University of California, Davis

# Take Home Project 2

Topic 1 Question 2: Hospital Beds & Length of Stay
Topic 2 Question 1: Athlete Red Blood Cell Count

Ana Boeriu & Julia Tien
STA 108
Dr. Erin Melcon
October 27, 2017

## TOPIC 1 QUESTION 2: Hospital Beds & Length of Stay

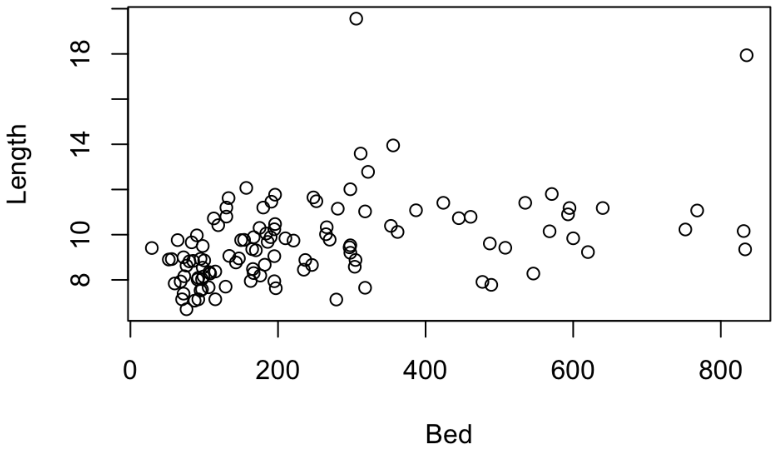
### INTRODUCTION:

Given data on our Y variable which is length of stay for a patient at the hospital in days and our X variable which is average number of beds in hospital during study period, the goal for this report is to transform the data so that it best fits the linear model.

### a. Original Plot & Diagnostic Plots

In this section we are performing diagnostics and testing for normality in the errors, constant variance in the errors, and linearity.

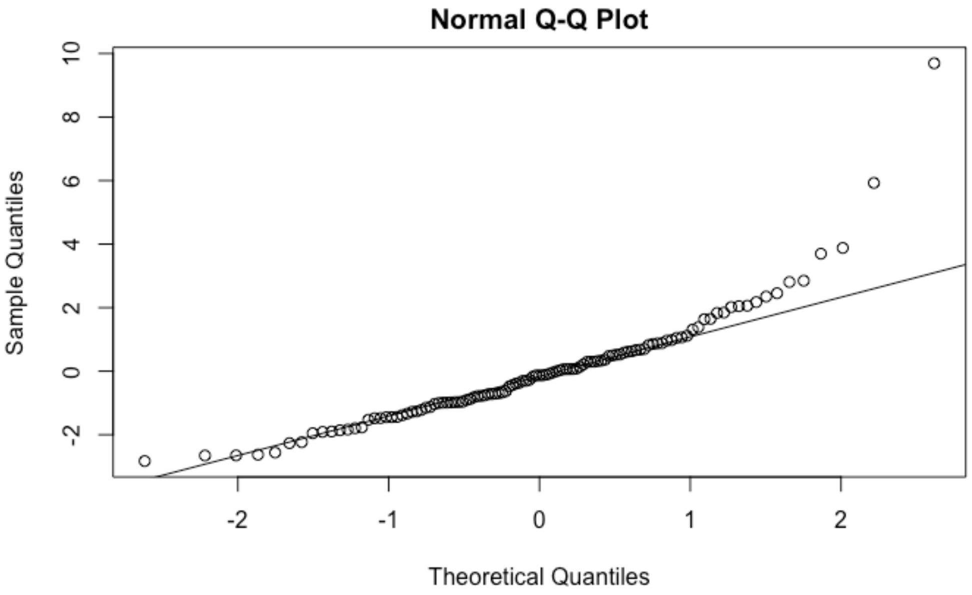
#### A) Plot of Original Data & Regression Line



Original Regression Line: $\widehat{Y} = 8.6253 + 0.004057X$

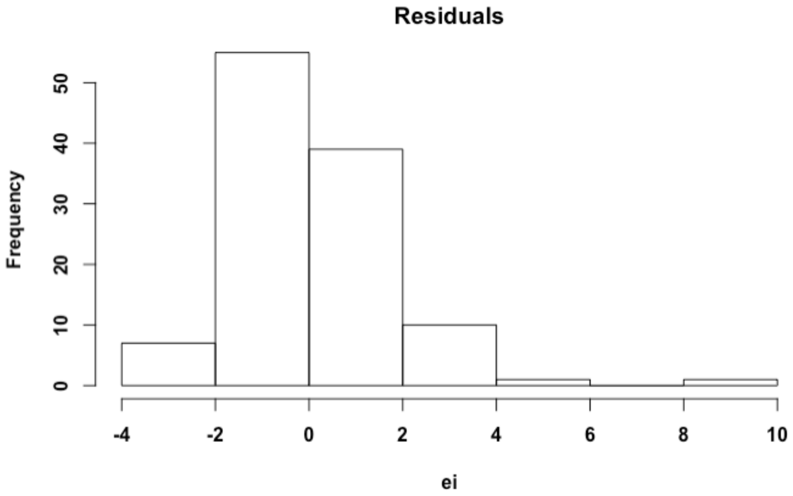
#### B) Testing for Normality:

1) **QQ Plot**



This data does not seem to be normally distributed. There are outliers and the points towards the end of the line are not where the points would lie if the data was totally normally distributed.

2) **Histogram of Errors**



The data does not seem to be normally distributed as the histogram of errors is positively skewed.

3) **Shapiro-Wilks Normality Test:3**

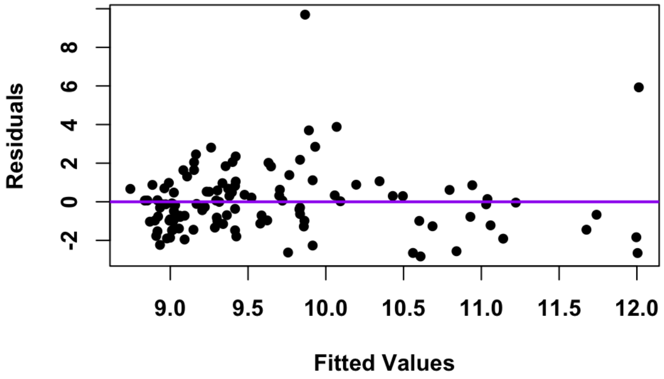
$Ho: Data\ is\ normal$ $\qquad$ $Ha: Data\ is\ non - normal$

P-value = 1.775e-08

Because this p-value is approximately 0 and less than any possible alpha value(1%,5%,10%), we reject the null and conclude that the data is not normally distributed.

### C) Testing for Constant Variance:

#### 1) Plot of Errors vs. Fitted Values



There are many more data points on the lower values of X and a couple of outliers, but the data points are somewhat in a “band” around 0.

**2) Fligner Killeen Test**

$Ho: Residuals\ have\ equal\ variance$ $Ha: Residuals\ do\ not\ have\ equal\ variance$

P-value = .03383

Because the p-value is less than alpha value of 0.05, there is evidence to reject the null so we conclude that errors do not have constant variance.

### D. Testing for Linear Relationship:

To make sure there is a significant linear relationship between X and Y which is part of the assumptions for linear regression, we perform a F-Test.
We are testing for a linear relationship: $Ho: \beta 1 = 0$ $Ha: \beta 1 \neq 0$

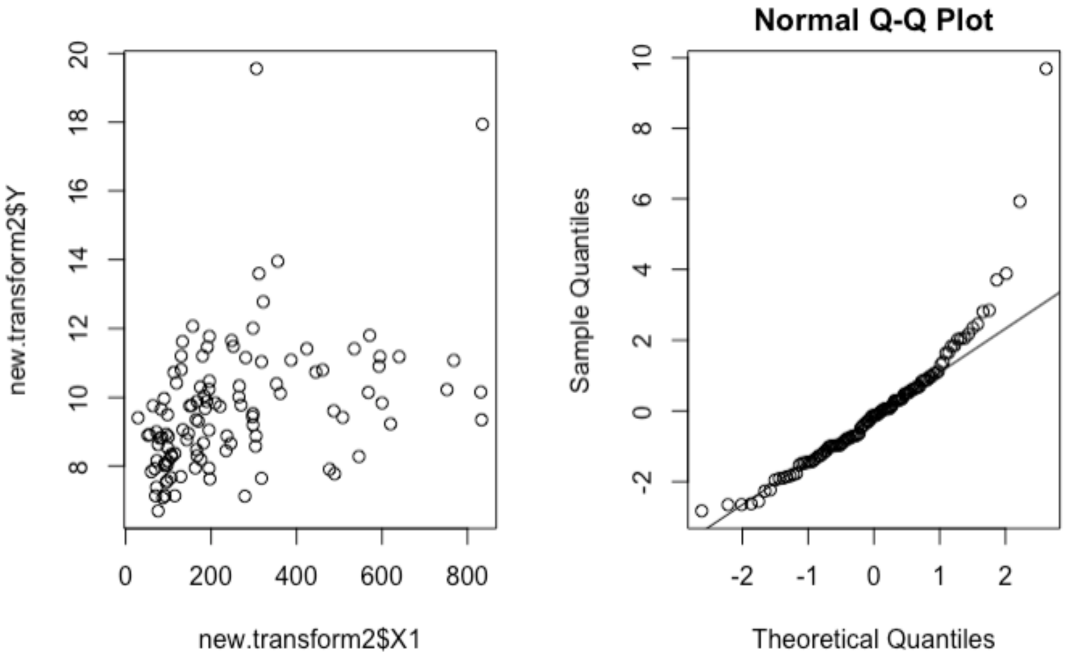
P-value = 6.765e-06

If in reality there is no linear relationship between length of stay and average number of beds in the hospital during the study period, the probability of observing our data or more extreme is 6.765e-06.
Because the p-value is less than alpha value of 0.05, reject the null that there is no significance and conclude there is a significant linear relationship between length of stay and average number of beds in the hospital so this assumption for linear regression is met.

