STA 108 Project 2, Prof. M. Pouokam

Group 8: Diya Jain, Anna Yeh, Eric Dong, Haley Bolanos, Jacqueline Gamez Hernandez, Kieran Sullivan, Nancy Chen, Ruihan Geng, Xiaoyi Fu, Nora Xia

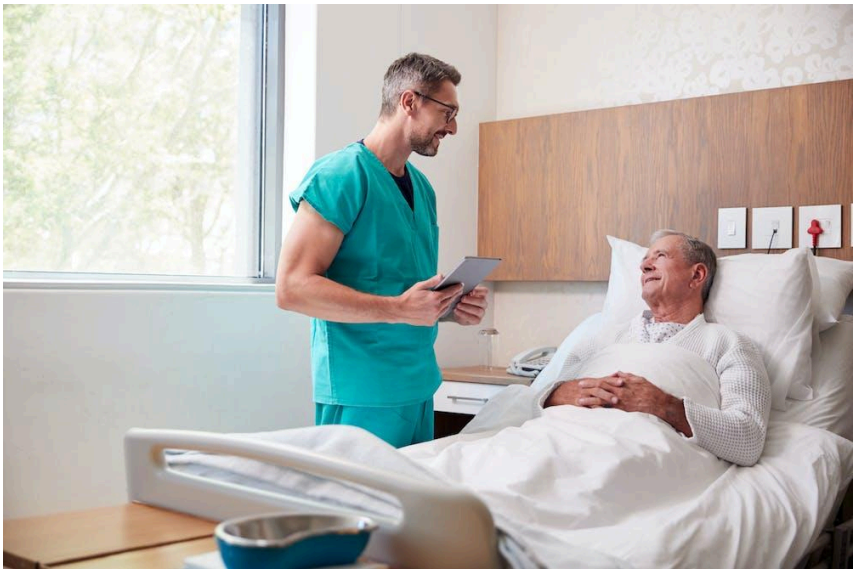


Figure 1: Patient in Hospital Bed

Figure 2: Australian athletes Peter Bol, Emma McKeon, and Matthew Dellavedova (track & field, swimming, and basketball, respectively)

3.1 Transformation of Variables (Topic I: Question 1)

I. Introduction

Healthcare officials at hospitals have the responsibility of ensuring, to the best of their ability, that patients leave in better condition than how they arrived. However, due to their nature, hospitals are densely contaminated with pathogens and prolonged visits may increase a patient's probability of contracting additional illnesses. This problem is a serious health concern as patients at a hospital are typically already immuno-compromised. Contracting hospital-acquired illnesses may be very intense for these immuno-compromised patients and may result in complications that prolong their recovery, and, at worst, death. Prolongment of a patient's recovery also “backs up” hospitals and prevents incoming patients from receiving adequate care.

Here, we will be analyzing the linear relationship between a patient's stay at a hospital in days and the ratio of number of cultures performed to number of patients without signs or symptoms of hospital-acquired infection, times 100. The larger the ratio, the better, as this means that only a small percentage of cultured patients have acquired an infection during their hospital stay. Our independent variable, x, will be the ratio, and our dependent variable, y, will be the duration of a patient's hospital stay.

Establishing a linear relationship between these two variables will serve to increase healthcare workers' understanding of the risk of keeping a patient at the hospital for a certain length of time. Healthcare officials can then assess the risk and determine if keeping a patient under their care is more beneficial to the patient's well-being than discharging them.

II. Summary of Data

II. A. Scatterplot of Data

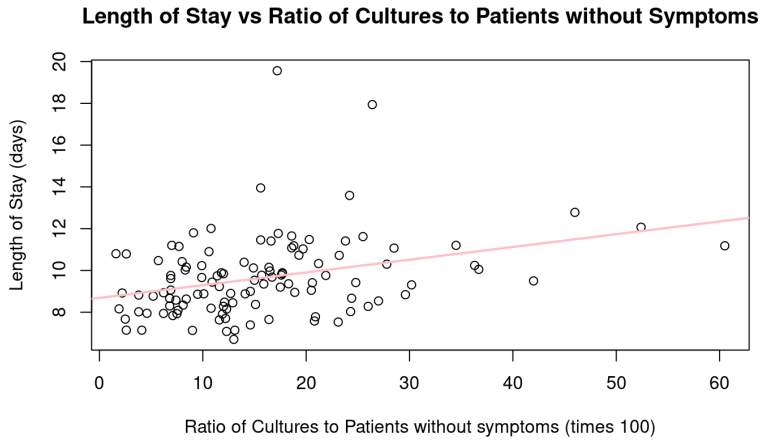


Figure 3.1 - Length of Stay vs. Ratio

Above is a scatterplot of our dataset where the independent variable, x, is the ratio of cultures to patients without symptoms of a hospital-acquired illness (times 100) and the dependent variable, y, is the length of a patient's hospital stay in days. The line running through the data is our estimated linear regression line.

The equation of our estimated linear regression line is:

$$\hat{Y} = 8.68475 + 0.0610\,X$$

Where 8.68475 is the length of a patient's hospital stay when the ratio of cultures to patients without symptoms is 0. 0.0610 is the slope of the line, demonstrating that with one unit increase of x, length of stay of a patient is estimated to increase by 0.0610, on average.

III. Diagnostics

III. A. Model and Assumptions

We will use a simple linear regression model to analyze our dataset. To run statistical tests on our constructed linear regression line, we will need to first verify the model's assumptions. The simple linear regression model is:

$$Y_i = \boldsymbol{\beta}_0 + \beta_1 X_i + \boldsymbol{\varepsilon}_i$$

Where i = 1,…, n. Y_i is the i-th value for the i-th subject. $\boldsymbol{\beta}_0$ is the unknown true y-intercept and $\boldsymbol{\beta}_1$ is the unknown true slope for our population, and $\boldsymbol{\varepsilon}_i$ accounts for the error as the simple linear regression model is not a perfect estimate of Y_i.

This model assumes:

I. (X_i, Y_i) are randomly sampled and independent

II. $\boldsymbol{\varepsilon}_i \sim N(0, \sigma^2_{\varepsilon})$

III. All X_i are treated as known constants (so we consider the distribution of interest Y |X)

III. B. Assessing Normality

The simple linear regression model assumes that the distributions of the residuals are normal. We can visualize and analyze the normality of the residuals by constructing a QQ plot. We can test for normality using the Shapiro-Wilks Hypothesis Test.

