A Predictive Model of Bodyfat

Kerem San

ES 56 – Probability & Statistics

Prof. Helen Suh

Spring 2020

Tufts University

Table of Contents

Introduction

The BodyFat dataset used in this report contains data on percent bodyfat measurements collected over a sample of 252 men. The dataset also includes various other body measurements such as age, density, and body weight, etc. This report explores different regression models that predict bodyfat values using these variables.

Methods

The “Pearson” correlation coefficient which measures the covariance between two samples, x and y, is computed using the formula: $r = \frac{1}{n-1} \Sigma \left(\frac{x - \bar{x}}{s_x}\right)\left(\frac{y - \bar{y}}{s_y}\right)$

This correlation coefficient r is between the values -1 and 1, and measures the linear relationship between two continuous random variables. The correlation between two variables are negative or positive depending on the sign of this coefficient.

Simple linear regression examines whether variability in a dependent variable can be explained in whole or in part using an independent variable. The regression line is calculated using the formula; $\hat{y} = \hat{\alpha} + \hat{\beta}x$

where, $\hat{\beta} = \frac{cov(x, y)}{Var(x)} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$ and $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$

Regression analysis is conducted by; checking assumptions regarding independence and checking for outliers, calculating the equation of the line, evaluating the estimated line to determine the strength of the relationship and validity of assumptions, and using the estimated line to make predictions. A hypothesis test for the slope and intercept tells whether the estimated line differs significantly from a line with a slope of zero. The R^2 value tells the percent variation explained by the regression line. This value is the squared value of the “Pearson” correlation coefficient, and can also be calculated by using the formula: $1 - \frac{\sum_i^n (y_i - \hat{y}_i)^2}{\sum_i^n (y_i - \bar{y})^2}$

The F-test is used to determine whether the linear regression model provides a better fit to the data than a model that contains no independent variables.

Multiple linear regression describes the relationship between a dependent variable and multiple, more than two, independent variables. It is calculated using the formula;

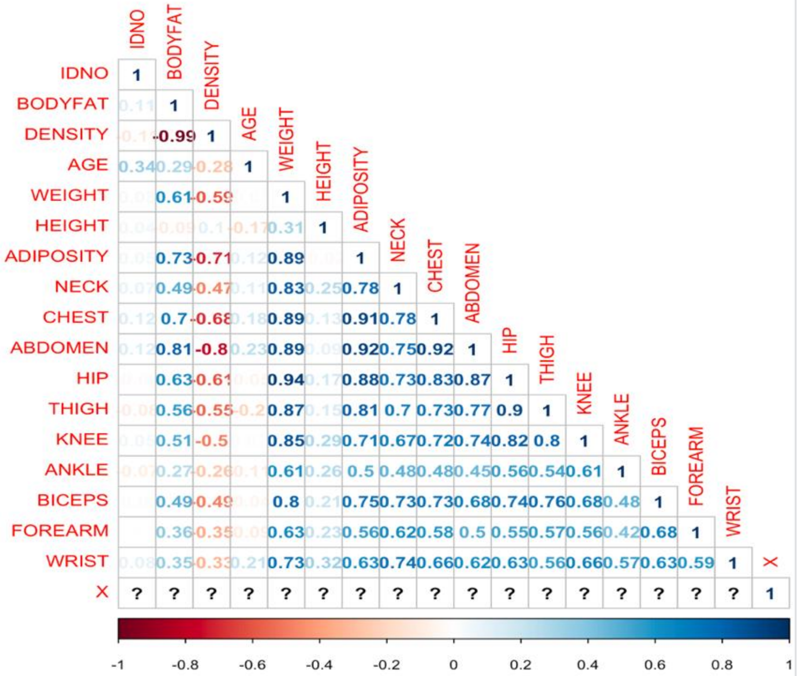
$$\hat{y} = \hat{\alpha} + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + \ldots + \hat{\beta}_q x_q$$

The slopes of each independent variable are used to determine the magnitude of prediction for each independent variable. The t-test is used to determine the statistical significance of each predictor. The R^2 coefficient of multiple determination is used to determine how much variance in the dependent variable is accounted for by the set of independent variables. The F-test is used to determine whether the set of independent variables collectively predicts the dependent variable. The adjusted R^2 value, which is always lower than the overall R^2, increases only if the new predictor improves the model more than what would be expected by chance. The best model is selected using a stepwise method, which starts with all predictors, and then sequentially drops worst predictors. Other variables that improve the model fit can be added after subtracting a variable.

Results

PART 1 – Explore Predictor Variables

Figure 1. Correlation Matrix



Code:

```
x <- cor(BodyFat)

corrplot(x, method = "number", type = "lower")
```

The figure above shows a correlation matrix of all the variables in the BodyFat dataset. It can be seen from the matrix that the three predictor variables that are the most correlated to bodyfat are adiposity, chest, and abdomen. Also, the density variable has a very strong negative correlation to bodyfat. However, using density as a predictor variable to bodyfat does not yield any significant results, as the scatterplot of density against bodyfat already has a regression line as seen below.

