# 

# 

# 

AMS 572 Data Analysis I

Group 8 Project

Members:

Stephanie Taboada

Christina Hartnett

Michael LaBarbera

Josh Foote

Ronak Shah

Introduction

Data providing detailed attributes for players registered in the latest edition of FIFA 19 database were utilized for this project. The data was downloaded from kaggle. The detailed attributes include the following : name, age, nationality, zone, club, value, wage, preferred foot, position on the pitch, height of the player, weight of the player. The dataset has more than 50 samples and more than 10 variables. It also has both categorical and continuous variables. Wages, for example, is a continuous variable.

## Objectives

The objective of this project is to investigate the relationship between many different attributes within FIFA soccer. First, we will test to see if the average wages of Manchester United Football Club differ from the average wages of the rest of the soccer clubs. Then on a smaller scale, we will investigate if there is a significant difference between the means of the wages of Manchester United Football Club and the Manchester City Football Club. We will then move on to testing the relationship of the position the individual plays and the player’s zone of origin. In this case we are testing whether the zone of origin has an effect on the position they play. We then used an ANOVA test to check if the mean age for at least one club differed significantly from the other clubs. We looked at 4 different clubs for this test. We then want to explore the relationship between wage of the individual and the zone of origin and wage of the individual preferred foot. From this dataset, we also removed a select portion data points in order to examine the effects of having missing data on our test.