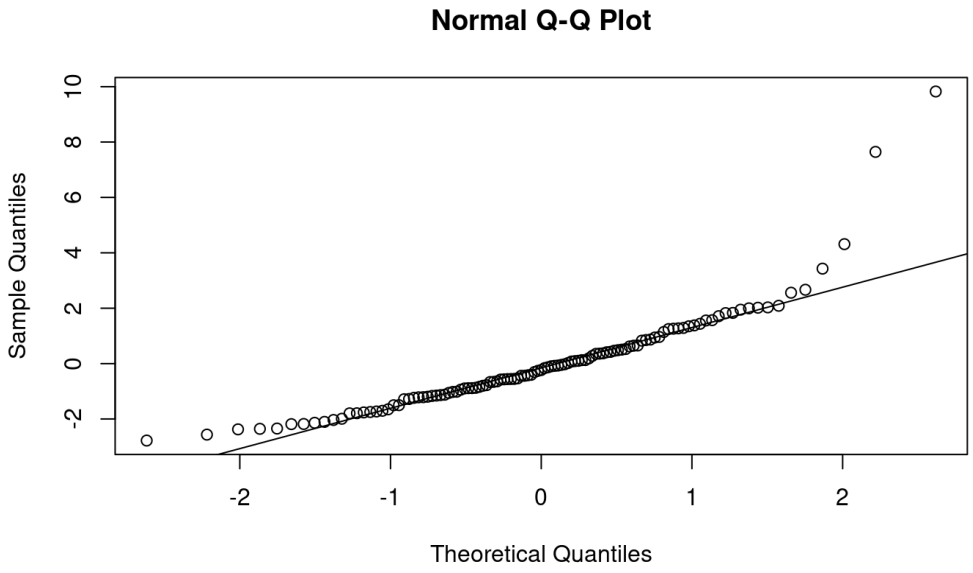


Figure 3.2 - QQ Plot

The QQ plot figure shows the residuals of our data plotted against a theoretical normal line. If the residuals plot on the normal line, the plot suggests normality of the residuals and our dataset. In Figure 3.2, most of the residuals within two quantiles of the center, fit on the normal line. Residuals further than two quantiles of the center, however, appear to deviate from the theoretical line, suggesting the presence of outliers which may be interfering with the normality of our dataset. It's important to note that the QQ plot does not serve as a statistical test for the normality of our residuals, and no definite conclusions can be made based on interpretations of the plot.

We will be performing a Shapiro-Wilks test to obtain statistically significant evidence on the normality of our dataset. The null and alternative hypotheses of this test are as seen below:

H0: The data is normally distributed

HA: The data is not normally distributed

After running the Shapiro-Wilks test, we obtained a p-value of 1.403e-09. Assuming that the residuals are normally distributed, the probability of observing our dataset or more extreme is 1.403e-09. At a statistical threshold of 0.05, we reject the null hypothesis. Therefore, we can conclude that our dataset is not normally distributed and one assumption of the simple linear regression model still needs to be met.

Our conclusion suggests that we may need to transform our variable y to obtain a normally distributed dataset. Without transformations, we must be cautious when analyzing and testing the linear relationship between the length of stay for a patient at the hospital, in days, and the ratio of the number of cultures performed to the number of patients without signs or symptoms of hospital-acquired infection, times 100.

III. C. Assessing Constant Variance

The simple linear regression model assumes constant variance of the residuals. We can first visualize and analyze the variance of the residuals by plotting the errors versus the fitted values. For statistical evidence of equal variance of the residuals, we will perform a Brown-Forsythe Hypothesis test.

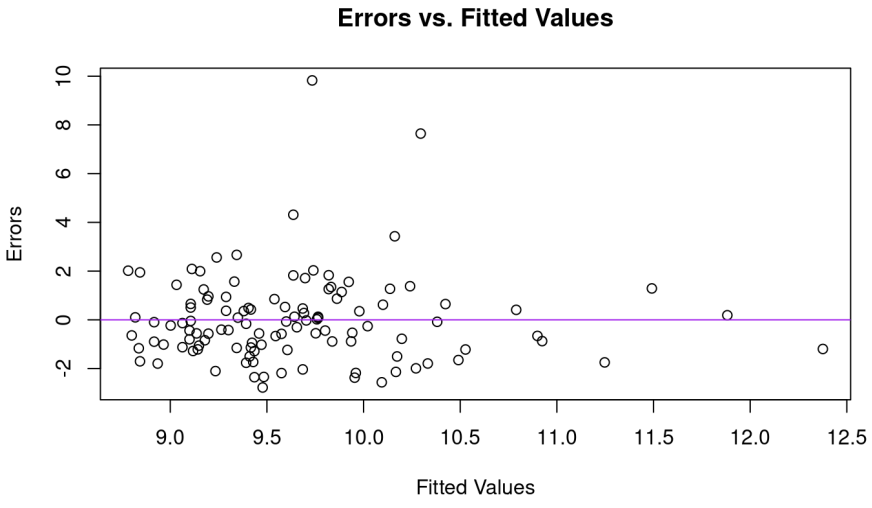


Figure 3.3 - Errors vs. Fitted Values

The Error vs. Fitted Value scatterplot shown above is a visualization of the error for each value of Y obtained from our estimated regression model. The line at ei = 0 serves to show how far our estimation of the length of stay for a patient is from the actual value at all values of x. If the equal variance assumption of the simple linear regression model is met, we should observe a constant spread of the errors from ei = 0. Looking at Figure 3.3, the residuals do not appear to be constant. The residuals appear to get smaller at higher values of Y. There also appear to be two outliers at approximately Y = 9.7 and 10.3.

However, interpretations of this figure are not reliable as they are purely subjective. We will conduct a Brown-Forsythe hypothesis test to obtain statistical evidence on the equal variance of our residuals. The null and alternative hypotheses of this test are as seen below:

H0: The variance of the residuals are equal

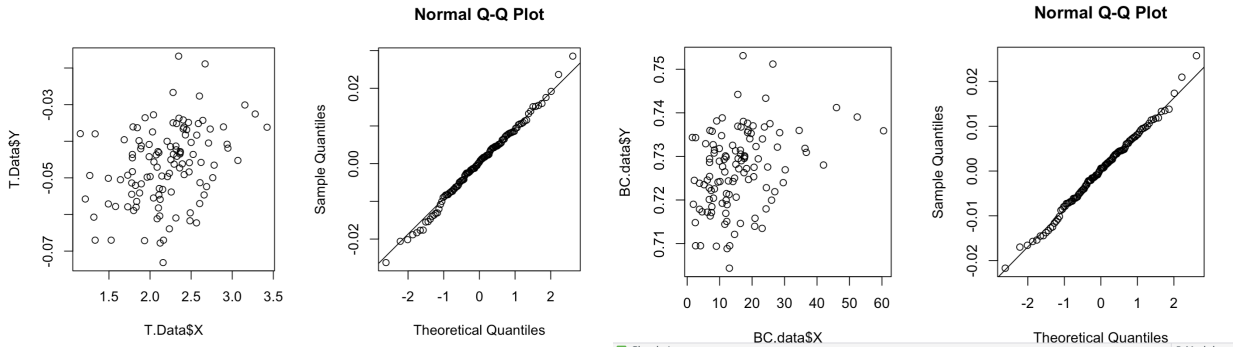
HA: The variance of the residuals are not equal

From the Brown-Forsythe test, we obtained a p-value of 0.119623. Assuming that the variance of the residuals is equal, the probability of observing our dataset or more extreme is 0.119623. At a statistical threshold of 0.05, we fail to reject the null hypothesis. Therefore, we can conclude that errors vary equally and the equal variance assumption of the simple linear regression model has been met.

IV. Removing outliers and transformation

We like to check the combination of transformations. To do this we will use the Tukey transformations and Box-Cox transformations, and compare which has the best combination of transformed variables.

