STA 108 Project II

Rue Ekpemiro

Professor Maxime Puoukam

Topic 1: Transformation of Variables (Question 2)

I. Introduction

The data consists of the average number of beds in a hospital during a study period as the explanatory variable and the length of stay for a patient at the hospital, in days, as the response variable. We will run diagnostics for a simple regression model on the original, outliers removed, transformed, transformation plus outliers removed versions of the data with an alpha of 0.05. Finally, we will discuss which is the best dataset to move forward with for a client.

II. Original Dataset

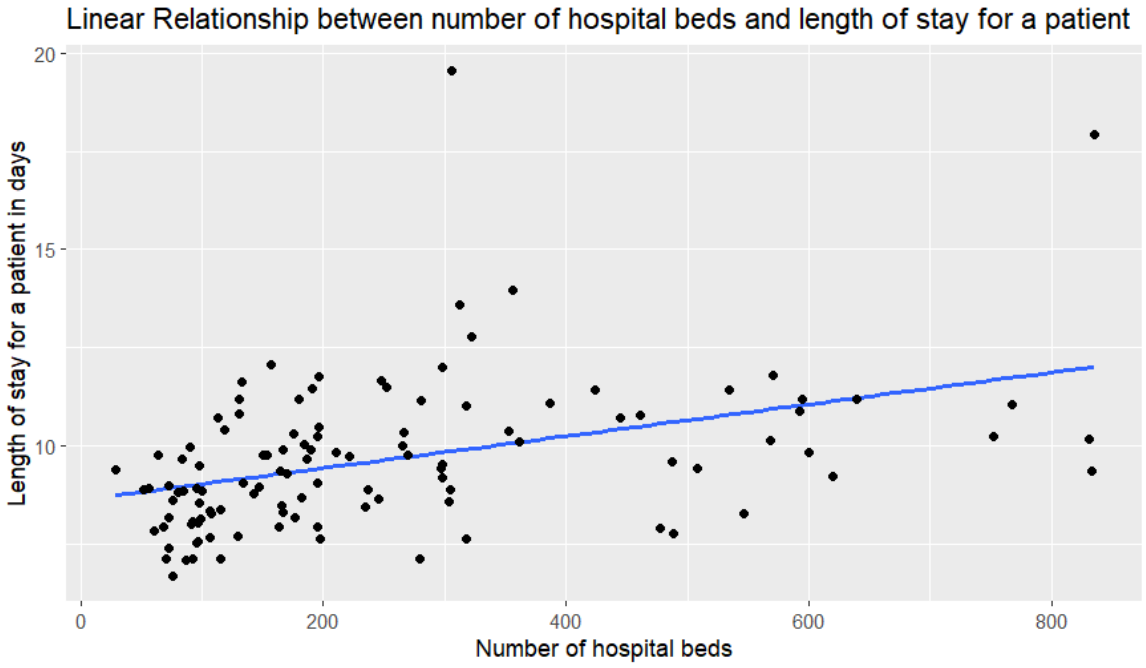
We will start with reporting the fitted regression line of the original dataset, then we will use diagnostic plots and tests to assess goodness of fit.

II.1 Regression Line

The fitted estimated regression line is:

$$\widehat{Y} = 8.625364 + 0.004057X_1$$

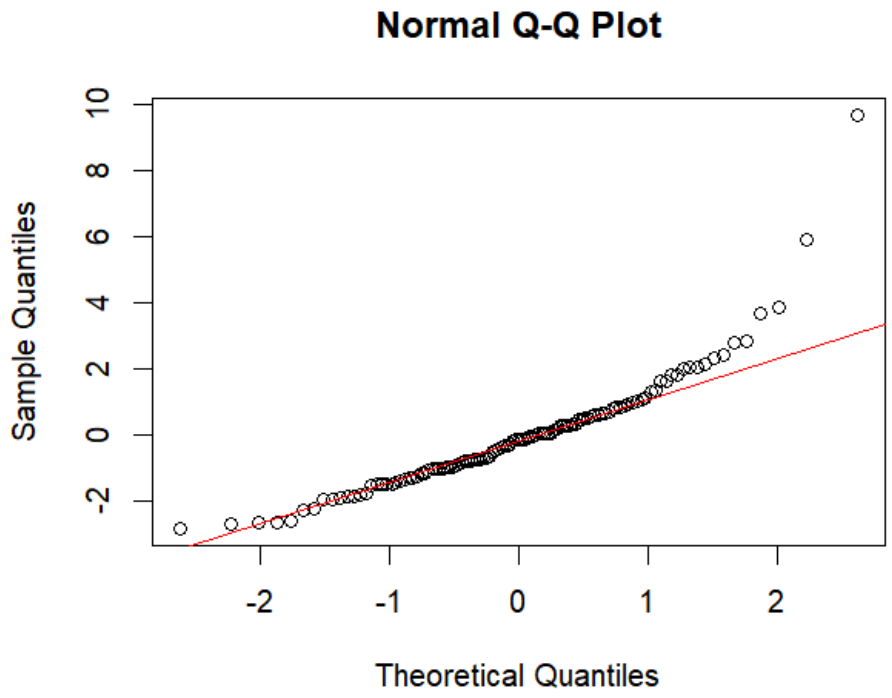
We can also fit the regression line on the scatter plot of our X and Y variables.