## b. Transformations & Outliers

In this section we transform the data to try correcting for non-normality and heteroskedasticity in the errors using the Box-Cox transformation and then Tukey for X, Y, and X and Y. We will then look for outliers as many times transforming the data first will also help adjust for outliers.

## A) Before Transformations



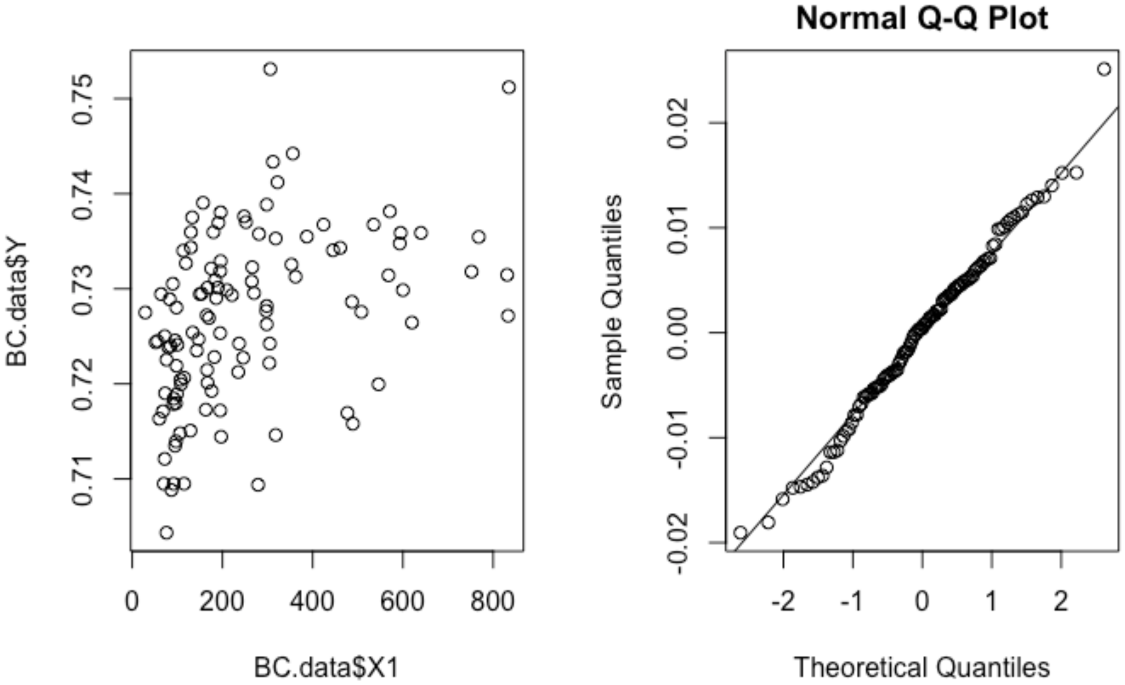
There are quite a few outliers in the QQ plot.

## B) Transformations

### 1) Box-Cox

Box-Cox transformations use $(X^{\lambda}-1)/\lambda$ to find the lambda that maximizes log-likelihood.

λ that maximizes log-likelihood = -1.3

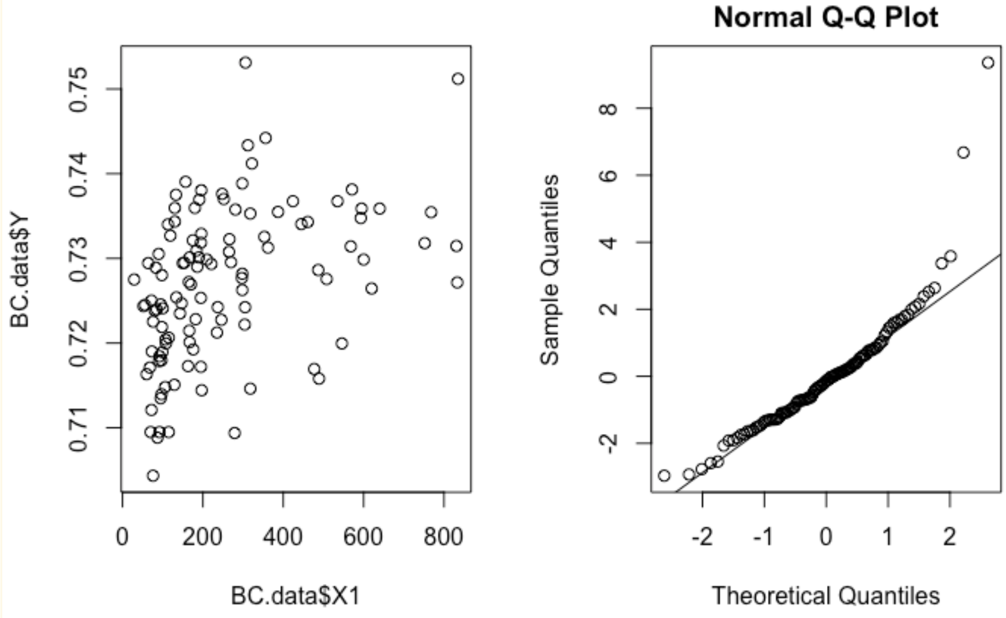


Outliers have been slightly fixed and the data looks a little more spread out.

2) **Tukey**

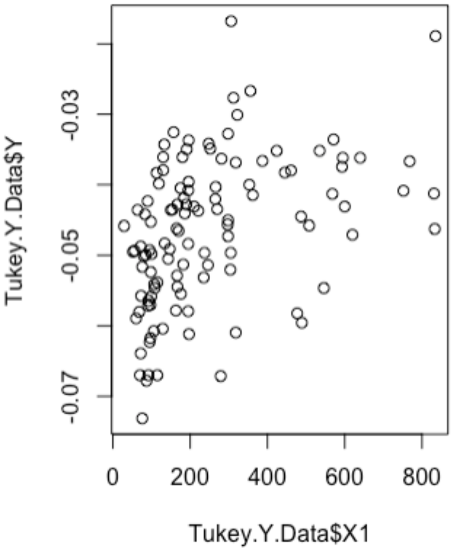
The Tukey transformation can transform X, Y, or X and Y by also choosing the lambda that maximizes log-likelihood.

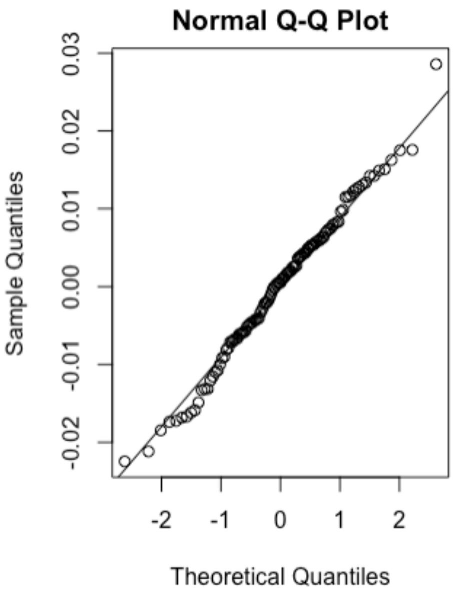
a) **Transform "Average number of beds" (X)**



Data looks more spread out but there are still a few outliers.

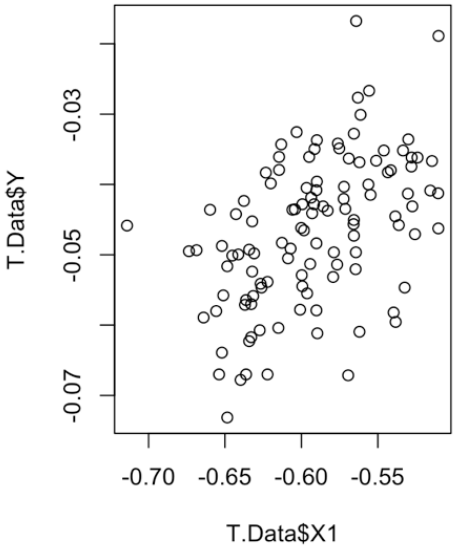
b) **Transform "Length of Stay" (Y)**

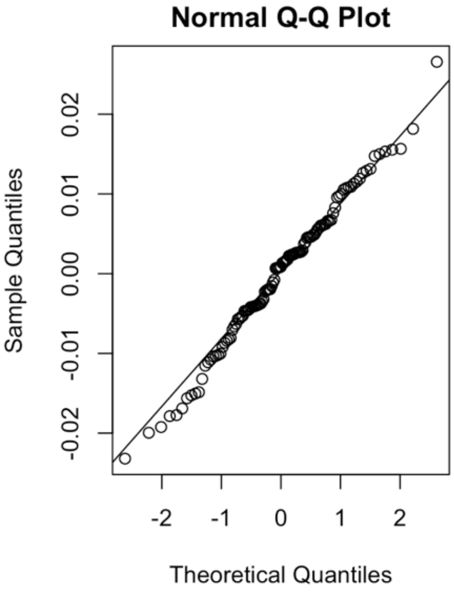




This transformation looks better than just transforming X.

### c) Transform “Average number of beds” & “Length of Stay” (X & Y)





This transformation seems to correct the data the best, as we discuss more in the following section.

