Linear Regression Report

Tyler Jackson

03/02/2021

# 1) Research Scenario and Question(s).

The Ultimate Fighting Championship (UFC) is an organization that hosts Mixed Martial Arts (MMA) fights at the highest caliber in the world and is largely responsible for the growing popularity of the sport today.  
As a practitioner and fan of the sport I have often wondered what causes certain fighters to win and certain fighters to lose (despite the Las Vegas odds).  
This is a very complex question, therefore I used an approach that takes a look at two explanatory variables and one response variable listed below:

**Explanatory1:** *Total Rounds Fought by the Fighter* (also referred to as *Rounds Fought*)  
**Explanatory2:** *Total Wins via Knock Out (KO) or Technical Knock Out (TKO)* (referred to as *Win.KO*)  
**Response:** *Total Number of Wins by the Fighter* (also referred to as *Total Wins*)

I used these 3 variables to pose the following questions:  
Is there evidence of a linear association between the response variable *Total Wins* and at least one of the two explanatory variables: *Rounds Fought*, *Win.KO*?  
If there is significant evidence that a linear association exists, what are the respective contributions from each variable?

# 2) Data Set

### Data Set Backgroun/Overview

The data set used in this project was accessed via the following link:  
<https://www.kaggle.com/mdabbert/ultimate-ufc-dataset>

Within the data set, there were over 100 columns and originally 4566 rows. The column headers had a nomenclature that assigned a prefix of either “R\_” or “B\_” indicating which corner the fighter was fighting out of (R\_ for red corner and B\_ for blue corner). This was followed by the name of the attribute that the variable was describing. For instance, **B\_total\_rounds\_fought** would denote that the **Blue** corner fighter had a total number of rounds in their UFC career of whichever integer was in that row/column.

This data set has data ranging all the way back to 3/21/2010 and has both female and male fighters from each weight class.

### Data Cleaning

The initial structure of the data set arranged each row to contain both the red and the blue corner fighter information side by side. In addition, there were dates associated with each bout (fight) which effectively made each combination of opponents along with the date of the bout a unique event.  
This is problematic because any fighter could have different *wins*, *Win.KO/TKO* and *total rounds fought* depending on which point in time the observation represents.  
As a result, the following actions were taken.  
First, I created two tibbles, one for the blue corner and one for the red corner.  
Next, I removed the unneeded columns from each tibble. After that, I changed the variable names to remove the prefixes (“B\_” and “R\_”) which left me with the columns “Fighter.Name”,“Rounds.Fought”, “Win.KO”,“Wins”. Once I had common column names, I performed a union on the two tibbles to give a single tibble with those four columns and data from both the original Red and Blue corners.  
The last preparation step taken was grouping the data by Fighter name and summarising the variables by the max of each fighter’s *Rounds.Fought*, *Win.KO/TKO* and *Wins*.  
These preparation steps allowed me to aggregate each fighter’s latest number of *Rounds.Fought*, *Win.KO/TKO* and *Wins* as of February 2021.

### Variable Overview

**Rounds.Fought** - This variable represents number of rounds fought by a fighter in the UFC as of 02/6/21.  
**Win.KO** - This variable represents the total number of Knockouts or Technical Knockouts that each UFC fighter has achieved as of 02/6/21.  
**Wins** - This variable represents number of wins achieved by a fighter in the UFC as of 02/6/21.

### Data Entry Error

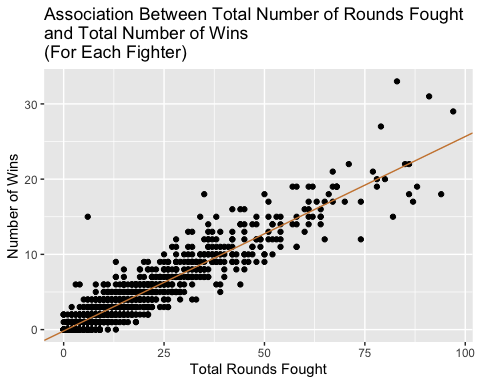
While analyzing the data from the original data source (linked below), I observed an abnormal data point for the *Total Number of Rounds Fought* by Brian Kelleher on 9/5/2020.  
Originally, the data set indicated that he fought **448** total rounds which was at least 5 times more than anyone else in UFC history. I did some investigation by checking other archives of fight data and discovered that the true number of rounds should have been: **18**.  
I fixed this error by updating the csv file directly because there was only one instance of the issue. If there were more, I would have written R code to correct the issue.