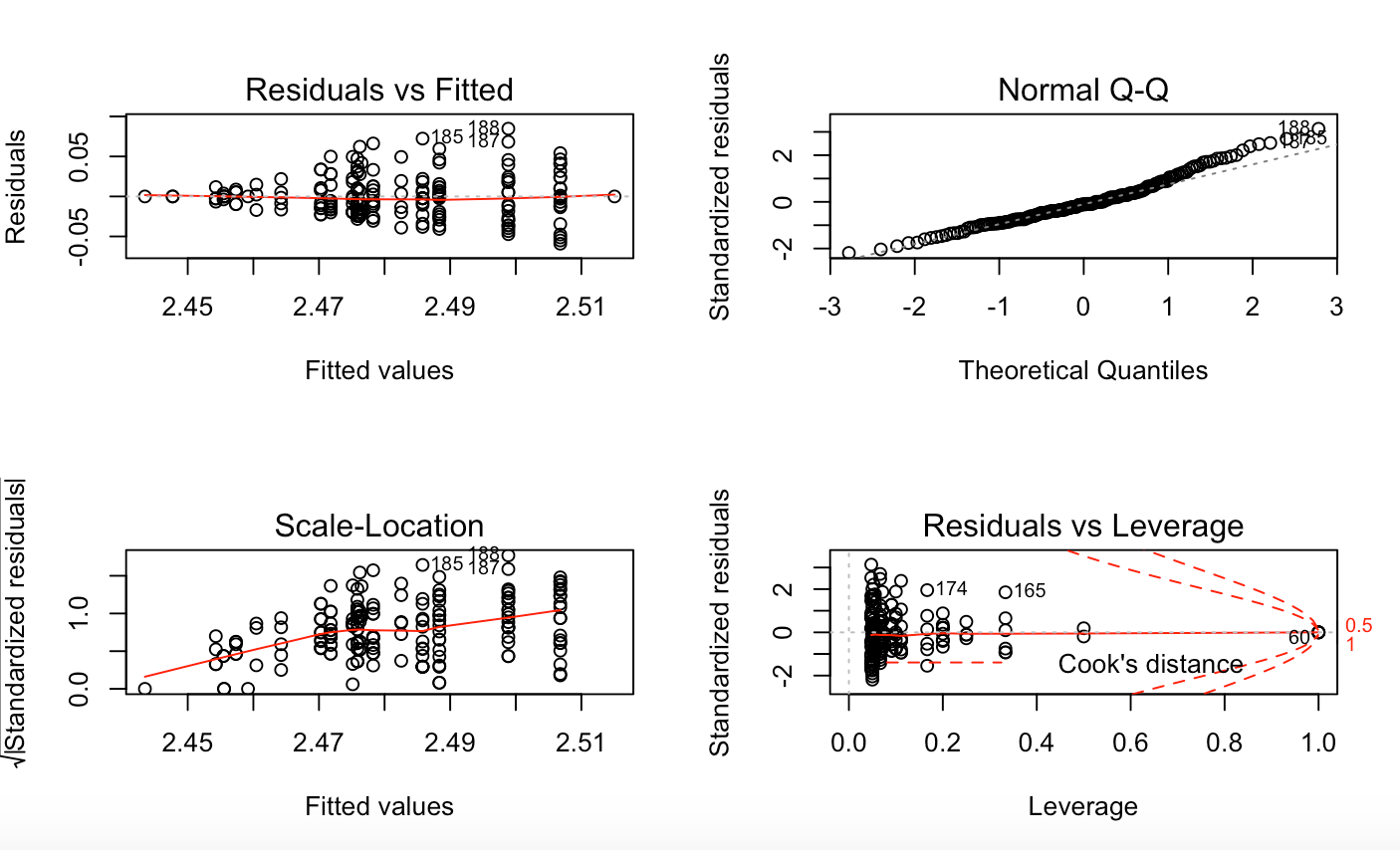
## Independent Two Sample T-test

### Introduction

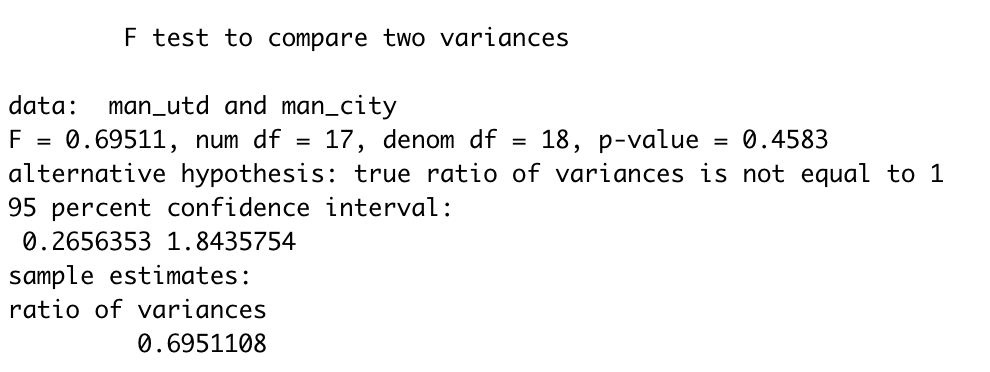
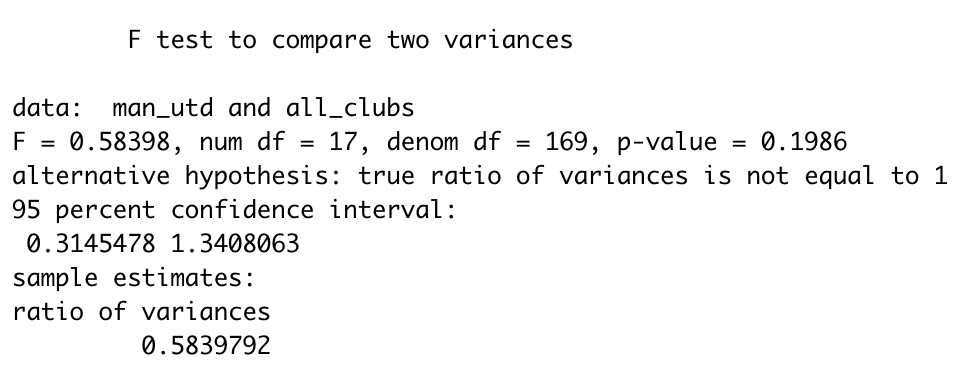
An independent two-sample t-test is used to determine if there is a significant difference between the means of two unrelated groups of sample. For this test, we analyzed whether there is a difference in the mean wages between the Manchester United Football Club and the rest of the clubs. We then studied whether there is a difference in the mean wages between Manchester United Football Club and Manchester City Football Club.

In order to use a t-test the following is assumed of the data we used:[[1]](#footnote-1)

* The data is continuous, not discrete. The data utilized for this project is continuous.
* The two samples are independent, meaning there is no relationship between the individuals in one sample as compared to the other. There is no relationship between the ages of the strikers and the goalkeepers.
* Both samples are randomly selected from their respective population, meaning each individual has an equal probability of being selected in the sample. Both samples for this t-test were randomly selected.
* When plotted the results show a normal distribution, bell-shaped distribution curve.



