = "")
abline(h =0)
```




```r
hist(resid(lm_powerlift_senior), xlab = "", main = "Histogram of Residuals")
```




```r
qqnorm(resid(lm_powerlift_senior))
qqline(resid(lm_powerlift_senior))
```



Summary

- We see that a linear model based on one explanatory variable doesn't explain the data well. The R^2 value is quite low which shows that the fitted model doesn't accuractely predict the values of Senior divisions competitiors bench press best based on their weight.