MATH1324 Introduction to Statistics Assignment 3

Aidan Cowie

3 June 2018

Importing And Cleaning The Data

Firstly, I chose to use mosaic package for the mathematics and statistics aspects, the readxl package to upload the data into R from excel, and Rcmdr for leveneTest statistics:

```
library(mosaic)
library(readxl)
library(Rcmdr)
Body<- read_excel("C:/Users/Aidan/Desktop/Body (2) Assignment 3 data.xlsx")
```

By checking the data types I realised that the sex was classified as an integer and had to be change to a factor and relabelled as such, and that there were a couple of body fat % less than 1% which is near impossible and not healthy the lowest ever recorded was near 0 who was a body builder and died due to his low body fat % most stop at 2%:

```
Body <- Body[!(Body$BFP_Brozek < 1),]
Body$Sex[Body$Sex==1]<- 'Male'
Body$Sex[Body$Sex==0]<- 'Female'
Body$Sex<- as.factor(Body$Sex)
```

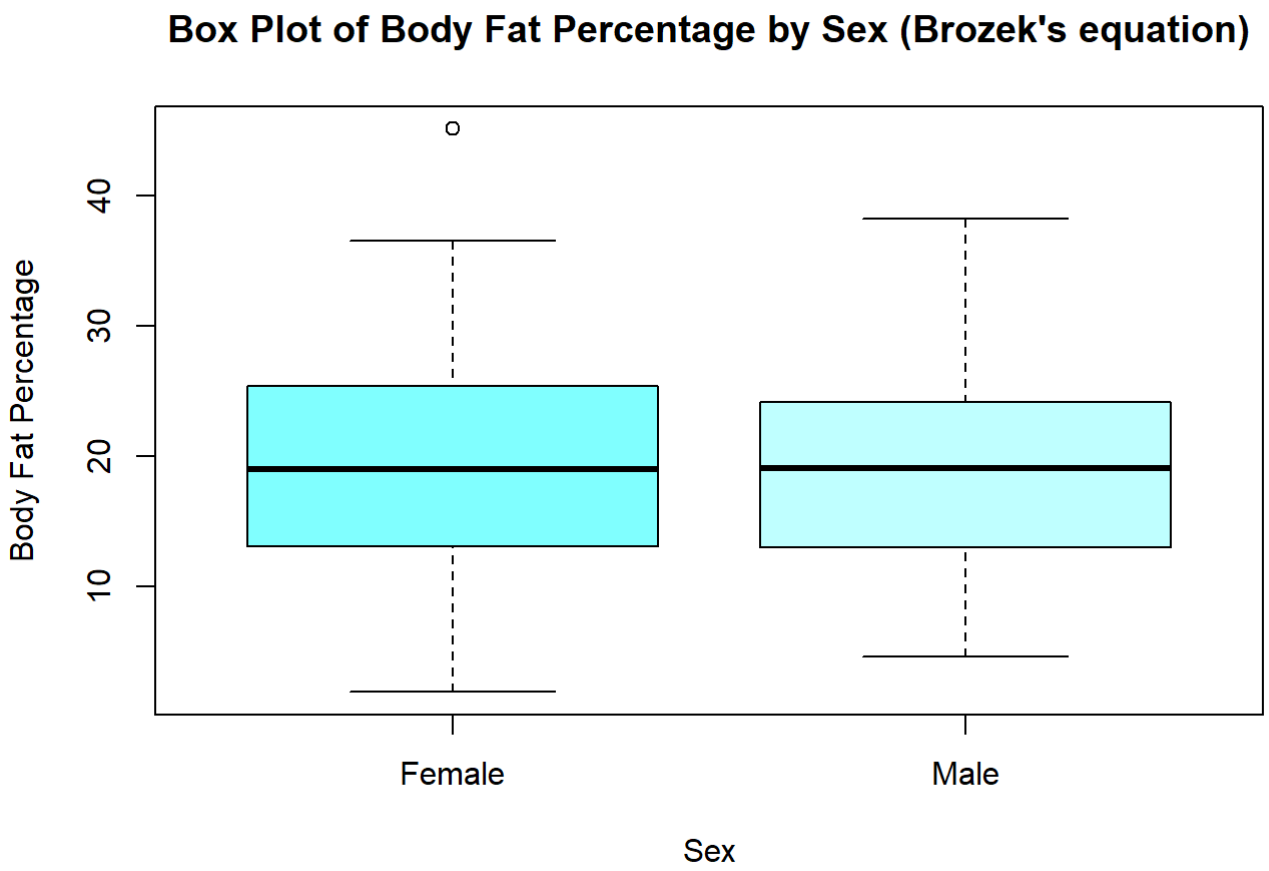
Since the serve outliers were removed, the data can be used for statistic evaluation. However, before this we want to split the data into male and female segregation shown below:

```
body_male<- subset(Body, Sex == 'Male')
body_female<- subset(Body, Sex == 'Female')
```