* The variances of the two populations are equal.



### Hypothesis

We conducted the two-sided t-tests that analyzed the following:

mean wage of the Manchester United Football Club versus the mean wage of the rest of the clubs are the same.

= mean wage of the Manchester United Football Club versus the mean wage of the rest of the clubs are not the same

*And*

mean wage of the Manchester United Football Club versus the mean wage of the Manchester City Football Club are the same.

= mean wage of the Manchester United Football Club versus the mean wage of Manchester City Football Club are not the same

### Equations

The Test Statistic for the Independent Two Sample T-test is the following:[[2]](#footnote-2)

Where and are the samples sizes, and are the sample means and and are the sample variances. If equal variances are assumed, then the formula reduces to the following, where is the pooled variance.

Reject the null hypothesis, if lTl >where v is equivalent to the following:

### 

### R Code

#### Check for normal distribution for wage

qqnorm(data$Wage1)

par(mfrow=c(2,2))

###plot for normal QQ looks fairly straight so assume normality is met

plot(lm(data$Wage1~data$Club))

#### Mean wage of Man. Utd vs Man. City using T-test

man\_utd <- data$Wage1[data$Club == "Manchester United"]

man\_city <- data$Wage1[data$Club == "Manchester City"]

#Equal Variances bc P-value is greater than alpha level

var.test(man\_utd,man\_city)

##fail to reject null hypothesis

t.test(x=man\_utd, y=man\_city,alternative = "greater", conf.level = 0.1)

#### Mean wage of Man. Utd vs all using T-test

man\_utd <- data$Wage1[data$Club == "Manchester United"]

all\_clubs <- data$Wage1[data$Club != "Manchester United"]

#Equal Variances bc P-value is greater than alpha level

var.test(man\_utd,all\_clubs)

##fail to reject null hypothesis

t.test(x=man\_utd, y=all\_clubs,alternative = "greater", conf.level = 0.1)

##Extra information about the data

mean(data$Wage[data$Club == "Manchester United"])

sd(data$Wage[data$Club == "Manchester United"])

min(data$Wage[data$Club == "Manchester United"])

max(data$Wage[data$Club == "Manchester United"])

mean(data$Wage[data$Club == "Manchester City"])

sd(data$Wage[data$Club == "Manchester City"])

min(data$Wage[data$Club == "Manchester City"])

max(data$Wage[data$Club == "Manchester City"])

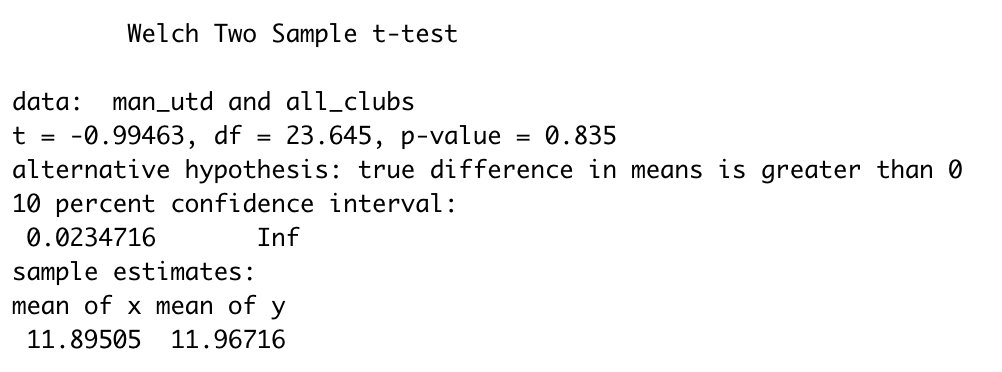
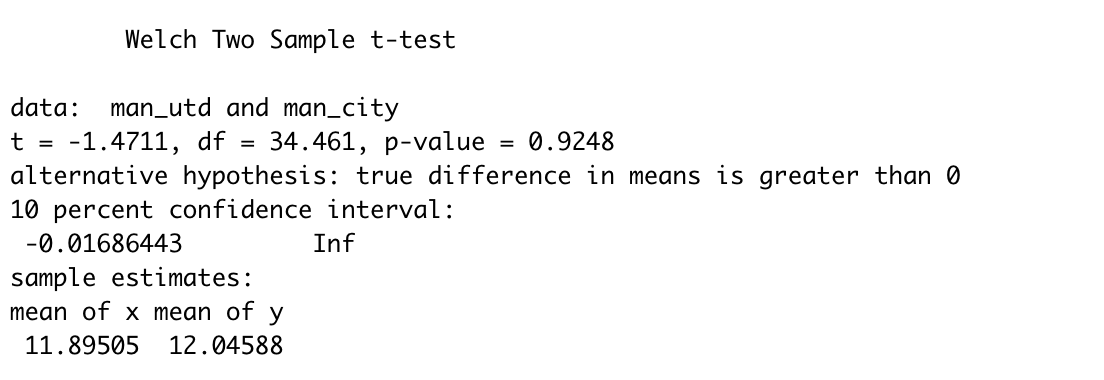
mean(data$Wage[data$Club != "Manchester United"])

