Week 11 Regression Analysis in R

Jonathan Hernandez

November 3, 2018

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 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.

- Get the data and examine a preview

```
library(data.table)
powerlift <- fread("openpowerlifting.csv")
head(powerlift, n=10)
```

```
##     MeetID                   Name Sex  Equipment Age    Division
##  1:      0       Angie Belk Terry   F      Wraps  47   Mst 45-49
##  2:      0            Dawn Bogart   F Single-ply  42   Mst 40-44
##  3:      0            Dawn Bogart   F Single-ply  42 Open Senior
##  4:      0            Dawn Bogart   F        Raw  42 Open Senior
##  5:      0           Destiny Dula   F        Raw  18  Teen 18-19
##  6:      0        Courtney Norris   F      Wraps  28 Open Senior
##  7:      0          Maureen Clary   F        Raw  60   Mst 60-64
##  8:      0          Maureen Clary   F        Raw  60 Open Senior
##  9:      0 Priscilla Sweat Pardue   F      Wraps  52       50-54
## 10:      0 Priscilla Sweat Pardue   F        Raw  52      Senior
##     BodyweightKg WeightClassKg Squat4Kg BestSquatKg Bench4Kg BestBenchKg
##  1:        59.60            60       NA       47.63       NA       20.41
##  2:        58.51            60       NA      142.88       NA       95.25
##  3:        58.51            60       NA      142.88       NA       95.25
##  4:        58.51            60       NA          NA       NA       95.25
##  5:        63.68          67.5       NA          NA       NA       31.75
##  6:        62.41          67.5   -183.7      170.10       NA       77.11
##  7:        67.31          67.5       NA      124.74       NA       95.25
##  8:        67.31          67.5       NA      124.74       NA       95.25
##  9:        65.95          67.5       NA      120.20       NA       54.43
## 10:        65.95          67.5       NA          NA       NA          NA
##     Deadlift4Kg BestDeadliftKg TotalKg Place  Wilks
##  1:          NA          70.31  138.35     1 155.05
##  2:          NA         163.29  401.42     1 456.38
##  3:          NA         163.29  401.42     1 456.38
##  4:          NA             NA   95.25     1 108.29
##  5:          NA          90.72  122.47     1 130.47
```

```
##  6:          NA         145.15  392.36     1 424.40
##  7:          NA         163.29  383.28     1 391.98
##  8:          NA         163.29  383.28     1 391.98
##  9:          NA         108.86  283.49     1 294.25
## 10:          NA         108.86  108.86     1 112.99
```

- Summary of the data

```
summary(powerlift)
```

```
##      MeetID         Name               Sex             Equipment
##  Min.   :   0   Length:386414      Length:386414      Length:386414
##  1st Qu.:2979   Class :character   Class :character   Class :character
##  Median :5960   Mode  :character   Mode  :character   Mode  :character
##  Mean   :5143
##  3rd Qu.:7175
##  Max.   :8481
##
##       Age           Division          BodyweightKg    WeightClassKg
##  Min.   : 5.00    Length:386414      Min.   : 15.88   Length:386414
##  1st Qu.:22.00    Class :character   1st Qu.: 70.30   Class :character
##  Median :28.00    Mode  :character   Median : 83.20   Mode  :character
##  Mean   :31.67                       Mean   : 86.93
##  3rd Qu.:39.00                       3rd Qu.:100.00
##  Max.   :95.00                       Max.   :242.40
##  NA's   :239267                      NA's   :2402
##     Squat4Kg       BestSquatKg        Bench4Kg       BestBenchKg
##  Min.   :-440.5   Min.   :-477.5   Min.   :-360.0   Min.   :-522.50
##  1st Qu.:  87.5   1st Qu.: 127.5   1st Qu.: -90.0   1st Qu.:  79.38
##  Median : 145.0   Median : 174.6   Median :  90.2   Median : 115.00
##  Mean   : 107.0   Mean   : 176.6   Mean   :  45.7   Mean   : 118.35
##  3rd Qu.: 212.5   3rd Qu.: 217.7   3rd Qu.: 167.5   3rd Qu.: 150.00
##  Max.   : 450.0   Max.   : 573.8   Max.   : 378.8   Max.   : 488.50
##  NA's   :385171   NA's   :88343    NA's   :384452   NA's   :30050
##   Deadlift4Kg     BestDeadliftKg      TotalKg          Place
##  Min.   :-461.0   Min.   :-410.0   Min.   :  11.0   Length:386414
##  1st Qu.: 110.0   1st Qu.: 147.5   1st Qu.: 272.2   Class :character
##  Median : 157.5   Median : 195.0   Median : 424.1   Mode  :character
##  Mean   : 113.6   Mean   : 195.0   Mean   : 424.0
##  3rd Qu.: 220.0   3rd Qu.: 238.1   3rd Qu.: 565.0
##  Max.   : 418.0   Max.   : 460.4   Max.   :1365.3
##  NA's   :383614   NA's   :68567    NA's   :23177
##      Wilks
##  Min.   : 13.73
##  1st Qu.:237.38
##  Median :319.66
##  Mean   :301.08
##  3rd Qu.:379.29
##  Max.   :779.38
##  NA's   :24220
```