Figure 2. Density vs Bodyfat

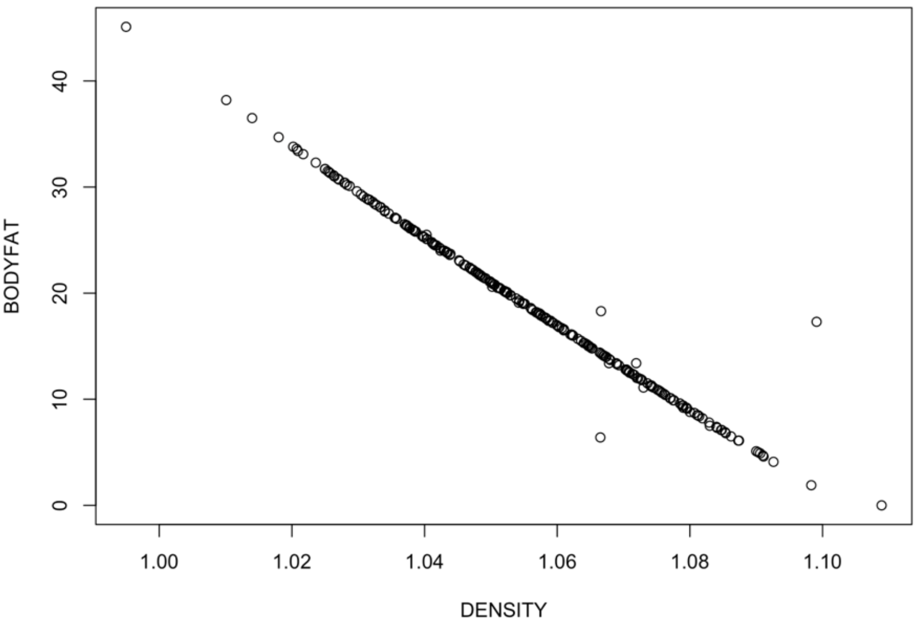


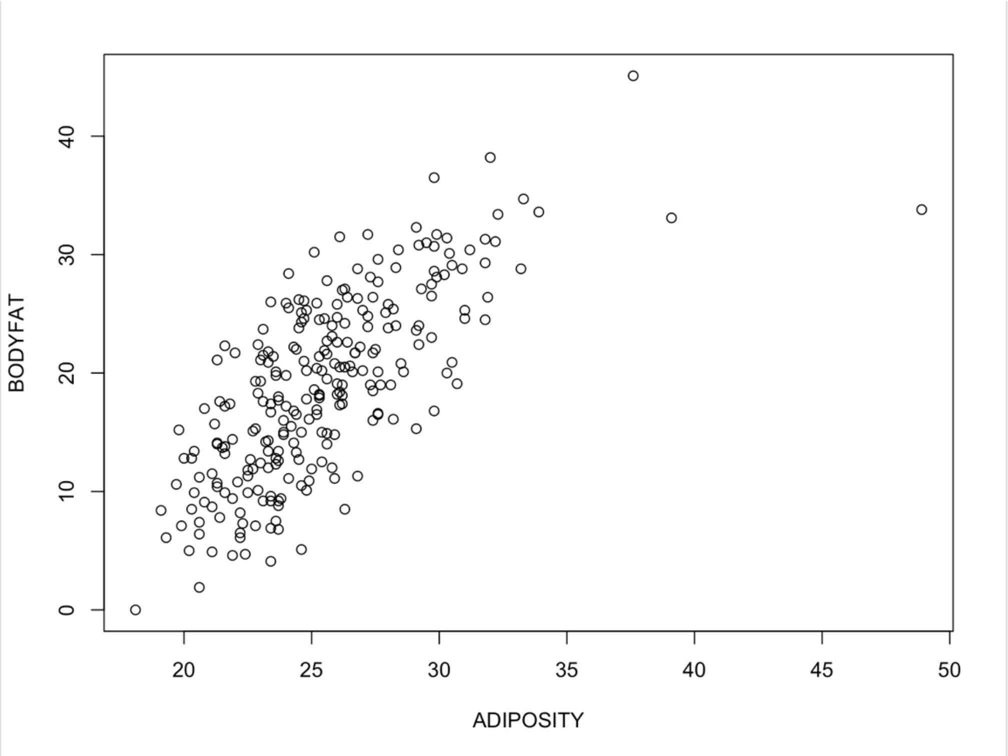
Table 1. Table of Correlation Coefficient of Predictor Variables with Interpretations

Predictor Variable	**Correlation Coefficient**	**Correlation Interpretation**
Density	-0.98	Very high negative
Abdomen	0.83	High positive
Adiposity	0.72	High positive
Chest	0.70	High positive
Hip	0.62	Moderate positive
Weight	0.61	Moderate positive
Thigh	0.56	Moderate positive
Knee	0.51	Moderate positive
Biceps	0.49	Low positive

Neck	0.49	Low positive
Forearm	0.36	Low positive
Wrist	0.35	Low positive
Age	0.29	Negligible
Ankle	0.27	Negligible
Height	0.09	Negligible

