Predicting Body Fat Percentage

CODE

Phil Steinke s3725547

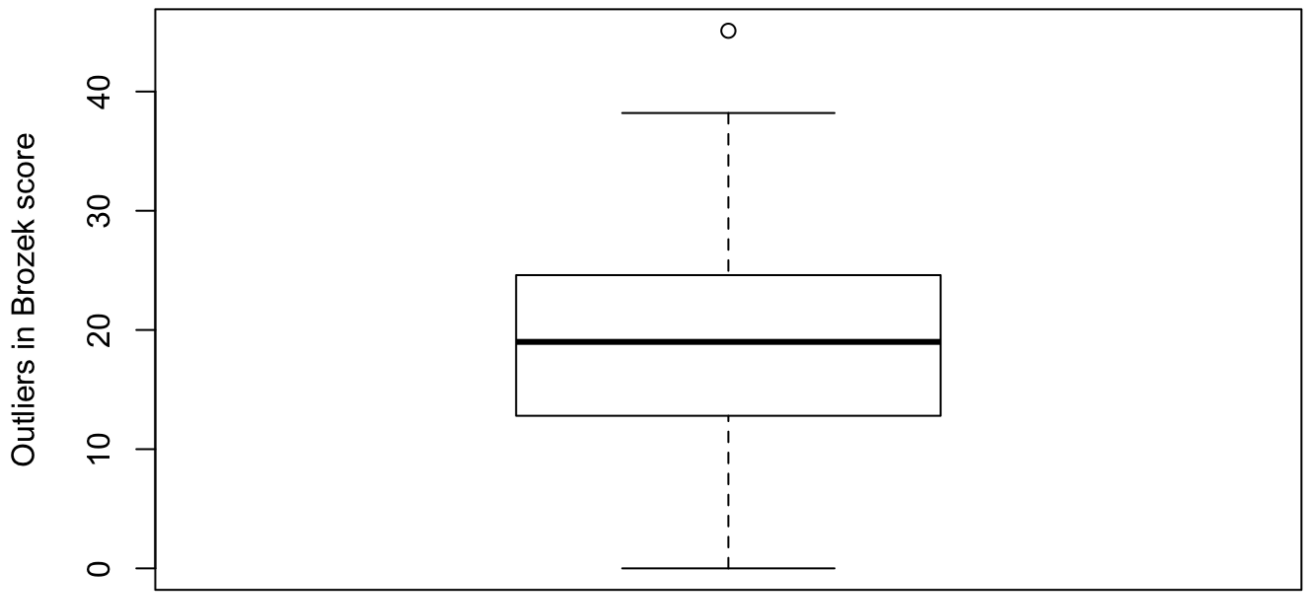
MATH1324 Introduction to Statistics Assignment 3 #### Executive Summary - This report investigates if there is a general, easy to determine, body circumference measurement that could be used as a general indicator for body fat percentage. - We have the goals to: 1. Establish a formula that can convert a body circumference measurement to a predicted body fat percentage and 2. Understand how well this prediction will hold

Data

- **Data**: A sample of 252 men and women from was obtained from JSE-DA (http://www.amstat.org/publications/jse/jse_data_archive.htm)
- **Factors**: *percentage* of body fat measured using the **Brozek** method - underwater weighing technique of density. Ten other body circumference measures (e.g. abdomen) are included as factors.
- The data which was collected was then visualised to find any potential outliers and form hypothesis to test.
- This report investigates if there is a general, easy to determine, body circumference measurement that could be used as a general indicator for body fat percentage.