Problem Statement Part 1

The aim of this statistics analysis is to determine the distribution between an individual's Body Fat Percentage and their sex. The null hypothesis, H_0: That males and females will have the same mean body fat percentage. The alternative hypothesis, H_A: That male and females will have a different mean of body fat percentages.

```
boxplot(Body$BFP_Brozek ~ Body$Sex, data=Body, main="Box Plot of Body Fat Percentage by Sex
 (Brozek's equation)", ylab="Body Fat Percentage", xlab = "Sex", col=cm.colors(5))
```




```
favstats(BFP_Brozek ~ Sex, data=Body)
```

```
##      Sex min    Q1 median    Q3  max     mean       sd   n missing
## 1 Female 1.9 13.05  19.00 25.40 45.1 19.63626 8.218279  91       0
## 2   Male 4.6 13.10  19.05 24.05 38.2 18.66000 7.348052 160       0
```

```
leveneTest(BFP_Brozek ~ Sex, data=Body)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group   1  1.6106 0.2056
##       249
```

The p-value for the Levene's test of equal variance for body fat % between males and females was p = 0.21. We find p > .05, therefore, we fail to reject H_0!. Thus, we are safe to assume equal variance.

```
t.test(BFP_Brozek ~ Sex, data=Body, var.equal=TRUE, alternative="two.sided")
```

```
##
##  Two Sample t-test
##
## data:  BFP_Brozek by Sex
## t = 0.96892, df = 249, p-value = 0.3335
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.008192  2.960720
## sample estimates:
## mean in group Female   mean in group Male
##             19.63626             18.66000
```

The two-sample t-test has the following statistical hypotheses:

where μ_1 and μ_2 refer to the population means of Females and Males respectively. The null hypothesis is simply that the difference between the two independent population means is 0. The difference between males and females estimated by the sample was 19.63626 - 18.660 = 0.969.

```
qt(0.025,df = 250)
```

```
## [1] -1.969498
```

As the test statistic t from the two-sample t-test assuming equal variance was t = 0.96892, which is less extreme than +/- 1.969498, we cannot reject H_0. According to the critical value method, there wasn't a statistically significant difference between male and female body temperature means.

The p-value of the two-sample t-test will tell us the probability of observing a sample difference between the means of 0.969, or one more extreme, assuming the difference was 0 in the population. The two-tailed p-value was reported to be p = 0.3335.

According to the p-value method, as $p = 0.334 > \alpha = 0.05$, we fail to reject H_0. Therefore, the results are not statistically significant.

Problem Statement Part 2

To estimate the 99% confidence interval for the mean body fat percentage in the population. We first have to assume that our sample size accurately captures the whole population and that the sample body fat percentage follows a standard distribution.

"100 (1− α)% CI, is an interval estimate for a population parameter, based on a given sample statistic, where if samples of a certain size n were repeatedly drawn from the population and a CI for each sample's statistic was calculated, 100(1−α)% of these intervals would capture the population parameter, whereas the other 100(α)% would not."

From here we can estimate that our sample size of 250 applicants mean and standard deviation represents the populations. Thus, the mean is 19.014 and SD is 7.673 respectively for our estimate of the 99% confidence interval.

As we are estimating two parameters from the sample, the population mean, μ, and standard deviation, σ, we need to consider the extra uncertainty or error associated with the estimation to ensure the expected coverage of the CI remains at the desired level. The family of t-distributions are used for this purpose.

```
confint(t.test(~BFP_Brozek, data=Body, conf.level = 0.99))
```

```
##   mean of x    lower    upper level
## 1  19.01394 17.75683 20.27106  0.99
```

From our previous estimate of the population the lower bound mean = 17.757 and the upper bound mean = 20.271. This estimation states that 99% of the random samples taken from the population, will have its mean value lie within the lower and upper bound calculated.