# 3) Description of Statistical Methods

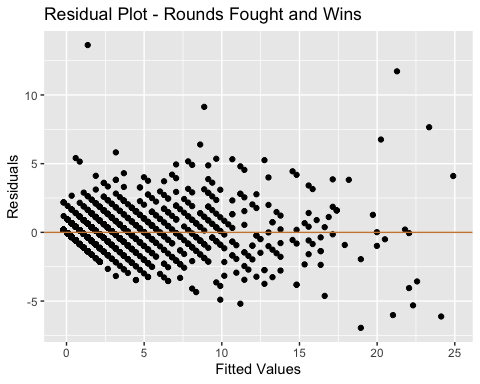
In order to investigate the research question presented in Part 1, I used a few statistical methods. First, I used Simple Linear Regression to analyze the relationship between *Total Rounds Fought* individually with *Total Wins* variable and identify if it has have a linear association. No formal 5-step test will be done for the Simple Linear Regression, only a summary of the p-value and R2 will be reported. Next, I looked at both explanatory variables together and their association with the response using Multiple Linear Regression.  
Finally, I use the F - test for Multiple Linear Regression (Global Test) to establish if **any** of the explanatory variables are associated with the response variable. Given those proved significant, I followed up with t-tests for Multiple Linear Regression to identify **which of the individual** contribution of each explanatory variable in the model.

# 4) Results Summary & Report

## Simple Linear Regression



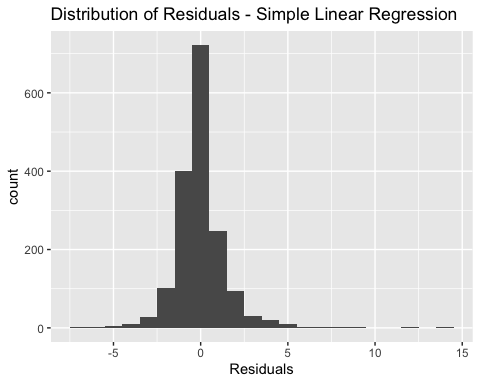
### Assumptions

 In order to make inference using this model, the following assumptions must hold:

**The true relationship is linear.**  
This relationship does appear to be linear when visually inspecting the residual plot above because there are no obvious patterns in the residual values.

**The observations are independent.** The data cleansing activities above ensured that each fighter only has one value and none of them are repeated or over represented in the data.

**The variation of the response variable around the regression line is constant.** This assumption does seem to be questionable in this case because there are less observations where the fighters have fought > 50 rounds however, the number of observations where fighters have fought > 50 times is still more than n=30 therefore, this should not impact the results in such a way that analysis cannot be done.



**The residuals are normally distributed.**  
The histogram above indicates that the Residual values are approximately normally distributed with only a few outliers.

### Simple Linear Regression - Summary

p-value: < 2.2e-16  
The p-value above is so small that it is represented as less than 2.2e-16 and indicates that we can reject H0. H0 states that there is not an association between *rounds fought* and *total wins*, which was rejected at the alpha = 0.05 level.

Multiple R-squared: 0.8868, Adjusted R-squared: 0.8867  
The R-Squared values above indicate the amount of variance that is explained by *rounds fought*. In this case, 88.6% of the variation is explained by the model using *rounds fought* as the explanatory variable.

## Multiple Linear Regression

### Formal Global Test - Multiple Linear Regression

#### 1. Set up the hypotheses and select the alpha level

H0: β1 = β2 = 0 (**Rounds Fought** and **Wins by KO/TKO** are not predictors of **Total Wins**)

H1: β1 ≠ 0 and/or β2 ≠ 0 (At least one of the slope coefficients is different than 0; **Rounds Fought** and/or **Wins by KO/TKO** are predictors/is a predictor of *Total Wins*)

α = 0.05

#### 2. Select the appropriate test statistic

In order to conduct the Global Test, the F test statistic can be used.

