ECON 421 Project: Some Exploratory Data Analysis

Derek Situ

March 9, 2022

Getting the data

```
# read in a csv with file path "path", keep only necessary columns and rename
#   for readability, add column "Subtotal" which is equal to the sum of
#   BestSquat, BestBench, and max(0, Deadlift1, Deadlift2)
read_results <- function(path) {
  read_csv(path, col_names = TRUE) %>%
    select(Name, Sex, Division,
           WeightClass = "WeightClassKg", Bodyweight = "BodyweightKg",
           Squat1 = "Squat1Kg", Squat2 = "Squat2Kg", Squat3 = "Squat3Kg",
           BestSquat = "Best3SquatKg",
           Bench1 = "Bench1Kg", Bench2 = "Bench2Kg", Bench3 = "Bench3Kg",
           BestBench = "Best3BenchKg",
           Deadlift1 = "Deadlift1Kg", Deadlift2 = "Deadlift2Kg",
           Deadlift3 = "Deadlift3Kg", BestDeadlift = "Best3DeadliftKg",
           Total = "TotalKg", Place) %>%
    mutate(Subtotal = BestSquat + BestBench +
             ifelse(Deadlift1 > 0 | Deadlift2 > 0,
                    ifelse(Deadlift2 > Deadlift1, Deadlift2, Deadlift1),
                    0)) %>%
    relocate(Subtotal, .before = Deadlift3)
}

# read in data
data_21 <- read_results("2021.csv")
data_19 <- read_results("2019.csv")
data_18 <- read_results("2018.csv")
data_17 <- read_results("2017.csv")
data_16 <- read_results("2016.csv")
data_15 <- read_results("2015.csv")
data_14 <- read_results("2014.csv")
data_13 <- read_results("2013.csv")
data_12 <- read_results("2012.csv")
```

What will lifters' beliefs about their opponents' third deadlifts be like? Can we create a regression model to estimate someone's third deadlift as suggested by the TA? Let's check if bodyweight is a good predictor of deadlift capabilities.

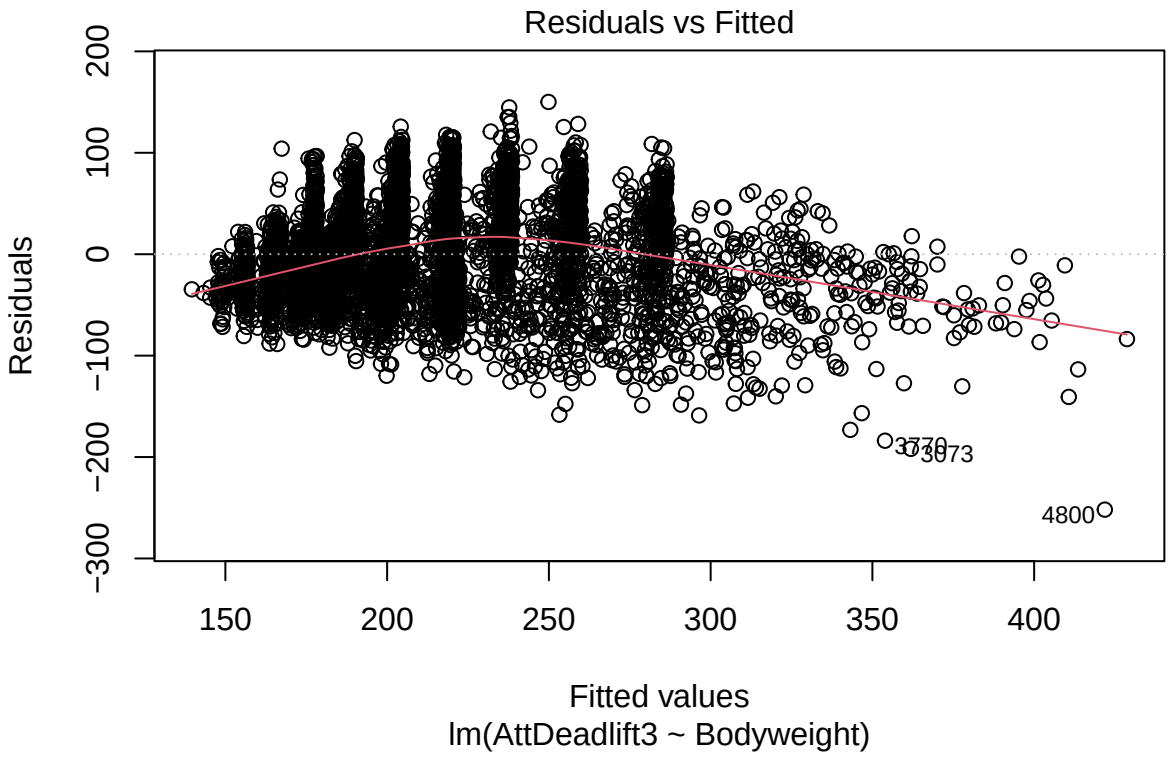
```
# merge all datasets so that we can run regressions with all of the data
all_data <- rbind(data_21, data_19, data_18, data_17, data_16, data_15,
                  data_14, data_13, data_12) %>%
```