sd(data$Wage[data$Club != "Manchester United"])

min(data$Wage[data$Club != "Manchester United"])

max(data$Wage[data$Club != "Manchester United"])

### Results



### Conclusion

For the first T-test at alpha level .1, (H0: μMU = μMC, Ha: μMU> μMC.), the null hypothesis is not rejected. There is not sufficient evidence to claim that the mean wage for Manchester United is greater than that of players from Manchester City.

For the second T-test at alpha level .1, (H0: μMU = μALL, Ha: μMU> μALL), the null hypothesis is not rejected. There is not sufficient evidence to claim that the mean wage for Manchester United Players is higher than players from all other clubs.

This is a significant finding because it is often mentioned that Manchester United over pays their players. So one would assume the average wage of Manchester United players would be higher than that of other players. We found the opposite appeared to be true.

## Chi-Square Test

### Introduction

A Chi-Square ( Test is used to test if there is a significant difference between the expected frequencies and the observed or actual frequencies. We used the Chi-Square test to test if there was a significant difference in the frequency of all positions between two zones of origin. Seperating players into the zones of the Americas and Europe/Africa. Europe and Africa were grouped as zone 2, and the Americas made up the zone 1 group. The following assumptions are met to be able to run a Chi-Square test on the data[[3]](#footnote-3)

The data in the cells should be frequencies, or counts of cases rather than percentages or some other transformation of the data.

* The levels (or categories) of the variables are mutually exclusive. That is, a particular subject fits into one and only one level of each of the variables.
* Each subject may contribute data to one and only one cell in the .
* The study groups must be independent. This means that a different test must be used if the two groups are related.
* There are 2 variables, and both are measured as categories, usually at the nominal level. However, data may be ordinal data. Interval or ratio data that have been collapsed into ordinal categories may also be used. While Chi-square has no rule about limiting the number of cells (by limiting the number of categories for each variable), a very large number of cells (over 20) can make it difficult to meet assumption #6 below, and to interpret the meaning of the results.
* The expected values should be 5 or more in at least 80% of the cells, and no cell should have an expected of less than one. This assumption is most likely to be met if the sample size equals at least the number of cells multiplied by 5. Essentially, this assumption specifies the number of cases (sample size) needed to use the for any number of cells in that .

### Hypothesis

the frequencies of positions of players do not differ between Zones of origin

= the frequencies of positions of players differ between the Zones of origin

### Equations

The test statistic for a Test is the following:

Where O is the observed value and E is the expected value and therefore:

Where is the probability for the corresponding .

Reject the null hypothesis, when

### R Code

Position<- data$New\_Position

Zone<- data$Zone

Zone1<- Zone[Zone== 1]

Zone2<- Zone[Zone== 2]

Z1<- Position[Zone== 1]

(length(Z1[Z1== "Midfield"])/length(Z1)) #percentage of Midfield from Z1

(length(Z1[Z1== "Defense"])/length(Z1)) #percentage of Defense from Z1

(length(Z1[Z1== "Forward"])/length(Z1)) #percentage of Forward from Z1

(length(Z1[Z1== "Keeper"])/length(Z1)) #percentage of Keeper from Z1

Z2<- Position[Zone== 2]