F = MS Reg / MS Res

df MS Reg = k = 2

df MS Res = n - k - 1 = 1678 - 2 - 1 = 1675

#### 3. State the decision rule

Decision Rule: Reject H0 if 𝑝 ≤ 𝛼. Otherwise, do not reject H0

#### 4. Compute the test statistic and the associated p-value

F−statistic: 9805 on 2 and 1675 DF

p−value: < 2.2e-16

#### 5. State your conclusion

Reject H0 since 𝑝 ≤ 𝛼 (2.2e-16 < 0.05). We have significant evidence at the 𝛼=0.05 level that *Rounds Fought* and *Wins via KO/TKO*, **when taken together** are predictive of *Total Wins*. In other words, there is evidence of a linear association between *Total Wins* and at least one of the two explanatory variables: *Rounds Fought* and *Wins via KO/TKO* (here, p < 2.2e-16).

### t-tests - Multiple Linear Regression

The following table summarizes the results from conducting t-tests to find p-values and slopes (ßi) for both *Rounds Fought* and *Wins by KO/TKO*.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Estimate | Std. Error | t-value | P-Value |
| Intercept | -0.0970606 | 0.0384286 | -2.525737 | 0.0116371 |
| Rounds Fought | 0.2072848 | 0.0026690 | 77.663181 | 0.0000000 |
| Wins KO/TKO | 0.5557522 | 0.0205028 | 27.106155 | 0.0000000 |

#### Rounds Fought

We **have** significant evidence at the 𝛼=0.05 level that βRounds ≠ 0 **after controlling** for *Wins by KO/TKO*.

The slope βRounds in this scenario represents the predicted change in the *number of wins* given the fighter has fought one additional *round*, after controlling for the other explanatory variable, *Wins KO/TKO*. Specifically, for every 1 additional *round fought* by a fighter, the predicted *wins* would be 0.2072848 wins higher, after controlling for the other independent variable listed above. This could be multiplied to be more meaningful to interpret a 5 *round* increase would mean the predicted increase in *wins* would be 1.036424 higher (about 1 more predicted win for 5 more rounds fought). Additionally, the slope is positive therefore, the association between *rounds fought* and predicted *wins* is a positive association.

#### Wins by KO/TKO

We **have** significant evidence at the 𝛼=0.05 level that βKO ≠ 0 **after controlling** for *Rounds Fought*.

The slope βKO in this scenario represents the predicted change in the *number of wins* given the fighter has one additional *win by knock out*, after controlling for the other explanatory variable, *Rounds Fought*. Specifically, for every 1 additional *KO victory* by a fighter, the predicted *wins* would be 0.5557522 wins higher, after controlling for the other independent variable listed above. Additionally, the slope is positive therefore, the association between *KO Victories* and predicted *wins* is a positive association.

### Assessing the Fit of the Multiple Linear Regression Model (R2)

Multiple R-squared: 0.9213  
Adjusted R-squared: 0.9212

Both the Multiple R-squared and Adjusted R-squared are ~0.92. The R2 value of ~0.92 or ~92% represents the proportion (percentage) of the variation in the *predicted wins* (response variable) explained by the multiple regression model (*Rounds Fought and KO/TKO Victories*). Because the R2 value was relatively high, it would be reasonable to say the regression fit well.

# 5) Conclusions & Limitations.

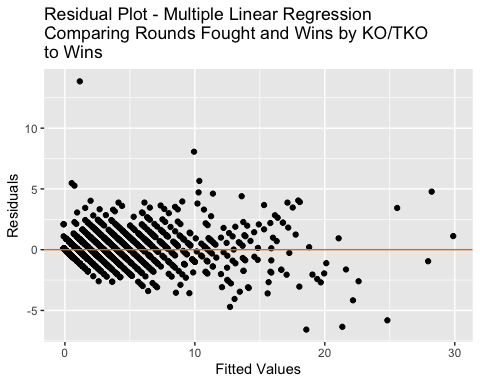
Overall, the simple linear regression model and the Multiple Linear Regression model had R2 values of 88.7% and 92% respectively. In other words, if the R2 value was 100%, then every point in the models would fall perfectly on the regression line. When comparing that to the two models, 88.7% and 92% of the variability was explained by the model and only 11.3% and 8% was not explainable by the two explanatory variables used.