Tukey transformations: Box-Cox transformations:



Since the diagonal reference line of QQ plot in Tukey transformations is straighter than Box-Cox transformations, and the dots spread more equally in the scatterplot in Tukey transformations. So Tukey is a better choice for this question.

Since there's still outliers, we can remove the outliers because that can skew the regression line, violate the normality, and violate the constant variables. So first, we need to identify the outliers, and remove it from the original data. To get the outliers, we calculate the standardized residuals, standardized residuals, and identifying outliers. We need to check which the absolute value of standardized residuals is larger than the cutoff, or which the absolute value of standardized residuals is larger than the cutoff, or which the absolute value of identifying outliers is larger than the cutoff. After that, we got the outliers for Y is 19.56 and 17.94, and for X is 12.2 and 26.4.

IV. A. Results for Original Data

From the Shapiro-Wilk's test for the original data, we obtained a p value of 1.403e-0.9, thus the data does not follow a normal distribution. Based on the results of the Fligner test, we can conclude that

there is no significant evidence to suggest that the variances in the length of stay for a patient at the hospital differ across the dataset. It suggests that the variance across different levels are constant.

Since the original data was not normal, we would have to normalize the data through either a Box-Cox or a Tukey transformation.

IV. A. a. Box-Cox transformation

The Box-Cox transformation transforms Y only. This means that we can no longer interpret the $\hat{Y}$ as the length of stay for a patient at the hospital, in days. We would have to revert back to the pre-transformation state, which can be extremely tedious. Thus, we will be mainly looking at the Tukey transformation via the conclusion above. As Tukey is clearly the better choice as seen from the two scatterplots.

IV. A. b. Tukey transformation

The Tukey transformation transforms both X and Y. After transforming the data, we observe that the Shapiro p value for the Tukey transformation for X and Y is 0.566 and 0.6976 respectively.

The transformation significantly improved the data. The significantly higher p values for X (the ratio of number of cultures performed to number of patients without signs or symptoms of hospital-acquired infection, times 100) and Y (the length of stay for a patient at the hospital, in days) compared to the pre-transformation state indicate a successful normalization of the data. Since the variance of the original dataset is constant, we would use the p value of X. With a p value of 0.566, which is well above the common alpha values of 0.20, 0.10, 0.05, etc., we fail to reject the null hypothesis of the Shapiro-Wilk test. This states that the data is normally distributed, which is an indication that Tuckey transformation effectively addresses the issue of non-normality in the dataset.

IV. A. c Removal of Outliers without transformation

Before removing the outliers, we obtained a p-value of 1.403e-09. After removing the outliers at (12.2, 19.56) and (26.4, 17.94), a second Shapiro-Wilk test is conducted, we obtain a p-value of 0.1011. Since the p value significantly increases , we must consider that outliers have a significant effect on the model. Since this p-value is greater than the common alpha values of 0.10, 0.05, etc., we would not have sufficient evidence to reject the null hypothesis. The difference in conclusions suggests that the removal of outliers is extremely significant, and we should include it as one of our possible transformations for normality.

IV. A. d Downsides

While the Tuckey transformation successfully normalizes the dataset, there are, however, several downsides and considerations to keep in mind. The Tucky transformation makes the interpretation of the data extremely difficult. The coefficients derived from the regression model of the transformed data relate to the transformed scale, which may be in different units. This makes it difficult to compare results with other studies or analyses conducted using the same dataset. It would require transforming the data back to its original form to understand its implications in the context of the original data. Another downside is selecting the optimal transformation. The parameters of our Tukey transformation are arbitrary and may require trial and error when selecting the optimal transformation.

While outlier removal may improve our model's performance and the accuracy of our results, There are also several downsides in removing outliers in the dataset. Outliers indicate critical occurrences that could be significant to understanding the dataset. The removal of outliers may lead to loss of valuable information about the dataset, which may negatively affect our conclusions on the dataset; therefore, we should only remove outliers if they significantly affect our p-value. Outliers also contribute to the variability in a dataset. Eliminating them can underestimate the true variability and spread of the data,

which may be misleading. Also if we remove the outliers to achieve a stronger linear relationship (a higher R^2 value), we may overfit the model to the remaining dataset. This means that the current model may perform well on the current dataset, but not so much on new datasets where outliers are significant.

V. Conclusion

Ultimately, I believe the transformed data is indeed a better fit, as the transformed data meets the requirements of both normality and homoscedasticity. This leads to more reliable results.

For a client who wants to use this data for simple linear regression, I would tell them to use a Tukey transformation and remove the two outliers. From only looking at the plots, the Tukey transformation is spread out evenly and doesn't seem to have an obvious trend or pattern within its plot. The Box-Cox transformation, consequently, has a cluster of data points towards the left of the plot, as well as visible outliers; this may indicate that the transformation has not effectively normalized the dataset. Since constant variance is already present throughout the dataset, the Tukey transformation appears to be the better option for normalizing the data.

3.2 Multiple Regression Modeling (Topic II: Question 1)

I. Introduction.

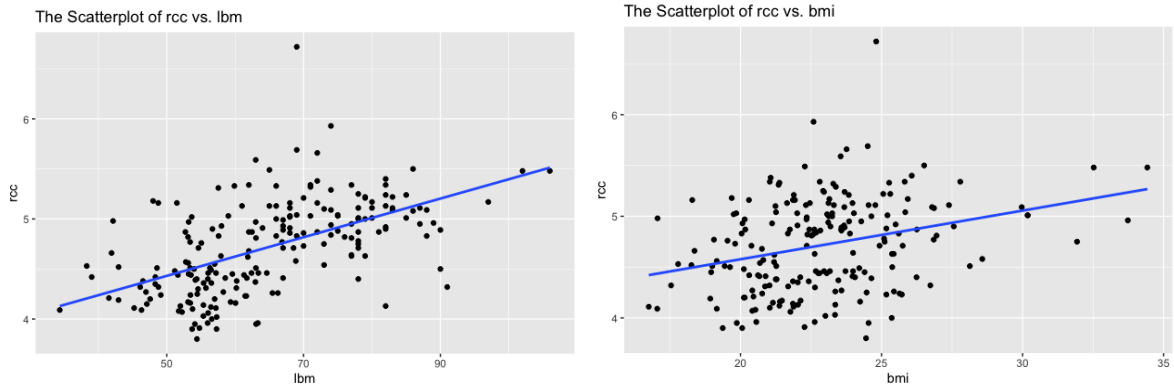
A random sample taken from Australian student athletes was used to model red blood cell count (per liter) based on other physical characteristics present. This model takes into account both qualitative and quantitative variables. The characteristics in the data include the athletes lean body mass (mass that is lean muscle in kilograms), body mass index (in kilograms), percent body fat, plasma ferritins (a measure of iron in the blood in nanograms), sex (male or female), and lastly, newsport (sports that involve a net, water, or primarily running). Our goal for this project is to test whether or not each one of these characteristics truly contributes to a "correct" model for the red blood cell count. In order to find a "correct" model, we will be using model selection criteria Bayesian Information Criteria (BIC) and forward subset selection using BIC. We will begin with an empty model and gradually incorporate each variable, calculating BIC each time and selecting the variables with the smallest BIC. One thing that is important to mention about our approach is the bigger penalization that comes as a result of BIC and the tendency to underfit that results from forward subset selection. This could potentially result in the penalization of variables that could be significant and essential to determining red cell count, and would not have been penalized under other model selection criteria.