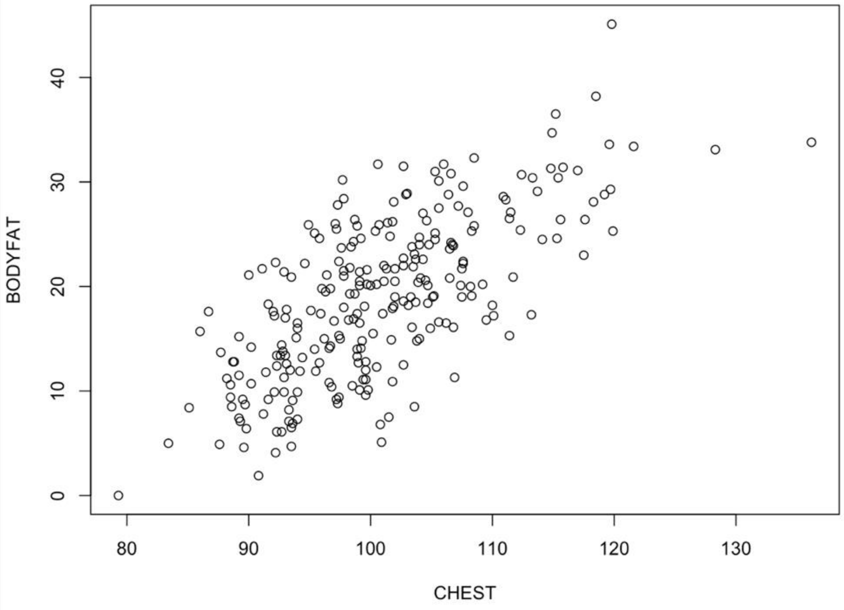
The table above shows, from high to low, each predictor variables' correlation coefficient to bodyfat as well as the interpretation of the correlation coefficient.

Figure 3. Adiposity vs Bodyfat



The plot above shows adiposity plotted against bodyfat. It can be seen from the scatterplot that there is a positive correlation between the two, disregarding the three outliers.

Figure 4. Chest vs Bodyfat



The plot above shows chest plotted against bodyfat. It can be seen from the scatterplot above that the two also have a positive correlation, but not as strong as that between adiposity and bodyfat.

Code:

```
plot(BODYFAT ~ DENSITY + AGE + WEIGHT + HEIGHT + ADIPOSITY +
    NECK + CHEST + ABDOMEN + HIP + THIGH + KNEE + ANKLE +
    BICEPS + FOREARM + WRIST, data = BodyFat)
```

PART 2 – Select Predictor Variables and Check for Their Suitability

Set 1: Knee, Biceps, Age

Set 2: Adiposity, Thigh, Ankle

Set 3: Chest, Adiposity, Forearm

Set 4: Adiposity, Chest, Abdomen

The four sets seen above are created by separating the correlation table, seen in Table 1, into thirds accordingly to correlation coefficients from high to low. Set 1 includes two predictor variables from the middle third and one predictor variable from the bottom third. Set 2 includes one predictor variable from the top third, one predictor variable from the middle third, and one predictor variable from the bottom third. Set 3 includes two predictor variables from the top third and one predictor variable from the middle third. Set 4 includes three predictor variables from the top third.

