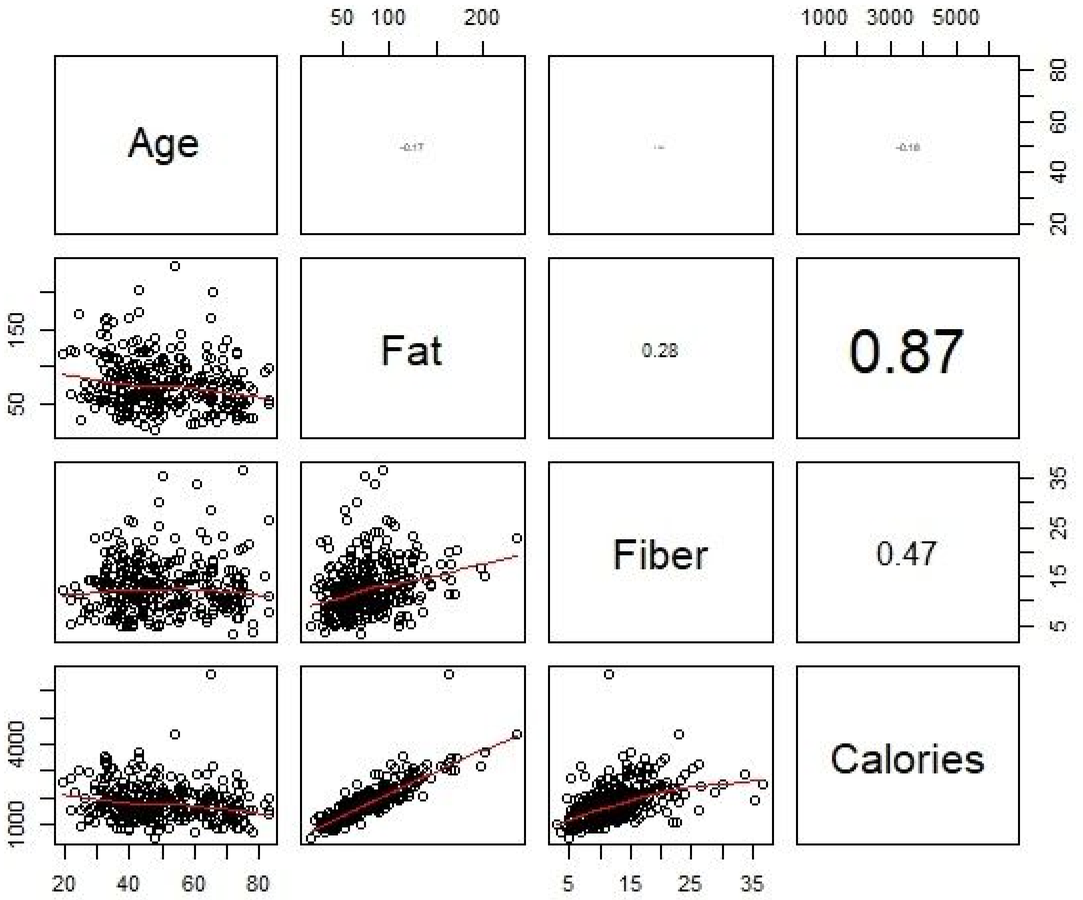
STAT210 Assignment 3

Tully McDonald
220196038

Q1.A) *Produce and interpret a pairs plot. What does the plot suggest as an appropriate model? Explain your response.*

Figure 1: Nutrition pairs plot (Age in years, Fat intake in grams, Fiber intake in grams).



In Figure 1, we have our three independent variables on top and our single dependent variable along the bottom row (Calories). Immediately the correlation coefficient between Fat and Calories appears to be most significant with a coefficient of 0.87, this means fat intake could be a very big indicator of calorie consumption which is not surprising. We also see a less (but still) significant correlation coefficient between Fiber and Calories of 0.47. There appears to be a minute multi-correlation between Fat and Fiber which may further reduce the usefulness of Fiber as a predictor. Finally, we look at Age which seems to be the least important predictor out of all our predictors, there is not very much that Age is contributing to this study.

Tully McDonald
220196038

Q1.B) *Fit a main effects model. Refer to the table on p. 349 of the text "Detecting Multicollinearity in the Regression Model." Produce relevant output that will allow you to check the four indicators of multicollinearity. Summarise your findings.*

Table 1: Correlation and P-value significance (cor.prob).

```
              Age        Calories    Fat        Fiber
Age           1.0000     0.00163     0.00255    4.28e-01
Calories     -0.1768     1.00000     0.00000    0.00e+00
Fat          -0.1695     0.87184     1.00000    6.19e-07
Fiber         0.0449     0.46548     0.27648    1.00e+00
```

Table 1 helps examine the first indicator of multicollinearity. As we can see, there are no significant correlations between independent variables.