II. Summary of the data

rcc	lbm	bmi	pcBfat	ferr	sex	newsport
Min. :3.800	Min. : 34.36	Min. :16.75	Min. : 5.630	Min. : 8.00	Length:202	Length:202
1st Qu.:4.372	1st Qu.: 54.67	1st Qu.:21.08	1st Qu.: 8.545	1st Qu.: 41.25	Class :character	Class :character
Median :4.755	Median : 63.03	Median :22.72	Median :11.650	Median : 65.50	Mode :character	Mode :character
Mean :4.719	Mean : 64.87	Mean :22.96	Mean :13.507	Mean : 76.88	NA	NA
3rd Qu.:5.030	3rd Qu.: 74.75	3rd Qu.:24.46	3rd Qu.:18.080	3rd Qu.: 97.00	NA	NA
Max. :6.720	Max. :106.00	Max. :34.42	Max. :35.520	Max. :234.00	NA	NA

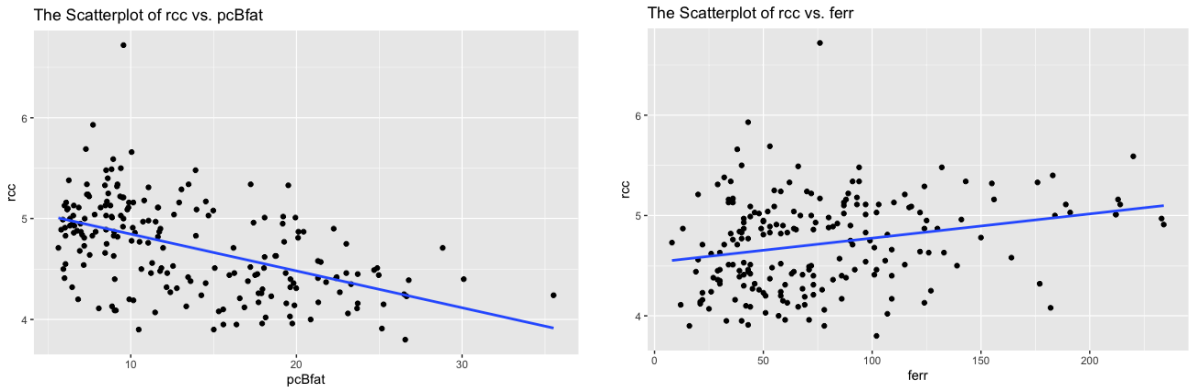
The table with a summary of our dataset's statistics is shown above. It includes the minimum, first quartile, median, mean, third quartile, and maximum of our Y and X variables. First we can look at

the Y variable rcc (Red blood cell count per liter). The mean of rcc is approximately 4.719 and the median is 4.755. The value of the mean is slightly lower than the median by 0.036, suggesting an approximately normal distribution of the red blood cell count per liter. Rcc has a range of 2.92, which might be of interest as it shows a large spread in red blood cell counts among the athletes. Then moving to the X1 variable lbm (lean body mass in kg). The athletes have an average lean body mass of approximately 64.87 kg, with values ranging from 34.36 to 106.00 kg. This wide range suggests a diverse group of athletes in terms of muscle mass. For the X2 variable, the average BMI is about 22.96 kg, which is within the normal weight range. The range from 16.75 to 34.42 indicates a wide spread of body weight categories among athletes (underweight or obese). The next variable X3, which is the Percent Body Fat, has a mean value of 13.507. While the range is wide, from 5.63% to 35.52%, which reflects the different body compositions present in the sample. Then we look at the X4 variable (Plasma ferritins), Plasma ferritin levels average at 76.88 ng, with a broad range from 8.00 to 234.00 ng. Sex and newsport are noted as categorical variables, so the summary does not provide the same statistics as other X variables.

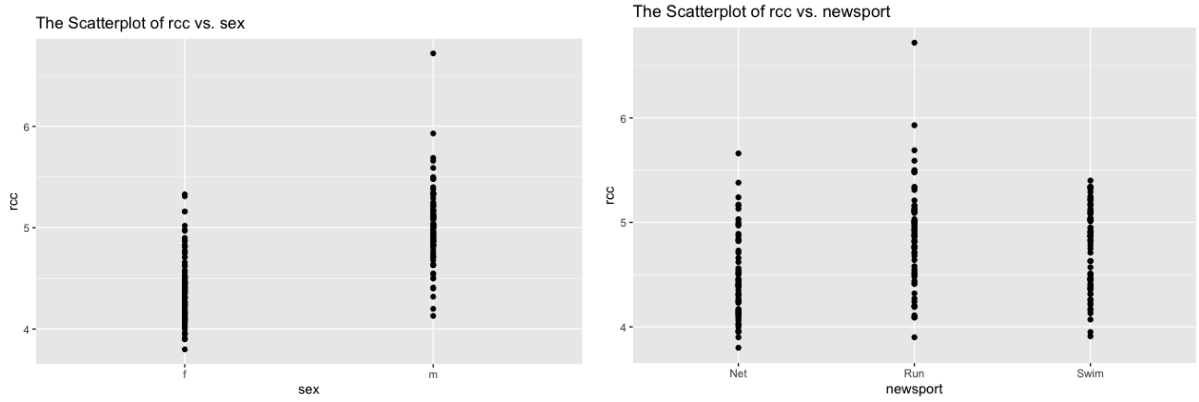


Taking a look at the plot of Y (rcc) vs. X1 (lbm) at left first. There appears to be a positive correlation between lean body mass and red blood cell count, indicated by the upward trend in the scatter plot. As lean body mass increases, red blood cell count tends to increase as well. This relationship is expected as a larger lean body mass may require more oxygen transport, and thus more red blood cells. However, there is a potential outlier at the central top area, lying far away from the data cluster. The scatterplot on the right is the plot of Y vs. X2. There's a slight positive correlation between bmi and rcc as

well. However, data points are more dispersed than in the plot on the left, which suggests that the relationship is not as strong or consistent as with lbm. Once again, a potential outlier is observed at the central top area.



The plot we have on the left is the scatterplot of Y (rcc) vs. X3 (pcBfat). It shows a negative correlation between rcc and pcBfat. As percent body fat increases, red blood cell count decreases. This would imply that athletes with lower body fat tend to have a higher concentration of red blood cells. According to this plot, we can still observe a potential outlier at the top left part. The right plot of rcc vs. ferr presents a slight positive correlation between red blood cell count and level of Plasma ferritins, with a potential outlier at top.