Figure 5. Adiposity Histogram

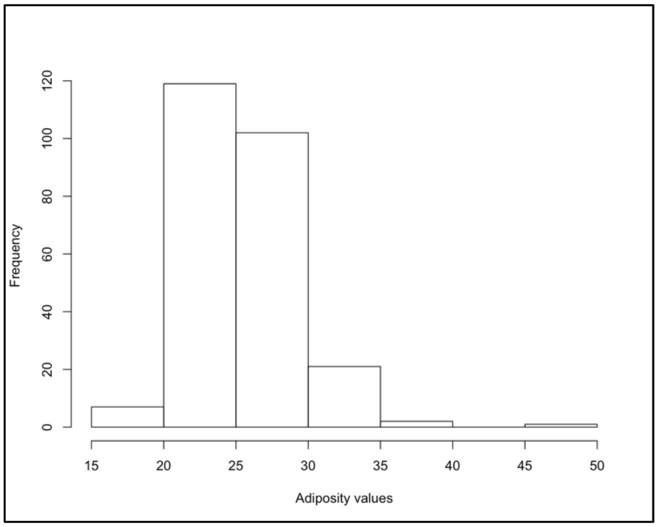


Figure 6. Adiposity Q-Q Plot

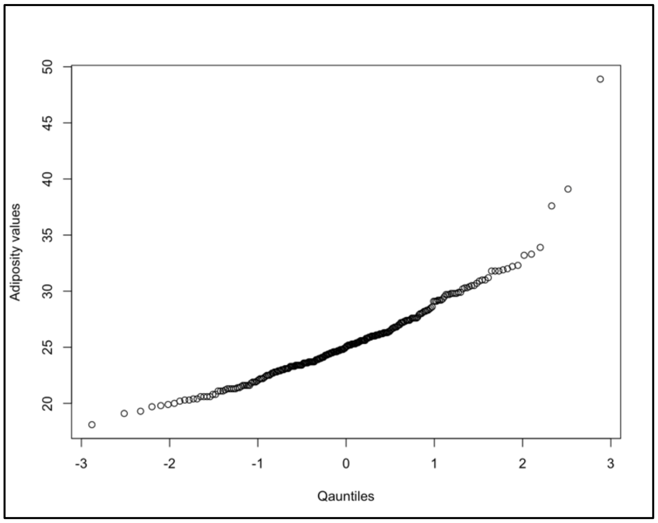


Figure 7. Thigh Histogram

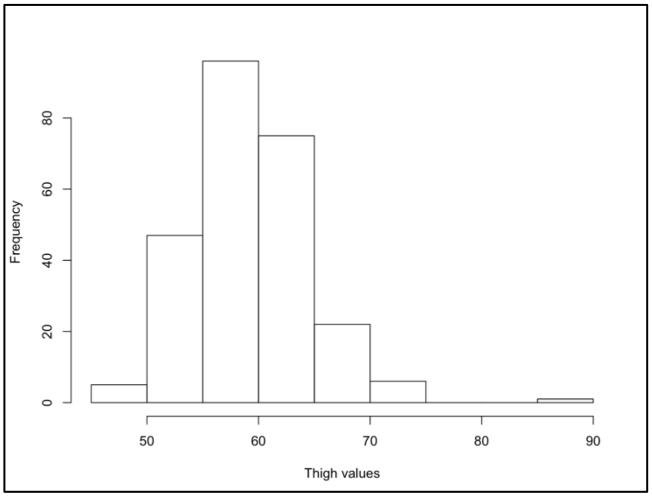


Figure 8. Thigh Q-Q Plot

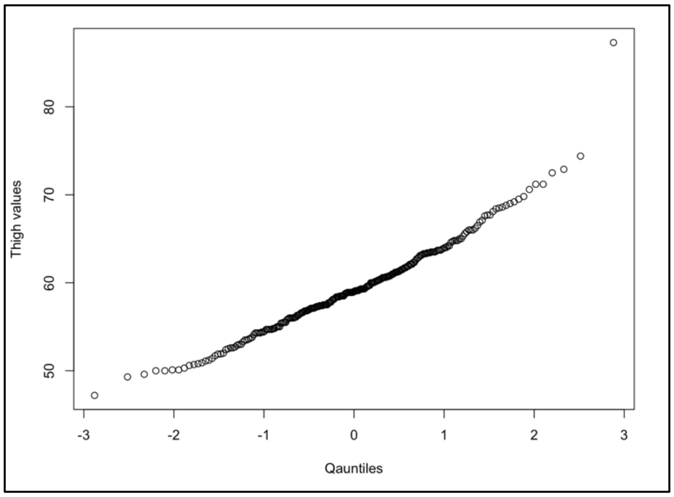


Figure 9. Ankle Histogram

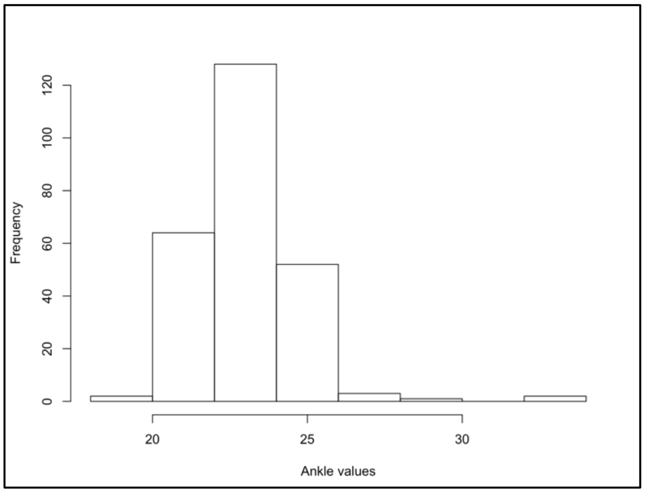
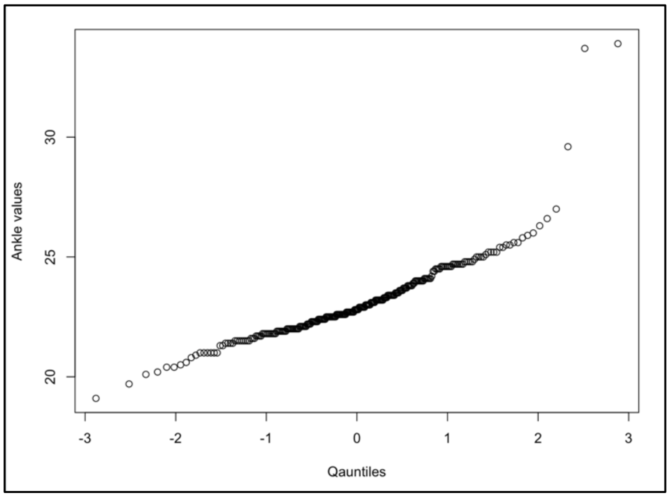


Figure 10. Ankle Q-Q Plot



The six figures above show histograms and Q-Q plots for the adiposity, thigh, and ankle predictor values used in Set 2. Examining the figures above, it is possible to conclude that all predictor variables are randomly sampled, in that every possible measurement that could be selected from these samples has a predetermined probability of being selected. This means that, the selection of one value from these samples is based on chance, and that every value in these samples has a known, non-zero probability of being selected. The histograms shown in figures 5, 7, and 9 all verify that the three predictor variables are randomly sampled, and also show different frequencies that result in different non-zero probabilities for any value selected from the samples. The three predictor variables are also independent observations as the occurrence of one observation in these samples provides no information about the occurrence of another observation. For example, an adiposity measurement of one man from the sample has no effect on another adiposity measurement of another man, which is also true for thigh and ankle measurements. The Q-Q plots for the three predictor variables, shown in figures 6, 8, 10, all portray a linear graph, disregarding any outliers. This shows that the residuals and the predictor variables have a linear relationship, and that all three predictor variables are normally distributed. As the adiposity, thigh, and ankle predictor variables are randomly sampled, are independent observations, and are normally distributed, they are suitable for a regression analysis.