Given the two values above, it appears that the Multiple Linear Regression Model fit the data better than the Simple Linear regression Model.

In order to interpret the MLR model, the following assumptions needed to be verified:

1)The true relationship is linear.  
2)The observations are independent.  
3)The variation of the response variable around the regression line is constant.  
4)The residuals are normally distributed.

Using the following two figures, the 4 assumptions above were investigated.



#### The true relationship is linear.

Using the residual plot above, I looked for any distinguishing patterns that may suggest that the relationship was non-linear but, I could not find any.

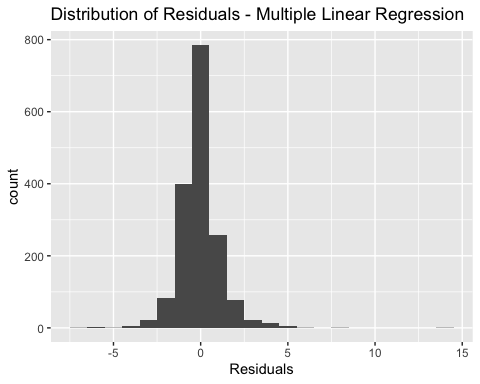
#### The observations are independent.

The steps described in the Data Cleaning section explain that each of the observations in the data set used in this model are independent of one another. Specifically, because each line represents an individual fighter and each duplicate was effectively removed because we took the max of each variable per fighter.

#### The variation of the response variable around the regression line is constant.

The variation of the MLR model has some what constant variance throughout the domain of the fitted values. There are a few outliers in the x-direction that may be causing inconsistent variation when the fitted values are greater than 20 but, overall, this assumption is satisfied.

#### The residuals are normally distributed.



Using the histogram above, the residuals are approximately normally distributed.

### Interpretation of the Slope (non-statistic explanation)

Given the assumptions above are true, next, the interpretation of the MLR model can be used to predict the number of *wins* given *Rounds Fought* and *Victories via KO/TKO*.

For example, if a fighter has fought 20 rounds in the UFC and has 2 wins by KO or TKO, then the following formula would be used to predict the number of wins expected for that fighter.

-0.0970606 + 0.2072848 \* (20) + 0.5557522 \* (2) = 5.16 wins (rounded to 5 wins)

### P-value (non-statistic explanation)

The p-values of each variable, *Rounds Fought* and *Wins by KO/TKO* were <2e-16. That means that if the probability that there was an association between *Rounds Fought* & *Wins by KO/TKO* and *Total FIghter Wins* could have been observed by chance is less than 0.00000000001% likely, therefore we can reject the idea that there is not an association between *Rounds Fought* & *Wins by KO/TKO* and *Total FIghter Wins* and accept the idea that there **is** a linear relationship between them.

### Limitations of Model

There are some limitations of the model worth noting. For example, as the *Rounds Fought* grew past 50 rounds, there were less fighters that have had that many rounds in the UFC, which would indicate that in a perfect scenario, I should gather more data from fighters with > 50 rounds of experience.

Another limitation to this data set were some of the outliers represented fighters that have fought lower amounts of rounds but have won most of them. This could have been due to the lack of experience of their opponents in their early fights and would not be representative to the fights later in their career when they reach the top 10 opponents. Although this is a potential limitation, there was such a large amount of data that the model was minimally affected by these points (R^2 1% change).

## Appendix: R Code

knitr::opts\_chunk$set(echo = FALSE, include = FALSE)  
# Won't show the code in the document  
library(ggplot2)  
library(knitr) # Need knitr for kable and the appendix  
library(plyr)  
library(car)  
library(lsmeans)  
library(tibble)  
library(tidyverse)  
library(aod)  
library(pROC)  
options(contrasts=c("contr.treatment","contr.poly"))  
  