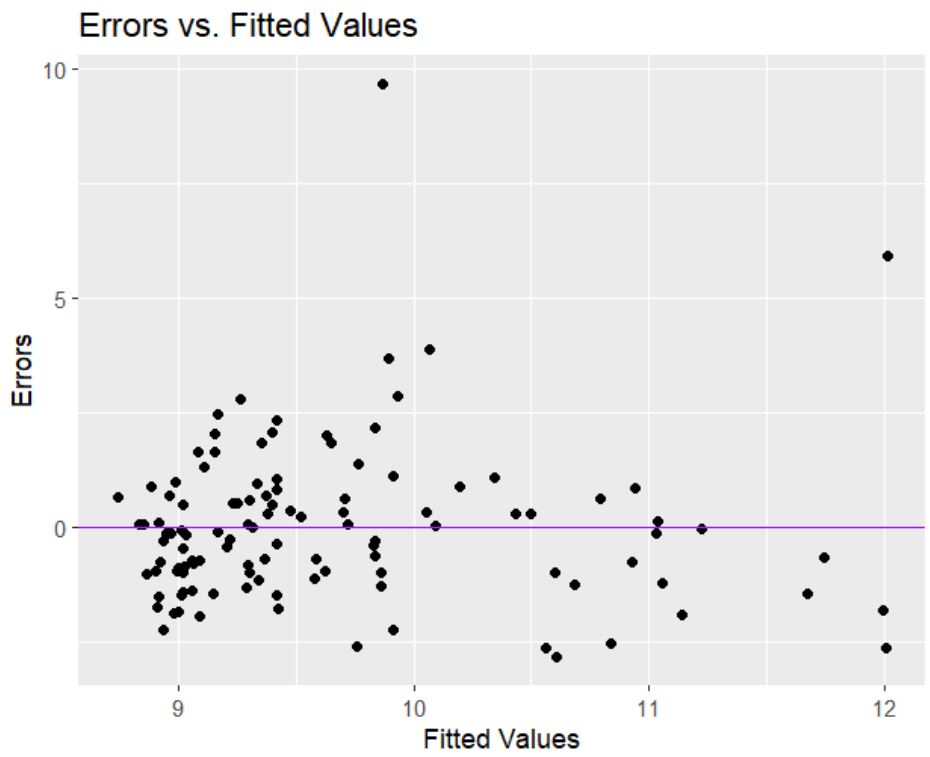
The scatterplot suggests a positive linear relationship between number of hospital beds and length of stay. It shows two potential outliers far from the regression line that may affect linearity, specifically the points where length of stay is over 17 days.

II.2 Diagnostic Plots

The QQ Plot and Errors vs. Fitted Values plots will be used to visually detect any potential violation to our model's assumptions of normality and constant variance.



The QQ Plot above shows excessive deviation of points from the line along the right-tail. Ideally the points would lay along the line if the residuals were perfectly normal. In reality, some deviation is expected, but our plot suggests that the residuals do not follow a normal distribution due to the large skew of points and outliers present. Next, we will look at the Errors vs. Fitted values plot below.



The errors vs. fitted values plot helps detect non-constant variance by looking at the vertical spread of the points. Ideally vertical spread should be consistent with no pattern present, but our plot suggests that the residuals do not have constant variance because of the plot's curved shape and two clear outliers that disrupt the spread.

II.3 Diagnostic Tests

To know if the non-normality and non-constant variance seen is significant we will run Shapiro Wilks and Fligner-Killeen tests in R. With the null and alternative hypotheses being:

SW, H_0: $The\ residuals\ follow\ a\ normal\ distribution,\ H_A$: $The\ residuals\ don't\ follow\ a\ normal\ distribution$

$FK,\ H_0$: $The\ residuals\ have\ constant\ variance,\ H_A$: $The\ residuals\ don't\ have\ constant\ variance$

For Shapiro-Wilks, we obtain a p-value of 1.775e-08. Thus, we reject the null and conclude that the residuals do not follow a normal distribution. For the Fligner-Killeen test, we obtain a

p-value of 0.03383. Thus, we reject the null hypothesis and conclude that the residuals do not have constant variance.

III. Data Transformation

For the outliers, we used the standardized/studentized residuals method to detect four rows with the cutoff of 1.98 and removed them. Since the original dataset violated normality and constant variance, our Y variable was transformed using a Box-Cox transformation. The optimal lambda was -1.3. Below is a table summarizing the results of the Shapiro Wilks and Fligner-Killeen tests for the original, outliers removed, transformed, outliers removed + transformed datasets.

Table 1: Diagnostic Test Results for each Dataset

	Original	Outliers Removed	Transformed	Outliers Removed + Transformed
Shapiro-Wilks	1.78E-08	0.2125	0.7264	0.2125
Fligner-Kileen	0.03383	0.2047	0.5572	0.1034

Based on the results in the table above, we fail to reject the previously stated null hypotheses and conclude that the errors are normally distributed and have constant variance across all manipulated datasets. Among these, the transformed data maximizes p-values for both normality and constant variance the most, followed by the dataset with outliers removed.

IV. Discussion

The transformation of Y did help to remedy the violation of our assumptions and provided the high p-values for both diagnostic tests. There are downsides to transforming our data; the units of length of hospital stay are no longer simply in days but in Box-Cox transformed units. This makes interpretation more difficult. Given the transformed dataset with outliers is

statistically the best fit, I would first recommend the client to use that for simple linear regression if changes to interpretation isn't an issue. If interpretation of Y in the original units is more important to the client, I would recommend the outliers removed dataset since it passed both tests of normality and constant variance.