## C) Outliers

We chose to use the transformed data that comes from using the Tukey transformation of X and Y because that one seems to make the data most linear, most normal, and has the most constant variance compared to no transformation or the other Box-Cox and Tukey transformations. Our next step was to look for and remove any outliers so that we could further meet the assumptions of linear regression, however we did not find any outliers when using an alpha of .05 with this new transformed data set. With an alpha of value of .5 there is an outlier, although using that alpha value is atypical. We also checked for leverage points, which are data points which have a large influence on the regression line. When using the diagonal of the hat matrix there are eight leverage values but when using cook's distance or difference in fitted values there were no leverage points, and so in the end we concluded to not remove any data points because the transformed data seems to have corrected for possible outliers from the original data.

## c. Discussion

Transforming the data did help. This table of p-values for different tests below provides further evidence that the Tukey transformation of X and Y helped make the data better meet the assumptions for linear regression.

| Shapiro-Wilks | Fligner-Killeen | Significant Linear Relationship (F-Test) |
|---|---|---|
| 0.7944 | 0.2113 | 8.334e-08 |

Transforming the data definitely helped adjust normality as the old p-value for the Shapiro-WIlks test was 1.775e-08 but now it is .7944 which clearly is greater than an alpha value of .05 so with the transformation we fail to reject the null and errors are approximately normal.
While the p-value for the Fligner-Killeen did decrease from .5572 to .2113, we still fail to reject the null with an alpha of .05 so we can still assume errors have constant variance.
The transformed data still produces a significant linear relationship between length of stay and average number of beds and in fact produced an even smaller p-value than the original 6.765e-06. Even though both probabilities round to 0, this is still a rather large decrease. The transformed data is a better fit as one can also seen when comparing the two graphs.

Downsides for transformations are that it can be costly especially  if the data set is very large as more time and resources are needed to keep measure and keep track of all the variables. It is also very hard to interpret results from transformations. With all of the evidence presented, we would recommend a client to use the Tukey transformation on X and Y. It helped to remove outliers and make the errors of the data normal and have constant variance.

**TOPIC 2 QUESTION 1: Athlete Red Blood Cell Count**

## I. Introduction

The goal of this project is to build a model which can model red blood cell count based off of other physical characteristics from Australian athletes. We are using the approach of multiple linear regression to build a more “correct” model which is more focused on accuracy as there will only be variables that have a significant relationship with the red blood cell count measured (liters), which is our Y. The explanatory variables we are analyzing are X1 which is lean body mass (kg), X2 which is body mass index (kg), X3 which is percent body fat, X4 which is plasma ferritins (ng), X5 which is a categorical variable of “Male” or “Female”, and X6 which is also a categorical variable that represents “Net” (sports that involve a net), “Swim”, and “Run”. Because there are so many different variables and models we could choose from, we will perform a series of model selection techniques to find the set of variables that will give us the best model that meets our goal of correctness. We will be using BIC as our model criteria and Forwards Backwards Selection as the subset selection. We will then perform diagnostics that assesses how well the model we choose meets the assumptions for multiple linear regression and use the model to calculate confidence intervals and predictions that provide insight about the true red blood cell count. Doctors or athletic trainers may be interested in this model to see how athletes can train better and safer in the future. Anybody who exercises may also be interested in using this model, especially those who have medical issues and must carefully monitor their red blood cell count.

## II. Summary of Data

### A. Summary Statistics

| Y | X1 | X2 | X3 | X4 | X5 | X6 |
|---|---|---|---|---|---|---|
| Min.: 3.80 | Min.: 34.36 | Min.: 16.75 | Min.: 5.630 | Min.: 8.00 | F: 100 | Net: 59 |
| 1st Qu.: 4.372 | 1stQu.: 54.67 | 1st Qu.: 21.08 | 1st Qu.: 8.545 | 1st Qu.: 41.25 | M: 102 | Run: 67 |
| Median: 4.755 | Median: 63.03 | Median: 22.72 | Median: 11.60 | Median: 65.5 | | Swim: 76 |
| Mean: 4.719 | Mean: 64.87 | Mean: 22.96 | Mean: 13.507 | Mean: 76.88 | | |
| 3rd Qu.: 5.030 | 3rd Qu.: 74.75 | 3rd Qu.: 24.46 | 3rdQu.: 18.00 | 3rd Qu.: 97.0 | | |
| Max.: 6.720 | Max.: | Max.: 34.42 | Max: 35.520 | Max.: 234.00 | | |

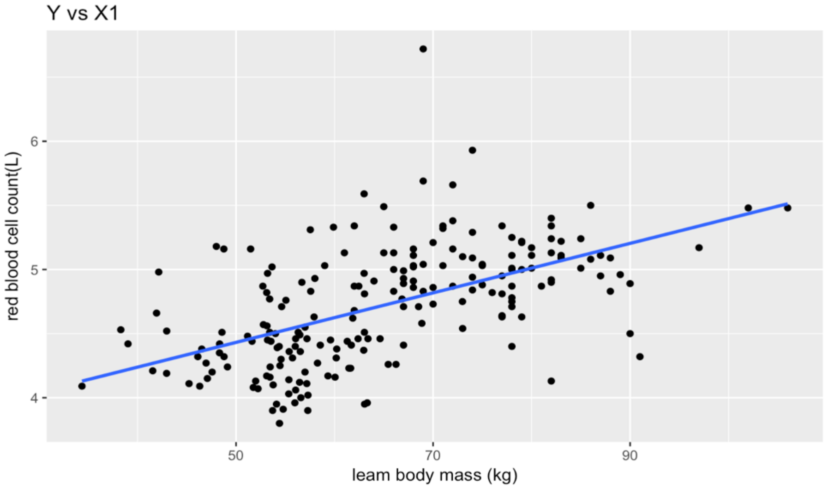
| | | | | |
|---|---|---|---|---|
| | 106.00 | | | |

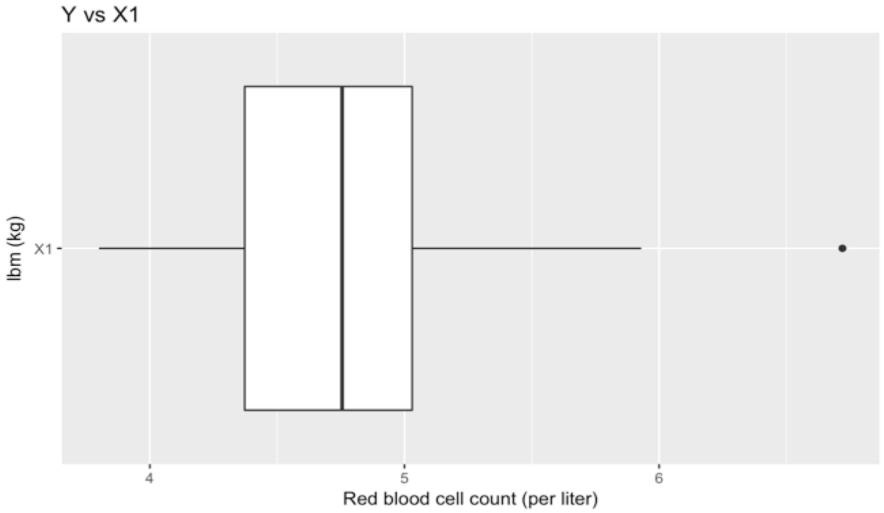
59 people play net sports, 67 play sports that involve running, and 76 people play sports that involve water. The rest of this information will be referenced below.

## B. Plots of Y vs. $X_i$

The scatterplots below show the linear relationships between the variables and red blood cell count. The box plots provide information on how the different variables interact with red blood cell count, such as the mean, amount of red blood cell count at different quartiles, and outliers.

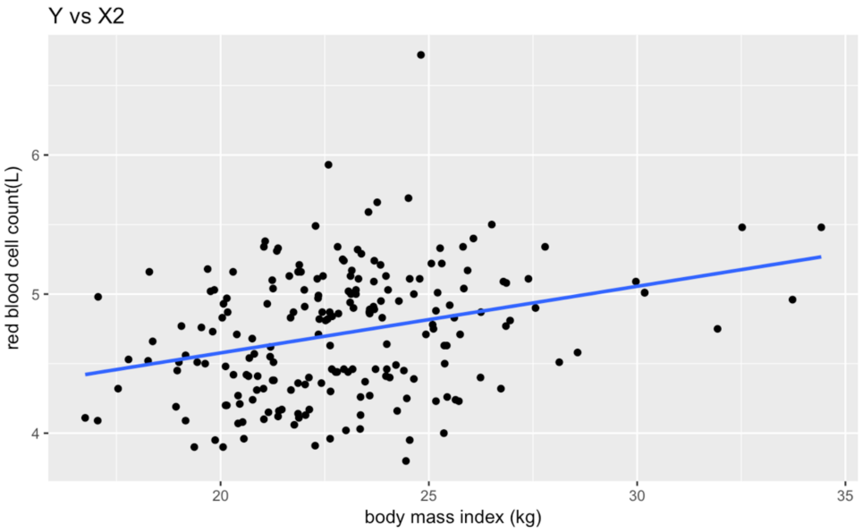
### 1. Red Blood Cell Count Vs. Lean Body Mass

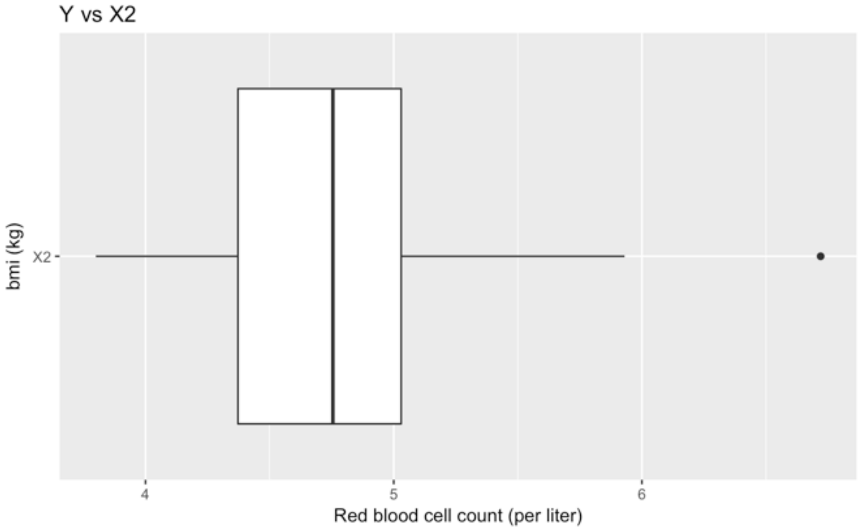




These plots shows the relationship between red blood cell count vs. lean body mass. Any red blood cell count value above 6 per liter and a lean body mass of 70 kilograms is an outlier. The mean is about 4.6 red blood cells per liter with a 1st quartile of 4.4 red blood cells per liter and a 3rd quartile of just above 5 red blood cells per liter.