The scatter plot of rcc vs. sex on the left reveals a noticeable difference in red blood cell counts between males and females, with males generally having a higher red blood cell count. Moving to the plot on the right (rcc vs. newsport), it does not show a clear trend distinguishing red blood cell counts among

different sports. However, it seems that there is some variation in red blood cell counts within each category due to different spreads. This suggests that the type of sport may not be as strong a predictor of red blood cell count as the physiological measures are. Once again, there is an outlier in both of the plots. There is an outlier on the top for "male" in the plot of sex that skews the distribution of data points, and an outlier for "run" in the plot of newsport as well. Therefore, removing outliers may need to be taken into consideration.

III. Model Selection

Our goal is to find the best, most correct model. We performed forward subset selection and used BIC as our criteria. In forward selection, we start with the empty model $Y_{Red\ Blood\ Cell\ Count} \sim 1$, then add one X to the model, record the BIC, and repeat these steps for every X not in the model. We only add the X which decreases BIC the most and repeat these steps until BIC no longer decreases, then this will be our final model. It is important to note that forward selection does not look at all possible models and tends to underfit.

We performed forward selection in R and the results of each iteration are listed in the following table.

Iteration	Selected model per iteration	BIC
0	$Y_{Red\ Blood\ Cell\ Count} \sim 1$	-311.19
1	$Y_{Red\ Blood\ Cell\ Count} \sim X_{Sex}$	-431.67
2	$Y_{Red\ Blood\ Cell\ Count} \sim X_{Sex} + X_{Newsport}$	-431.72

The empty model $Y_{Red\ Blood\ Cell\ Count} \sim 1$ has a starting BIC of -311.19. In the first iteration, we find that the variable sex minimizes BIC the most (BIC = -431.67), so the first iteration resulted in the regression $Y_{Red\ Blood\ Cell\ Count} \sim X_{Sex}$. In the second iteration, we find that adding the 3-level categorical variable

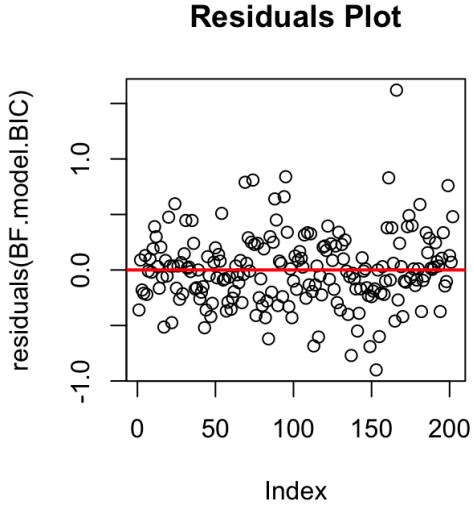
newsport further minimizes BIC the most compared to the other independent X variables (BIC = -431.72), so the second iteration resulted in the regression $Y_{Red\ Blood\ Cell\ Count} \sim X_{Sex} + X_{Newsport}$. In the third iteration, BIC begins to increase when we try to add other Xs. Therefore, our final estimated regression with BIC = -431.72 is:

$$Y_{Red\ Blood\ Cell\ Count} = 4.3203 + 0.5801 * X_{1_{Sex,\ Male}} + 0.1998 * X_{2_{Newsport,\ Run}} + 0.1040 * X_{2_{Newsport,\ Swim}}$$

where $\beta_0 = 4.3203$; $\beta_1 = 0.5801$; $\beta_2 = 0.1998$; $\beta_3 = 0.1040$.

IV. Diagnostics

A. Assessing constant variance



A residual plot is a graphical tool used to visually assess the goodness of fit of a regression model. It displays the residuals, which are the differences between the observed values of the dependent variable and the values predicted by the regression model, against the corresponding values of the independent variable or the predicted values. For our model, shown above, the residual plot follows a random distribution suggesting that our assumption of linearity and homoscedasticity are met. The spread of the residuals remains relatively consistent across the range of predicted values. There is no noticeable pattern of increasing or decreasing spread, which suggests that the assumption of homoscedasticity is met.

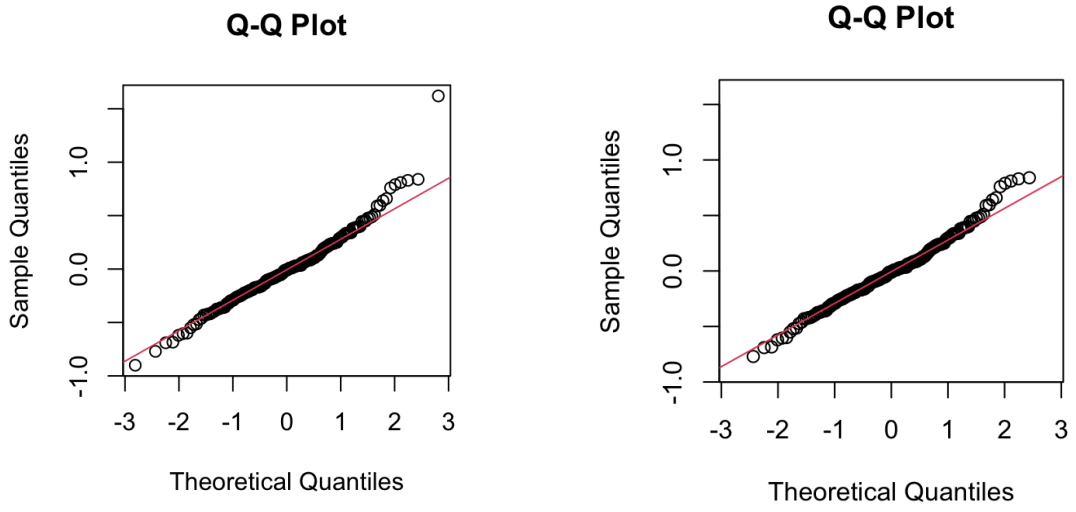
Further, since the residuals are randomly scattered around the line y=0, which suggests that the relationship between the predictor variables included in our models and the red blood cell count is adequately captured by the linear regression model. To further test this, we utilize the Brown-Forsythe test.

H0: The variance of the residuals are equal

HA: The variance of the residuals are not equal

Our p-value is 0.317, using an alpha of 0.05, we fail to reject the null hypothesis. Therefore,we do not have enough evidence to conclude that the variances of the residuals are significantly different across the range of predicted values, which is consistent with the assumption of homoscedasticity. This finding supports the adequacy of the linear regression model in capturing the relationship between the predictor variables and the red blood cell count.

B. Assessing Normality



A QQ plot is a graphical tool used to assess whether a given data set follows a normal distribution. The quantiles of the sample data are plotted against the quantiles of a theoretical distribution. In the graph on the left, is our original data set, as we can see there are two outliers present, which are points deviating from the line. When removing those outliers, as seen from the graph on the right, we see

that the majority of our points lie directly on the line; this indicates that the empirical distribution of our data closely matches the theoretical normal distribution, meaning that our normality assumption is met. To further confirm, we can perform a Shapiro-Wilks test, using the null and alternative hypothesis listed below.

H0: The data is normally distributed

HA: The data is not normally distributed

Through running this test, we get a p-value 0.0001147. Using the significance level is 0.05, we reject the null hypothesis of normality of the residuals, showing that our data is not normally distributed and our data does not meet the normality assumption. When moving forward with our study, it is important to keep this in mind when analyzing our results.