1. Test whether the mean body fat percentage for males and females are the same (two-sample t-test)

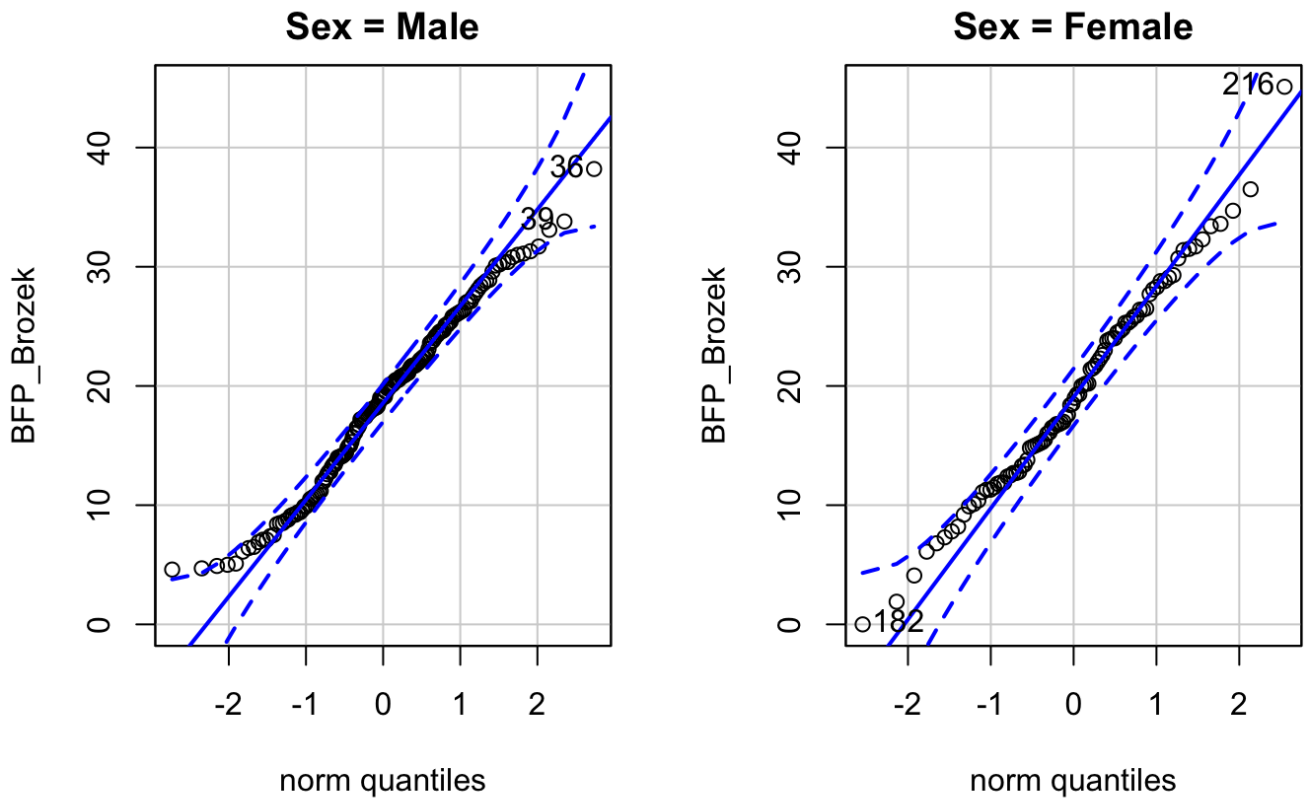
```
boxplot(body$BFP_Brozek, ylab = "Outliers in Brozek score")
```



There's an outlier at `> 40` and a minimum score in `BFP_Brozek` of `0`

Test the assumption of normality

```
qqPlot(BFP_Brozek ~ Sex, data = body, dist = "norm")
```



There is a slight s shape to the distribution, which may warrant further investigation

Homogeneity Levene's test

- H_0: $\sigma^2_1 = \sigma^2_2$
- H_a: $\sigma^2_1 \neq \sigma^2_2$

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

```
Levene's Test for Homogeneity of Variance (center = median)
       Df F value Pr(>F)
group   1  2.1974 0.1395
      250
```

- H_0, p > .05
- H_a, p < .05

- p = 0.1395
- p > .05

The p-value for the Levene's test of equal variance for body weight between males and females was greater than `.05`
We fail to reject the null hypothesis

Assuming Equal Variance

- H_0: $\mu_{male} - \mu_{female} = 0.0$

- H_a: $\mu_{male} - \mu_{female} \neq 0.0$

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

```
    Two Sample t-test

data:  BFP_Brozek by Sex
t = -0.75154, df = 250, p-value = 0.453
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.761898  1.236246
sample estimates:
  mean in group Male mean in group Female
            18.66000             19.42283
```

The mean of male `(18.66)` is not the same as female `(19.42)`, so fail to reject the null hypothesis

2. Confidence Interval

Estimate the `99% confidence interval` for the mean body fat percentage in the population.