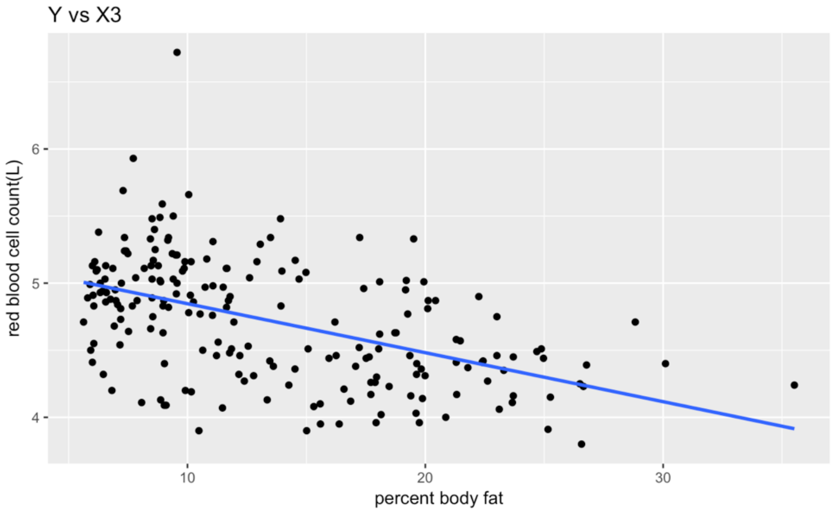
### 2. Red Blood Cell Count Vs. Body Mass Index

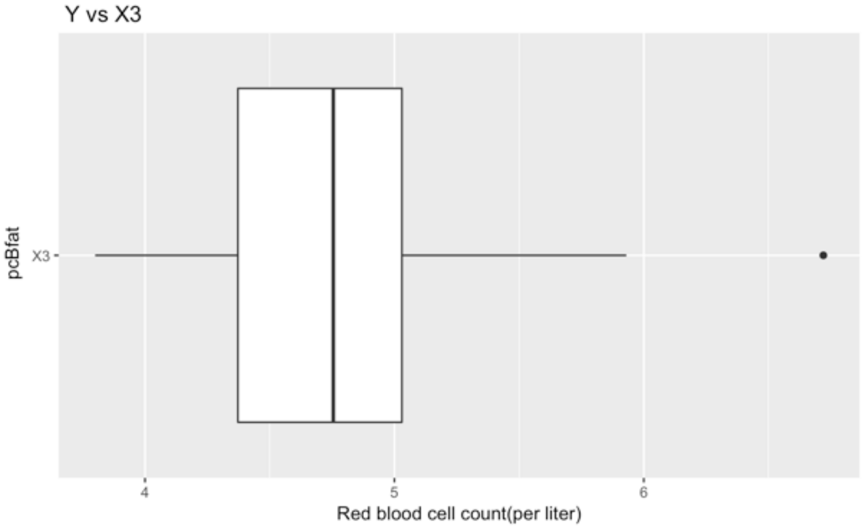




These plots shows the relationship between red blood cell count vs. body mass index Any red blood cell count above 6 per liter and body mass index of 24 kilograms is an outlier. The mean is about 4.6 red blood cells per liter with a 1st quartile of 4.3 red blood cells per liter and a 3rd quartile of just above 5 red blood cells per liter.

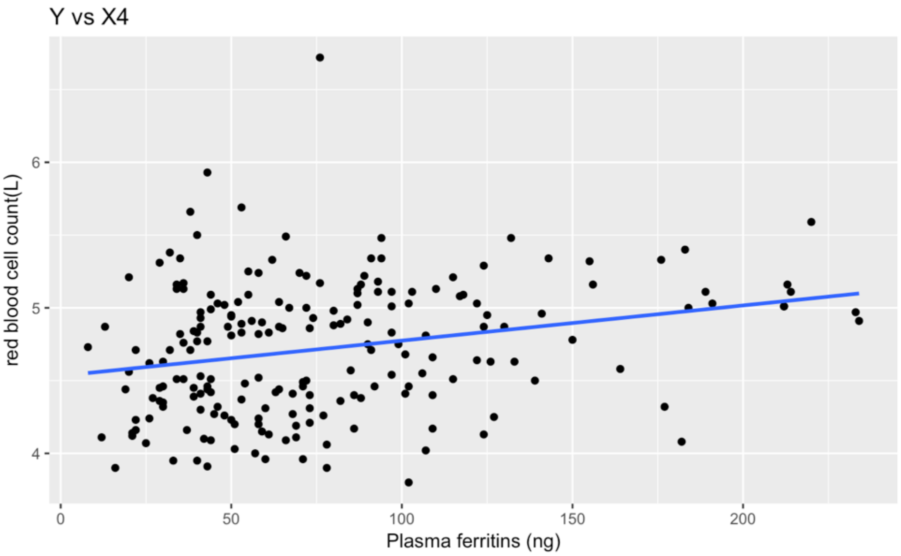
### 3. Red Blood Cell Count Vs. Percent Body Fat

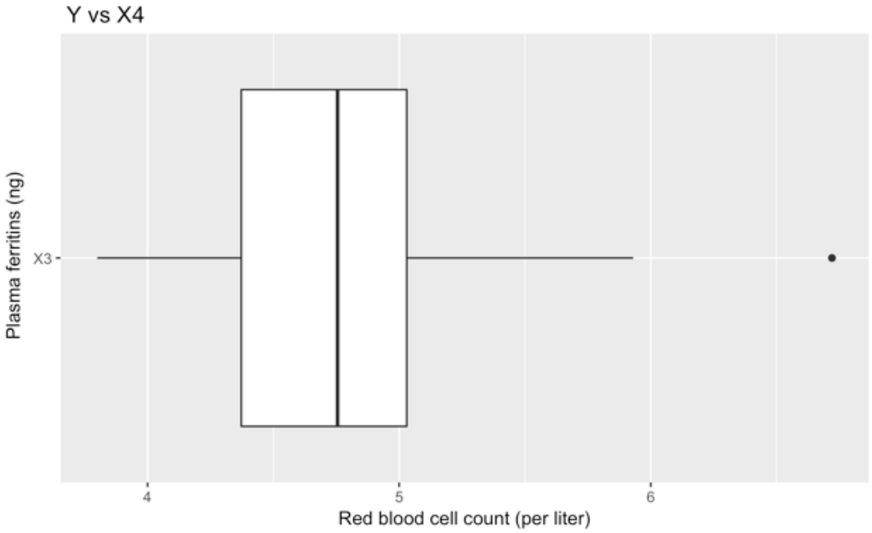




These plots shows the relationship between red blood cell count versus percent body fat. The scatterplot shows us that anything above a red blood cell count of 6 per liter and a 9% body fat is an outlier. The mean is about 4.55 red blood cells per liter with a 1st quartile of 4.35 red blood cells per liter and a 3rd quartile of just above 5 red blood cells per liter.

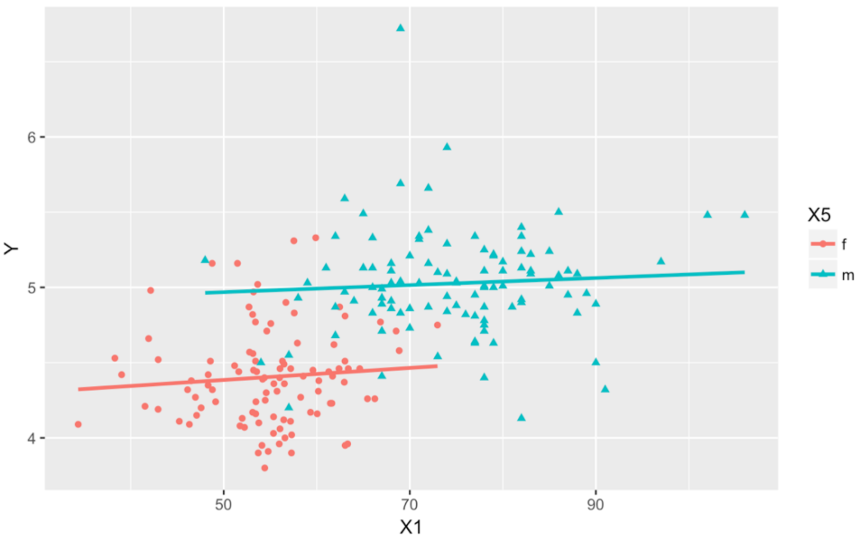
### 4. Red Blood Cell Count Vs. Plasma Ferritins