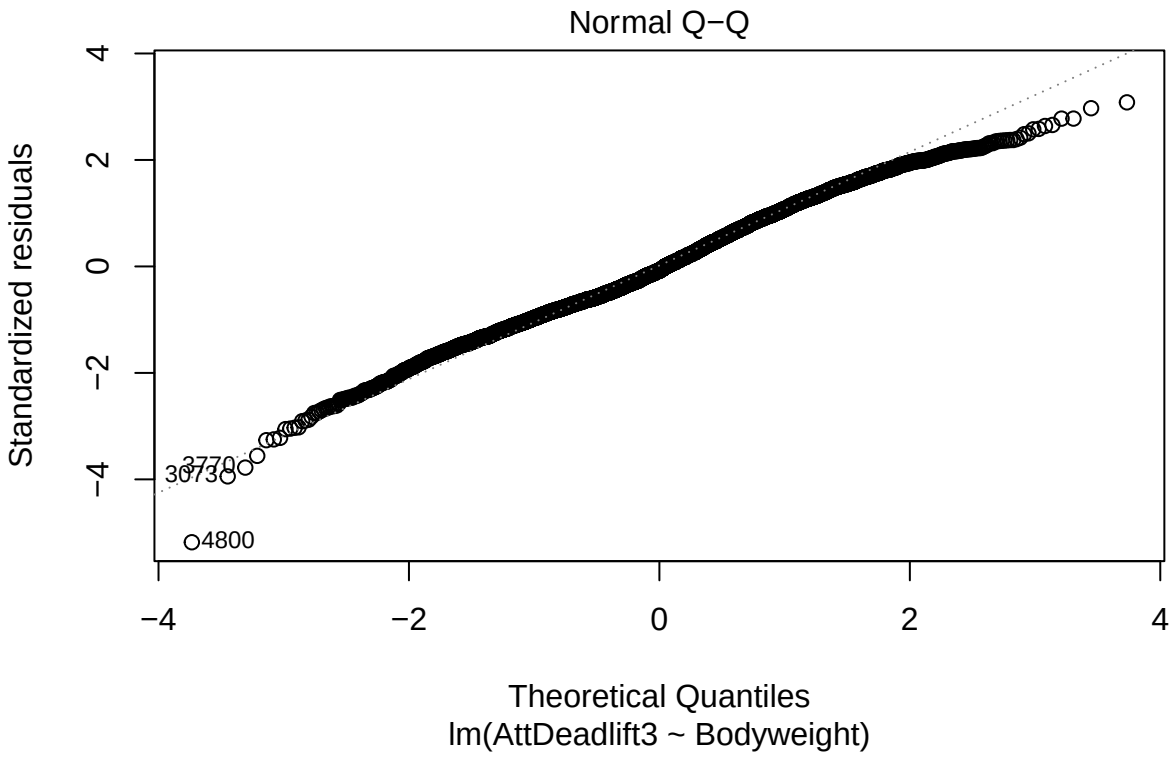
```
  select(Name, WeightClass, Bodyweight, Squat3, Deadlift2, Deadlift3) %>%
  mutate(AttSquat3 = abs(Squat3),
         AttDeadlift2 = abs(Deadlift2),
         AttDeadlift3 = abs(Deadlift3),
         Percent_Increase = (AttDeadlift3 - AttDeadlift2) / AttDeadlift2,
         D3_Success = ifelse(Deadlift3 > 0, 1, 0)) #%>%
  #filter(!(WeightClass %in% c("84+", "105", "120", "120+"))) ############
```

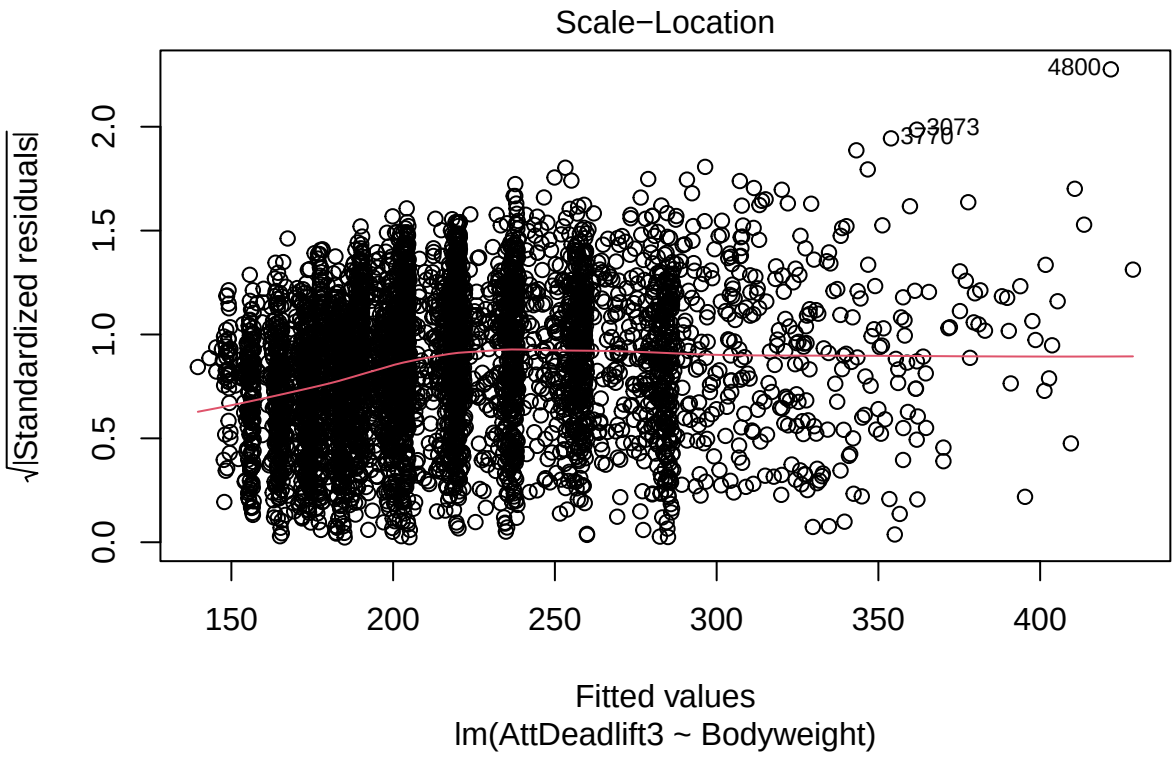
```
lm_fit <- lm(AttDeadlift3 ~ Bodyweight, data = all_data)
summary(lm_fit)
```