Code:

```
adiposity <- BodyFat$ADIPOSITY

hist(adiposity, main = "", xlab = "Adiposity values")

qqnorm(adiposity, main = "", xlab = "Qauntiles", ylab = "Adiposity values")
```

Table 2. Collinearity between the three predictor variables

Predictor Variables	Correlation Coefficient
Adiposity - Thigh	0.81
Adiposity - Ankle	0.50

Thigh - Ankle	0.54

Multicollinearity occurs when independent variables in a regression model are correlated.[1] The table shown above displays the correlation between the three predictor variables used in Set 2. It can be seen that there is a high positive correlation between adiposity and thigh values, and a moderate positive correlation between adiposity and ankle values, and between thigh and ankle values. This result shows that the three predictor variables are collinear. Our regression analysis enables us to isolate the relationship between each of our predictor variables to our dependent variable, bodyfat. However, our predictor variables being collinear makes it difficult for our model to estimate the relationship between each predictor variable to bodyfat independently, as changes in one predictor variable are also associated with changes in another predictor variable. This means that the precision of the regression coefficient estimates in our model, which represent the mean change in the dependent variable, bodyfat, for each unit change in a predictor variable while holding other predictor variables constant, will be affected.

Code:

```
cor(adiposity, thigh)

cor(adiposity, ankle)

cor(thigh, ankle)
```

[1] https://statisticsbyjim.com/regression/multicollinearity-in-regression-analysis/

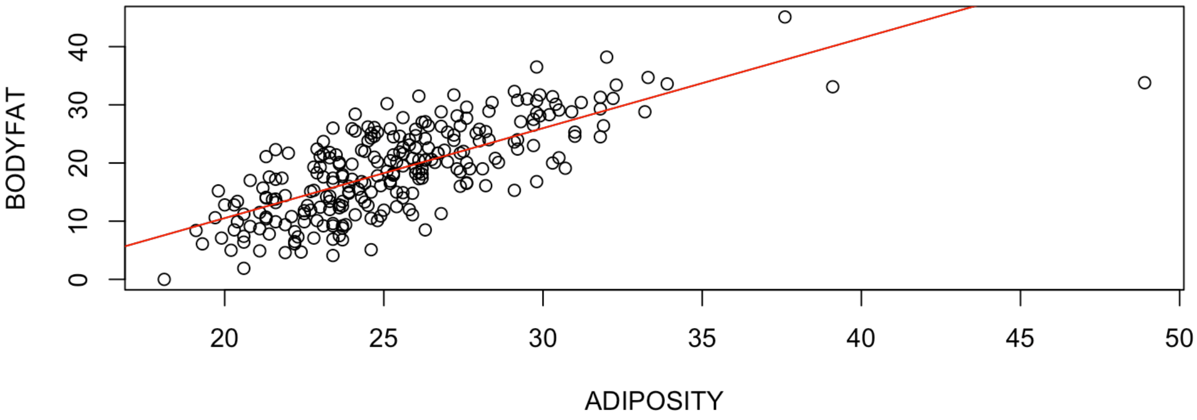
PART 3 – Create Linear and Multiple Linear Regression Models

Table 3. Linear Regression of three predictor variables in Set 2 to Bodyfat

Predictor Variable	**R^2 value**	**Slope**	**Y-Intercept**
Adiposity	0.52	1.54	-20.40
Thigh	0.31	0.82	-30.29
Ankle	0.06	1.22	-9.24

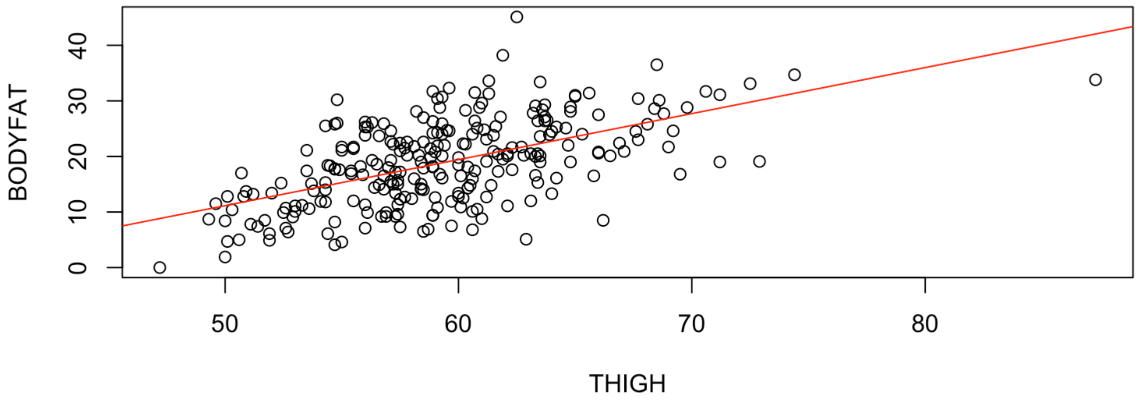
The table above shows R^2 values, slopes, and y-intercepts, which are the results of the linear regression analysis on each of the three predictor variables in Set 2 to bodyfat. The R^2 values are the squared values of the correlation coefficients show in Table 1.