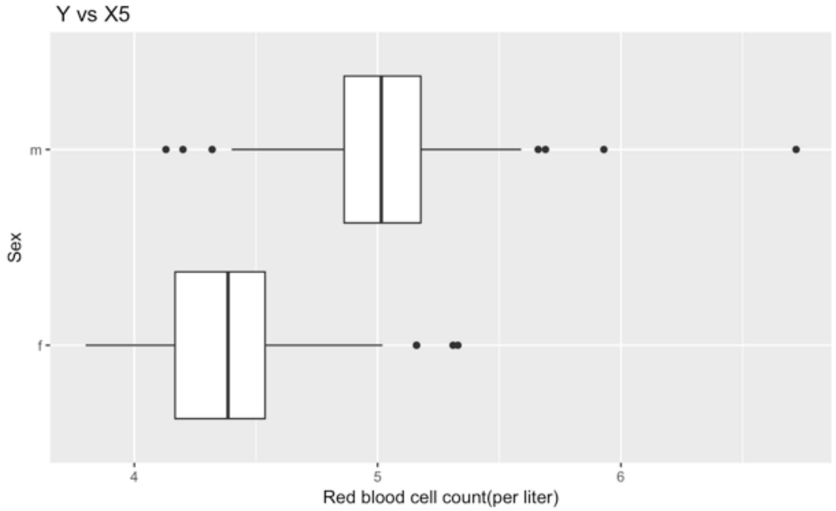


These plots shows the relationship between red blood cell count vs. plasma ferritins. Anything above a red blood cell count of 6 per liter and plasma ferritins value of 76 nanograms is an outlier. The mean is about 4.65 red blood cells per liter with a 1st quartile of 4.37 red blood cells per liter and a 3rd quartile of about 5.1 red blood cells per liter.

## 5. Red Blood Cell Count vs. Sex

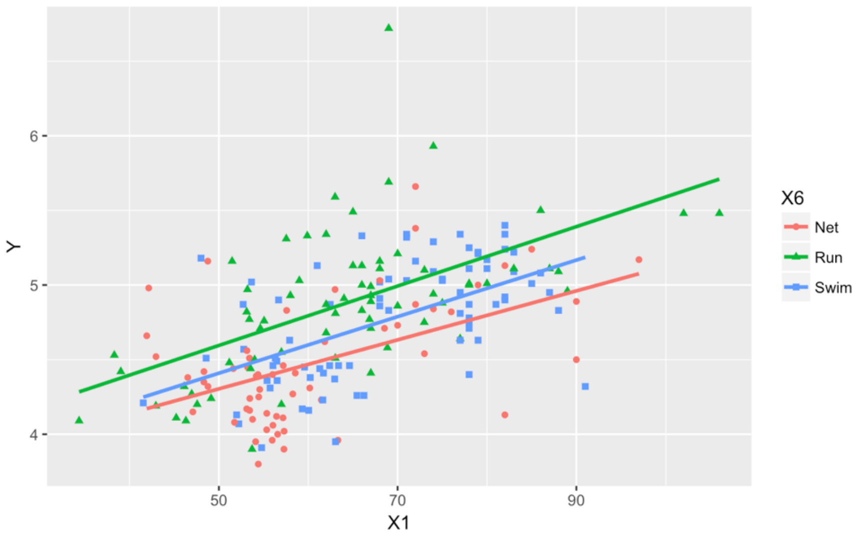


This scatterplot is between a numerical and categorical variable so we can see red blood cell count for both males and females.

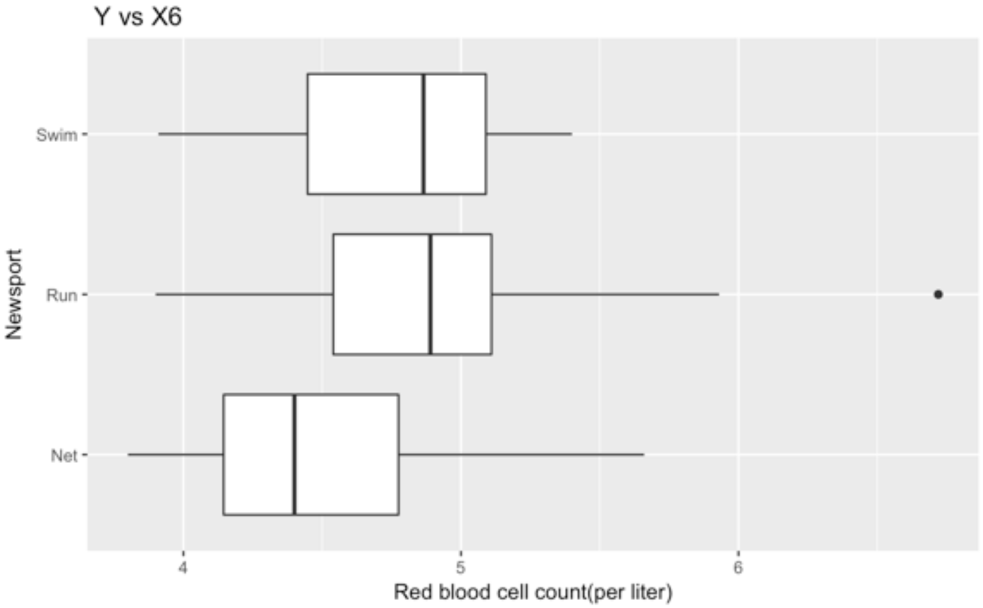


Because sex is a categorical variable, this is a grouped box plot which shows information of red blood cell count for males and females. One can clearly see that males on average have higher red blood cell counts than females.

### 6. Red Blood Cell Count Vs. New Sports



Because this variable is also a categorical variable, the plot shows red blood cell count for athletes who play a sport that involves running, water, or a net. The red blood cell count for athletes who do a sport that involves running seems to be on average slightly higher than those who do a sport that involves water. The red blood cell count for athletes who do a sport that involves water or running on average is higher for those who do a sport involving a net.



This grouped box plot confirms that athletes who do a sport that involves running has a higher red blood cell count on average than those who do a sport involving a net or water.

**C. Equation of the line:**

Y^=4.0408763 + 0.0028985X1+ 0.0072302X2 - 0.0006854X3 - 0.0004318X4 + 0.5207096X5m    0.2043172X6Run + 0.0942896X6Swim

Where:
Y = **rcc :** red blood cell count per liter
X1 = **lbm :** lean body mass (mass that is lean muscle) in kg,
X2 = **bmi** : Body mass index (kg),
X3 = **pcBfat** : Percent body fat,
X4 = **ferr** : Plasma ferritins, ng (a measure of iron in the blood),
X5 = **sex** : male or female
X6 = **newsport** : Net (sports that involve a net), Swim (sports that involve water), and Run (sports that involve running)

## III. Model Selection

In this section we are using stepwise regression for all subsets and a specific subset selection process to see which combination of explanatory variables will produce the best model based off of the criteria we have chosen to use which is BIC. There are many variables in our model and we must make sure the ones in our model are significant. We looked at the model chosen when using all subsets and then Forwards Backwards Selection.

### A. Model Criteria

Because our goal is to have a correct model, we are choosing only explanatory variables which have a significant relationship with red blood cell count. This model may be smaller than ones which have the goal of prediction. There are many different types of model criteria such as Adjusted $R^2$, Predictive Sum of Squares, Akaike Information Criteria, Bayesian Information Criteria and CP Mallows which measure how “correct” or good of a predictor the model is. For the purposes of this class AIC or BIC are often used as model selection criteria for correctness as they penalize large models. AIC may overfit correct models and BIC penalizes large models even more, so we chose to focus on BIC as we imagine if somebody is trying to see a person’s red blood cell count to prescribe a medication for example they will really need the result to be correct. We choose the model that lowers BIC the most.

### B. All Subsets

When there are a fewer number of predictor variables all possible models and the corresponding model criteria can be calculated. We looked at both AIC and BIC, however since we decided to stick with BIC the model highlighted below is the one we will be focusing on. This was the model with the lowest BIC when looking at all possible models.

| Best model with AIC | Best model with BIC |
|---|---|

| | |
|---|---|
| $Y = 4.066031891 + 0.004611036_{X1} + 0.487963168_{X5M} + 0.215025151_{X6Run} + 0.094965164_{X6\ Swim}$ | $Y = 4.3203209 + 0.5800667X_{5m} + 0.1997876_{X6Run} + 0.1039817_{X6Swim}$ |

### C. Subset Selection

Subset selection does not evaluate all possible models, however it is faster and does not cost as much because not all subsets are calculated.

We chose to use Forwards Backwards Selection because we want a correct model which means having a smaller one. Of course underfitting a model is not ideal however between underfitting and overfitting, underfitting the model is more likely to achieve the smaller and often more correct model which Forwards Selection or Forwards Backwards Selection is more likely to do. Forwards Backwards Selection does not underfit as much as Forwards Selection though so we used Forwards Backwards as our subset selection. The model highlighted below is the one that lowers BIC the most when using Forwards Backwards.

| Model | P (# of $\beta$) | BIC |
|---|---|---|
| Y,X5m | 1 | -115.16397 |
| Y,X5m,X6Run | 2 | -117.40019 |
| Y,X1,X5m,X6Run | 3 | -115.33944 |
| Y,X1,X5m,X6Run,X6Swim | 4 | -112.65189 |
| Y,X1,X4,X5m,X6Run,X6Swim | 5 | -107.79419 |
| Y,X1,X2,X4,X5m,X6Run,X6Swim | 6 | 102.68345 |
| Y,X1,X2,X3,X4,X5m,X6Run,X6Swim | 7 | 97.38223 |

The models from all subsets regression and and subset selection match even though the model from Forwards Backwards Selection is seemingly shorter. This is actually because X6 is one of the categorical variables and it is saying sports which involve water or net are just not as important in predicting red blood cell count as sports that involve running. Those two variables would ideally be combined into one category in the real world. However for this project since this is a categorical variable we cannot just drop parts of it and we will keep type of sport as an important variable.