V. Analysis

V. A. Building the Estimated Linear Regression Line

Recall the final estimated regression line in model selection section,

$$Y_{Red\ Blood\ Cell\ Count} = 4.3203 + 0.5801 * X_{1_{Sex,\ Male}} + 0.1998 * X_{2_{Newsport,\ Run}} + 0.1040 * X_{2_{Newsport,\ Swim}}$$

The goal of the estimated regression line is to model for Y = red blood cell count (per liter) of Australian athletes, based on two categorical X predictors: X_1 = sex (male or female) and X_2 = type of newsport (Net (sports that involve a net), Swim (sports that involve water), and Run (sports that primarily involve running). For b_0, it means that we expect the athlete's red blood cell count per liter is 4.3203 when the athlete is female and plays a sport that involves a net. For b_1, it means that we expect male athletes to have a higher average red blood cell count per liter by 0.5801 holding the type of sport they play constant. For b_2, it means that we expect athletes who play sports that primarily involve running have a higher average red blood cell count per liter by 0.1998 compared to those who play sports involving a net,

holding the sex of athletes constant. For b_3, it means that we expect athletes who play sports that involve water have a higher average red blood cell count per liter by 0.1040 compared to those who play sports involving a net, holding the sex of athletes constant. When both $X_{Newsport,\,Run}$ and $X_{Newsport,\,Swim}$ equals to 0, it means that the athlete plays a sport that involves a net.

V. B. Checking the Model Fit

To determine if our estimated regression line fits well with our data, we will use two ways to check. Firstly, we are going to test if the reduced model:

$$Y_{i_{Red\ Blood\ Cell\ Count}} = \beta_0 + \beta_1 * X_{1_{Sex,\,Male}} + \beta_2 * X_{2_{Newsport,\,Run}} + \beta_3 * X_{2_{Newsport,\,Swim}}$$

is a better fit compared to the full model:

$$Y_{i_{Red\ Blood\ Cell\ Count}} = \beta_0 + \beta_1 * X_{1_{Sex,\,Male}} + \beta_2 * X_{2_{Newsport,\,Run}} + \beta_3 * X_{2_{Newsport,\,Swim}} + \beta_4 * X_{3_{LBM}} + \beta_5 * X_{4_{BMI}} + \beta_6 * X_{5_{PcBFat}} + \beta_7 * X_{6_{Ferr}}$$

Let alpha, which is the significance level, be 0.05, we have the following hypotheses:

$$H_0\ (null\ hypothesis)\colon\ The\ reduced\ model\ is\ a\ better\ fit\ (\beta_4 = \beta_5 = \beta_6 = \beta_7 = 0)$$

$$H_\alpha\ (alternative\ hypothesis)\colon\ The\ reduced\ model\ is\ not\ a\ better\ fit\ (at\ least\ 1\ of\ \beta_4, \beta_5, \beta_6, \beta_7\ is \neq 0)$$

Then, we have the test statistic comparing the reduced and full models:

$$F_s = \frac{SSE_R - SSE_F}{df(SSE_R) - df(SSE_F)} / MSE_F = \frac{21.45476 - 21.09658}{198 - 194} / 0.1087 = 0.8234.$$

Using this result, $p\ value$ at $df(num) = 4$ and $df(denom) = 194$ is $P(F > 0.8234)$ which is bigger than 0.20. 0.20 is bigger than alpha = 0.05, therefore, fail to reject H_0 and support that the reduced level is a better fit.

Secondly, we are going to calculate Partial R^2 to determine the reduction in error. We have the following equation:

$$R^2\{X_3, X_4, X_5, X_6 | X_1, X_2\} = \frac{(error\ before)-(error\ after)}{error\ before} = \frac{21.4548 - 21.0966}{21.4548} = 0.01669 .$$

According to our result, the error for a model with X_1 and X_2 in it is reduced when we add the rest of the Xs (X_3, X_4, X_5, X_6) by 1.669%, which is incontestably insignificant.

V. C. The Confidence Interval for $\beta_0, \beta_1, \beta_2, \beta_3$

Recall our estimated linear regression line:

$$Y_{i_{Red\ Blood\ Cell\ Count}} = \beta_0 + \beta_1 * X_{1_{Sex,\ Male}} + \beta_2 * X_{2_{Newsport,\ Run}} + \beta_3 * X_{2_{Newsport,\ Swim}},$$

after our estimated linear regression line is proved to be reliable, we want to have and estimation for the real β's of our estimated regression line by using simultaneous confidence intervals. Let the confidence level be 95%, we have $\alpha = 0.05$ and the equation is: $b_i \pm t_{\alpha/2,\ n-p} s\{b_i\}$. Before we actually calculate the simultaneous confidence interval, we have to choose the appropriate multipliers between Bonferroni, Working Hotelling, and Scheffe. Using R, we have the following results:

Bonferroni	Working-Hotelling	Scheffe
2.415	3.986	2.820

We are going to choose Bonferroni since it is the smallest multiplier among all three and will give the narrowest interval. Thus, we get our result for all the β's. For β_0, we are overall 95% confident that the athlete's red blood cell count per liter is between 4.2073 and 4.4333 when the athlete is female and plays a sport that involves a net. For β_1, we are overall 95% confident that the average red blood cell count per liter for male athletes is higher than female athletes by between 0.4576 and 0.7025, holding the

type of sport they play constant. For β_2, we are overall 95% confident that athletes who play sports that primarily involve running have a higher average red blood cell count per liter by between 0.0459 and 0.3537 compared to those who play sports involving a net, holding the sex of athletes constant. For β_3, we are overall 95% confident that athletes who play sports that involve water have a higher average red blood cell count per liter between -0.0453 and 0.2532 compared to those who play sports involving a net, holding the sex of athletes constant. However, because zero is involved in the confidence interval of β_3, it is reasonable to say that β_3 is insignificant. Since it is one of the category of $X_{2\ Newsport,\ Swim}$, we are still going to keep it while knowing that sports that involve water would not make a big difference in athletes' red blood cell count per liter holding all other variables constant.

VI. Interpretation:

In the model selection section, we used BIC as the model selection criterion to predict the final model of red blood cell count (Y) in Australian athletes and we got an estimated regression line. We have two approaches to test whether the estimated regression model fits the data or not. The first one is comparing it to a reduced model. The Null hypothesis(H_0) is the reduced model can explain the variation in red blood cell count and is a better fit (the variables do not significantly improve the model's fit) The alternative hypothesis(H_α) is the full model that can provide a significantly better fit than the reduced model(the variable significantly improves the model's fit). Then we use the formula $F_s = \frac{SSE_R - SSE_F}{df(SSE_R) - df(SSE_F)} / MSE_F$ to calculate the range of p-value, and the result is that p-value is greater than 0.2, which is also greater than normal alpha 0.05. As a result, we fail to reject the null and conclude that the reduced model is a better choice to fit the data than the full model(a subset of the $\beta's$ is 0). This

means a reduced model without the difference in sex and newsport can also predict red blood cell count among Australian athletes.