(length(Z2[Z2== "Midfield"])/length(Z2)) #percentage of Midfield from Z2

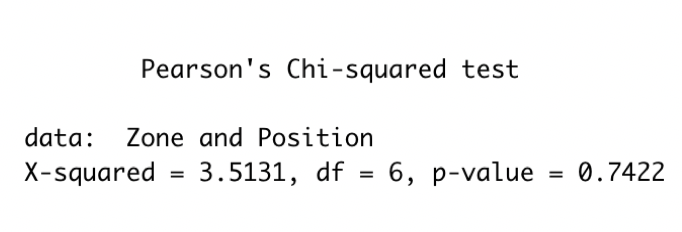
(length(Z2[Z2== "Defense"])/length(Z2)) #percentage of Defense from Z2

(length(Z2[Z2== "Forward"])/length(Z2)) #percentage of Forward from Z2

(length(Z2[Z2== "Keeper"])/length(Z2)) #percentage of Keeper from Z2

chisq.test(Zone,Position)

### Results



### Conclusion

For the at alpha level .1, we will fail to reject the Null Hypothesis. Claiming there is not enough evidence to say that the frequency or proportion for positions of players differs among zone of origin.

## 

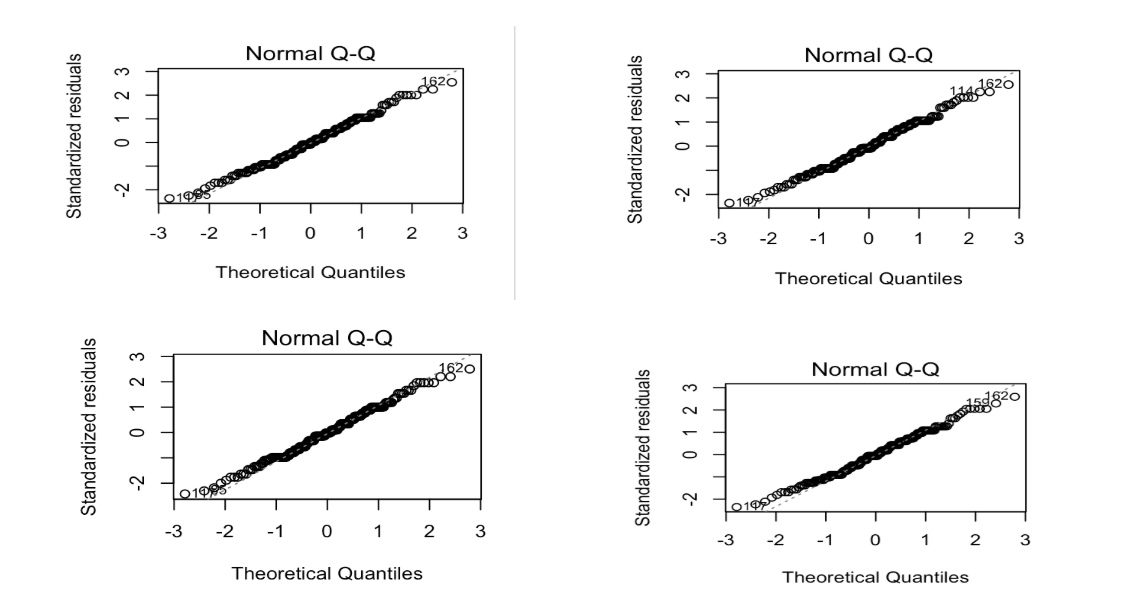
## Analysis of Variance test

### Introduction

### An Analysis of Variance test (ANOVA) is used to test if the mean values of different smaller populations are equal. We used the Analysis of Variance test to test if the mean weight for at least one club differs from the others. The clubs we tested are Arsenal vs. Chelsea vs. FC Barcelona vs. Juventus. We will be using the unbalanced test for Variance (Scheffé) because the sizes of our samples do not equal.

● Using this test we assume equal but unknown variances

● The study groups must be independent. This means that a different test must be used if the two groups are related.