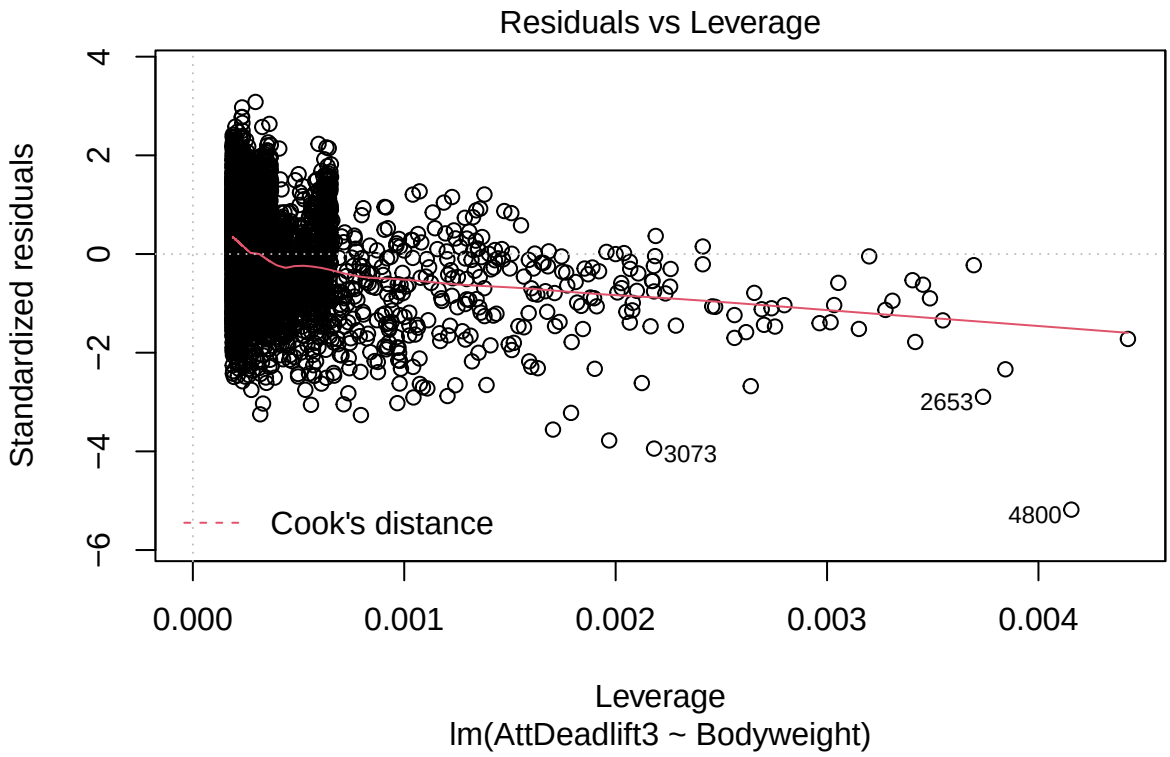
```
##
## Call:
## lm(formula = AttDeadlift3 ~ Bodyweight, data = all_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -251.845  -34.184   -3.733   35.913  150.116
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 73.39340    2.22659   32.96   <2e-16 ***
## Bodyweight   1.77645    0.02647   67.12   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 48.71 on 5301 degrees of freedom
##   (198 observations deleted due to missingness)
## Multiple R-squared:  0.4594, Adjusted R-squared:  0.4593
## F-statistic:  4505 on 1 and 5301 DF,  p-value: < 2.2e-16
```

```
plot(lm_fit)
```


Residuals vs Fitted
Residuals
200
100
0
−100
−200
−300
150
200
250
300
350
400
3770
3073
4800
Fitted values
lm(AttDeadlift3 ~ Bodyweight)

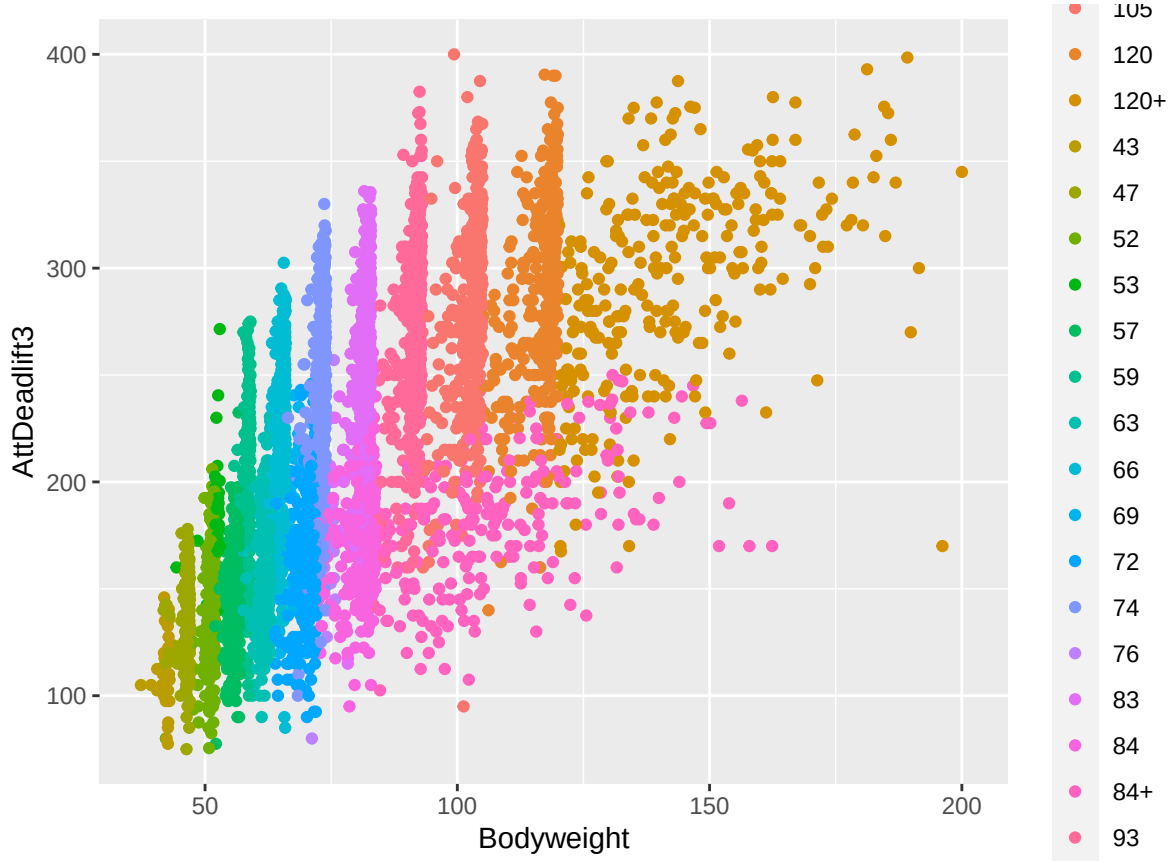

Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(AttDeadlift3 ~ Bodyweight)
3770
3673
4800
−4
−2
0
2
4