- Dimensions of the data and its structure

```r
dim(powerlift)
```

```
## [1] 386414     17
```

```r
str(powerlift)
```

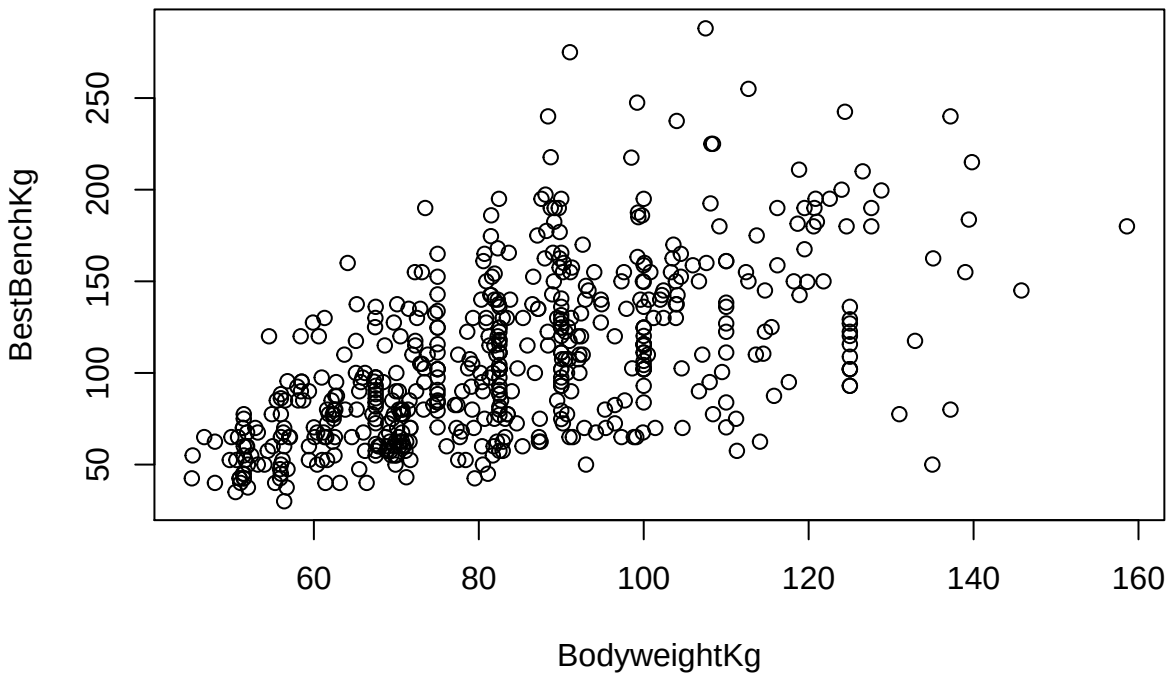
```
## Classes 'data.table' and 'data.frame':   386414 obs. of  17 variables:
##  $ MeetID        : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Name          : chr  "Angie Belk Terry" "Dawn Bogart" "Dawn Bogart" "Dawn Bogart" ...
##  $ Sex           : chr  "F" "F" "F" "F" ...
##  $ Equipment     : chr  "Wraps" "Single-ply" "Single-ply" "Raw" ...
##  $ Age           : num  47 42 42 42 18 28 60 60 52 52 ...
##  $ Division      : chr  "Mst 45-49" "Mst 40-44" "Open Senior" "Open Senior" ...
##  $ BodyweightKg  : num  59.6 58.5 58.5 58.5 63.7 ...
##  $ WeightClassKg : chr  "60" "60" "60" "60" ...
##  $ Squat4Kg      : num  NA NA NA NA NA ...
##  $ BestSquatKg   : num  47.6 142.9 142.9 NA NA ...
##  $ Bench4Kg      : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ BestBenchKg   : num  20.4 95.2 95.2 95.2 31.8 ...
##  $ Deadlift4Kg   : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ BestDeadliftKg: num  70.3 163.3 163.3 NA 90.7 ...
##  $ TotalKg       : num  138.3 401.4 401.4 95.2 122.5 ...
##  $ Place         : chr  "1" "1" "1" "1" ...
##  $ Wilks         : num  155 456 456 108 130 ...
##  - attr(*, ".internal.selfref")=<externalptr>
```