Figure 11. Adiposity vs Bodyfat Scatterplot



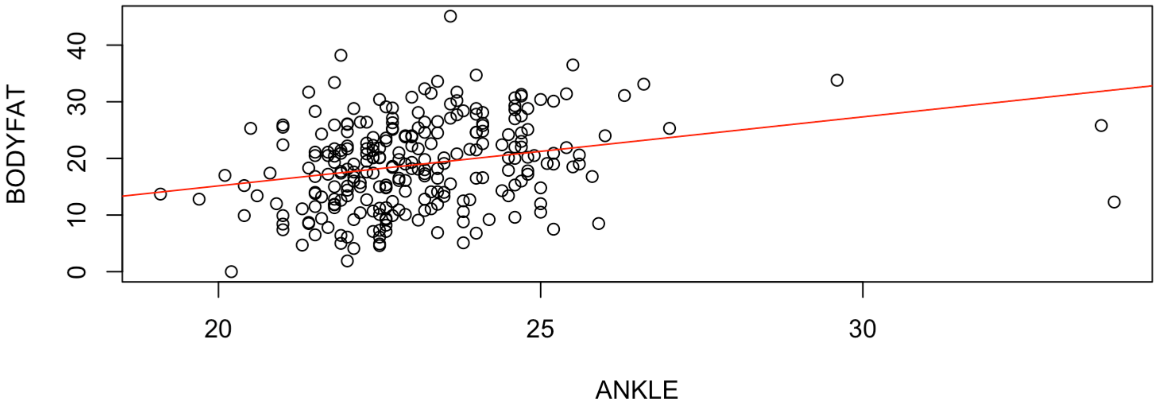
The figure above shows a scatterplot of adiposity values plotted against bodyfat, with the red line showing the regression line. As it can be seen from the slope of the regression line, there is a significant association between adiposity and bodyfat. The R^2 value of adiposity to bodyfat in Table 3 also confirms this result and shows that there is a strong relationship between adiposity values and bodyfat.

Figure 12. Thigh vs Bodyfat Scatterplot



The figure above shows a scatterplot of thigh values plotted against bodyfat, with the red line showing the regression line. It can be seen from the regression line that there is a relationship between thigh values and bodyfat, however, this relationship is not as significant as that between adiposity and bodyfat, as this regression line seems to have a lesser slope than that of the former. The R^2 value of thigh values to bodyfat in Table 3 shows that there is moderately strong relationship between the two.

Figure 13. Ankle vs Bodyfat Scatterplot



The figure above shows a scatterplot of ankle values plotted against bodyfat, with the red line showing the regression line. As it can be seen from the scatterplot, the regression line has a lesser slope and is significantly flatter compared to that of the previous two scatterplots, indicating that there is not a significant association between bodyfat and ankle measurements. The R^2 value of ankle to bodyfat in Table 3 is also quite small, showing that the relationship between ankle values and bodyfat is negligible.

Table 4. Overall and Adjusted R^2 values for the four sets

Set	**Overall R^2**	**Adjusted R^2**
Set 1	0.38	0.37
Set 2	0.54	0.53
Set 3	0.57	0.56
Set 4	0.67	0.67

The table above shows the overall and adjusted R^2 values of a multiple regression analysis for the four different sets mentioned in the previous part. It can be seen from the table above that Set 4 has the highest R^2 value.

Table 5. Y-Intercept and Slope Comparisons of predictor variables for Set 4

Y-Intercept	-27.97	
Predictor Variables	**Multiple Regression Slope**	**Simple Linear Regression Slope**
Chest	-0.22	0.64
Abdomen	0.78	0.58
Adiposity	-0.11	1.54

The table above shows the y-intercept of the Set 4 model along with the slopes for each predictor variable in the model. The column on the right also shows the slopes for each predictor variable from their simple linear regression to bodyfat. These two slope values for each of the three predictor variables differ as, whereas only the predictor variable contributes to the slope of the regression line for a simple linear regression analysis, other predictor values also affect the slope for a multiple regression analysis. In regression analysis, the slope value analyzes the proportion of variance in independent variables that covaries with the dependent variable, meaning that it is the mean amount of a change in the dependent variable when an independent variable increases by one unit. As the predictor variables are collinear, a change in a predictor variable affects both the independent variable, bodyfat, and also other predictor variables. This results in predictor variables having differing slopes in the model compared to their simple linear regression.

Table 4 shows the overall and adjusted R^2 values for all four sets. Whereas the overall R^2 value increases simply by adding new variables, the adjusted R^2 value increases only if the new variable improves the model more than would be expected by chance. This means that the adjusted R^2 value is always less than the overall R^2 value and is a more accurate measure of correlation as it addresses the problem of the overall R^2 value. Examining the adjusted R^2 values for the four sets in Table 4, it can be seen that the values are different for all the four sets, and that Set 4 has the largest value, followed by Set 3, Set 2, and then Set 1. This result is expected since all four sets have different predictor variables that have different correlation coefficients to bodyfat. Since all three predictor variables of Set 4 are from the top third of the correlation table, seen in Table 1, and are the most correlated variables to bodyfat compared to the other variables of the three other sets, it makes sense that the model has the highest correlation coefficient to bodyfat and is the best fit model. The overall fit of the model shows that 67 percent of the variance in bodyfat is account for by the chest, abdomen, and adiposity values.

The significance of this model can be determined using a F-test. The F-test, which evaluates the null hypothesis that all regression coefficients are equal to zero versus the alternative that at least one is not, tells whether the regression model provides a better fit to the data than a model that contains no independent variables. Examining the p-value for the model, which is much less than an alpha level of 0.05 using a 0.05 significance level, it is possible the reject the null hypothesis. This result of the F-test shows that the model is significant.

My Model: Abdomen and weight predictor variables to predict bodyfat.