#Import CSV File  
setwd("~/Documents/Masters Degree/Data Analytics and Visualization/Assignments/Final Project")  
data <- read.csv("DataSetTermProject.csv")  
as\_tibble(data)  
#(2) Select Appropriate Columns   
##Select Each Corner's Tibble  
data.b<-select(data, starts\_with("B\_"))  
data.r<-select(data, starts\_with("R\_"))  
##Remove unneeded columns  
data.b<-data.b[,c(1,14,19,22)]  
data.r<-data.r[,c(1,14,19,22)]  
##Rename columns to drop "R\_" and "B\_"  
names(data.b)<-c("Fighter.Name","Rounds.Fought", "Win.KO","Wins")  
names(data.r)<-c("Fighter.Name","Rounds.Fought", "Win.KO","Wins")  
##Union the two tibbles  
data.b.r<-union(data.r,data.b)  
##Group by Fighter  
data.final<-data.b.r %>%   
 group\_by(Fighter.Name) %>%  
 summarise(  
 Rounds.Fought = max(Rounds.Fought),  
 Win.KO = max(Win.KO),  
 Wins = max(Wins))  
##Removed Fighter's Name  
data.final<-data.final %>%  
 select(!Fighter.Name)  
  
#(4) Generate a Scatterplot - Rounds Fought  
model.slinr.roundsFought <- lm(data.final$Wins~data.final$Rounds.Fought)  
intercept.roundsFought<-model.slinr.roundsFought$coefficients[[1]]  
slope.roundsFought<-model.slinr.roundsFought$coefficients[[2]]  
ggplot(data = data.final, mapping = aes(x=Rounds.Fought, y=Wins)) +  
 geom\_point() +   
 ggtitle("Association Between Total Number of Rounds Fought\nand Total Number of Wins\n(For Each Fighter)") +   
 xlab("Total Rounds Fought") +  
 ylab("Number of Wins") +  
 geom\_abline(slope = slope.roundsFought, intercept = intercept.roundsFought , col = "tan3")  
#(4) Generate Residual Plot - Rounds Fought  
ggplot(data = model.slinr.roundsFought, mapping = aes(x=fitted(model.slinr.roundsFought), y=resid(model.slinr.roundsFought))) +  
 geom\_point() +   
 ggtitle("Residual Plot - Rounds Fought and Wins") +   
 xlab("Fitted Values") +  
 ylab("Residuals") +  
 geom\_abline(intercept = 0, slope = 0 , col = "tan3")  
ggplot(data = model.slinr.roundsFought, mapping = aes(x=resid(model.slinr.roundsFought))) +  
 geom\_histogram(binwidth = 1) +   
 ggtitle("Distribution of Residuals - Simple Linear Regression") +   
 xlab("Residuals")   
#(4)  
summary(model.slinr.roundsFought)  
#(4) Generate a Scatterplot - Rounds Fought & Win by KO/TKO  
model.mlinr <- lm(data.final$Wins~data.final$Rounds.Fought+data.final$Win.KO)  
summary(model.mlinr)  
mlr.table<-cbind(summary(model.mlinr)[[4]][1:3],  
 summary(model.mlinr)[[4]][4:6],  
 summary(model.mlinr)[[4]][7:9],  
 summary(model.mlinr)[[4]][10:12])  
colnames(mlr.table)<-c("Estimate", "Std. Error", "t-value", "P-Value")  
rownames(mlr.table)<-c("Intercept","Rounds Fought", "Wins KO/TKO")  
kable(mlr.table)  
data.outlier.rm<-data.final[!(data.final$Wins==15 & data.final$Rounds.Fought==6),]  
model.slinr.outlier.rm <- lm(data.outlier.rm$Wins~data.outlier.rm$Rounds.Fought)  
summary(model.slinr.outlier.rm)  
  
#(4) Generate Residual Plot - MLR (Rounds Fought and Wins via KO/TKO)  
ggplot(data = model.mlinr, mapping = aes(x=fitted(model.mlinr), y=resid(model.mlinr))) +  
 geom\_point() +   
 ggtitle("Residual Plot - Multiple Linear Regression\nComparing Rounds Fought and Wins by KO/TKO\nto Wins") +   
 xlab("Fitted Values") +  
 ylab("Residuals") +  
 geom\_abline(intercept = 0, slope = 0 , col = "tan3")  
ggplot(data = model.mlinr, mapping = aes(x=resid(model.mlinr))) +  
 geom\_histogram(binwidth = 1) +   
 ggtitle("Distribution of Residuals - Multiple Linear Regression") +   
 xlab("Residuals")  
summary(model.mlinr)  
write.csv(data.final,file = "DataSetAfterCleaning.csv")