Topic 2: Multiple Regression Modeling (Question 1)

I. Introduction

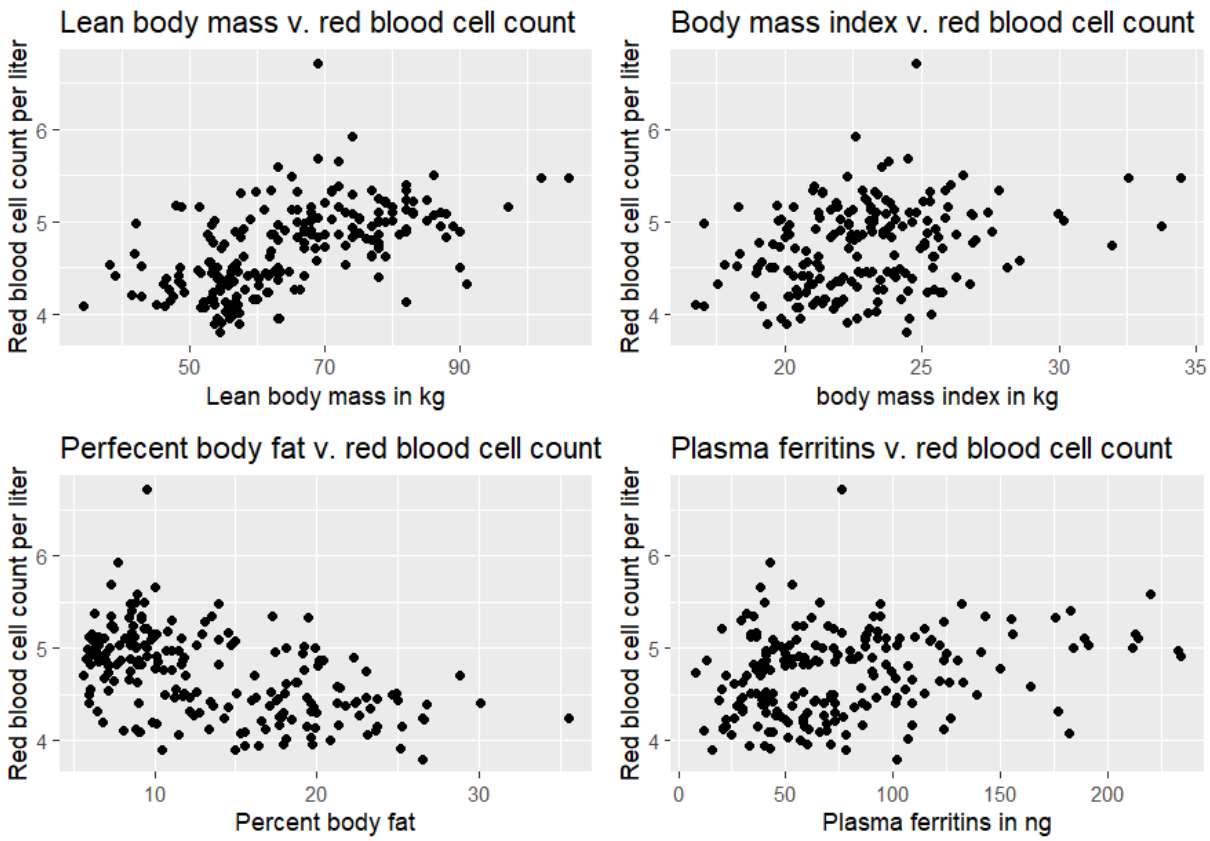
The data was a random sample taken from Australian athletes to model red blood cell count based on other physical characteristics. It consists of red blood cell count (per liter) as the response variable and six predictor variables. The goal of our analysis is to determine the "correct" model so we can answer the question: which predictors truly affect red blood cell count? To do this we will select the multiple regression model that best fits the data for interpretation. All analysis will be done with an alpha of 0.05.

II. Summary of Data

In this section we will look at scatter plots of the numerical predictors to assess linearity of the variables, and then boxplots of the categorical predictors to see the relationship between groups. Finally, we will use a correlation matrix to look at the correlation between predictors.