Scale−Location
4800
3073
3770
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
150
200
250
300
350
400
Fitted values
lm(AttDeadlift3 ~ Bodyweight)



From the residuals plot, we see that when we estimate Deadlift3 to be 300kg or higher, we are usually overestimating. The next plot shows why:

```
all_data %>%
  ggplot(aes(x = Bodyweight, y = AttDeadlift3, colour = WeightClass)) +
  geom_point()
```



From plotting bodyweight against attempted Deadlift3, we see that there is a linear relationship!... until bodyweight reaches about 100kg. So when we fit our regression line through the data, it fits nicely until we get to about 100kg, and then we tend to overestimate. It appears that after reaching 100kg, there is no benefit in gaining bodyweight in terms of deadlift. This is a well-known and discussed phenomenon in powerlifting because people who weigh 100kg+ tend to be so large that they cannot physically position themselves to take advantage of optimal leverages and physics. This negative trade-off seems to offset the gains from being heavier at a 1-to-1 ratio.

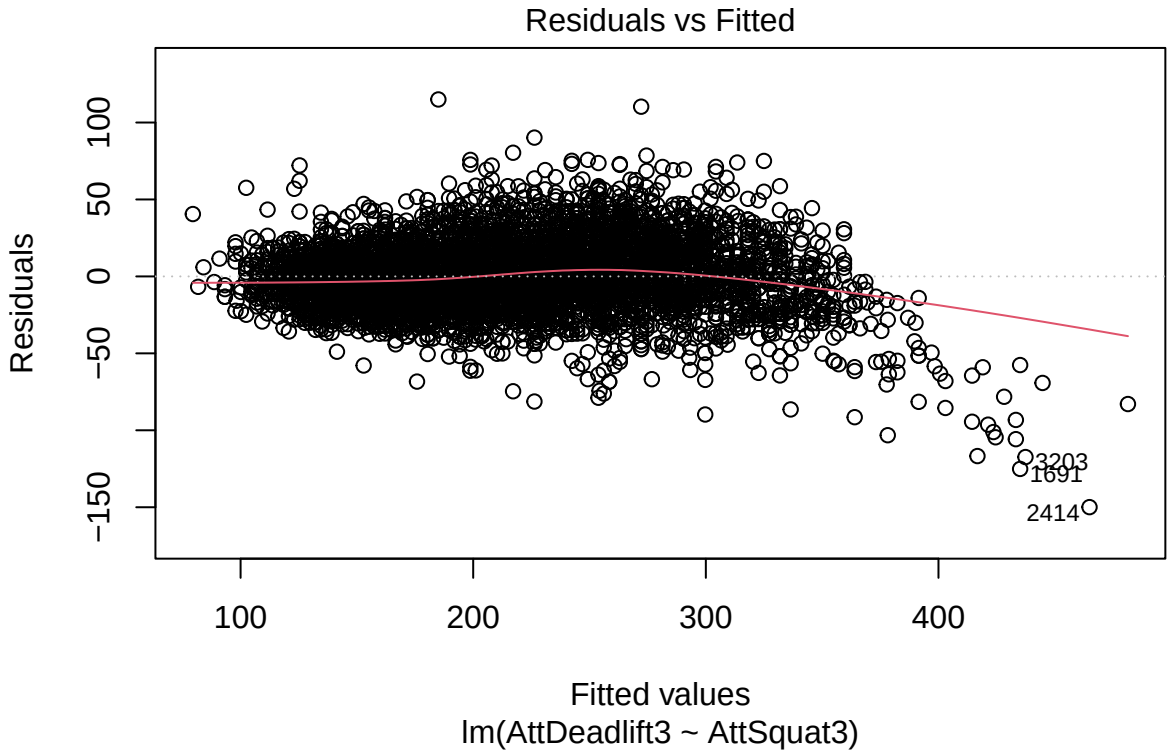
Okay, we can't easily fit a linear model with bodyweight as a regressor. What about using attempted Squat3 to predict attempted Deadlift3? They both heavily involve the leg muscles.