● We must also note that the data must be normally distributed for the populations

### Hypothesis

μA = μC=μB = μJ

= At least one mean weight does not equal

### Equations for Scheffé

The test statistic is:

Reject the null hypothesis if

### R Code

plot(lm(data$Weight~data$Club== c("Arsenal")))

plot(lm(data$Weight~data$Club== c("Chelsea")))

plot(lm(data$Weight~data$Club== c("FC Barcelona")))

plot(lm(data$Weight~data$Club== c("Juventus")))

WeightA<- data$Weight[data$Club== c("Arsenal")]#Weights for Arsenal players

WeightC<- data$Weight[data$Club== c("Chelsea")] #Weights for Chelsea players

WeightB<- data$Weight[data$Club== c("FC Barcelona")] #Weights for FC Barcelona players

WeightJ<- data$Weight[data$Club== c("Juventus")] #Weights for Juventus players

mean\_amt <- c(mean(WeightA),mean(WeightB),mean(WeightC),mean(WeightJ))

Club\_name <- c('Arsenal','Chelsea', 'FC Barcelona', 'Juventus')

plot(mean\_amt, xlab = "Club", ylab = "Mean amount", xaxt='n', ann=FALSE)

axis(1, at= 1:4, labels=Club\_name)

A<- length(WeightA) #how many players for Arsenal

C<- length(WeightC) #how many players for Chelsea

B<- length(WeightB) #how many players for FC Barcelona

J<- length(WeightJ) #how many players for Juventus

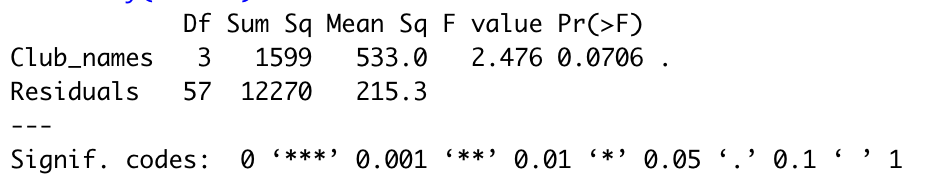
Weights<- c(WeightA, WeightC, WeightB, WeightJ)

Club\_names <- rep(c('Arsenal','Chelsea', 'FC Barcelona','Juventus'),c(A,C,B,J))

fitAOV<- aov(Weights~Club\_names)

summary(fitAOV)

### Results



### Conclusions

### For this test at alpha level of .1, we will reject the null hypothesis in favor of the alternative. claiming there is enough evidence to say that at least one mean for weight differs between Arsenal, Chelsea, FC Barcelona, Juventus.

## Generalized Linear Models

### Introduction

Generalized Linear Models (GLMs) refers to a large class of models that extends normal theory regression to several useful distributions, including the gamma, poisson, and binomial. There are three major components to any GLM:[[4]](#footnote-4)

* An exponential family of probability distributions
* A linear predictor ( is the quantity which incorporates the information about the independent variables into the model. It is related to the expected value of the data through the link function.
* Link function provides the relationship between the linear predictor and the mean of the distribution function. The link function is denoted by the following equation:

Assumptions of the GLM include the following:

* The data are independently distributed
* The dependent variable does not need to be normally distributed, but assumes a distribution such as binomial, poisson, multinomial, gamma, etc.
* A linear relationship exists between the transformed response in terms of the link function and explanatory variables
* A linear relationship between the dependent variable and the independent variables is not assumed
* Homogeneity of variance does not need to be satisfied.
* Errors need to be independent but not normally distributed
* Maximum likelihood estimation (MLE) is used to estimate the parameters

### Hypothesis

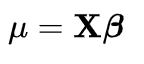
The GLM tested the following:

Wages of the players are independent on their zone of origin and their preferred foot

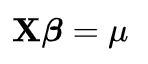
= Wages of the players are dependent on their zone of origin and their preferred foot

### Equations

The mean of Normal distribution



The link function for the Normal distribution is the following:



### R Code

attach(data)

Zone[which(Zone==2)]= "Europe"

Zone[which(Zone==1)] = "Americas"

Zone.new = Zone[which(Zone!=3)]

preferred.foot.new = Preferred.Foot[which(Zone!=3)]

wage.new = Wage[which(Zone!=3)]

fit1 = glm(wage.new~Zone.new + preferred.foot.new)

summary(fit1)

### Results

### 

### Conclusions

Our conclusion is to reject the null hypothesis in favor of the alternative for zone, meaning that wage is dependent on zone. We will fail to reject the null hypothesis for preferred foot, claiming that the wage does not depend on the preferred foot of the player.