II.1 Scatter plots

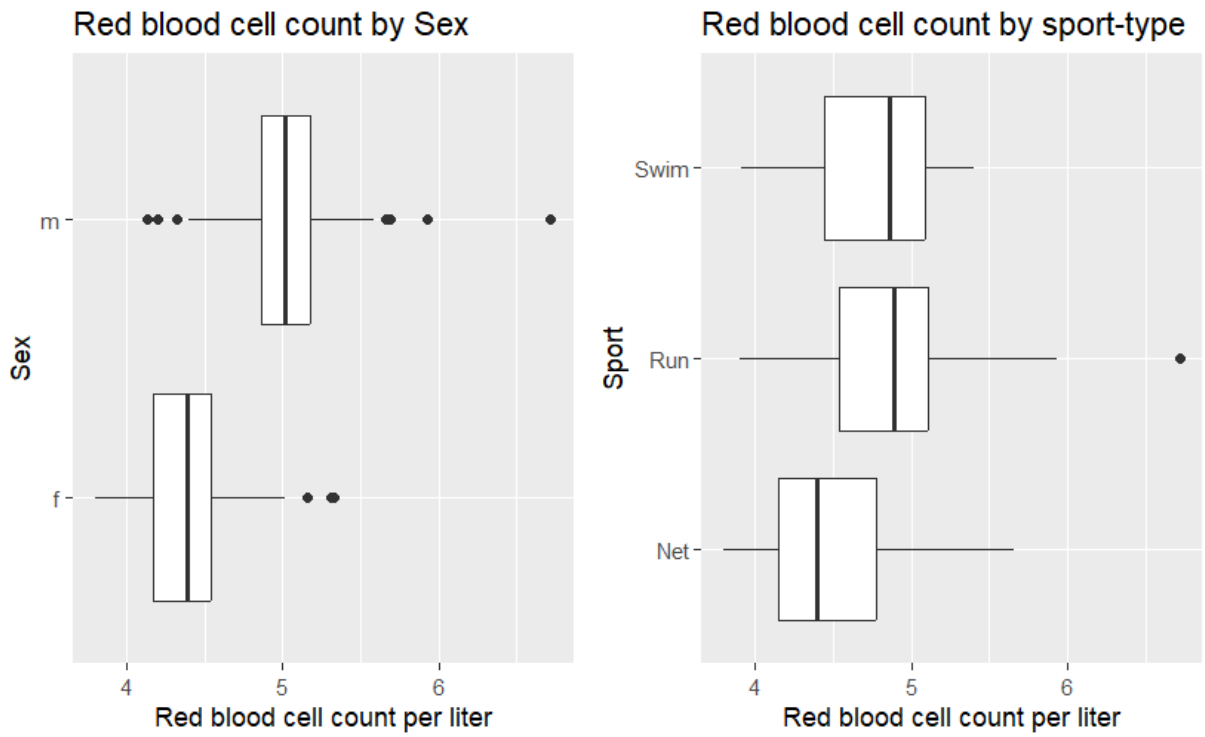
The scatter plots will be used to assess if there are linear trends between each of the numerical predictors and red blood cell count.



For the lean body mass and body mass index plots, we can see a suggested positive linear trend between those predictors and red blood cell count. The percent body fat plot suggests a negative linear trend, while plasma ferritins plot suggests that there is no linear relationship present due to its curved shape. All plots above also appear to have at least a few outliers present.

II.2 Boxplots

For the categorical predictors, we will use the boxplots below to analyze their spread.



First looking at red blood cell count by sex, we can see males typically have a higher red blood cell count than females with their median count at 5 per liter and females at about 4.4 per liter. The range of male red blood cell count is also greater and with more outliers at both ends present. In the red blood cell count by sport-type boxplots, it shows sports with running having the highest median count followed by swimming and net sports. The outlier at over 6 per liter is also clearly shown again here as previously mentioned in scatter plots.

II.3 Correlation Matrix

The correlation matrix will tell us if any predictors are linearly correlated using a gradient scale from -1 to 1.

From the matrix above, we can see that body max index and lean body mass is highly correlated with a coefficient of 0.71. When body mass increases, lean body mass tends to increase. This indicates that the matrix could be unstable and suggests multicollinearity.

III. Model Selection

In this section we will fit and discuss the full model before performing forward stepwise regression to select our final model using BIC criteria. Finally, we will report our final model.

III.1 Full Model

First, we will fit the full regression model to our data which is defined as:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_{p-1} X_{p-1} + \varepsilon$$

Where we have a total of $(p - 1)$ predictors that will estimate p total $\beta's$; the ε that accounts for deviation in our observations from the line. Since we are dealing with a sample we will fit the

linear model and derive the estimated regression line that will assist with our interpretations.

$$\widehat{Y} = 4.0408 + 0.0029X_1 + 0.0072X_2 - 0.0007X_3 - 0.0004X_4 + 0.5207X_{5,M} + 0.2043X_{6,Run} + 0.0942X_{6,Swim}$$

Here, $X_{5,M}$ is a dummy variable equal to 1 if the athlete is male and 0 if female. $X_{6,Run}$ is 1 if the athlete participates in running, $X_{6,Swim}$ is 1 for swimming, and 0 if in net sports.

Table 1: Summary of Full Regression Model

Term	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.0409	0.1954	20.6834	0.0000
lbm	0.0029	0.0049	0.5880	0.5572
bmi	0.0072	0.0195	0.3703	0.7115
pcBfat	−0.0007	0.0083	−0.0823	0.9345
ferr	−0.0004	0.0006	−0.7617	0.4472
sexm	0.5207	0.1045	4.9839	0.0000
newsportRun	0.2043	0.0774	2.6397	0.0090
newsportSwim	0.0943	0.0624	1.5121	0.1321

The table above tells us the estimates, standard error, test statistic, and p-value of each predictor. Based on the p-values in the summary table, the predictors $X_{5,M}$ and $X_{6,Run}$ are highly significant in the model. By fitting the model we also obtain the F statistic of 27.67 and p-value of less than 2.2e-16. Which confirms that at least one of the predictors is significantly related to red blood cell count. Additionally, we obtained an adjusted R-squared of 0.4815 meaning that the predictor variables explain 48.15% of the variance in red blood cell count.

III.2 Forward Stepwise Regression: BIC

Now we will perform forward selection using BIC criteria since this tends to underfit and our goal is to select a smaller model for interpretation. After computing in R, the lowest BIC achieved is -434.18 and we get a final model containing just $X_{5,M}$, $X_{6,Run}$, and $X_{6,Swim}$.

