---
title: "Regression"
author: "Aidan Case"
output: html_notebook

Data was taken from [kaggle](https://www.kaggle.com/datasets/dansbecker/powerlifting-database).

Linear regression is a method of generating a linear graph to approximate data in a dataset. It works by taking a bunch of data and finding weights for each of the variables, creating slopes for each. The main strength of linear regression is that it is not computationally intensive. It's main weakness is it's tendency to under or overfit the data depending on what it can represent (eg, it could be a polynomial graph, and linear regression would not be able to represent that).

Load data.

```{r}
ds <- read.csv("openpowerlifting.csv", header=TRUE)[c(1,2,3,5,7,8,10,12,14)]
ds$Sex <- factor(ds$Sex)
# disqualified values are < 0
ds <- ds[ds$BestDeadliftKg > 0,]
ds <- ds[ds$BestBenchKg > 0,]
ds <- ds[ds$BestSquatKg > 0,]
ds <- ds[complete.cases(ds),]
```

Split the data.

```{r}
fds <- ds[ds$Sex == "F",]
fsplit <- sample(1:nrow(fds), .8*nrow(fds), replace=FALSE)
ftrain <- fds[fsplit,]
ftest <- fds[-fsplit,]
mds <- ds[ds$Sex == "M",]
msplit <- sample(1:nrow(mds), .8*nrow(mds), replace=FALSE)
mtrain <- mds[msplit,]
mtest <- mds[-msplit,]
```

Visualize data.

```{r}
head(ds)
tail(ds)
print("Female: ")
summary(fds)
cor(fds[c(4,5,7:9)], use="com")
pairs(fds[c(4,5,7:9)])
print("Male: ")
summary(mds)
cor(mds[c(4,5,7:9)], use="com")
pairs(mds[c(4,5,7:9)])
par(mfrow=c(1,2))
plot(ds$BodyweightKg, ds$BestDeadliftKg,
col=ifelse(ds$Sex=="F","pink","blue"))
plot(ds$Age, ds$BestDeadliftKg, col=ifelse(ds$Sex=="F","pink","blue"))
```

Simple linear regression model.

```{r}
# male model
mlm <- lm(BestDeadliftKg~BestSquatKg,data=mtrain)
```

```
print("\nMale Summary:")
summary(mlm)
par(mfrow=c(2,2))
plot(mlm)
# female model
flm <- lm(BestDeadliftKg~BestSquatKg,data=ftrain)
print("\nFemale Summary:")
summary(flm)
par(mfrow=c(2,2))
plot(flm)
```

Both the female and male residuals follow a very slight curve on the fitted graph, meaning that they correlate pretty well. A strange offshoot on the negative part of the male graph might be possible to attribute to other factors. Normal Q-Q for both has points follows the line fairly well in the middle, but does not represent the extremes properly. Scale-Location seems to indicate no overfitting for each. The leverage indicates that there is isn't any extreme outlier.

Multiple linear regression model.

```{r}
# male model
mmlm <-
lm(BestDeadliftKg~BestSquatKg+BestBenchKg+BodyweightKg+Age,data=mtrain)
print("\nMale Summary:")
summary(mmlm)
par(mfrow=c(2,2))
plot(mmlm)
# female model
fmlm <-
lm(BestDeadliftKg~BestSquatKg+BestBenchKg+BodyweightKg+Age,data=ftrain)
print("\nFemale Summary:")
summary(fmlm)
par(mfrow=c(2,2))
plot(fmlm)
```
Modified multiple linear regression.

```{r}
# male model
mmmlm <- lm(BestDeadliftKg~log(BestSquatKg)+log(BestBenchKg)
+log(BodyweightKg)+Age,data=mtrain)
print("\nMale Summary:")
summary(mmmlm)
par(mfrow=c(2,2))
plot(mmmlm)
# female model
fmmlm <- lm(BestDeadliftKg~log(BestSquatKg)+log(BestBenchKg)
+BodyweightKg+Age,data=ftrain)
print("\nFemale Summary:")
summary(fmmlm)
par(mfrow=c(2,2))
plot(fmmlm)
```

The difference between the two is that the first multi model has less correlation than the modified multimodel because the outliers in the male bodyweight, bench, and squat were so high/low that they messed up the regession. The third model created seems to be a lot better due to the accounting for said outliers, and the first model doesn't account for physical body difference.

Preds:

```{r}
test <- function(lm, test, whoami) {
  pred <- predict(lm, newdata=test)
  corre <- cor(pred, test$BestDeadliftKg)
  mse <- mean((pred-test$BestDeadliftKg)^2)
  rmse <- sqrt(mse)
  print(paste(whoami,":::"))
  print(paste("correlation:", corre))
  print(paste("mse:",mse))
  print(paste("rmse:",rmse))
  print("")
}
# men
test(mlm, mtest, "Men's single parameter")
test(mmlm, mtest, "Men's multi parameter")
test(mmmlm, mtest, "Men's multi modded parameter")
# women
test(flm, ftest, "Women's single parameter")
test(fmlm, ftest, "Women's multi parameter")
test(fmmlm, ftest, "Women's multi modded parameter")
```

This whole difference is because there are very extreme outliers in weight lifting (ie, heavyset lifters), so using more parameters and then restricting those parameters significantly improved the model.