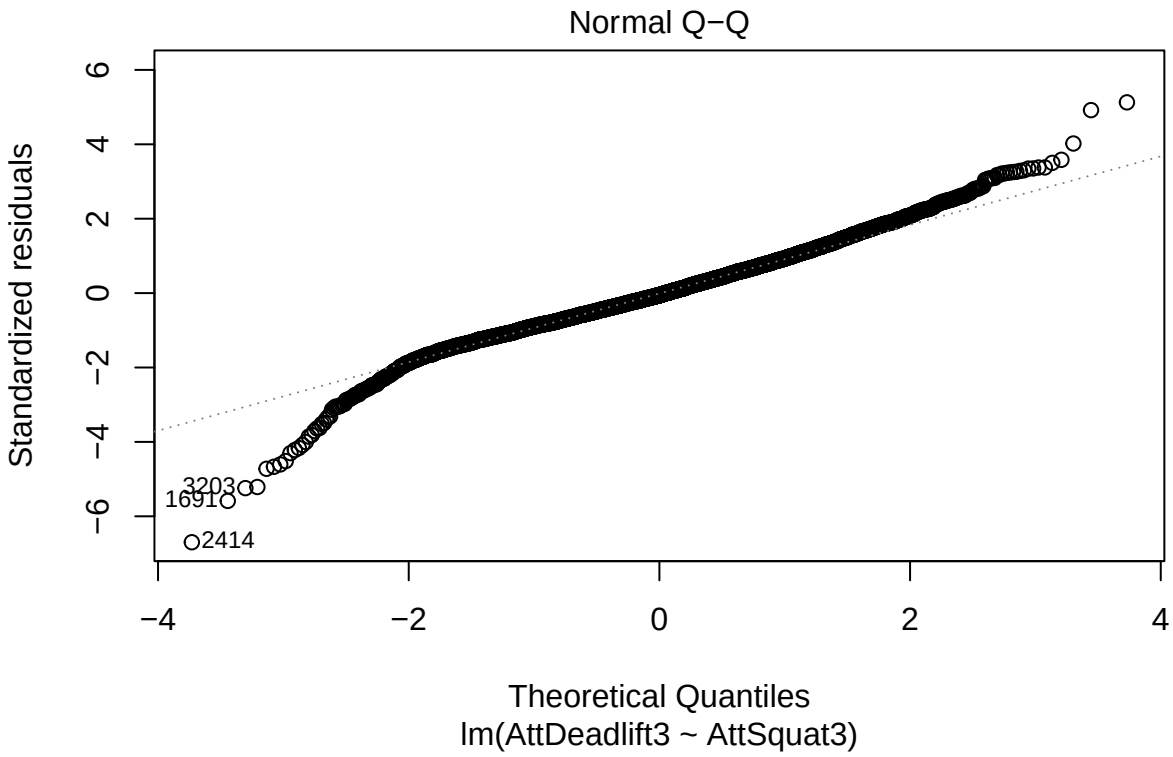
```
lm_fit <- lm(AttDeadlift3 ~ AttSquat3, data = all_data)
summary(lm_fit)
```

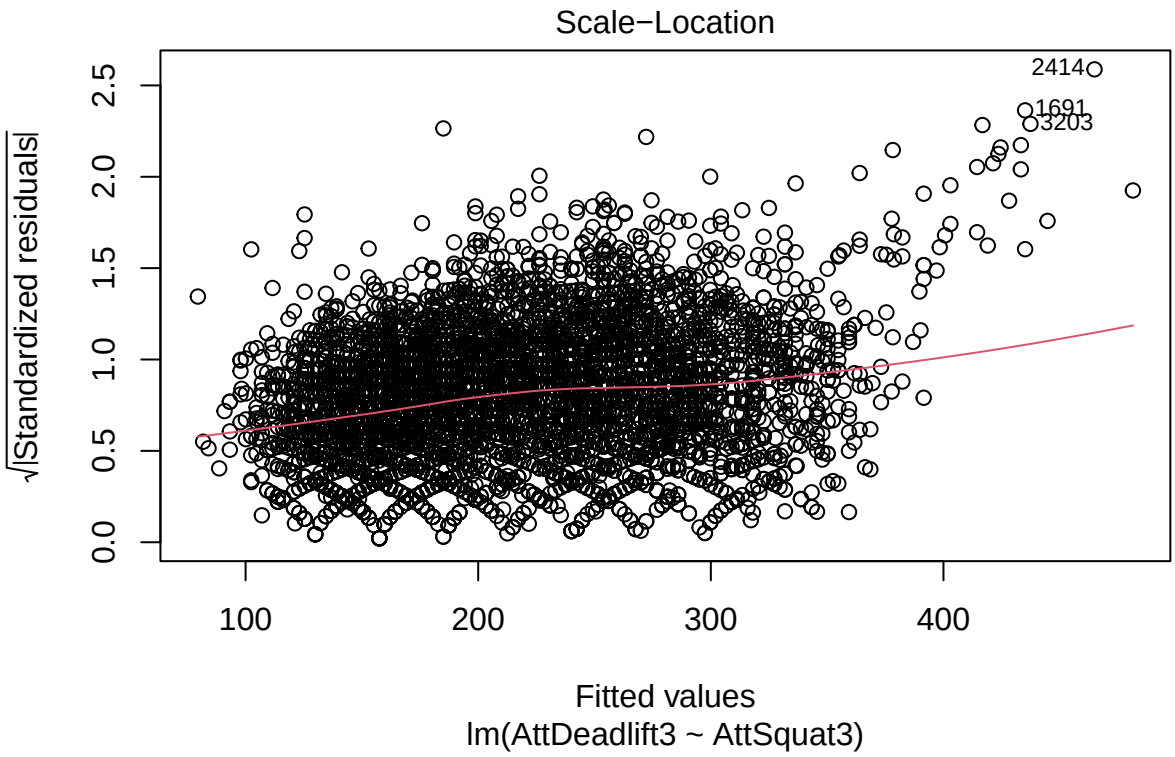
```
##
## Call:
## lm(formula = AttDeadlift3 ~ AttSquat3, data = all_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -149.927  -14.167   -1.255   13.713  114.979
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 42.773957   0.914002    46.8   <2e-16 ***
## AttSquat3    0.917724   0.004569   200.9   <2e-16 ***
```