- H_0: $\alpha = 0$
- H_a: $\alpha \neq 0$

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

mean of x	lower	upper	level
<dbl>	<dbl>	<dbl>	<dbl>
18.93849	17.97689	19.9001	0.95

1 row

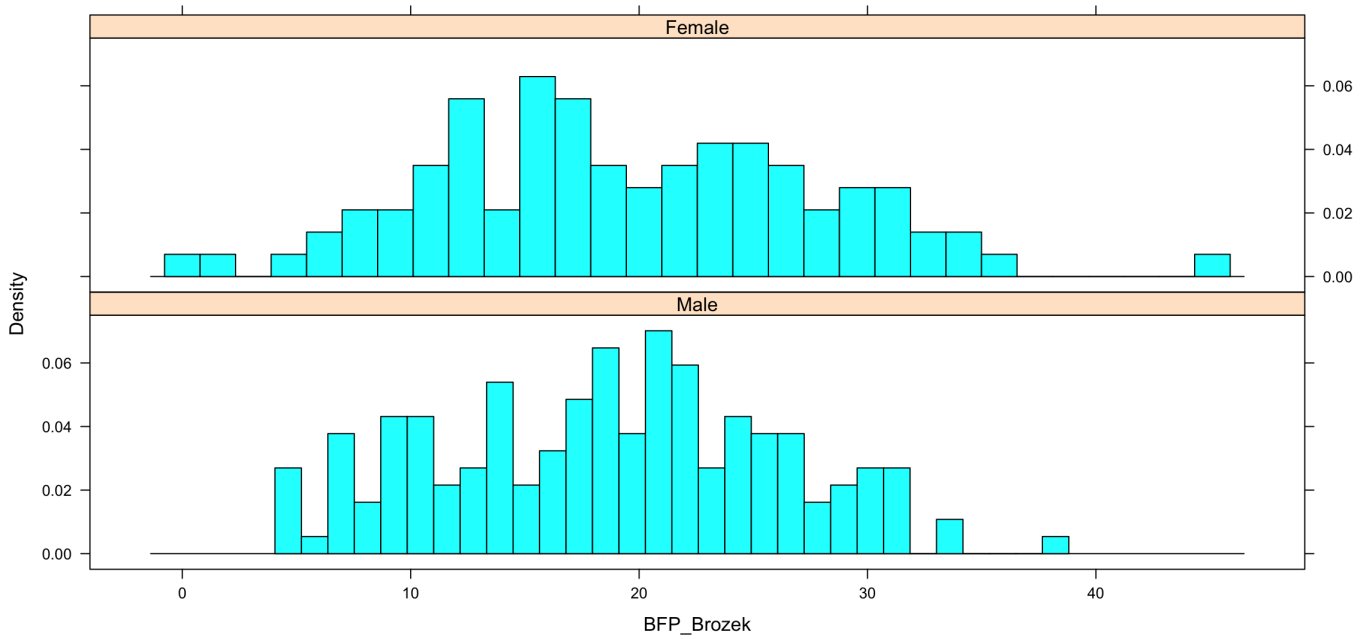
α is not captured `α < 18.93849 < 19.9001` So is not captured between the upper and lower confidence interval, therefore we reject the null hypothesis.

```
favstats(~ BFP_Brozek | Sex, data = body)
```

Sex	min	Q1	median	Q3	max	mean	sd	n	missing
<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<int>	<int>
Male	4.6	13.100	19.05	24.05	38.2	18.66000	7.348052	160	0
Female	0.0	12.775	18.75	25.35	45.1	19.42283	8.425499	92	0

2 rows

```
histogram(~ BFP_Brozek | Sex, data = body, bins=100, nint=30, layout = c(1, 2))
```




```
confint(t.test(~ BFP_Brozek, data = subset(body, subset = (Sex == "Male")), conf.level = 0.99))
```

mean of x	lower	upper	level
<dbl>	<dbl>	<dbl>	<dbl>
18.66	17.1455	20.1745	0.99

1 row

```
binom.approx(44, 109, conf.level = 0.99)
```

x	n	proportion	lower	upper	conf.level
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
44	109	0.4036697	0.282621	0.5247185	0.99

1 row

3. Researchers believe that average body fat percentage is `less than 12.5`. Test this claim.

One sample T Test - One tail

- H_0: $\mu = 12.5$
- H_a: $\mu < 12.5$

```
t.test(~ BFP_Brozek, data = body, mu = 12.5, alternative = "less")
```

```
    One Sample t-test

data:  BFP_Brozek
t = 13.187, df = 251, p-value = 1
alternative hypothesis: true mean is less than 12.5
95 percent confidence interval:
     -Inf 19.74458
sample estimates:
mean of x
 18.93849
```