## 

## Missing Values Analysis

### Generalized Linear Models

We ran the GLM distribution test on the same set of data with random missing values. For the test, 20% of the player values were set to “NA”. The missing values were randomly selected. Two tests were run, first replacing the “NA” values with the mean of the wages and the second test the “NA” values were replaced with the median wage. This method was used in order for the degrees of freedom to be equal in both the missing value analysis and the original GLM method.

#### Hypotheses

The GLM distribution tested the following:

Wages of the players are independent on their zone of origin and their preferred foot

= Wages of the players are dependent on their zone of origin and their preferred foot

#### R Code

#### “NA” values replaced with mean:

Wage.Missing = data$Wage.Missing

Wage.Missing = impute(Wage.Missing, mean)

Wage.Missing = Wage.Missing[which(Zone!=3)]

fit2 <- glm(Wage.Missing~Zone.new + preferred.foot.new)

summary(fit2)

“NA” values replaced with median:

Wage.Missing = data$Wage.Missing

Wage.Missing = impute(Wage.Missing,median)

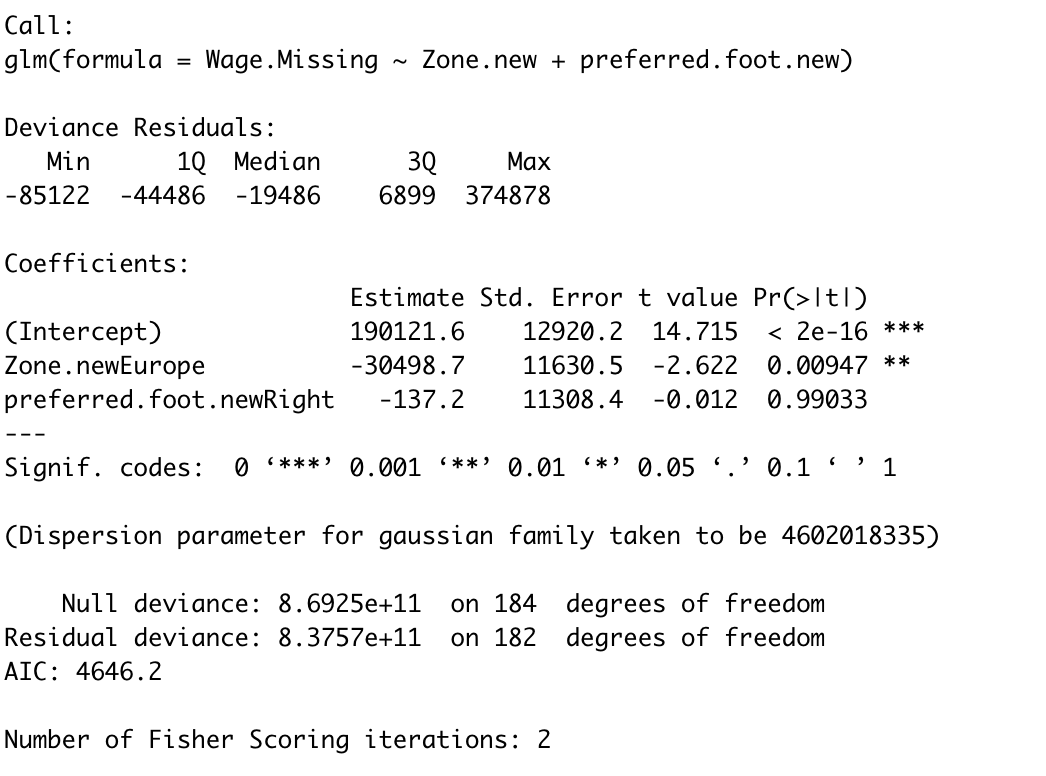
Wage.Missing = Wage.Missing[which(Zone!=3)]