Problem Statement Part 3

Researchers believe that average body fat percentage is less than 12.5. Thus $H_0 < 12.5$ and $H_A > 12.5$, for this experimental test.

```
t.test(~BFP_Brozek, data=Body ,mu = 12.5, alternative="greater")
```

```
##
##  One Sample t-test
##
## data:  BFP_Brozek
## t = 13.45, df = 250, p-value < 2.2e-16
## alternative hypothesis: true mean is greater than 12.5
## 95 percent confidence interval:
##  18.21435      Inf
## sample estimates:
## mean of x
##  19.01394
```

A greater one-tailed test was used to determine if the mean body fat percentage readings were significantly different from the previously assumed population mean of less than 12.5%. The 0.05 level of significance was used. The sample's mean oral body temperature was M = 19.014%, SD = 7.673%. The results of the one-sample t-test found the mean body fat percentage to be statistically significantly greater than the population oral mean temperature, t(250) = 13.45, p < .001, 95% CI [18.21435, Inf].

Problem Statement Part 4

Find the single best predictor of body fat percentage (Brozek method) using the body circumference data. Write a report that explains your method for identifying the single best predictor. Use the best predictor to determine a model that can convert a person's body circumference measurement to an estimated body fat percentage. Ensure you test the model parameters and any assumptions. Critique the predictive ability of the model and draw an overall conclusion to help the investigators

```
bfp_neck<- lm(BFP_Brozek ~ Neck, data = Body)
bfp_chest<- lm(BFP_Brozek ~ Chest, data = Body)
bfp_abdomen<- lm(BFP_Brozek ~ Abdomen, data = Body)
bfp_hip<- lm(BFP_Brozek ~ Hip, data = Body)
bfp_thigh<- lm(BFP_Brozek ~ Thigh, data = Body)
bfp_Knee<- lm(BFP_Brozek ~ Knee, data = Body)
bfp_Ankle<- lm(BFP_Brozek ~ Ankle, data = Body)
bfp_Biceps<- lm(BFP_Brozek ~ Biceps, data = Body)
bfp_Forearm<- lm(BFP_Brozek ~ Forearm, data = Body)
bfp_Wrist<- lm(BFP_Brozek ~ Wrist, data = Body)
```

```
msummary(bfp_abdomen)
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -34.76949    2.48522  -13.99   <2e-16 ***
## Abdomen       0.58051    0.02665   21.79   <2e-16 ***
##
## Residual standard error: 4.51 on 249 degrees of freedom
## Multiple R-squared:  0.6559, Adjusted R-squared:  0.6545
## F-statistic: 474.6 on 1 and 249 DF,  p-value: < 2.2e-16
```

Body Fat Percentage vs Abdomen circumference (cm) had the highest R^2 value at 0.6545 which reflects the proportion of variability in the dependent variable that can be explained by a linear relationship with the predictor variable. Therefore, Abdomen circumference (cm), explained 65.6% of the variability in final Body Fat Percentage readings. The R2 is a measure of goodness of fit for linear regression. The better the line fits the data (i.e. the closer the data points sit on the line) the higher R2 will be.

```
coef(summary(bfp_abdomen))
```