- `P = 1`
- We do not reject the null hypothesis

We could do a further test by: - a two tailed test - the H_a being $\mu > 12.5$ - testing a different mu based on further research - If we test the upper tail: `t.test(~ BFP_Brozek, data=body, mu = 12.5, alternative = "greater")` - Other tests for normal standard distribution

We could be testing for problems with the data:

- Siri and Density are also included, making the initial dataset multivariate
- Includes a minimum of `0` in `BFP_Brozek`

```
favstats(~ BFP_Brozek, data = body)
```

min	Q1	median	Q3	max	mean	sd	n	missing
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<int>	<int>
0	12.8	19	24.6	45.1	18.93849	7.750856	252	0

1 row

```
filter(.data = body, BFP_Brozek == 0)
```

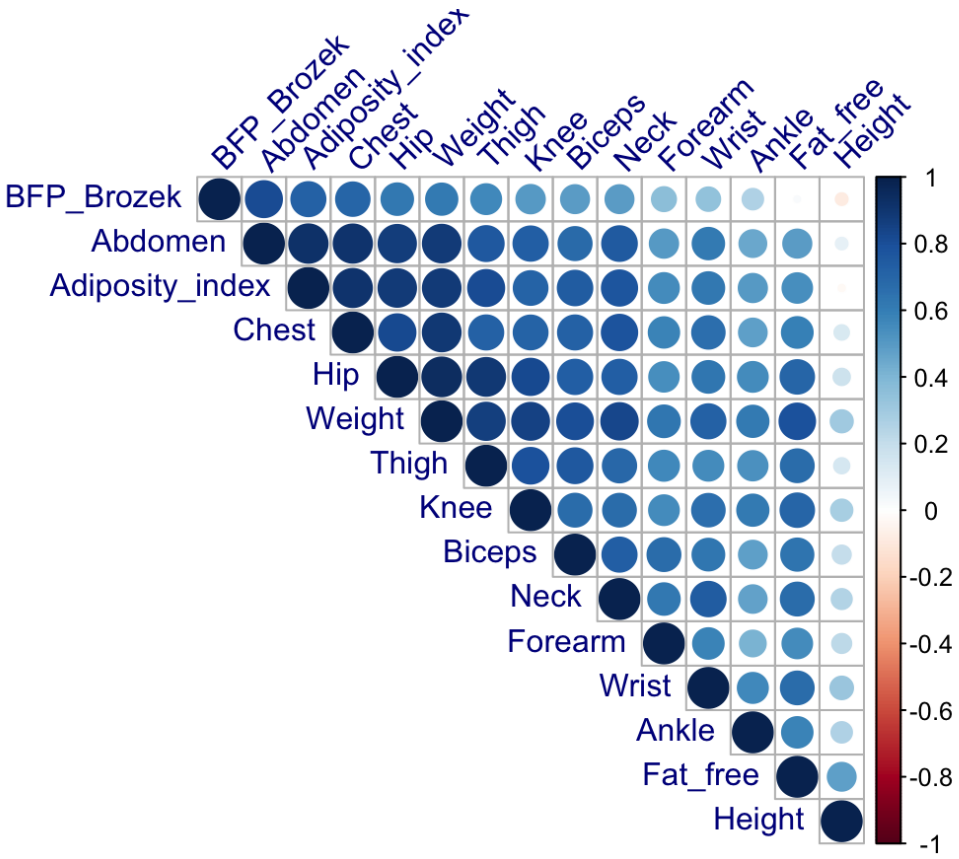
```
package 'bindrcpp' was built under R version 3.4.4
```

C...	BFP_Brozek	BFP_Siri	Density	...	Weight	Height	Adiposity_index	Fat_free	N...
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
182	0	0	1.1089	40	118.5	68	18.1	118.5	33.8