The final mode we will use throughout the rest of this report is:

$$\boxed{Y = 4.3203209 + 0.5800667X_{5m} + 0.1997876_{X6Run} + 0.1039817_{X6Swim}}$$

## IV. Diagnostics

Diagnostics are performed to evaluate how well the data meets the assumptions needed for multiple regression, which are that observations are independent, errors have constant variance and they are normally distributed. We will look at different tests for these assumptions, and find and remove outliers. Typically we would transform data to correct for non-normality, non-linearity, or non-constant variance if necessary but for the purposes of this project we will just remove outliers and then perform diagnostics.

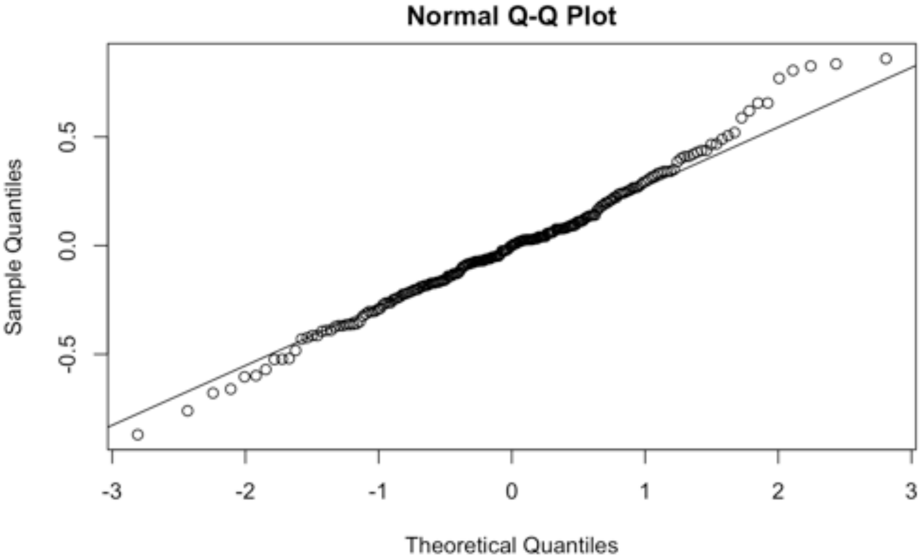
### A. Outliers

| Y | $X_1$ | X2 | X3 | X4 | X5 | X6 |
|---|---|---|---|---|---|---|
| 6.72 | 69 | 24.81 | 9.56 | 76 | m | Run |

Thus, the outlier is any male that chose running as a sport and who has a plasma ferritins of 76 nanograms, 9.56% body fat, a body mass index of 24.84, a lean body mass of 69 kg and a red blood cell count of 6.72 per liter. In order to further improve our model we will remove this outlier so that it does not skew any data. We looked for leverage points as those have an especially large influence over the model and can signify an outlier, however there were none.

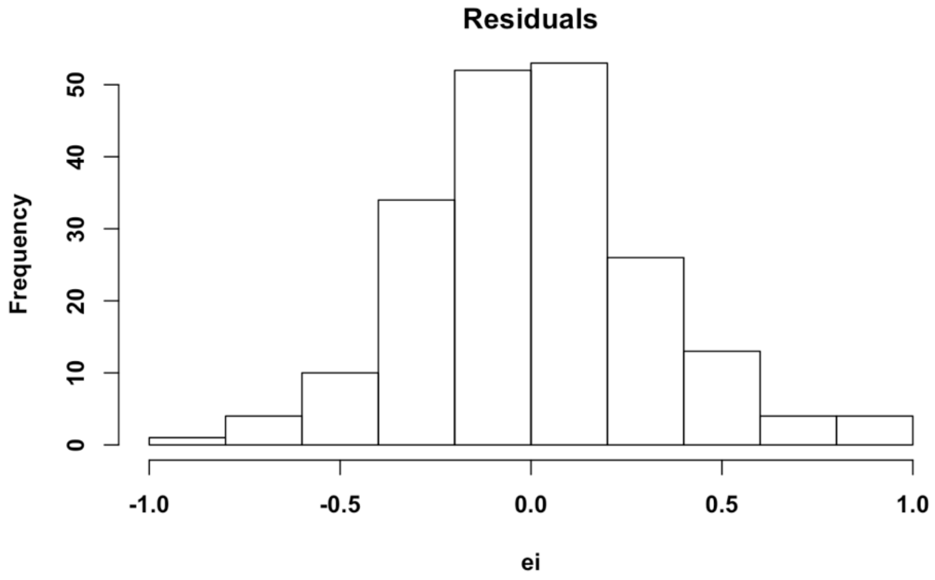
### B. Testing for Normality

#### 1. QQ Plot



This QQ plot has quite a few data points not on the Y=X line which could signify non-normal errors; however, it is rather subjective judgement.

#### 2. Histogram of Errors



This histogram definitely looks approximately normal.

**3. Sharpiro-Wilkes Normality Test**

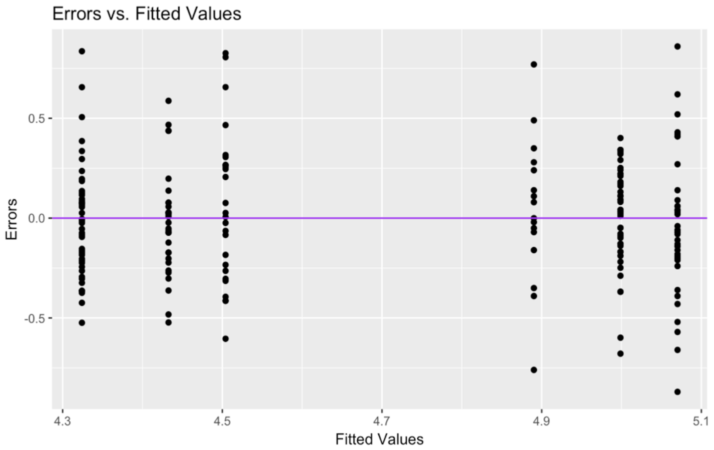
$Ho: Data\ is\ normal \qquad Ha: Data\ is\ non-normal$

P-value = 0.1484

Because this p-value is relatively large and greater than an alpha value of 0.05, we fail to reject the null hypothesis and conclude errors are approximately normally distributed.

## C. Testing for Constant Variance

### 1. Plot of Errors Vs Fitted Values



### 2. Fligner Killeen Test

$Ho: Residuals\ have\ equal\ variance \qquad Ha: Residuals\ do\ not\ have\ equal\ variance$

P-value = .6818

Because the p-value is greater than alpha value of 0.05, there is evidence to fail to reject the null hypothesis so we can conclude that errors have constant variance.

### D. Testing for Significant Linear Relationship

Here we want to test for a significant linear relationship between sex and the type of sport and red blood cell count. It is also testing to see if we can drop X5 and X6 from the full model.

$$Ho: \beta 5 = \beta 6 = 0 \qquad Ha: \beta 5 \neq \beta 6 \neq 0$$

P-value: 2.2e-16

Since our p-value is approximately zero we can conclude that there is a significant linear relationship between sex and type of sport and the Y variable of red blood cell count so this assumption for multiple regression is met. This also shows that the variables cannot be dropped. If in reality there was no linear relationship between sex and type of sport played and red blood cell count, then the probability of observing our sample data or more extreme is 2.2e-16.

All assumptions for multiple regression seem to be met according to the diagnostics.

## V. Analysis & Interpretations

### A. Partial $R^2$

0.438543

When we add sex to model that already has type of sport played we reduce our error by 43.85%.

0.04765198

When we add type of sport played to a model that has sex we reduce our error by 4.765%
Sex of the athlete seems to be the variable that affects red blood cell count more.

### B. Coefficient of Determination

0.5075

50.75% of the total variation in red blood cell count was explained by the linear relationship with the variables.

**C. Test Statistics**

| Estimate | Estimate | Std error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| Intercept | 4.32410 | 0.04206 | 102.805 | 2e-16 |
| X5m | 0.56613 | 0.04564 | 12.405 | 2e-16 |
| X6Run | 0.17991 | 0.05739 | 3.135 | 0.00198 |
| X6Swim | 0.10846 | 0.05554 | 1.953 | 0.05228 |

The test statistics gives us information regarding the number of standard deviations for $b_i$ (where i=5,6) is from the null value of zero. Thus we can conclude the test statistic for X5 is 12.405 standard deviations from the null hypothesis value of zero. This is a very large distance from zero. We can see that as we add more X's the t value decreases.

**D. Testing Individual Variables**

We perform F-tests on each individual variable to further confirm they are individually significant in explaining red blood cell count:

$$Ho: \beta 5 = 0 \qquad Ha: \beta 5 \neq 0$$

P-value: 2.2e-16

Since our p-value is small and would be less than any significant alpha (1%, 5%, 10%) we reject the null and conclude that sex is significant and cannot be dropped from the full model. If in reality there was no linear relationship between sex and red blood cell count, then the probability of observing our sample data or more extreme is 2.2e-16.

$$Ho: \beta 6 = 0 \qquad Ha: \beta 6 \neq 0$$

P-value: 0.008154

We reject the null at any significant alpha since our p-value is small and conclude that New sport (X6) is significant and cannot be dropped from the full model. If in reality there was no linear relationship between type of sport and red blood cell count, then the probability of observing our sample data or more extreme is 0.008154.