- Grab rows such that the Best Bench Press > 0 (those competitors who were most likley disqualified) and Division as Seniors and extract just the 2 attributes in question

```r
library(dplyr)
powerlift_senior <- powerlift %>% filter(BestBenchKg > 0 & Division %like% 'Senior') %>%
  select(BodyweightKg, BestBenchKg)
```

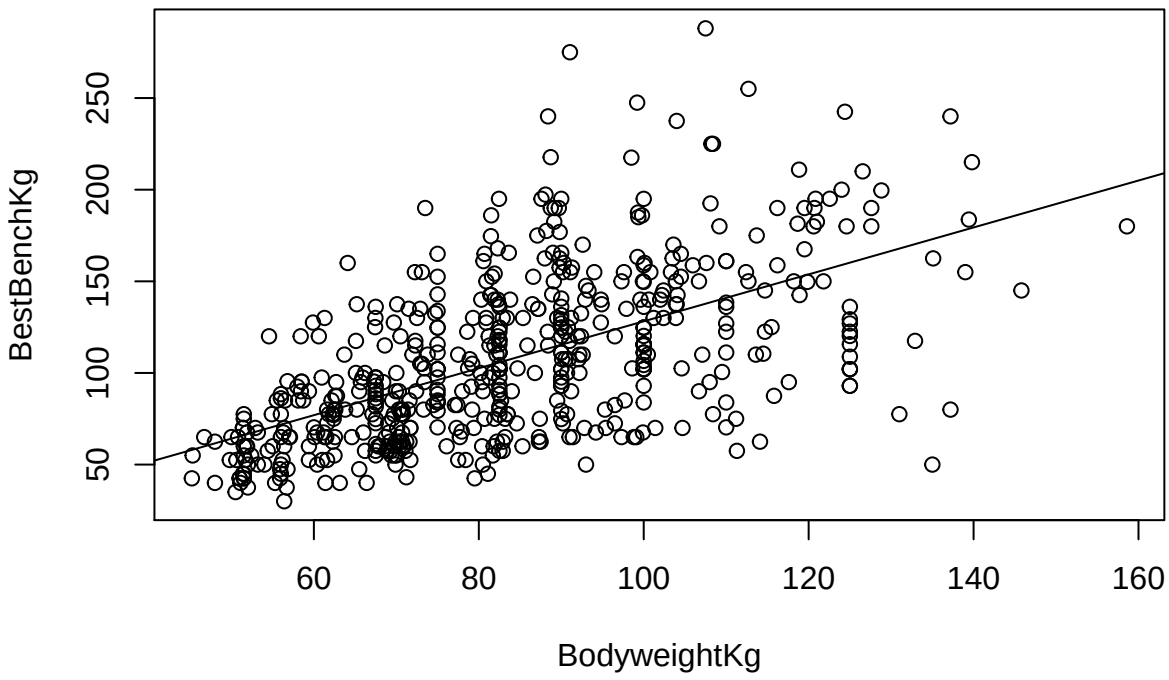
Visualize the data (EDA)

```r
with(powerlift_senior, plot(BodyweightKg, BestBenchKg))
```



Residual Analysis

```
lm_powerlift_senior <- lm(BestBenchKg ~ BodyweightKg, data = powerlift_senior)
with(powerlift_senior, plot(BodyweightKg, BestBenchKg))
abline(lm_powerlift_senior)
```




```
summary(lm_powerlift_senior)
```

```
##
## Call:
## lm(formula = BestBenchKg ~ BodyweightKg, data = powerlift_senior)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -122.985  -24.352   -3.433   20.126  158.286
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.13762    6.42581   0.021    0.983
## BodyweightKg  1.28035    0.07453  17.178   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 36.41 on 553 degrees of freedom
##   (10 observations deleted due to missingness)
## Multiple R-squared:  0.348,  Adjusted R-squared:  0.3468
## F-statistic: 295.1 on 1 and 553 DF,  p-value: < 2.2e-16
```

- Residual plots

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