Table 6. Y-Intercept, R^2, and Slope value for my model

Y-Intercept	-41.35

Adjusted R^2 value	0.72
Predictor Variable	**Slope**
Abdomen	0.91
Weight	-0.13

The table above shows the y-intercept, the adjusted R^2 value, and the slope values for the predictor variable for my model. It can be seen that this model has a higher R^2 value compared to that of Set 4, seen in Table 5, meaning that it is a better fit to explain the variability in bodyfat values. While creating my model, I was aiming the create the best model to predict bodyfat values, so I started with the top five predictor variables that are the most correlated to bodyfat in Table 1. This initial model included the abdomen, adiposity, chest, hip, and weight variable. I then performed a stepwise model selection by comparing AIC values, an estimator of out-of-sample prediction error, to find the best model. The initial model having five variables had an AIC value of 722.34, whereas the final model containing the two predictor variables abdomen and weight had an AIC value of 717.45. Conducting an ANOVA analysis on the two models shows that the simple model is better than the complex initial model, as the p-value of the F-test is not sufficiently low.

PART 4 – Body Mass Index

To create a model using body mass index, I added the body mass index variable to the other two predictor variables, abdomen and weight, of my previous model shown in Table 6. To consider issues related to collinearity and multicollinearity, I computed the correlation coefficient of the body mass index variable to the abdomen and weight variables, and these values were 0.41 and

0.39 respectively. Both these correlation coefficients fall into to the ‘low positive correlation’ category, which minimizes the effect of multicollinearity to the model.

Table 7. Values for a Predictive Model for Bodyfat using BMI

Y-Intercept	-41.05
Adjusted R^2 value	0.72
Predictor Variable	**Slope**
Abdomen	0.91
Weight	-0.13
Body Mass Index	0.04

The table above shows the results of a multiple regression analysis of adding the body mass index variable to the previous model, shown in Table 6. It can be seen that although this model has a R^2 value slightly higher than that of the previous model, the body mass index does not contribute much to predict bodyfat values in this model as it has a very low slope value of 0.04. Also, the p-value from a t-test for body mass index in this model is 0.13 which is higher than alpha level of 0.05 using a 0.05 significance level. This shows that there is not a significant linear relationship between the body mass index variable and the dependent bodyfat variable for this model. However, this model is still significant as its p-value from a F-test is still significantly small. When the body mass index variable is added to the four sets shown in Table 4, the result of the t-test is still the same for each of these four model and the body mass index variable does not offer a significant effect in explaining variability in bodyfat values for these four models as well. This result is expected, as the simple linear regression coefficient, the R^2 value, of body mass index to bodyfat is 0.13, which shows that the body mass index has a negligible effect in explaining variability in bodyfat values.

Table 8. WHO BMI weight classifications.

BMI	**Classification**
< 18.5	Underweight
18.5 – 24.99	Normal weight
25 – 29.99	Overweight
>= 30	Obese

When performing a similar regression analysis using the same predictor variables shown in Table 7 for each class of BMI, shown in the table above, it did not make sense to fit a predictive model for the underweight category since only 1 sample out of the 252 samples of the BodyFat dataset fell into the underweight BMI category.

Table 9. Values for a Predictive Model using Normal Weight BMI

Y-Intercept	9.20
Adjusted R^2 value	0.58
Predictor Variable	**Slope**
Abdomen (Normal weight)	0.43
Weight (Normal weight)	-0.16
Body Mass Index (Normal weight)	-0.13

The table above shows the y-intercept, the adjusted R^2 value, and the slopes of the predictor variables for a model for the normal weight BMI class. It can be seen that this model does not have a better fit compared to the overall model, shown in Table 7, as it has a lower R^2. Although the body mass index variable seems to have a larger slope value, in magnitude, its t-test still computes a p-value lower than 0.05 meaning that the body mass index variable for the normal weight category does not have a significant linear relationship to normal weight bodyfat values. That being said, the F-test for the normal weight model produces a significantly small p-value which shows that this normal weight model is significant.

Table 10. Values for a Predictive Model using Overweight BMI

Y-Intercept	73.94
Adjusted R^2 value	0.63
Predictor Variable	**Slope**
Abdomen (Overweight)	1.44
Weight (Overweight)	-0.54
Body Mass Index (Overweight)	-3.60

The table above shows the y-intercept, the adjusted R^2 value, and the slopes of the predictor variables for a model for the normal weight BMI class. It can be seen that although this model does not have a better fit than the overall model, shown in Table 7, since it has a lower R^2 value, it has a higher R^2 value than that of the normal weight model, shown in Table 9, meaning that is has a better fit than the normal weight model. Also, it can be seen from the table above that the body mass variable for the overweight category has a higher slope compared to that of the normal weight model. Furthermore, the t-test for the overweight body mass index variable produces a significantly small p-value, which shows that there is in fact a significant linear relationship between the overweight body mass index values and the overweight bodyfat values. The F-test for the overweight model also produces a significantly small p-value which shows that this overweight model is significant.

Table 11. Values for a Predictive Model using Obese BMI

Y-Intercept	-8.03
Adjusted R^2 value	0.87
Predictor Variable	**Slope**
Abdomen (Obese)	-0.15
Weight (Obese)	0.17
Body Mass Index (Obese)	0.07

The table above shows the y-intercept, the adjusted R^2 value, and the slopes of the predictor variables for a model for the obese BMI class. It can be seen from the table above, that the obese body mass index model explains 87 percent of the variation in obese bodyfat values. This R^2 value is higher than those of the overall, normal weight, and the overweight body mass index models, shown in tables 7, 9, and 10 respectively. Although having a higher R^2 value, it may not be easy to conclude that this obese body mass index model has a better fit and significance compared to the previous models, considering it has a very small sample size. The obese body mass index category has a sample size of 25, whereas the sample size for the normal weight, overweight, and the overall models were 124, 102, and 252 respectively. This small sample size of the model may have impacted the validity of the results of the regression analysis and the fit for predicting bodyfat values for the obese body mass index category. That being said, The F-test for this obese model produces a significantly small p-value which shows that the obese body mass index model is significant.