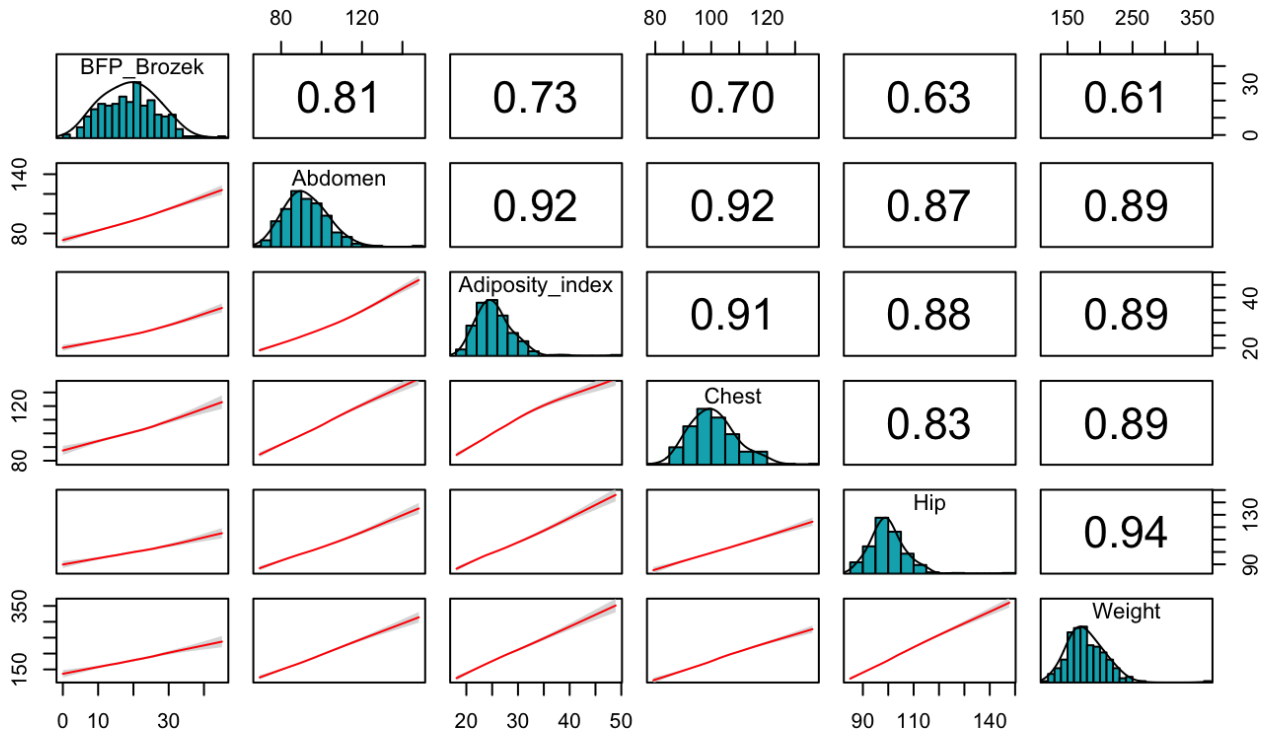
1 row | 1-10 of 20 columns

4. Regression modelling - Find the single best predictor of body fat percentage (Brozek method) using the body circumference data.

```
bodyCorrelationRcorrUnordered <- bodyCorrelation %>% as.matrix() %>% rcorr(type = "pearson") # Unordered bivariate
# Order our factors by Pearson correlation with with BFP_Brozek
bodyCorrelationFactorOrdered <- bodyCorrelationRcorrUnordered$r %>% as.data.frame() %>% .[1] %>% order(decreasing = T)
bodyCorrelation <- bodyCorrelation[bodyCorrelationFactorOrdered]
bodyCorrelationOrderedByPearson <- bodyCorrelation %>% as.matrix() %>% rcorr(type = "pearson") # Ordered R Correlation
#bodyCorrelationOrderedByPearson %>% colnames() # <- Pearson correlation in order
bodyCorrelationOrderedByPearson$r %>% corrplot(method = "circle", type = "upper", tl.col = "darkblue", tl.srt = 45,  p.mat = bodyCorrelationOrderedByPearson$p, sig.level = 1, insig = "blank")
```




```
pairs.panels(bodyCorrelation[1:6], method = "pearson", hist.col = "#00AFBB", density = T, ellipses = F, cor = T, ci = T, digits = 2, rug = F, breaks = 20, stars = F, show.points = F)
```



Write a report that explains your method for identifying the single best predictor.

Abdomen was found to be the best predictor for body fat percentage via the Brozek method, with an `0.81*` correlation. It was found using the Pearson correlation using the `rcorr()` function. The top 5 correlating factors are shown above on the pairplot diagram

```
bodyAbdomenMaxModel <- lm(Abdomen ~ BFP_Brozek, data = body)
msummary(bodyAbdomenMaxModel)
```