The second method is to calculate the Partial R^2 to find out whether the variable affects the prediction of the model. We found that $R^2\{X_3, X_4, X_5, X_6 \mid X_5 + X_6\}$ is equal to 0.01669, which means the error for a model with X_5 and X_6 when we add the full model variable except them(X_4, X_5, X_6, X_7) has a reduction of 1.669%, which is extremely low and we can indicate that the X_5 (sex) and X_6(newsport) are insignificant to the difference in red blood cell count among Australian athletes.

Next, we calculated the simultaneous confidence interval through equation $b_i \pm t_{\alpha/2,\, n-p} s\{b_i\}$. (We choose Bonferroni multipliers) Under 95% confidence interval($\alpha = 0.05$):

The confidence interval for β_0 is [4.2073, 4.4333], which is the range of the red blood cell count for female athletes who play net sports.

The confidence interval for β_1 is [0.4576, 0.7025], indicating that while all other factors are equal, male athletes are 0.4576 to 0.7025 higher in red blood cell count than female athletes.

The confidence interval for β_2 is [0.0459, 0.3537], indicating that while sex remains the same, athletes involved in running sports are 0.0459 to 0.3537 higher in red blood cell count than those in net sports.

The confidence interval for β_3 is [-0.0453, 0.2532], since the interval contains 0, so we conclude that playing water sports does not significantly influence red blood cell count compared to playing net sports, when sex and other factors are held constant.

VII. Conclusion:

In conclusion, by using forward subset selection and BIC, we found that the overall best model to predict the red blood cell count per liter of these Australian athletes was one that focused on the two categorical variables provided: sex of the athlete (male or female), and the type of sport they play (involving water, running, or a net). When testing to see if the assumptions of a linear model were met, our residual plot showed randomly distributed residuals, indicating that the assumptions of linearity and homoscedasticity are met. Our Q-Q plot, after removing outliers, shows points nicely arranged on the line, indicating that the data is normally distributed. A Shapiro-Wilks test further emphasized this finding.

After testing our regression line against the reduced model, we found it to be a strong fit for our data. Using a Bonferroni multiplier, we found the confidence intervals for all of our parameters, with one significant finding being that β_3 is likely insignificant, meaning that an athlete playing a sport involving water would likely not have a strong effect on their red blood cell count. However, we kept it as it was one of the three possible categories for X_6 and did not want to remove all athletes who this applied to. Further research could be conducted, including athletes from other disciplines, to see the effect of playing those sports on red blood cell count.

R Code Appendix
#3.1 a

Plot of Dataset and Estimated Linear Regression Equation

```{r}
transform <- read.csv("/cloud/project/Transform1.csv")
head(transform)
the.model = lm(Y ~ X, data = transform)
the.model$coefficients
plot(transform$X, transform$Y, main = "Y vs X", xlab = "X", ylab = "Y")
abline(the.model,col = "pink",lwd = 2)
```

QQplot

```{r}
qqnorm(the.model$residuals)
qqline(the.model$residuals)
```

Shapiro Wilks

```{r}
ei = the.model$residuals
yhat = the.model$fitted.values
the.SWtest = Shapiro.test(ei)
the.SWtest
```

ei vs. fitted values

```{r}
library(ggplot2)
plot(yhat, ei, main = "Errors vs. Fitted Values",xlab = "Fitted Values",ylab = "Errors")
abline(h = 0,col = "purple")
```

BF test

```{r}
install.packages("carData")
library(car)
Group = rep("Lower",nrow(transform))
Group[transform$X > median(transform$X)] = "Upper"
Group = as.factor(Group)
```

```
        transform$Group = Group

        the.BFtest = leveneTest(ei~Group, data=transform, center=median)
        p.val = the.BFtest[[3]][1]
        p.val
```
Estimated Linear Regression Line Equation

$\hat Y = 8.68475+(0.0610)X_1$

##3.1(b)

```{r setup, include=FALSE}
Transform1 <- read.csv("~/Downloads/Transform1.csv", stringsAsFactors=TRUE)
```

Tukey Transformations

```{r, pressure = FALSE}
par(mfrow = c(1,2))

new.state = data.frame(days = Transform1[,1], beds = Transform1[,2])
summary(new.state)
names(new.state) = c("Y","X")
full.model = lm(Y~ X,data = new.state)
small.model = lm(Y ~ X, data = new.state)
plot(new.state$X, new.state$Y)
qqnorm(small.model$residuals)
qqline(small.model$residuals)

library(rcompanion)
tukeyY = transformTukey(new.state$Y,plotit = FALSE)
tukeyX = transformTukey(new.state$X, plotit = FALSE)

par(mfrow = c(1,2))
T.Data = data.frame(Y = tukeyY,X = tukeyX)
T.model = lm(Y ~ X, data = T.Data)
plot(T.Data$X, T.Data$Y)
qqnorm(T.model$residuals)
qqline(T.model$residuals)
```

#Box-Cox Transformations

```{r, pressure = FALSE}
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.state$Y^lambda - 1)/lambda
BC.data = data.frame(Y = BC.Y, X = new.state$X)
par(mfrow = c(1,2))
BC.model = lm(Y ~ X, data = BC.data)
plot(BC.data$X, BC.data$Y)
qqnorm(BC.model$residuals)
qqline(BC.model$residuals)
```
#outliers
```{r, pressure = FALSE}
full.model = lm(Y ~ X, data = Transform1)
empty.model = lm(Y ~ 1, data = Transform1)
n= nrow(new.state)
BF.model.BIC = stepAIC(full.model,  scope = list(lower = empty.model, upper=
full.model), k = log(n),trace=FALSE,direction = "both")
best.model= BF.model.BIC
alpha = 0.1 ; n = nrow(new.state); p = length(best.model$coefficients)
cutoff = qt(1-alpha/(2*n), n -p )
cutoff.deleted = qt(1-alpha/(2*n), n -p -1 )
ei.s = best.model$residuals/sqrt(sum(best.model$residuals^2)/(nrow(new.state) -
length(best.model$coefficients)))
ri = rstandard(best.model)
ti = rstudent(best.model)
outliers = which(abs(ei.s)> cutoff | abs(ri) > cutoff | abs(ti) > cutoff.deleted)
new.state[outliers,]

new.data = Transform1[-outliers,]
new.data
```
#3.1(C)

```{r}
library(rcompanion)
tukeyY = transformTukey(Transform1$Y,plotit = TRUE)
tukeyX = transformTukey(Transform1$X, plotit = TRUE)
```

```
```{r}
#REMOVE TWO OUTLIERS THEN CALCULATE SUBSEQUENT P VALUE
Transform1 <- Transform1[-c(47, 112), ]
the.model = lm(Y ~ X, data = Transform1)
the.model$coefficients
ei = the.model$residuals
yhat = the.model$fitted.values
the.SWtest = shapiro.test(ei)
the.SWtest
```

Importing datasets
```{r}
Transform1 <- read.csv("C:/Users/nache/OneDrive - University of California, Davis/Classes1/STA 108 (Win 24)/Project 2/Transform1.csv")
athelete <- read.csv("C:/Users/nache/OneDrive - University of California, Davis/Classes1/STA 108 (Win 24)/Project 2/athelete.csv")
```
Installing packages
```{r}
install.packages("leaps")
install.packages("rcompanion")
install.packages("MPV")
```
Creating function for partial R^2
```{r}
Partial.R2 = function(small.model,big.model){
  SSE1 = sum(small.model$residuals^2)
  SSE2 = sum(big.model$residuals^2)
  PR2 = (SSE1 - SSE2)/SSE1
  return(PR2)
}
```
Creating criteria
```{r}
All.Criteria = function(the.model){
  p = length(the.model$coefficients)
  n = length(the.model$residuals)
  the.LL = logLik(the.model)
  the.BIC =  -2*the.LL + log(n)*p
  the.AIC =  -2*the.LL + 2*p
```
```

```
  the.PRESS = PRESS(the.model)
  the.R2adj = summary(the.model)$adj.r.squared
  the.results = c(the.LL,p,n,the.AIC,the.BIC,the.PRESS,the.R2adj)
  names(the.results) = c("LL","p","n","AIC","BIC","PRESS","R2adj")
  return(the.results)
}
```