Conclusion

In conclusion, this report creates a predictive model for bodyfat values in the BodyFat dataset by conduction regression analysis using different independent variables from the dataset. Part 1 of the report explores the correlation of these independent variables and lists them from high to low accordingly to their correlation coefficient to bodyfat. This part also plots some predictor variables against bodyfat to get a feel for the dataset. Part 2 creates four sets using three predictor variables and examines if the three predictor variables of Set 2 are suitable for a regression analysis. This part also explores multicollinearity and collinearity between these three predictor variables. Part 3 performs a simple linear regression analysis of these three variables of set 2 to the dependent variable, bodyfat. This part also does multiple regression analysis on the four sets against bodyfat, and also creates another model to fit the bodyfat values the best. Part 4 creates a predictive model for bodyfat using an additional body fat index variable and examines four different models using the four BMI weight classifications.

Code:

```
## PART 1
remove(list = ls())
install.packages("corrplot")
library(corrplot)
BodyFat <- read.csv("BodyFat.csv")

x <- cor(BodyFat)
corrplot(x ,method = "number", type = "lower")

plot(BODYFAT ~ DENSITY + AGE + WEIGHT + HEIGHT + ADIPOSITY +
     NECK + CHEST + ABDOMEN + HIP + THIGH + KNEE + ANKLE +
     BICEPS + FOREARM + WRIST, data = BodyFat)

## PART 2
adiposity <- BodyFat$ADIPOSITY
hist(adiposity, main = "", xlab = "Adiposity values")
qqnorm(adiposity, main = "", xlab = "Qauntiles", ylab = "Adiposity values")

thigh <- BodyFat$THIGH
hist(thigh, main = "", xlab = "Thigh values")
qqnorm(thigh, main = "", xlab = "Qauntiles", ylab = "Thigh values")

ankle <- BodyFat$ANKLE
hist(ankle, main = "", xlab = "Ankle values")
qqnorm(ankle, main = "", xlab = "Qauntiles", ylab = "Ankle values")

cor(adiposity, thigh)
cor(adiposity, ankle)
cor(thigh, ankle)

## PART 3
# simple linear regression
model1 <- lm(BODYFAT ~ ADIPOSITY, data = BodyFat)
```

```
summary(model1)
plot(BODYFAT ~ ADIPOSITY, data = BodyFat)
abline(model1, col = "red")

model2 <- lm(BODYFAT ~ THIGH, data = BodyFat)
summary(model2)
plot(BODYFAT ~ THIGH, data = BodyFat)
abline(model2, col = "red")

model3 <- lm(BODYFAT ~ ANKLE, data = BodyFat)
summary(model3)
plot(BODYFAT ~ ANKLE, data = BodyFat)
abline(model3, col = "red")

# multiple regression
m1 <- lm(BODYFAT ~ KNEE + BICEPS + AGE, data = BodyFat)
summary(m1)

m2 <- lm(BODYFAT ~ ADIPOSITY + THIGH + ANKLE, data = BodyFat)
summary(m2)

m3 <- lm(BODYFAT ~ ADIPOSITY + CHEST + WRIST, data = BodyFat)
summary(m3)

m4 <- lm(BODYFAT ~ ADIPOSITY + ABDOMEN + CHEST, data = BodyFat)
summary(m4)

# my model
library(MASS)
fit <- lm(BODYFAT ~ ABDOMEN + ADIPOSITY + CHEST + HIP + WEIGHT, data = BodyFat)
step <-stepAIC(fit, direction="both")
step$anova

fit2 <- lm(BODYFAT ~ ABDOMEN + WEIGHT, data = BodyFat)
summary(fit2)
```

```
anova(fit, fit2)


## PART 4
pounds <- BodyFat$WEIGHT
inches <- BodyFat$HEIGHT
bmi <- pounds * 703 / (inches^2)

abdomen <- BodyFat$ABDOMEN
cor(bmi, pounds)
cor(bmi, abdomen)

bmi_fit <- lm(bodyfat ~ abdomen + pounds + bmi, data = BodyFat)
summary(bmi_fit)

summary(lm(bodyfat ~ bmi))

# BMI classes
under <- bmi[bmi < 18.5]
under

bmi_normal <- bmi[bmi >= 18.5 & bmi <= 24.99]
abd_normal <- abdomen[bmi_normal]
wght_normal <- pounds[bmi_normal]
bfat_normal <- bodyfat[bmi_normal]

fit_normal <- lm(bfat_normal ~ bmi_normal + abd_normal + wght_normal)
summary(fit_normal)

bmi_over <- bmi[bmi >= 25 & bmi <= 29.99]
abd_over <- abdomen[bmi_over]
wght_over <- pounds[bmi_over]
bfat_over <- bodyfat[bmi_over]

fit_over <- lm(bfat_over ~ bmi_over + wght_over + abd_over)
summary(fit_over)
```

```
bmi_obese <- bmi[bmi >= 30]
abd_obese <- abdomen[bmi_obese]
wght_obese <- pounds[bmi_obese]
bfat_obese <- bodyfat[bmi_obese]

fit_obese <- lm(bfat_obese ~ bmi_obese + wght_obese + abd_obese)
summary(fit_obese)
```