III.3 Final Model

The final model is summarized in the following table:

Table 2: Summary of Final Regression Model

Term	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.3203	0.0448	96.3628	0.0000
sexm	0.5801	0.0486	11.9421	0.0000
newsportRun	0.1998	0.0611	3.2722	0.0013
newsportSwim	0.1040	0.0592	1.7563	0.0806

Notably sex is the most significant factor with a p-value of ≈ 0, then sport-type when comparing running to net sports. When comparing swimming to to net sports, the predictor is not shown to be significant in our table, but since this is one dummy variable for our categorical X_6 it is kept in the model.

The final model can be written as:

$$RBCC = 4.0408 + 0.5207 \times Sex(Male) + 0.2043 \times Newsport(Run) + 0.0942 \times Newsport(Swim)$$

$[Red\ Blood\ Cell\ Count = RBCC]$.

IV. Diagnostics

In this section we will assess if our final selected model violated any of the multiple regression model assumptions and remove outliers if needed. The assumptions are as follows: all

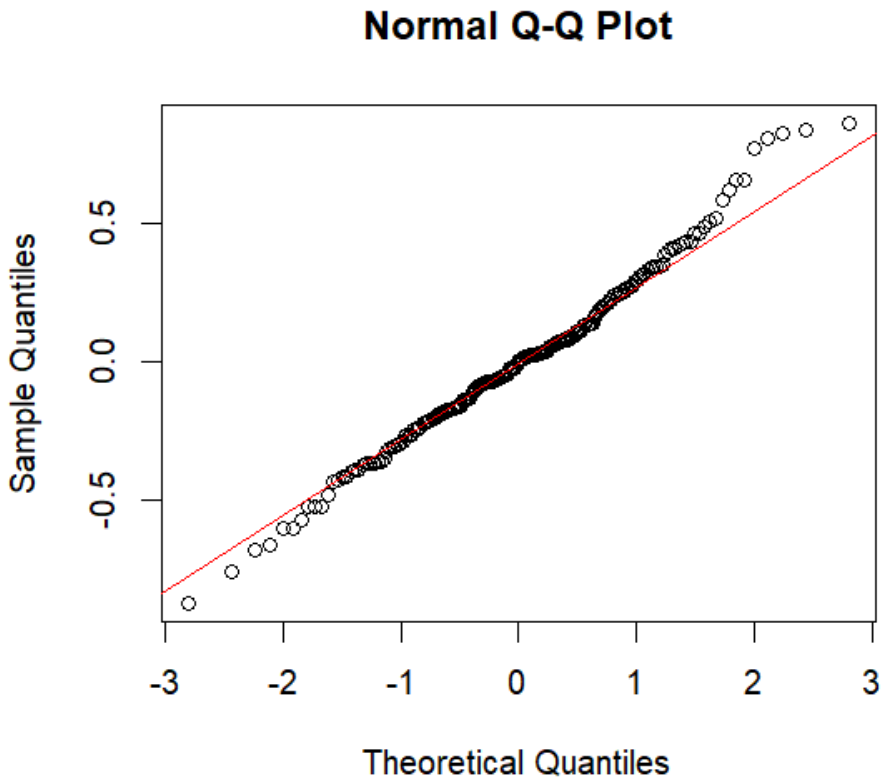
observation vectors are independent, X_1,..., X_{p-1} are not random variables, and the errors follow a normal distribution with constant variance. Here we will focus on the assumptions of normality and constant variance.

IV.1 Outlier Detection and Removal

The studentized/standardized residual method with Bonferroni correction was used to obtain the t-cutoff of 3.54. Using the cutoff we detected 1 outlier and will remove it from our dataset as it may harm the assumptions that we will be testing. No leverage points were also detected.

IV.2 Assessing Normality

To check if our assumption of normality has been met, we will use a QQ Plot and a Shapiro Wilks test.



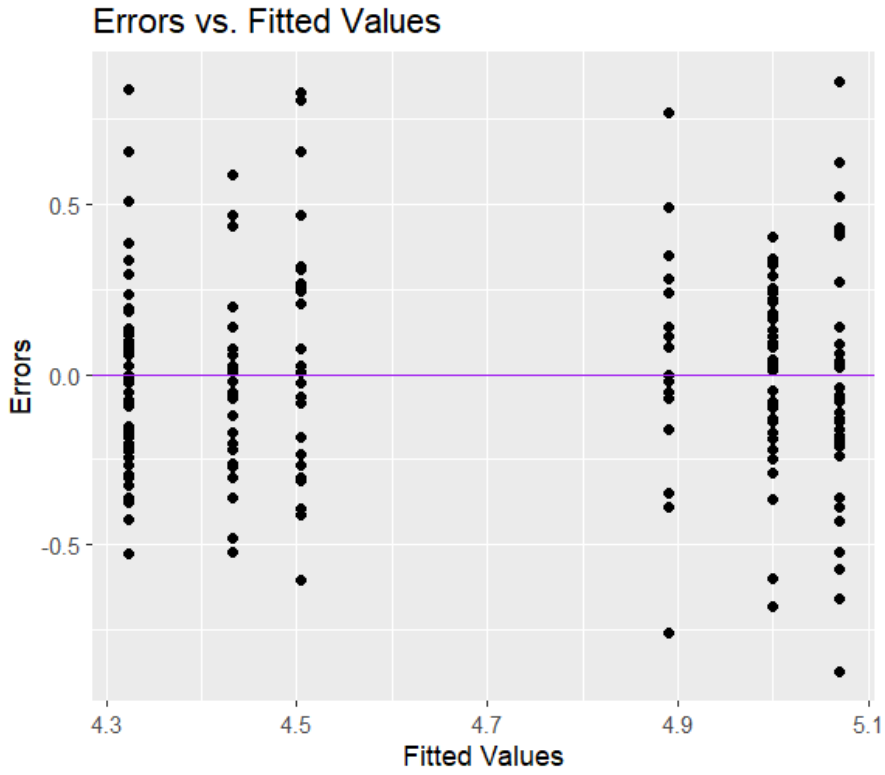
The QQ Plot above suggests that the residuals are approximately normal with only minimal deviation of the points at both tails present. To formally test this observation we will run a Shapiro-Wilks test with the null and alternative hypotheses being:

H_0: *The residuals follow a normal distribution,* H_A: *The residuals don't follow a normal distribution*

From the test we get a p-value of 0.1484. Therefore, we fail to reject the null and conclude that the residuals follow a normal distribution.

IV.3 Assessing Homoscedasticity

To check if our assumption of constant variance has been met we will use an Error v. Fitted values plot and the Brown Forsythe test since our residuals are normal.



The above plot suggests that the residuals do have constant variance due to the similar vertical spread of the points across the plot. To formally test this we will run the Brown-Forsythe test with the null and alternative hypotheses being:

H_0: $The\ residuals\ have\ constant\ variance,\ H_A$: $The\ residuals\ do\ not\ have\ constant\ variance$

We obtain the p-value of 0.6416. Thus, we fail to reject the null and conclude that the residuals have constant variance.

V. Analysis & Interpretation

Since our goal was to identify the model with predictors that truly affect red blood cell count, we will now interpret the final model and estimate the effects of these predictors. Simultaneous confidence intervals for b_i will then be constructed to further confirm our findings.

V.1 Model Interpretation

To recall, our selected final model is:

$$RBCC = 4.0408 + 0.5207 \times Sex(Male) + 0.2043 \times Newsport(Run) + 0.0942 \times Newsport(Swim)$$

The beta estimates are interpreted as:

b_0 = the average red blood cell count of female athletes that play net sports is 4.0408 per liter.

b_1 = male athletes have a higher red blood cell count than female athletes on average by 0.5207 per liter (holding sport-type constant).

b_2 = Athletes involved in running sports have on average a higher red blood cell count than those who do net sports by 0.2043 per liter (holding gender constant).

b_3 = Athletes involved in swimming sports have on average a higher red blood cell count than those who do net sports by 0.0942 per liter (holding gender constant).

The regression lines for each sport-type and gender, multiple regression lines can also be plotted in 2 plots.

V.3 Confidence Intervals

The simultaneous confidence intervals for each beta estimated is formulated as:

$$b_i \pm C^* s\{b_i\}$$

We will use the computed Bonferroni correction of 2.415 for the narrowest interval. The intervals are summarized in the table below.

Table 3: Simultaneous Confidence Intervals

Term	0.833 %	99.167 %
(Intercept)	4.2225	4.4257
sexm	0.4559	0.6763
newsportRun	0.0413	0.3185
newsportSwim	−0.0257	0.2426

To interpret,

b_0 = We are overall 95% confident that the average red blood cell count of female athletes that play net sports is between 4.225 and 4.4257 per liter.

b_1 = We are overall 95% confident that male athletes have a higher red blood cell count than female athletes on average by between 0.4559 and 0.6763 per liter, holding sport-type constant.

b_2 = We are overall 95% confident that athletes involved in running sports have on average a higher red blood cell count than those who do net sports by between 0.0.0413 and 0.3185 per liter, holding gender constant.

b_3 = We are overall 95% confident that the average difference in average red blood cell count between athletes involved in swimming and those in net sports is between −0.0257 and 0.2426 per liter, holding gender constant. Since this interval includes zero, we cannot conclude that

swimming is associated with a significantly higher or lower red blood cell count compared to net sports.

VI. Conclusion

Based on our model criteria and selection, our true model was the model including only the variables sex and type of sport to predict red blood cell count. Specifically, our model found that male athletes tend to have higher red blood cell count than female athletes. Those who participated in running sports also tend to have higher red blood cell count when compared to those who play net sports. While swimming also showed a small positive association, it was not statistically significant when compared to net sports.

This final model was made to favor simplicity and interpretability. Diagnostic checks confirmed that this model met the assumptions of normality and constant variance. Simultaneous confidence intervals for the estimated betas supported our findings made during model selection. In summary, red blood cell count appears to be most strongly influenced by biological sex and sport type, with the predictor sex having the largest effect.