**D. Confidence intervals**

**1. Multiple CI's for betas**

| | Lower Bound | Upper Bound |
|---|---|---|
| Intercept | 4.21807164 | 4.4301311 |
| X5m | 0.45107808 | 0.6811742 |
| X6Run | 0.03523075 | 0.3245954 |

| | | |
|---|---|---|
| X6 Swim | -0.03155869 | 0.2484697 |

We are 95% overall confident that the red blood cell count for a male rather than a female will increase by between 0.451 and 0.6812 liters on average, holding type of new sport constant.
We are 95% overall confident that the red blood cell count increases by between .0352 and .325 liters on average for those who participate in a sport that involves running rather than a net, holding sex constant.
We are not able to interpret X6 Swim because the confidence interval contains 0. This further builds on what we said in Part IIIC that not all of the categories in the categorical variable of "type of sport" are significant in determining red blood cell count, however we cannot just drop parts of this categorical variable.

**2. Prediction Intervals**

To practice using our model we calculated and interpreted intervals for a male with a BMI of 19 who runs, a female with BMI of 24 who Swims and a female with a BMI of 31 who plays sports that involve a net.

| Bonferonni | WH | Scheffe |
|---|---|---|
| 2.415 | 3.110 | 2.820 |

Because we want to find the prediction intervals we will look to compare Bonferroni multiplier with Scheffe multiplier. We chose to use the Bonferroni multiplier because it is the smallest when compared to Scheffe.

| BMI | Sex | New sport | estimate | 95% lower | 95% Upper |
|---|---|---|---|---|---|
| 19 | m | Run | 5.070141 | 4.317569 | 5.822712 |
| 24 | f | Swim | 4.432557 | 3.679146 | 5.185968 |
| 31 | f | Net | 4.324101 | 3.571525 | 5.076678 |

We are 95% overall confident that for a male patient who runs and has a BMI of 19, his red blood cell count is between 4.32 and 5.82 red blood cells/liter. The actual estimated value is 5.07 liters which falls in that range.
We are 95% simultaneously confident that a female who swims and has a BMI of 24 her red blood cell count is between 3.679 and 5.186 red blood cells per liter. The actual estimated value of 4.43 confirms the interval is in the correct range.
We are family-wise 95% confident that for a female who just started playing net sports and that has a BMI of 31 her red blood cell count is between 3.57 and 5.08. The actual estimated value of 4.32 fall well within the specified range.

## VII. Conclusion

In this project we started with a dataset of Australian athletes and decided to build a model geared towards correctness rather than prediction which often will produce a smaller model. Through scatterplots and boxplots for each variable we found different statistics and also used general summary statistics to analyze the data and fit a regression line of the original data. However because data is often messy and variables may be present when they really aren't significant, we proceeded with a model selection technique of stepwise regression to find which variables did not need to be in our model. We decided to use BIC as the model criteria and Forwards Backwards Selection subset selection as this help achieve our goal of building a model that is more correct in predicting red blood cell count. This showed us that the significant variables are "sex" and "type of sport" so these were the two variables we used in our revised model. With this new model, we then removed an outlier and performed diagnostics to see how they meet the assumptions for multiple linear regression, which they did. We also analyzed the data by evaluating importance of the variables in the model and calculating test statistics, P-values, and various confidence intervals to see how the model works in practice. We found that both sex and type of sport played are rather important in determining red blood cell count however one of the categories in type of sport are not as useful. Doctors, athletic trainers and any athlete who want to monitor their red blood cell count could hopefully find this information useful.

# R APPENDIX

## Topic 1

**PART A**

**A) PLOT**

```
transform2 <- read.csv("/Users/Julia/Downloads/Transform2.csv")
plot(TransformTEMP)
the.model$coefficients
```

**B) Testing for Normality**

**1) QQ Plot**

```
qqnorm(the.model$residuals)
qqline(the.model$residuals)
```

**2) Histogram of Errors**

```
hist(the.model$residuals, main = "Residuals", xlab = "ei",pch = 19,font = 2,font.lab = 2,cex = 1.25)
```

**3) Shapiro-Wilkes**

```
the.model = lm(Y ~ X1,data = transform2)
ei = the.model$residuals
the.SWtest = shapiro.test(ei)
the.SWtest
```

**C) Testing for Constant Variance**

**1) Errors vs. Fitted Values**

```
plot(transform2$X,the.model$residuals, main = "Scatterplot",pch = 19,font = 2,font.lab = 2,cex = 1.25, ylab = "ei values",xlab = "X variable")
abline(h = 0, lwd = 2, col = "purple")
```

**2) Fligner-Killeen**

```
Group = rep("Lower",nrow(BC.data))
Group[BC.data$Y > median(BC.data$Y)] = "Upper"
Group = as.factor(Group)
BC.data$Group = Group
BC.data$ei = BC.model$residuals
the.FKtest= fligner.test(BC.data$ei, BC.data$Group)
the.FKtest
```

**D) Testing for Linear Relationship**

```
smaller.model = lm(Y ~ 1, data = transform2)
anova.small = anova(smaller.model)
larger.model = lm(Y ~ X1, data = transform2)
anova.large = anova(larger.model)
anova(smaller.model,larger.model)
```

**E) R Squared and Adjusted R Squared**

```
the.R2adj = summary(the.model)$adj.r.squared
the.R2adj
```

**PART B**

**A) Before Removing Outliers**

```
        par(mfrow = c(1,2))
        small.model = lm(Y ~ X1, data = new.transform2)
        plot(new.transform2$X1, new.transform2$Y)
        qqnorm(small.model$residuals)
        qqline(small.model$residuals)
```

**B) Transformations Before Removing Outliers**

**1) Box Cox**

```
        library(MASS)
        BC = boxcox(small.model,lambda = seq(-6,6,0.1),plotit = FALSE)
        lambda = BC$x[which.max(BC$y)]
        lambda
        BC.Y = (new.transform2$Y^lambda - 1)/lambda
        BC.data = data.frame(Y = BC.Y, X1 = new.transform2$X1)
        par(mfrow = c(1,2))
        BC.model = lm(Y ~ X1, data = BC.data)
        plot(BC.data$X1, BC.data$Y)
        qqnorm(BC.model$residuals)
        qqline(BC.model$residuals)
```

**2) Tukey**

**a) Transform X**

```
Tukey.X.Data = data.frame(Y = new.transform2$Y,X1 = tukeyX)
par(mfrow = c(1,2))
Tukey.X.model = lm(Y ~ X1, data = Tukey.X.Data)
```

```
plot(BC.data$X1, BC.data$Y)
qqnorm(Tukey.X.model$residuals)
qqline(Tukey.X.model$residuals)
```

**b) Transform Y**

```
Tukey.Y.Data = data.frame(Y = tukeyY,X1 = new.transform2$X1)
par(mfrow = c(1,2))
Tukey.Y.model = lm(Y ~ X1, data = Tukey.Y.Data)
plot(Tukey.Y.Data$X1, Tukey.Y.Data$Y)
qqnorm(Tukey.Y.model$residuals)
qqline(Tukey.Y.model$residuals)
```

**c) Transform X & Y**

```
library(rcompanion)
par(mfrow = c(1,2))
small.model = lm(Y ~ X1, data = new.transform2)
plot(new.transform2$X1, new.transform2$Y)
qqnorm(small.model$residuals)
qqline(small.model$residuals)
tukeyY = transformTukey(new.transform2$Y, plotit = FALSE)
tukeyX = transformTukey(new.transform2$X1, plotit = FALSE)
par(mfrow = c(1,2))
T.Data = data.frame(Y = tukeyY,X1 = tukeyX)
T.model = lm(Y ~ X1, data = T.Data)
plot(T.Data$X1, T.Data$Y)
qqnorm(T.model$residuals)
qqline(T.model$residuals)
```

**C) Outliers**

```
install.packages("leaps")
install.packages("MPV")
library(leaps)
library(MPV)

best.model = T.model

ei.s = best.model$residuals/sqrt(sum(best.model$residuals^2)/(nrow(new.transform2) - length(best.model$coefficients)))
ei.s
ri = rstandard(best.model)
ti = rstudent(best.model)

alpha = 0.1 ; n = nrow(new.transform2); p = length(best.model$coefficients)
cutoff = qt(1-alpha/(2*n), n -p )
cutoff.deleted = qt(1-alpha/(2*n), n -p -1 )

outliers = which(abs(ei.s)> cutoff | abs(ri) > cutoff | abs(ti) > cutoff.deleted)
new.transform2[outliers,]

alpha = 0.5 ; n = nrow(new.transform2); p = length(best.model$coefficients)
cutoff = qt(1-alpha/(2*n), n -p )
cutoff.deleted = qt(1-alpha/(2*n), n -p -1 )

outliers = which(abs(ei.s)> cutoff | abs(ri) > cutoff | abs(ti) > cutoff.deleted)
new.transform2[outliers,]
```

**PART C**

**Shapiro-Wilkes**

```
the.model.Tdata = lm(Y ~ X1,data = T.Data)
ei.Tdata = T.model$residuals
the.SWtest.Tdata = shapiro.test(ei.Tdata)
the.SWtest.Tdata
```

**Fligner-Killeen**

```
BC.data.Tdata = data.frame(Y = T.Data$Y, X1 = T.Data$X1)
BC.model.Tdata = lm(Y ~ X1, data = BC.data.Tdata)

Group = rep("Lower",nrow(BC.data.Tdata))
Group[BC.data.Tdata$Y > median(BC.data.Tdata$Y)] = "Upper"
Group = as.factor(Group)
BC.data.Tdata$Group = Group
BC.data.Tdata$ei = BC.model.Tdata$residuals
the.FKtest.Tdata= fligner.test(BC.data.Tdata$ei, BC.data.Tdata$Group)
the.FKtest.Tdata
```