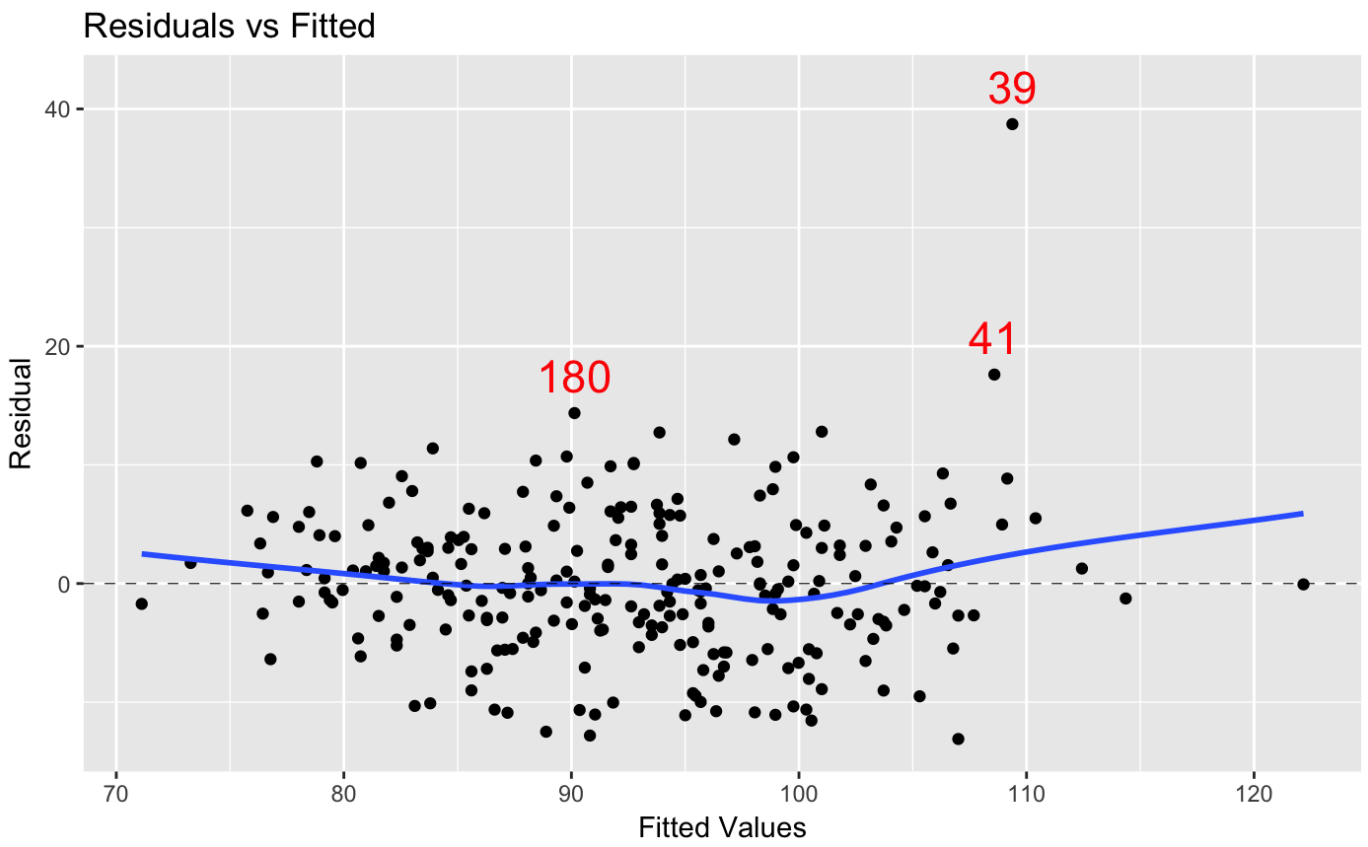
```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 71.11688    1.04630   67.97   <2e-16 ***
BFP_Brozek   1.13204    0.05115   22.13   <2e-16 ***

Residual standard error: 6.28 on 250 degrees of freedom
Multiple R-squared:  0.6621,    Adjusted R-squared:  0.6608
F-statistic: 489.9 on 1 and 250 DF,  p-value: < 2.2e-16
```

- H_0: α = 0
- $H_{A/sub>}$: α ≠ 0
- P- value to test if α =0.0 is <2e-16*** We reject the null hypothesis