Appendix: R Script

```
Transform2 <- read.csv("C:/Users/oekpe/OneDrive/Desktop/STA 108/projects/Transform2.csv")

the.model <- lm(Y~X,Transform2)
summary(the.model)
#QQ Plot
qqnorm(the.model$residuals)
qqline(the.model$residuals, col='red')
hist(the.model$residuals, main = "Residuals", xlab = "ei",pch = 19,font = 2,font.lab = 2,cex = 1.25)
#Shapiro-Wilks
ei = the.model$residuals
SWtest = shapiro.test(ei)
SWtest
library(ggplot2)
#ei vs fitted values
Transform2$ei = the.model$residuals
Transform2$yhat = the.model$fitted.values

qplot(yhat, ei, data = Transform2) + ggtitle("Errors vs. Fitted Values") + xlab("Fitted Values") +
ylab("Errors") + geom_hline(yintercept = 0,col = "purple")
#FK Test
new.Group = rep("Lower",nrow(Transform2)) #Creates a vector that repeats "Lower" n times
new.Group[Transform2$Y > median(Transform2$Y)] = "Upper" #Changing the appropriate values to "Upper"
new.Group = as.factor(new.Group) #Changes it to a factor, which R recognizes as a grouping variable.
Transform2$new.Group = new.Group
new.FKtest= fligner.test(Transform2$ei, Transform2$new.Group)
new.FKtest
#Outlier Detection and Removal
alpha = 0.05
n = nrow(Transform2)
t.cutoff= qt(1-alpha/(2), n-2)
t.cutoff

rij = rstandard(the.model)
CO.rij = which(abs(rij) > t.cutoff)
CO.rij

new.data <- Transform2[-CO.rij,]
new.model <- lm(Y~X,new.data)
#Shapiro-Wilks
new.data$yhat = new.model$fitted.values

n.ei = new.model$residuals
n.SWtest = shapiro.test(n.ei)
n.SWtest
#FK test
new.data$n.ei = new.model$residuals
new.Group = rep("Lower",nrow(new.data)) #Creates a vector that repeats "Lower" n times
new.Group[new.data$Y > median(new.data$Y)] = "Upper" #Changing the appropriate values to "Upper"
new.Group = as.factor(new.Group) #Changes it to a factor, which R recognizes as a grouping variable.
new.data$new.Group = new.Group
```

```r
new.FKtest= fligner.test(new.data$n.ei, new.data$new.Group)
new.FKtest
#Transformation
library(MASS)
BC = boxcox(the.model,lambda = seq(-6,6,0.1),plotit = FALSE)
lambda = BC$x[which.max(BC$y)]
lambda

BC.Y = (Transform2$Y^lambda - 1)/lambda
BC.data = data.frame(Y = BC.Y, X1 = Transform2$X)

par(mfrow = c(1,2))
BC.model = lm(Y ~ X1, data = BC.data)
plot(BC.data$X1, BC.data$Y)
#Shapiro Wilks
t.ei = BC.model$residuals
t.SWtest = shapiro.test(t.ei)
t.SWtest
#FK test
BC.data$t.ei = BC.model$residuals
new.Group = rep("Lower",nrow(BC.data)) #Creates a vector that repeats "Lower" n times
new.Group[BC.data$Y > median(BC.data$Y)] = "Upper" #Changing the appropriate values to "Upper"
new.Group = as.factor(new.Group) #Changes it to a factor, which R recognizes as a grouping variable.
BC.data$new.Group = new.Group
new.FKtest= fligner.test(BC.data$t.ei, BC.data$new.Group)
new.FKtest
#Outlier removal of transformed dataset
r.data <- BC.data[-CO.rij,]
r.model <- lm(Y~X1,r.data)
#Shapiro Wilks
r.ei = new.model$residuals
r.SWtest = shapiro.test(r.ei)
r.SWtest
#FK test
r.data$r.ei = r.model$residuals
new.Group = rep("Lower",nrow(r.data)) #Creates a vector that repeats "Lower" n times
new.Group[r.data$Y > median(r.data$Y)] = "Upper" #Changing the appropriate values to "Upper"
new.Group = as.factor(new.Group) #Changes it to a factor, which R recognizes as a grouping variable.
r.data$new.Group = new.Group
r.FKtest= fligner.test(r.data$r.ei, r.data$new.Group)
r.FKtest
#regression line
the.model

ggplot(Transform2,aes(X,Y)) + geom_smooth(method='lm',se = FALSE)+geom_point(shape = 19) + ggtitle("Numb
```

```r
athelete <- read.csv("C:/Users/oekpe/OneDrive/Desktop/STA 108/projects/athelete.csv")
#Scatter plots

library(patchwork)
library(ggplot2)
a <- ggplot(athelete,aes(lbm,rcc)) + geom_point(shape = 19) + ggtitle("Lean body mass v. red blood cell
b <- ggplot(athelete,aes(bmi,rcc)) + geom_point(shape = 19) + ggtitle("Body mass index v. red blood cel
c <- ggplot(athelete,aes(pcBfat,rcc)) + geom_point(shape = 19) + ggtitle("Perfecent body fat v. red blo
d <- ggplot(athelete,aes(ferr,rcc)) + geom_point(shape = 19) + ggtitle("Plasma ferritins v. red blood c
(a+b)/(c+d)
#Correlation Matrix

n.data <- subset(athelete, select = -c(6,7))

library(corrplot)
datamatrix <- cor(n.data)
corrplot(datamatrix,method = "number")

# Boxplots
e <- ggplot(athelete, aes(y=rcc, x = sex))+ geom_boxplot() + ylab("Red blood cell count per liter")+ xla
