---
title: "Regression"
output:
  pdf_document: default
  html_document: default
date: "2023-02-18"
name: Imad Siddiqui

In this code block I divide the data into 80% train and 20% test. I then performed data exploration on 5 different columns of the dataset.

```{r}
set.seed(1234)
i <- sample(1:nrow(male_teams), nrow(male_teams)*0.8, replace=FALSE)
train <- male_teams[i,]
test <- male_teams[-i,]

mean(train$overall)
median(train$overall)
range(train$overall)
mean(train$attack)
mean(train$defence)

```

I made 2 graphs, one for attack vs overall and one for defense vs overall
```{r}
plot(train$overall~train$attack, xlab="attack", ylab="overall")
abline(lm(train$overall~train$attack), col="red")

plot(train$overall~train$defence, xlab="defence", ylab="overall")
abline(lm(train$overall~train$defence), col="blue")
```

I made a simple linear regression mode, took the summary, and plotted the residuals. The summary tells me that the predictor does at good job. The p value is low, and the minimum and maximum residuals are also pretty low. But, it can still be better by adding in a few more predictors.

I also plotted the residuals in the same code block. The residual vs fitted shows a almost horizontal line with all residuals being even distributed around the line. This is a good thing. The Normal Q-Q graph is also a good indication since the 2 lines intersect through most of the graph. They only leave each other near the sides of the graph. The scale-location graph is also a good indication since all the residuals are spread evenly. The last graph is the only tricky one. There are some outliers shown although they are not significant enough to be removed from the dataset entirely. Most of the residuals have a low leverage with some outliers being common towards areas of greater leverage.
```{r}
lm1 <- lm(overall~attack, data=train)
summary(lm1)
plot(lm1)

```

I made 2 more linear regression models using multiple predictors and compared the 3 models using anova(). The third model is defintely the best since it has a lower RSS than the rest. Although, the first two are also amazing.I believe the third model is the best simply because it has the most predictors and those predictors are directly tied to the prediction.
```{r}
lm2 <- lm(overall~attack+defence, data=train)
lm3 <- lm(overall~attack+defence+midfield, data=train)
```

````
anova(lm1, lm2, lm3)

```
````

In these next three code blocks, I took the correlation, mse, and rmse of the models using the test data. The second one was better than the first and the third one was the best. As expected, the third model performed the best, and the second model performed second best. This is simply because it has more of the predictors that are required to make an accurate prediction. The overall rating is a combination of the ratings of the attack, defense, and midfield players.

````
```{r}
pred1 <- predict(lm1, newdata=test)
cor1 <- cor(pred1, test$overall)
mse1 <- mean((pred1-test$overall)^2)
rmse1 <- sqrt(mse1)

print(paste(cor1))
print(paste(mse1))
print(paste(rmse1))

```

```{r}
pred2 <- predict(lm2, newdata=test)
cor2 <- cor(pred2, test$overall)
mse2 <- mean((pred2-test$overall)^2)
rmse2 <- sqrt(mse2)

print(paste(cor2))
print(paste(mse2))
print(paste(rmse2))

```

```{r}
pred3 <- predict(lm3, newdata=test)
cor3<- cor(pred3, test$overall)
mse3 <- mean((pred3-test$overall)^2)
rmse3 <- sqrt(mse3)

print(paste(cor3))
print(paste(mse3))
print(paste(rmse3))

```
````
````