**F-Test**

```
smaller.model.Tdata = lm(Y ~ 1, data = T.Data)
anova.small.Tdata = anova(smaller.model.Tdata)
larger.model.Tdata = lm(Y ~ X1, data = T.Data)
anova.large.Tdata = anova(larger.model.Tdata)
anova(smaller.model.Tdata,larger.model.Tdata)
```

## TOPIC 2

**A. Summary Statistics**

```
summary(new.athlete)
```

**B. Scatterplots/ Boxplots:**

i. **Red blood cell count vs lean body mass**

```
ggplot(new.athlete,aes(X1,Y)) + geom_point(shape = 19) + geom_smooth(method='lm',se= FALSE) + ggtitle("Y vs X1") + ylab("red blood cell count(L)") + xlab("leam body mass (kg)")
ggplot(new.athlete, aes(y=Y, x = factor("X1")))+ geom_boxplot() + ylab("Red blood cell count") + xlab("lbm")+ coord_flip() + ggtitle("Y vs X1")
```

ii. **Red blood cell count vs body mass index**

```
ggplot(new.athlete,aes(X2,Y)) + geom_point(shape = 19) + geom_smooth(method='lm',se= FALSE) + ggtitle("Y vs X2") + ylab("red blood cell count(L)") + xlab("body mass index (kg)")
ggplot(new.athlete, aes(y=Y, x = factor("X2")))+ geom_boxplot() + ylab("Red blood   cell     count") + xlab("")+ coord_flip() + ggtitle("")
```

iii. **Red blood cell count vs percent body fat**

```
ggplot(new.athlete,aes(X3,Y)) + geom_point(shape = 19) + geom_smooth(method='lm',se= FALSE) + ggtitle("Y vs X3") + ylab("red blood cell count(L)") + xlab("percent body fat")
ggplot(new.athlete, aes(y=Y, x = factor("X3")))+ geom_boxplot() + ylab("Red blood cell count") + xlab("")+ coord_flip() + ggtitle("")
```

iv. **Red blood cell count vs plasma ferritins**

```
ggplot(new.athlete,aes(X4,Y)) + geom_point(shape = 19) + geom_smooth(method='lm',se= FALSE) + ggtitle("Y vs X4") + ylab("red blood cell count(L)") + xlab("Plasma Ferritins (ng)")
ggplot(new.athlete, aes(y=Y, x = factor("X4")))+ geom_boxplot() + ylab("Red blood cell count") + xlab("")+ coord_flip() + ggtitle("Y vs X4")
```

v. **Red blood cell count vs sex**

```
ggplot(new.athlete, aes(y=Y, x = factor(X5)))+ geom_boxplot() + ylab("Red blood cell count") + xlab("Sex")+ coord_flip() + ggtitle(" Y vs X5")
```

vi. **Red blood cell vs new sports**

```
ggplot(new.athlete, aes(y=Y, x = factor(X6)))+ geom_boxplot() + ylab("Red blood cell count") + xlab("Newsport")+ coord_flip() + ggtitle(" Y vs X6")
```

**C. Equation of the line**

```
new.athlete = data.frame(rcc = athlete[,1], lbm = athlete[,2], bmi = athlete[,3], pcBfat =  athlete[,4], ferr=   athlete[,5],sex=athlete[,6], newsport=athlete[,7])

  names(new.athlete) = c("Y","X1","X2","X3","X4","X5","X6")
  new.athlete.model = lm(Y ~ X1 + X2 + X3 + X4 + X5 + X6, data =new.athlete)
  new.athlete.model
```

**III. Model Selection**

**A. Stepwise regression**

```
full.model = lm(Y ~X1+X2+X3+X4+X5+X6, data =new.athlete)
empty.model = lm(Y ~ 1, data =new.athlete)
n = nrow(new.athlete)
```

```
forward.model.AIC = stepAIC(empty.model, scope = list(lower = empty.model, upper= full.model), k = 2,direction =
"forward",trace = FALSE)
forward.model.BIC = stepAIC(empty.model,  scope = list(lower = empty.model, upper= full.model), k =
log(n),trace=FALSE,direction = "forward")
forward.model.AIC$coefficients
forward.model.BIC$coefficients

backward.model.AIC = stepAIC(full.model, scope = list(lower = empty.model, upper= full.model), k = 2,direction =
"backward",trace = FALSE)
backward.model.BIC = stepAIC(full.model,  scope = list(lower = empty.model, upper= full.model), k =
log(n),trace=FALSE,direction = "backward")

FB.model.AIC = stepAIC(empty.model, scope = list(lower = empty.model, upper= full.model), k = 2,direction = "both",trace
= FALSE)
FB.model.BIC = stepAIC(empty.model,  scope = list(lower = empty.model, upper= full.model), k =
log(n),trace=FALSE,direction = "both")

BF.model.AIC = stepAIC(full.model, scope = list(lower = empty.model, upper= full.model), k = 2,direction = "both",trace =
FALSE)
BF.model.BIC = stepAIC(full.model,  scope = list(lower = empty.model, upper= full.model), k =
log(n),trace=FALSE,direction = "both")

forward.model.AIC$coefficients
backward.model.AIC$coefficients
FB.model.AIC$coefficients
BF.model.AIC$coefficients

forward.model.BIC$coefficients
backward.model.BIC$coefficients
FB.model.BIC$coefficients
BF.model.BIC$coefficients

best.model=FB.model.BIC
best.model
```

## IV. Diagnostics.

### A. Outliers

```
cutoff = 6.70
outliers = which(new.athlete$Y > cutoff |  new.athlete$Y < -cutoff)
new.removed.outliers.data = new.athlete[-outliers,]
new.removed.outliers.data
newfinaldata.model = lm(Y ~ X5 + X6, data = new.removed.outliers.data)
new.removed.outliers.data$ei = newfinaldata.model$residuals
new.removed.outliers.data$yhat = newfinaldata.model$fitted.values
new.removed.outliers.data$ei
new.removed.outliers.data$yhat
```

### Leverage Points

```
all.values = influence.measures(best.model)$infmat
colnames(all.values)
lev.hat = which(all.values[,"hat"] >2*p/n)
new.removed.outliers.data[lev.hat,]
lev.DF = which(all.values[,"dffit"] >1)
new.removed.outliers.data[lev.DF,]
lev.DF = which(all.values[,"cook.d"] >qf(0.50,4,197))
new.removed.outliers.data[lev.DF,]
```

### B. Diagnostic Plots

#### i. Error vs Fitted values

```
qplot(yhat, ei, data = new.removed.outliers.data) +  ggtitle("Errors vs. Fitted    Values") + xlab("Fitted Values") +
ylab("Errors") + geom_hline(yintercept = 0,col = "purple")
```

#### ii. QQ plot

```
qqnorm(newfinaldata.model$residuals)
```

```
qqline(newfinaldata.model$residuals)
```

### C. Shapiro Wilks normality test

```
ei = newfinaldata.model$residuals
the.SWtest = shapiro.test(ei)
the.SWtest
```

### D. Fligner-Killeen test

```
Group = rep("Lower",nrow(new.removed.outliers.data))
Group[new.removed.outliers.data$Y > median(new.removed.outliers.data$Y)] = "Upper"
Group = as.factor(Group)
new.removed.outliers.data$Group = Group
the.FKtest= fligner.test(new.removed.outliers.data$ei, new.removed.outliers.data$Group)
the.FKtest
```

### E. F-test

#### i. Drop X5 from the full model

```
smaller.model = lm(Y ~ X6,  data = new.removed.outliers.data)
anova.small = anova(smaller.model)
larger.model = lm(Y ~ X5 + X6 ,data = new.removed.outliers.data)
anova.large = anova(larger.model)
anova(smaller.model,larger.model)
```

#### ii. Drop X6 from full model

```
smaller.model = lm(Y ~ X5,  data = new.removed.outliers.data)
anova.small = anova(smaller.model)
larger.model = lm(Y ~ X5 + X6 ,data = new.removed.outliers.data)
anova.large = anova(larger.model)
anova(smaller.model,larger.model)
```

## IV. Analysis

### A. Partial $R^2$

```
small.model=lm(Y ~ X6, data = new.removed.outliers.data)
big.model = lm(Y~X5 + X6, data = new.removed.outliers.data )
Partial.R2(small.model, big.model)
small.model=lm(Y ~ X5, data = new.removed.outliers.data)
big.model = lm(Y~X5 + X6, data = new.removed.outliers.data )
Partial.R2(small.model, big.model)
```

### B. Coefficient of Determination

```
a=summary(newfinaldata.model)
a
```

### C. Confidence Intervals(Simultaneous)

#### 1. Multiple CI’s for Betas

```
alpha =0.05
SCI =confint(newfinaldata.model,level = 1-alpha/4)
SCI
```

#### 2. Prediction

```
all.of.them = mult.fun(nrow(new.removed.outliers.data), length(newfinaldata.model$coefficients), 3, 0.05)
all.of.them
all.of.them = mult.fun(nrow(salary), length(salary.model$coefficients), 3, 0.01)
xs = data.frame(X2=c(19,24,31),X5 = c("m","f","f"), X6=c("Run","Swim", "Net"))
Mul=mult.fun(nrow(new.removed.outliers.data),length(newfinaldata.model$coefficients), 3, 0.05)
Mul
keep = Mul[1]
all.the.CIs = mult.CI(all.of.them[1], xs,newfinaldata.model,0.05,"prediction")
cbind(xs,all.the.CIs)
cbind
```