f <- ggplot(athelete, aes(y=rcc, x = newsport))+ geom_boxplot() + ylab("Red blood cell count per liter")
e+f
#full model fit
full.model <- lm(rcc~.,athelete)
full.model

#Summary Table of full model
library(gt)
library(webshot2)
z <- coef(summary(full.model))
z_df <- as.data.frame(z)
z_df$Term <- rownames(z_df)
rownames(z_df) <- NULL
gt_table <- gt(z_df) |>
  cols_move_to_start(columns = Term) |>
  tab_header(
    title = "Table 1: Summary of Full Regression Model"
  )|>
  fmt_number(columns = where(is.numeric), decimals = 4)
gtsave(gt_table, "table.png")

summary(full.model)
full.model = lm(rcc~., athelete)
empty.model = lm(rcc~1, athelete)

n = nrow(athelete)
library(MASS)

forward.model.BIC = stepAIC(empty.model, scope = list(lower = empty.model, upper= full.model), k = log(n
forward.model.BIC$coefficients
final.model <- lm(rcc~sex+newsport, athelete)

#Summary Table of final model
```

```r
w <- coef(summary(final.model))
w_df <- as.data.frame(w)
w_df$Term <- rownames(w_df)
rownames(w_df) <- NULL
gt_table <- gt(w_df) |>
  cols_move_to_start(columns = Term) |>
  tab_header(
    title = "Table 2: Summary of Final Regression Model"
  )|>
  fmt_number(columns = where(is.numeric), decimals = 4)
gtsave(gt_table, "table2.png")

summary(full.model)
#Outlier Detection
alpha = 0.05
n = nrow(athelete)
t.cutoff= qt(1-alpha/(2), n-2)
t.cutoff

rij = rstandard(final.model)
CO.rij = which(abs(rij) > t.cutoff)
CO.rij
new.data <- athelete[, c("rcc", "sex", "newsport")]

best.model <- final.model
ei.s = best.model$residuals/sqrt(sum(best.model$residuals^2)/(nrow(new.data) - length(best.model$coeffi

ri = rstandard(best.model)

ti = rstudent(best.model)

par(mfrow =c(2,2))

hist(ei.s,main ="Semi-studentized/standardized residuals")
hist(ri,main="Studentized/Standardized residuals")
hist(ti, main="Deleted Residuals")

# identify outliers

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

cutoff.deleted = qt(1-alpha/(2*n), n -p -1 )
cutoff.deleted

outliers = which(abs(ei.s)> cutoff | abs(ri) > cutoff | abs(ti) > cutoff.deleted)
outliers

n.data <- new.data[-outliers,]
n.model <- lm(rcc~sex+newsport, n.data)
#qqplot
qqnorm(n.model$residuals)
```

```r
qqline(n.model$residuals, col ='red')
#Shapiro-Wilk test
ei = n.model$residuals
the.SWtest = shapiro.test(ei)
the.SWtest

n.data$ei = n.model$residuals
n.data$yhat = n.model$fitted.values

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

ei = n.model$residuals
Group = rep("Lower",nrow(n.data)) #Creates a vector that repeats "Lower" n times
Group[n.data$rcc > median(n.data$rcc)] = "Upper"
Group = as.factor(Group) #Changes it to a factor, which R recognizes as a grouping variable.
n.data$Group = Group
#the.FKtest= fligner.test(data.Hair$ei, data.Hair$Group)
#the.FKtest
library(car)
the.BFtest = leveneTest(ei~Group, data=n.data, center=median)
the.BFtest
mult.fun = function(n,p,g,alpha){
  bon = qt(1-alpha/(2*g), n-p)
  WH = sqrt(p*qf(1-alpha,p,n-p))
  Sch = sqrt(g*qf(1-alpha,g,n-p))
  all.mul = c(bon,WH,Sch)
  all.mul = round(all.mul,3)
  names(all.mul) = c("Bon","WH","Sch")
  return(all.mul)
}

mult.CI = function(C.star,x.stars,the.model,alpha,the.type = "confidence"){
  all.preds = predict(the.model,x.stars)
  if(the.type == "confidence"){
    all.se = predict(the.model,x.stars,interval = the.type,se.fit = TRUE)$se.fit
  } else if(the.type == "prediction"){
    all.se = predict(the.model,x.stars,interval = the.type,se.fit = TRUE)$se.fit
    MSE = sum(the.model$residuals^2)/(length(the.model$residuals) - length(the.model$coefficients))
    all.se = sqrt(all.se^2 + MSE)
  }
  LB = all.preds - C.star*all.se
  UB = all.preds + C.star*all.se
  all.CIs = cbind(LB,UB)
  colnames(all.CIs) = paste((1-alpha)*100, "%",c(" Lower"," Upper"), sep = "")
  results = cbind(all.preds,all.CIs)
  colnames(results)[1] = "Estimate"
  return(results)
}
all.of.them = mult.fun(nrow(n.data), length(n.model$coefficients), 3, 0.05)
all.of.them
alpha =0.05
SCI =confint(n.model,level = 1-alpha/3)
```

```r
SCI
SCI_df <- as.data.frame(SCI)

# Optional: Add rownames as a column
SCI_df$Term <- rownames(SCI_df)
SCI_df <- SCI_df[, c("Term", names(SCI_df)[1:2])]  # Reorder columns if needed

# Make the table
table_gt <- gt(SCI_df) %>%
  tab_header(title = "Table 3: Simultaneous Confidence Intervals") %>%
  fmt_number(columns = 2:3, decimals = 4)
gtsave(table_gt, filename = "SCI_table.png")
```