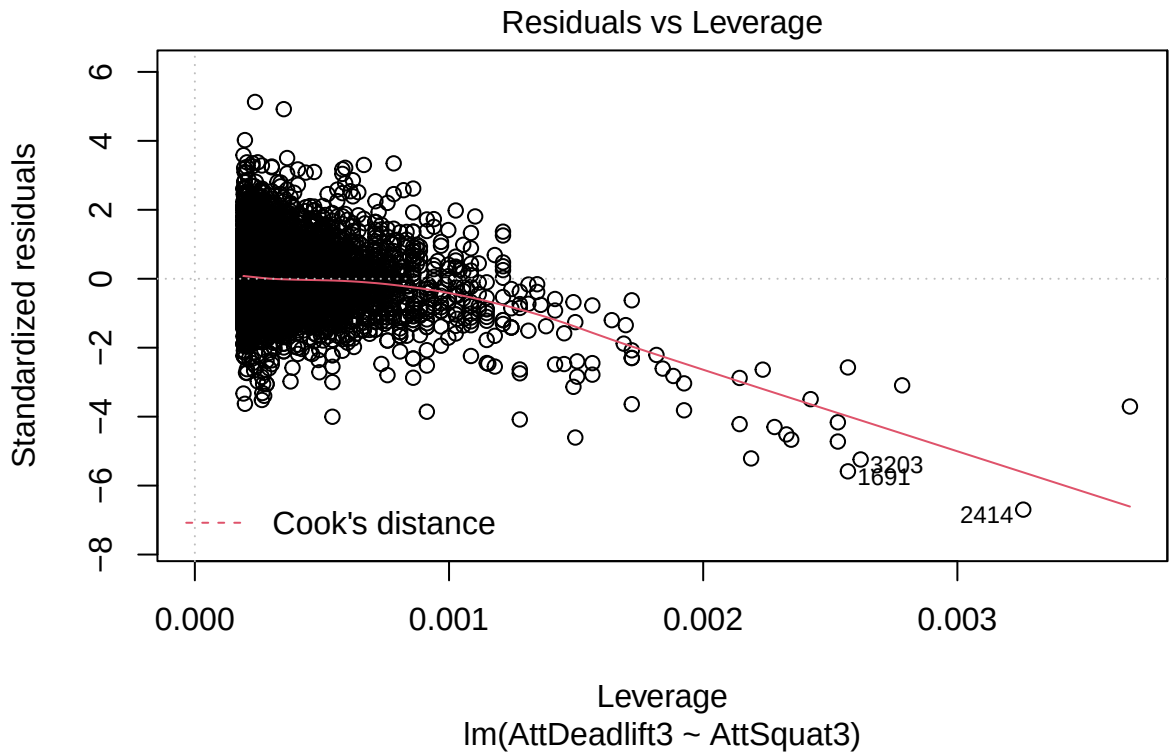
```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.42 on 5228 degrees of freedom
##   (271 observations deleted due to missingness)
## Multiple R-squared:  0.8853, Adjusted R-squared:  0.8853
## F-statistic: 4.034e+04 on 1 and 5228 DF,  p-value: < 2.2e-16
```

```
plot(lm_fit)
```



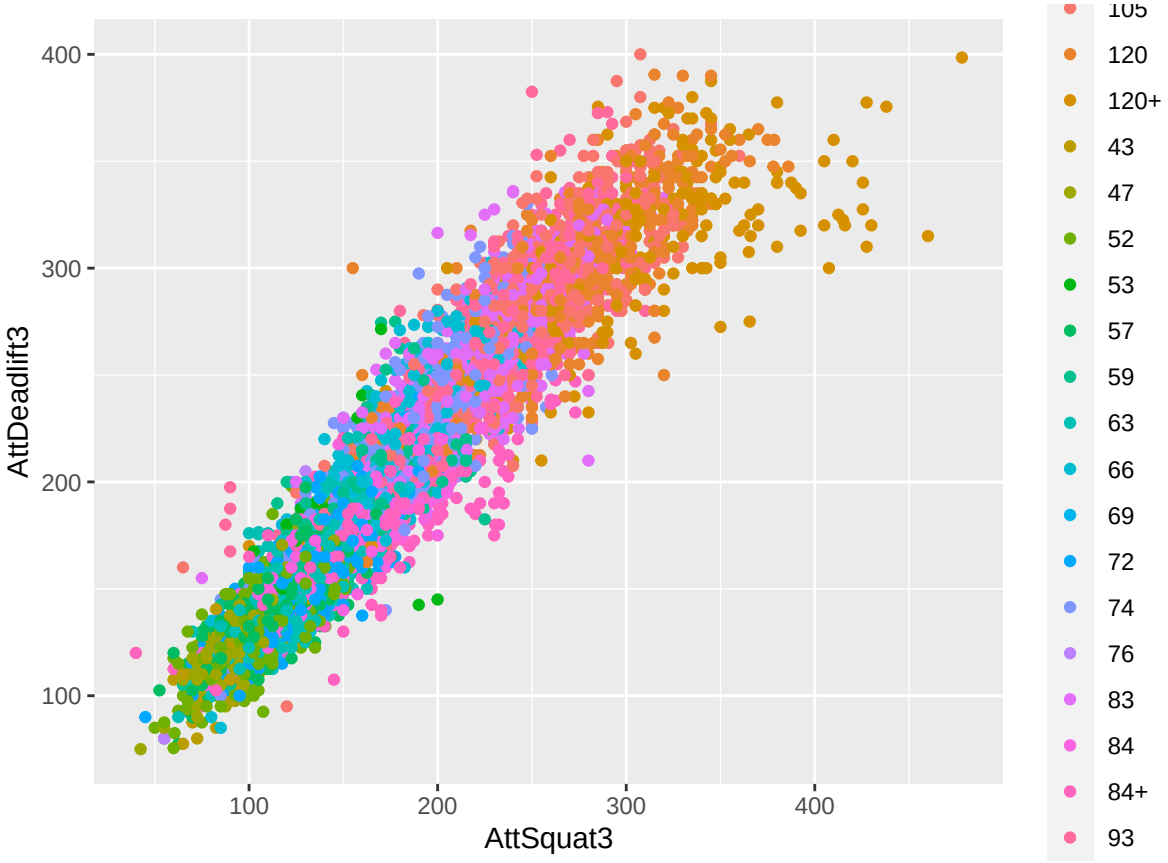

Normal Q−Q
Standardized residuals
6
4
2
0
−2
−4
−6
3203
1691
2414
−4
−2
0
2
4
Theoretical Quantiles
lm(AttDeadlift3 ~ AttSquat3)


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.5
2414
1691
3203
100
200
300
400
Fitted values
lm(AttDeadlift3 ~ AttSquat3)



Again, whenever we predict large values of attempted Deadlift3 (350kg+), we are overestimating. Let's investigate by looking at the plot of attempted Squat3 vs. attempted Deadlift3:

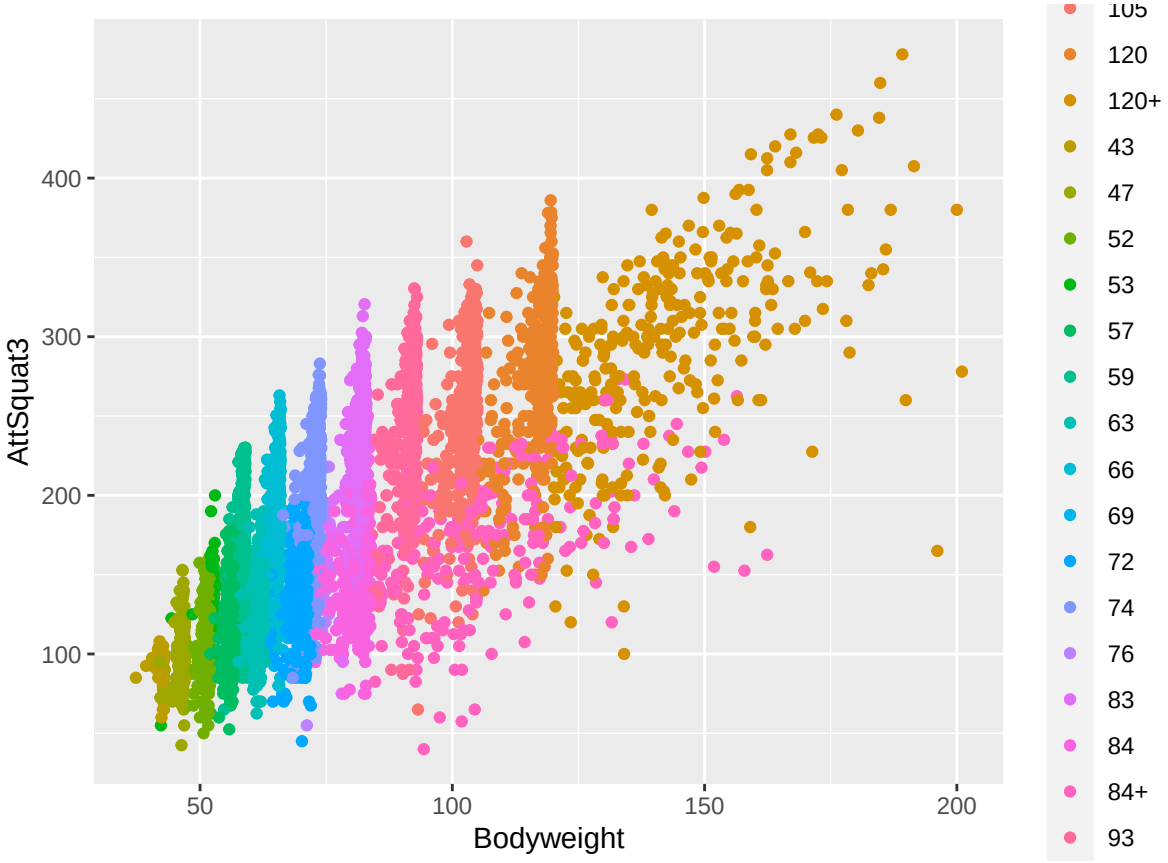
```
all_data %>%
  ggplot(aes(x = AttSquat3, y = AttDeadlift3, colour = WeightClass)) +
  geom_point()
```



Similarly to the bodyweight case, there is a linear relationship between attempted Deadlift3 and attempted Squat3... until Squat3 reaches about 300kg. After reaching a Squat3 of 300kg, there seems to be no correlation between a higher squat and a higher deadlift.

Actually, what we can guess from these two results is that bodyweight is a great predictor of Squat3. Large (100kg+) powerlifters don't have the same problem utilizing optimal leverage in the squat like they do in the deadlift, so squat just increases with bodyweight. Here's a plot to confirm:

```
all_data %>%
  ggplot(aes(x = Bodyweight, y = AttSquat3, colour = WeightClass)) +
  geom_point()
```



So it seems like we can't easily fit a simple linear model using bodyweight nor squat as regressors. We could take a big non-game-theory-related detour to come up with a more sophisticated prediction procedure that maybe uses a non-linear model...

OR, here's a simple way to predict what a lifter is reasonably capable of doing for Deadlift3, which is and can be actually be used by lifters in a real competition setting: Deadlift2 tends to be about 96% of Deadlift3.

```
mean((all_data$AttDeadlift2 / all_data$AttDeadlift3), na.rm = TRUE)
```

```
## [1] 0.9645791
```

This holds up empirically! In other words, Deadlift3 $\approx$ 1.04Deadlift2.

All of this whole analysis is just to say that when we model uncertainty of the success of a lifter's Deadlift3, we can consider the fact that typically, lifters think they are reasonably capable of doing 104% of their Deadlift2 for their Deadlift3.