Quick summary of the data

```{r}
summary(athelete)
```

Part II:

Summary of data

```
athelete <- read.csv("~/Desktop/UCD作业/STA108/project 2/athelete.csv")
# summary of statistics
library(knitr)
sum_athelete <- summary(athelete)
kable(sum_athelete)
# plot of Y vs. X1
library(ggplot2)
model_X1 = lm(rcc ~ lbm, data = athelete)
ggplot(athelete,aes(lbm,rcc)) + geom_point(shape = 19) + geom_smooth(method='lm',se=
FALSE) + ggtitle("The Scatterplot of rcc vs. lbm") + ylab("rcc") + xlab("lbm")
# plot of Y vs. X2
model_X2 = lm(rcc ~ bmi, data = athelete)
ggplot(athelete,aes(bmi,rcc)) + geom_point(shape = 19) + geom_smooth(method='lm',se=
FALSE) + ggtitle("The Scatterplot of rcc vs. bmi") + ylab("rcc") + xlab("bmi")
# plot of Y vs. X3
model_X3 = lm(rcc ~ pcBfat, data = athelete)
ggplot(athelete,aes(pcBfat,rcc)) + geom_point(shape = 19) + geom_smooth(method='lm',se=
FALSE) + ggtitle("The Scatterplot of rcc vs. pcBfat") + ylab("rcc") + xlab("pcBfat")
# plot of Y vs. X4
model_X4 = lm(rcc ~ ferr, data = athelete)
ggplot(athelete,aes(ferr,rcc)) + geom_point(shape = 19) + geom_smooth(method='lm',se=
FALSE) + ggtitle("The Scatterplot of rcc vs. ferr") + ylab("rcc") + xlab("ferr")
# plot of Y vs. X5
model_X5 = lm(rcc ~ sex, data = athelete)
ggplot(athelete,aes(sex,rcc)) + geom_point(shape = 19) + geom_smooth(method='lm',se=
FALSE) + ggtitle("The Scatterplot of rcc vs. sex") + ylab("rcc") + xlab("sex")
# plot of Y vs. X6
model_X6 = lm(rcc ~ newsport, data = athelete)
ggplot(athelete,aes(newsport,rcc)) + geom_point(shape = 19) + geom_smooth(method='lm',se=
FALSE) + ggtitle("The Scatterplot of rcc vs. newsport") + ylab("rcc") + xlab("newsport")
```

Model selection (Goal: correctness)

```
```{r message=FALSE}
```{r message=FALSE}
library(MPV)
#athelete$newsport_swim <- ifelse(athelete$newsport == 'Swim',1, 0)
#athelete$newsport_run  <- ifelse(athelete$newsport == 'Run', 1, 0)
full.model = lm(rcc ~ lbm + bmi + pcBfat + ferr + sex + newsport_swim + newsport_run,data =
athelete)
```

All.Criteria(full.model)
```

```
Creating possible models of interest
```{r}
library(leaps)
all.models = regsubsets(rcc ~., data = athelete)
some.stuff = summary(all.models)
names.of.data = c("Y",colnames(some.stuff$which)[-1])
n= nrow(athelete)
K = nrow(some.stuff$which)
nicer = lapply(1:K,function(i){
  model = paste(names.of.data[some.stuff$which[i,]],collapse = ",")
  p = sum(some.stuff$which[i,])
  BIC = some.stuff$bic[i]
  CP = some.stuff$cp[i]
  results = data.frame(model,p,CP,BIC)
  return(results)
})
nicer = Reduce(rbind,nicer)
nicer
```
Forward selection
```{r}
full.model = lm(rcc ~ lbm + bmi + pcBfat + ferr + sex + newsport, data = athelete)
empty.model = lm(rcc ~ 1, data = athelete)
n = nrow(athelete)
library(MASS)
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),direction = "forward", trace=FALSE)

forward.model.AIC$coefficients
forward.model.BIC$coefficients
```
Backward selection
```{r}
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),direction = "backward",trace = FALSE)

backward.model.AIC$coefficients
backward.model.BIC$coefficients
```
Forward-Backward selection
```{r}
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),direction = "both",trace = FALSE)

FB.model.AIC$coefficients
FB.model.BIC$coefficients
```
Backward-Forward selection
```{r}
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),direction = "both",trace = FALSE)

BF.model.AIC$coefficients
BF.model.BIC$coefficients
```

V. Analysis
```{r}
Partial.R2 = function(small.model,big.model){
  SSE1 = sum(small.model$residuals^2)
  SSE2 = sum(big.model$residuals^2)
  PR2 = (SSE1 - SSE2)/SSE1
  return(PR2)
}
```
```{r}
new.athelete = data.frame(rcc = athelete[,1], lbm = athelete[,2], bmi = athelete[,3], pcBfat = athelete[,4], ferr = athelete[,5], sex = athelete[,6], newsport = athelete[,7])
names(new.athelete) = c("Y", "X1","X2","X3","X4","X5","X6")
```
```{r}
full.model = lm(Y ~ X1 + X2 + X3 + X4 + X5 + X6, data = new.athelete)
```

full.model
```
```{r}
small.model = lm(Y ~ X5 + X6, data = new.athelete)
small.model
```
```{r}
PR2 = Partial.R2(small.model, full.model)
PR2
```
```{r}
sse.full <- sum((fitted(full.model) - athelete$rcc)^2)
sse.full
the.table = anova(full.model)
the.table
MSE = sse.full/194
MSE
```
```{r}
sse.x5x6 <- sum((fitted(small.model) - athelete$rcc)^2)
sse.x5x6
the.tablex5x6 = anova(small.model)
the.tablex5x6
```
```{r}
x = (sse.x5x6 - sse.full)/4
Fs = x/MSE
Fs
```
```{r}
pr = (SSE1-SSE2)/SSE1
pr
SSE1 = sum(small.model$residuals^2)
SSE1
SSE2 = sum(full.model$residuals^2)
SSE2
```
```{r}
all.of.them = mult.fun(202, 8, 3, 0.05)
all.of.them
```
```{r}
alpha =0.05
```
```

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