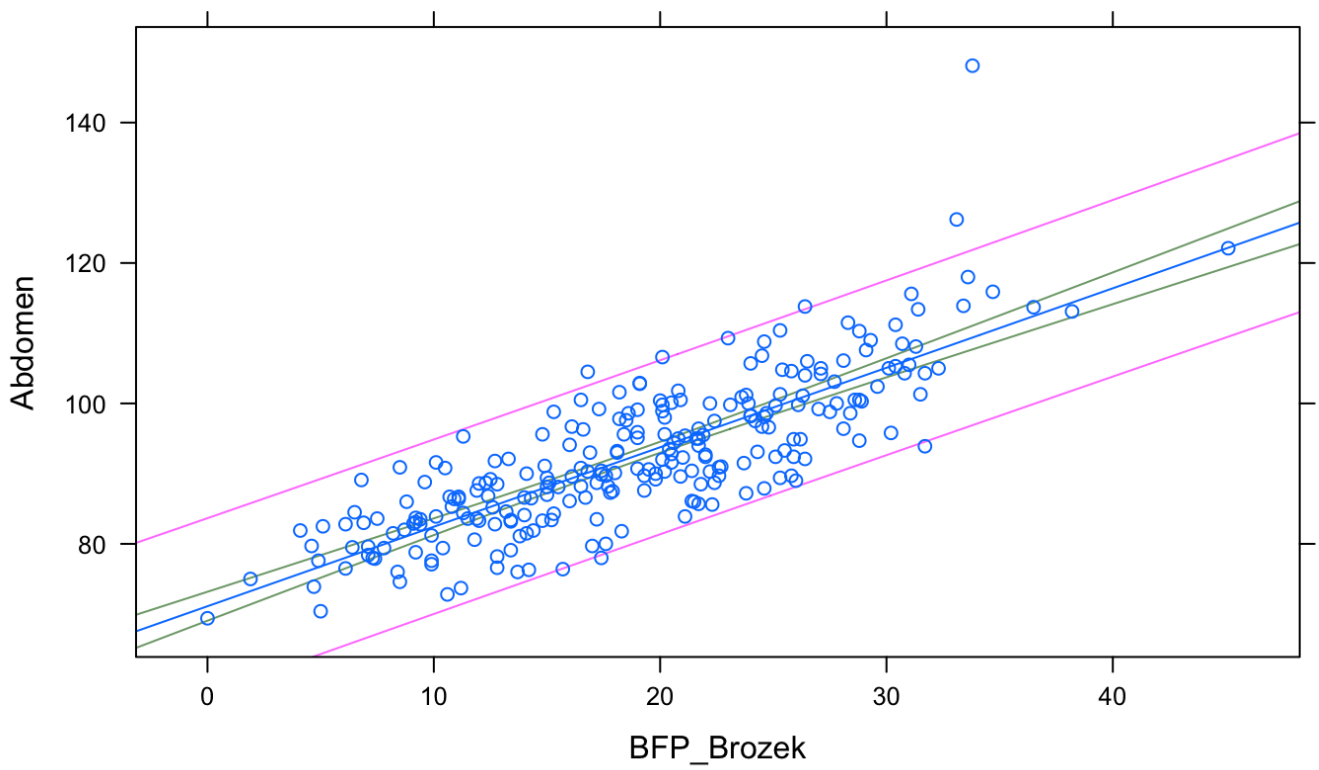
- P-value to test if β =0.0 <2e-16 ***
- H_0: β = 0
- $H_{A/sub>}$: β ≠ 0 We reject the null hypothesis

Testing Assumption of Homoscedasticity

```
mplot(bodyAbdomenMaxModel, 1)
```




```
xyplot(Abdomen ~ BFP_Brozek, data = body, ylab = "Abdomen", xlab = "BFP_Brozek", panel=panel.lmbands)
```



```
pf(489.9,1,250,lower.tail = FALSE ) # 7.71020869865819e-61
```