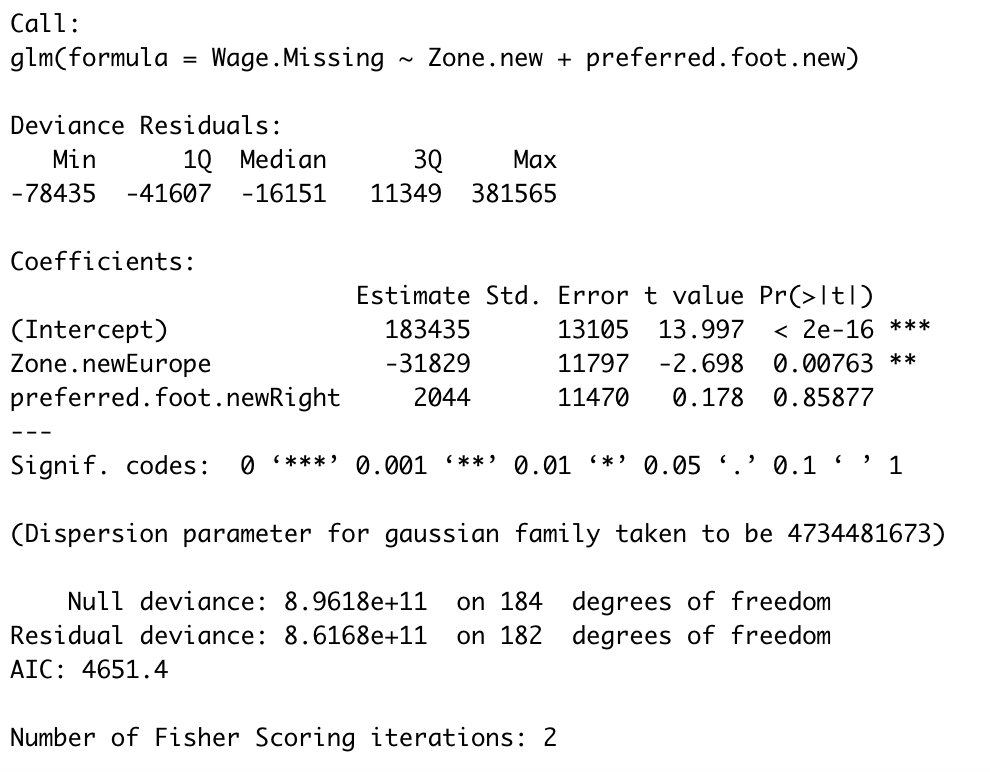
fit3 <- glm(Wage.Missing~Zone.new+preferred.foot.new)

summary(fit3)

Results

“NA” values replaced with mean:



“NA” values replaced with median:

### Conclusion

Our conclusion for both is to reject the null hypothesis in favor of the alternative for zone, meaning that wage is dependent on zone. Again, we will fail to reject the null hypothesis for both mean and median on preferred foot, claiming that the wage does not depend on the preferred foot of the player. This is the same conclusion as we saw with the original GLM but the zone p-value has decreased significantly in both mean and median and the preferred foot value has about doubled in both.

### Missing Value Analysis

In many instances there will be missing data points when collecting data from a large population for various different reasons. In the case of this data, missing data points may be seen in wage due to insecurities of players with wages lower than the wages of their teammates/competitors. There are a few ways to handle missing data when the data is missing at random. One way to handle this is to treat the missing data points as if they replace the missing points with central values like the median to preserve the sample size. This method can limit the amount of bias seen. Another way to deal with missing data is to just disregard the datapoint even exists. This is used more often when there is not a large amount of missing points (ex. Less than 5% of the data).

On the other hand, we can not use these methods or any method for the case of nonignorable missing data. In this case if we carry on with our analysis the measures of center will be meaningless as we will be missing key components to the data set. The omitting of these data points will make our analysis extremely misleading. These data points should always be included in the analysis in order to get an accurate read on the set.[[5]](#footnote-5)

1. <https://ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Two-Sample_T-Test.pdf> [↑](#footnote-ref-1)
2. <https://www.itl.nist.gov/div898/handbook/eda/section3/eda353.htm> [↑](#footnote-ref-2)
3. <https://newonlinecourses.science.psu.edu/stat504/node/216/> [↑](#footnote-ref-3)
4. <https://newonlinecourses.science.psu.edu/stat504/node/216/> [↑](#footnote-ref-4)
5. <https://www.statisticssolutions.com/missing-values-in-data/> [↑](#footnote-ref-5)