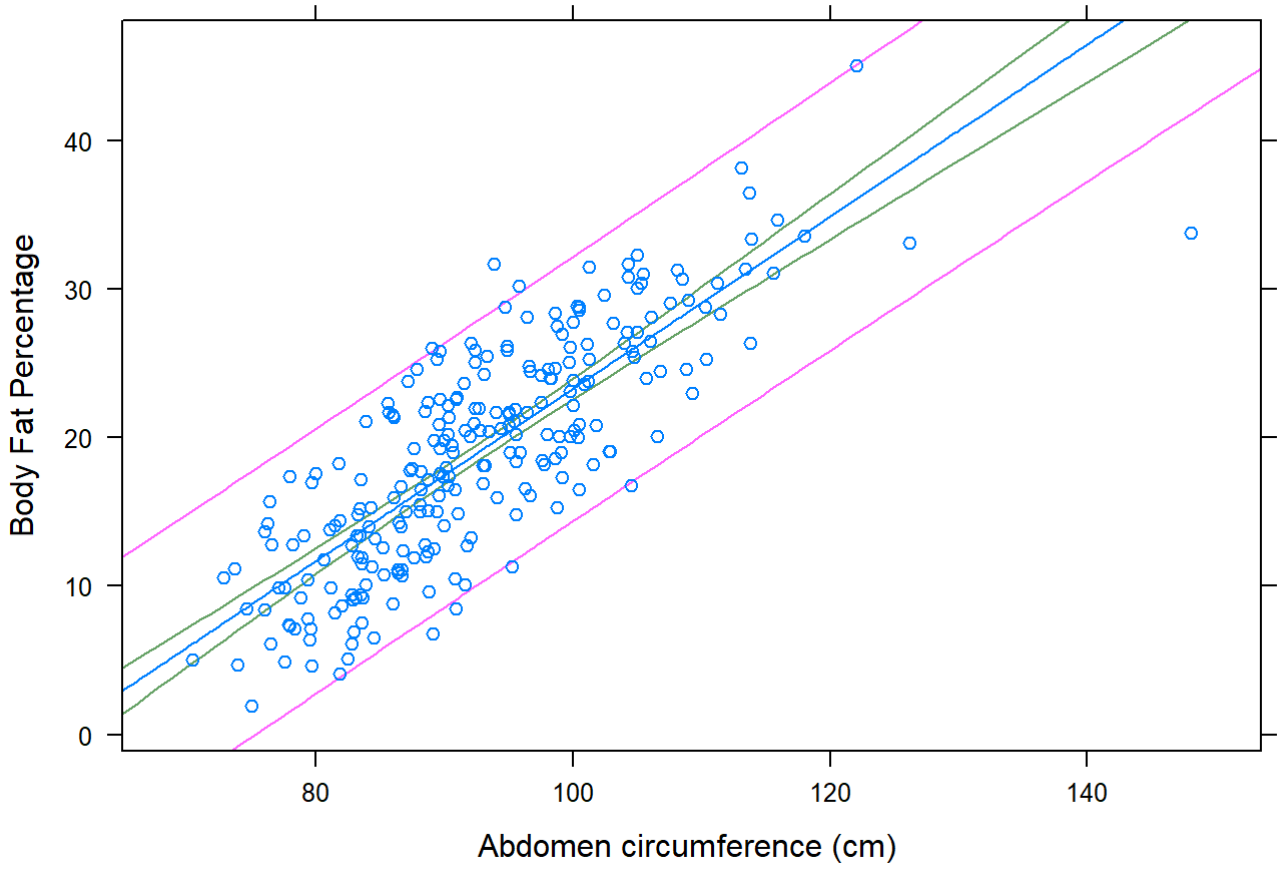
```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) -34.7694865 2.48522364 -13.99049 3.294667e-33
## Abdomen       0.5805124 0.02664773  21.78468 1.304381e-59
```

```
confint(bfp_abdomen)
```

```
##                   2.5 %      97.5 %
## (Intercept) -39.6642261 -29.8747469
## Abdomen       0.5280287   0.6329961
```