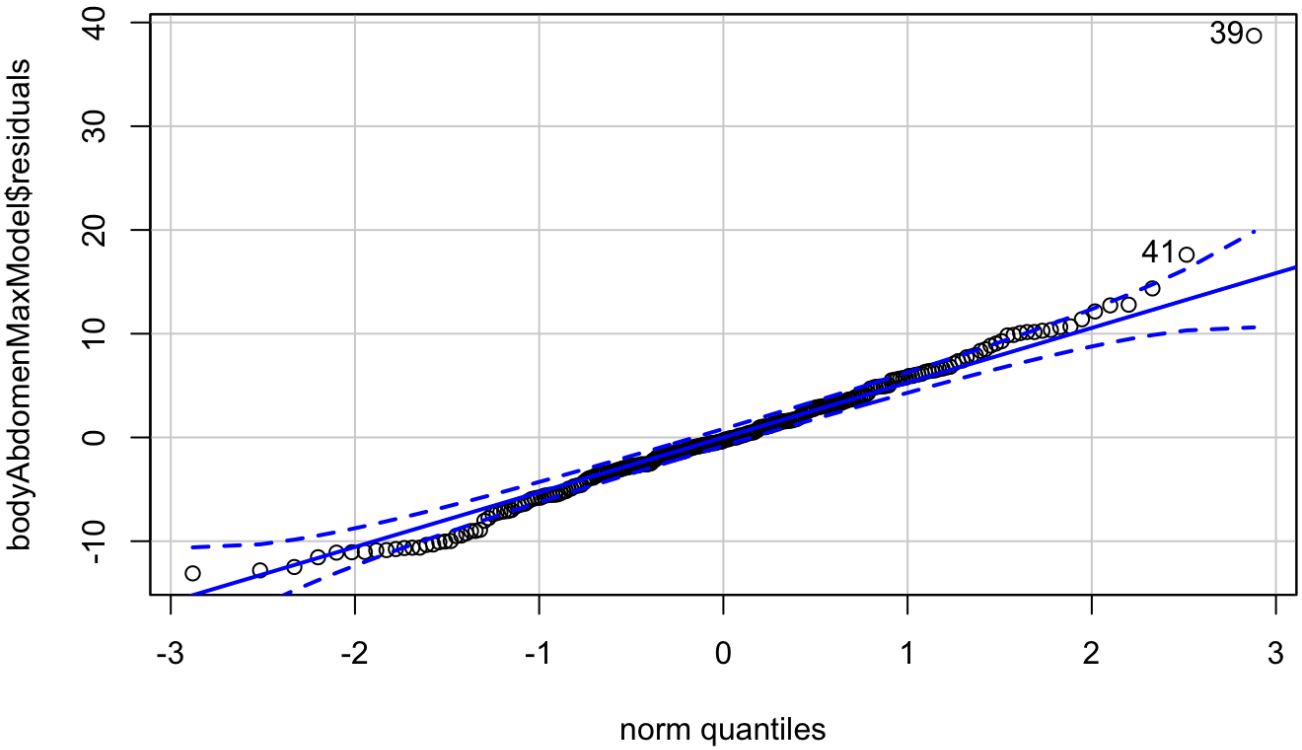
```
[1] 7.710209e-61
```

p (7.71020869865819e-61) is less than the Confidence Level Therefore we can reject the null hypothesis

Testing the Residuals

```
qqPlot(bodyAbdomenMaxModel$residuals, dist="norm")
```

```
[1] 39 41
```



Confidence Interval

```
confint(bodyAbdomenMaxModel, level = .99)
```

```
                0.5 %    99.5 %
(Intercept) 68.4010649 73.832686
BFP_Brozek   0.9992825  1.264792
```

Test Intercept (α)

- H_0: α = 0
- H_a: α ≠ 0

The 99% Confidence Interval (CI) for α is [68.4010649, 73.832686] H_0: α = 0 is not captured between this interval, so we reject H_0.

Testing the Slope (β)

- H_0: β = 0
- H_A: β ≠ 0

The 99% Confidence Interval (CI) for β is [0.9992825, 1.264792] H_0: β = 0 is not captured between this interval, so we reject H_0.

```
coef(summary(bodyAbdomenMaxModel))
```

```
             Estimate Std. Error  t value      Pr(>|t|)
(Intercept) 71.116875  1.0462981 67.97000 3.282118e-163
BFP_Brozek   1.132037  0.0511453 22.13375  7.706457e-61
```

Critique the predictive ability of the model and

The initial outlier at `> 40` and a minimum score in `BFP_Brozek` of `0` may have skewed the data

Draw an overall conclusion to help the investigators

- Prior to fitting the regression, a scatterplot assessing the bivariate relationship between `BFP_Brozek` and `abdomen` was inspected.
- The scatterplot demonstrated evidence of a `positive` linear relationship.
- The overall regression model was **statistically significant**, `F(1, 250) = 489.9, p < .001`
- The results show that `Abdomen` explains 66.21% of the variability in BFP_Brozek, R2 = 0.6621. The estimated regression equation was Abdomen = 1.132037 * BFP_Brozek
- The positive slope for `abdomen` was statistically significant,
  `b =1.132 , t(250) = 22.13, p < .001, 99% CI [68.4010649 ,73.832686].`
- Final inspection of the residuals supported normality and homoscedasticity.

```
save.image()
rm(list = ls())
```