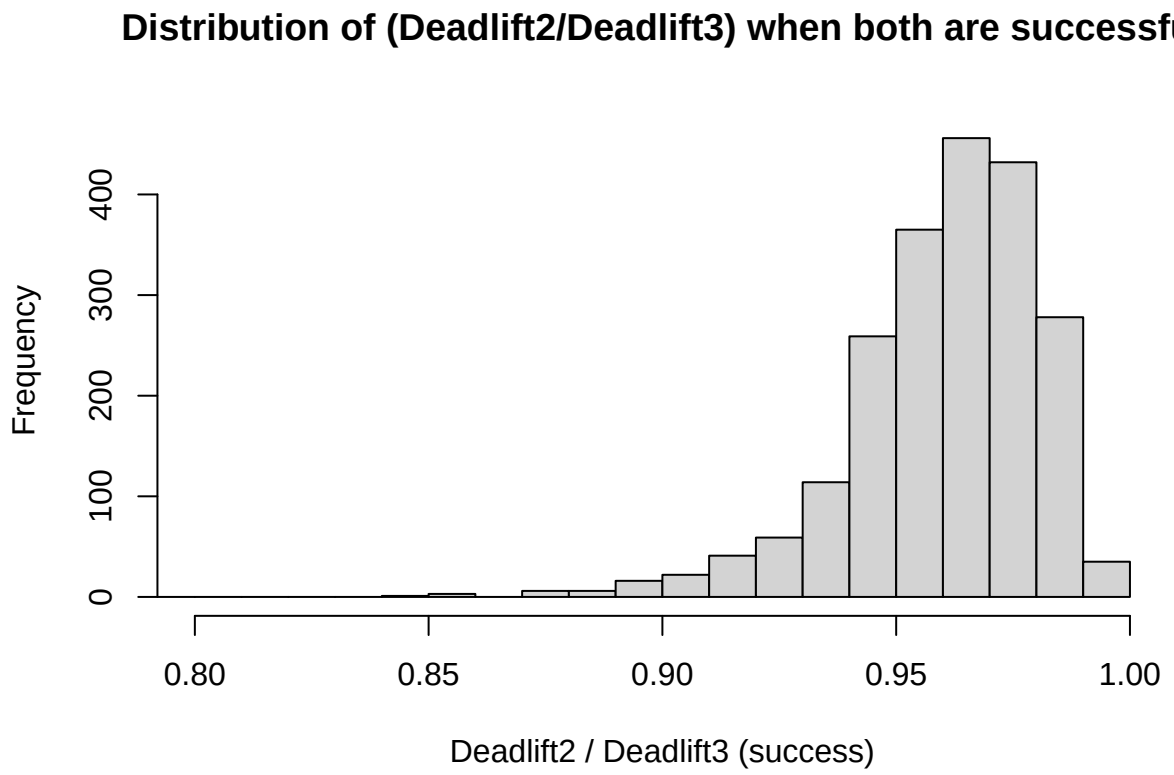
```
# filter for only those with successful Deadlift2
d2_success <- all_data %>%
  filter(Deadlift2 > 0)

# mean and median ratio (deadlift2 / attempted deadlift3)
mean(d2_success$Deadlift2 / d2_success$AttDeadlift3, na.rm = TRUE) # 0.96
```

```
## [1] 0.9583924
```

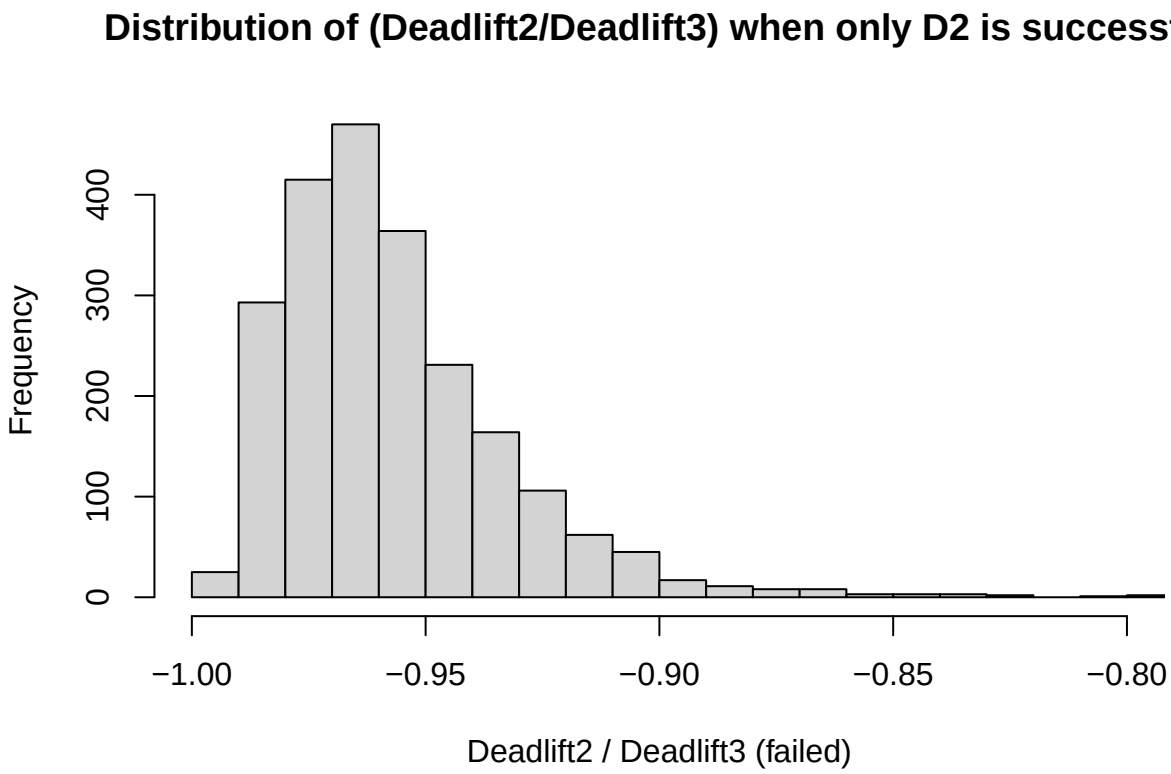
```r
median(d2_success$Deadlift2 / d2_success$AttDeadlift3, na.rm = TRUE) #0.96
```

```
## [1] 0.962963
```

```r
# distribution of (deadlift2 / attempted deadlift3) when deadlift3 is successful
# note: Deadlift3 is equal to -n when weight n is attempted and failed
hist(d2_success$Deadlift2 / d2_success$Deadlift3, breaks = 200,
     main = "Distribution of (Deadlift2/Deadlift3) when both are successful",
     xlab = "Deadlift2 / Deadlift3 (success)",
     xlim = c(0.8, 1))
```




```r
hist(d2_success$Deadlift2 / d2_success$Deadlift3, breaks = 200,
     main = "Distribution of (Deadlift2/Deadlift3) when only D2 is successful",
     xlab = "Deadlift2 / Deadlift3 (failed)",
     xlim = c(-1, -0.8))
```


Distribution of (Deadlift2/Deadlift3) when only D2 is successful
Frequency
0
100
200
300
400
−1.00
−0.95
−0.90
−0.85
−0.80
Deadlift2 / Deadlift3 (failed)