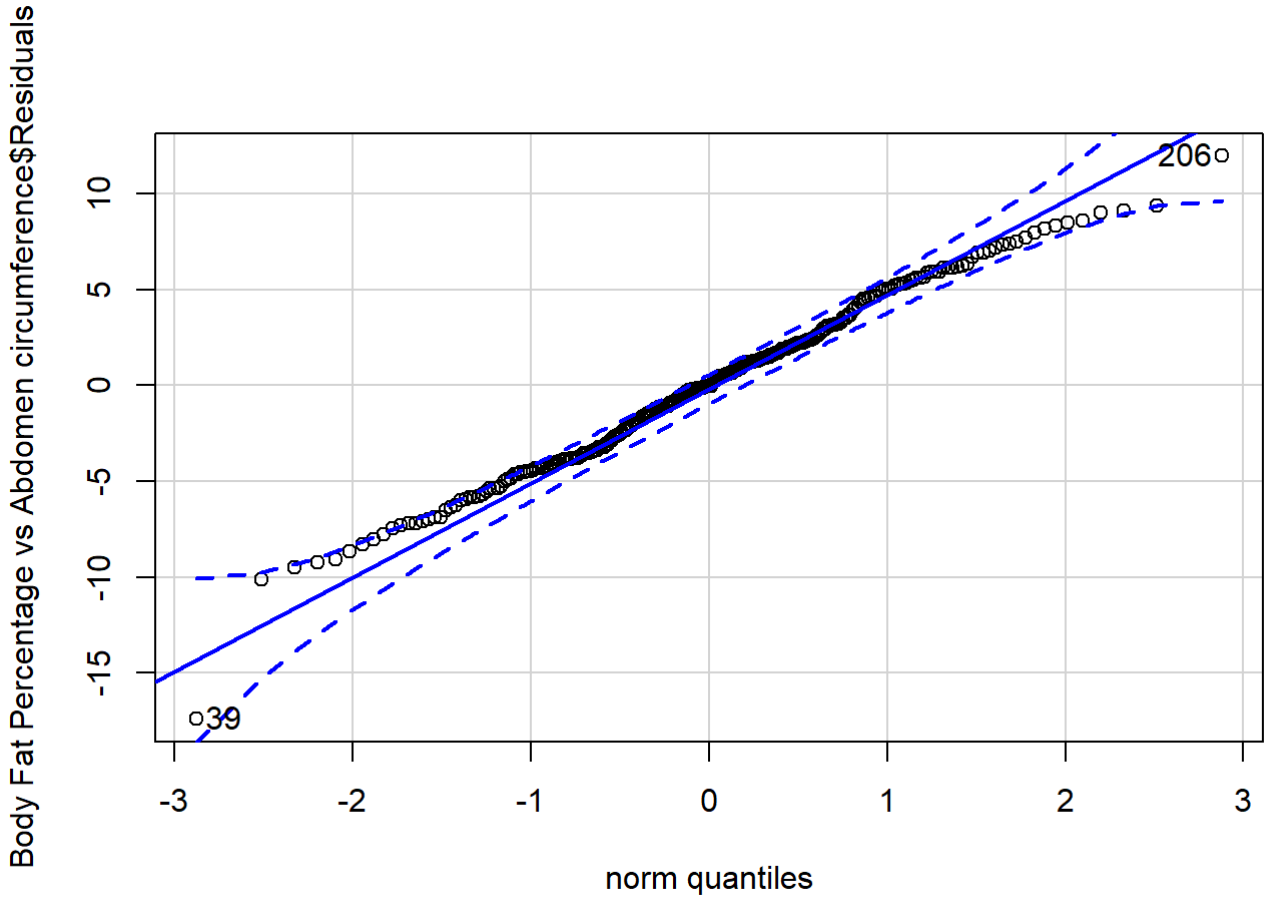
The intercept/constant is reported as a = 2106.864. The constant, or intercept, is the average value for y when x = 0. In this example, this value represents the average V02 max score when OUES is equal to 0. Given that an OUES of 0 is impossible (assuming you’re alive) the constant typically has no meaningful interpretation. To test the statistical significance of the constant, we set the following statistical hypotheses:

```
xyplot(BFP_Brozek ~ Abdomen, data = Body, ylab ="Body Fat Percentage",xlab="Abdomen circumfer
ence (cm) ",panel=panel.lmbands)
```



The data exhibits a positive linear trend. The blue line is the line of best fit for the linear regression. The green bands represent the 95% CI of mean Body Fat Percentage % readings for the regression line. The pink outer lines are the prediction intervals. The prediction intervals are where 95% of the data will fall assuming the residuals are normally distributed. There are a few outliers to this exception shown in the figure above.

```
qqPlot(bfp_abdomen$residuals, dist="norm", ylab ="Body Fat Percentage vs Abdomen circumferenc
e$Residuals")
```



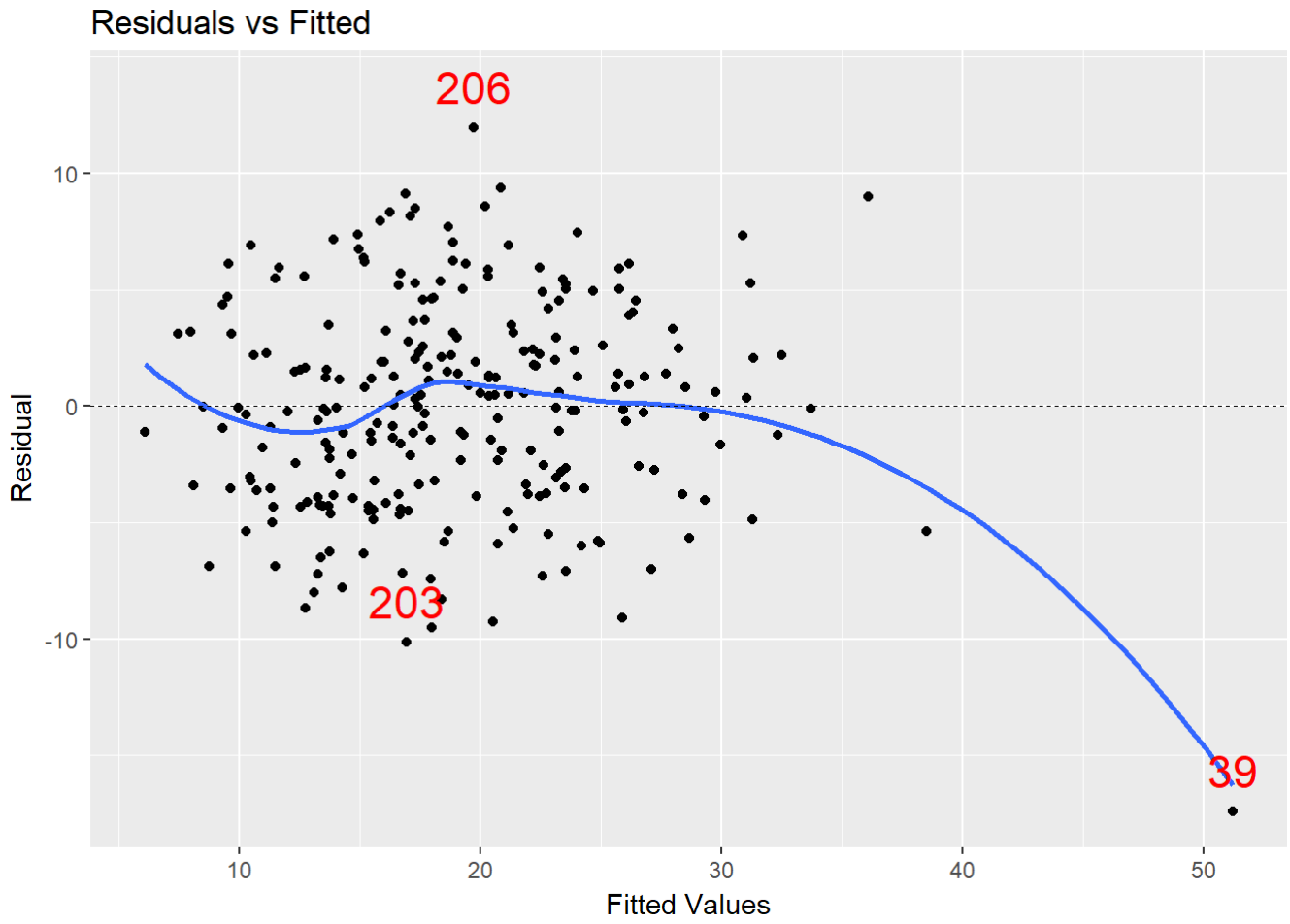

```
## [1]  39 206
```

The plot above suggests there are no major deviations from normality. It would be safe to assume the residuals are at least approximately normally distributed.

```
mplot(bfp_abdomen, 1)
```

```
## [[1]]
```

```
## `geom_smooth()` using method = 'loess'
```



A linear regression model was fitted to predict the dependent variable, Body Fat Percentage %, using measures of Abdomen circumference (cm) as a single predictor. Prior to fitting the regression, a scatterplot assessing the bivariate relationship between Body Fat Percentage and Abdomen circumference was inspected. The scatterplot demonstrated evidence of a positive linear relationship. Other non-linear trends were ruled out. The overall regression model was statistically significant, F(1, 249) = 474.6, p < .001, and explained 65.6% of the variability in Body Fat Percentage, R2 = .656 The estimated regression equation was Body Fat Percentage = -34.7694865 + 0.581*Abdomen circumference. The positive slope for Abdomen circumference was statistically significant, b = 0.581, t(250) = -13.99049, p < .001, 95% CI [0.5280287, 0.6329961]. Final inspection of the residuals supported normality and homoscedasticity with a few outliers.