Table 2: Summary table of lm.

```
Call:
lm(formula = Calories ~ Fat + Fiber + Age, data = nutrition_no_id)

Residuals:
   Min     1Q Median    3Q    Max
  -648   -138    -21     95   3556

Coefficients:
              Estimate    Std. Error    t value    Pr(>|t|)
(Intercept)   285.810     79.792        3.58       0.0004
Fat           15.971      0.516         30.92      <2e-16
Fiber         31.693      3.234         9.80       <2e-16
Age           -2.489      1.153        -2.16       0.0317

Residual standard error: 292 on 311 degrees of freedom
Multiple R-squared:  0.817,    Adjusted R-squared:  0.816
F-statistic:  464 on 3 and 311 DF,  p-value: <2e-16
```

Looking at the p-values in the Coefficients of Table 2, we see that none of them are breaching the 0.05 cut-off point. Also, our p-value at the bottom of the table ($2 * 10^{-16}$) confirms that our model is useful in general.

Now looking at both Tables 1 & 2 we check for opposite signs. Observe between the tables, we see no opposite signs as highlighted with colour-coding. Which indicates no multicollinearity. Finally, we will check the VIF value which returns simply:

```
  Fat    Fiber  Age
 1.12   1.09   1.04
```

The general cut-off for indications of multicollinearity in VIF is 10, and as we can see, all our values are much less than 10. Which is another indication of no multicollinearity.

To summarise these tests, none of them show any significant indications of multicollinearity. So we can safely move forward without worrying about it being an issue.

Q1.C) *Determine an adequate model and justify your choice of model terms with reference to forward stepwise model selection. The "upper" model should include all possible interactions. Include all relevant output and the summary table for the final model.*

Table 3: Stepwise table outputs with upper of formula(~ Age+Fat+Fiber).

```
Start:  AIC=4110
Calories ~ 1

        Df Sum of Sq      RSS      AIC
+ Fat    1  1.10e+08 3.49e+07   3663
+ Fiber  1  3.15e+07 1.14e+08   4035
+ Age    1  4.54e+06 1.41e+08   4102
<none>               1.45e+08   4110

Step:  AIC=3663
Calories ~ Fat

        Df Sum of Sq      RSS      AIC
+ Fiber  1   7926709 26939732   3583
<none>               34866442   3663
+ Age    1    125933 34740508   3663

Step:  AIC=3583
Calories ~ Fat + Fiber

       Df Sum of Sq      RSS     AIC
+ Age   1    397473 26542259  3581
<none>              26939732  3583

Step:  AIC=3581
Calories ~ Fat + Fiber + Age
```

Table 3 output has most of the work done for us. The final model will be Calories ~ Fat + Fiber + Age, because our Stepwise function has calculated the differences between each possible model and compared them to determine which one has the lowest AIC. As you can see in the output, our model has the lowest score when all our predictors are included.

Table 4: Summary of nutrition final model.

```
Call:
lm(formula = Calories ~ Fat + Fiber + Age, data = nutrition_no_id)

Residuals:
   Min     1Q Median     3Q    Max
  -648   -138    -21     95   3556

Coefficients:
              Estimate   Std. Error  t value    Pr(>|t|)
(Intercept)   285.810    79.792      3.58       0.0004
Fat           15.971     0.516       30.92      <2e-16
Fiber         31.693     3.234       9.80       <2e-16
Age           -2.489     1.153       -2.16      0.0317

Residual standard error: 292 on 311 degrees of freedom
Multiple R-squared:  0.817,    Adjusted R-squared:  0.816
F-statistic:  464 on 3 and 311 DF,  p-value: <2e-16
```

Table 5: Confidence intervals for nutrition.

```
              Estimate   Std. Error  2.5 %     97.5 %
(Intercept)     285.81     79.792    128.81    442.81
Fat             15.97      0.516     14.96     16.99
Fiber            31.69     3.234      25.33     38.06
Age              -2.49     1.153     -4.76     -0.22
```

Q1.D) *From the final model in (c), what is the estimated coefficient for Fat in the model when a person eats 25 grams of fiber in a day? Interpret this value.*

Depending on their calorie intake, it may vary anywhere from E(Fat) = 5.68 + 0.047*1338 -1.05*25 = 42g to E(Fat) = 78g (based on the median 50th percentile calorie intake), with an exact estimate of 64g.

Q1.E) *Write a concise (one paragraph), informative conclusion based on your analysis and results.*

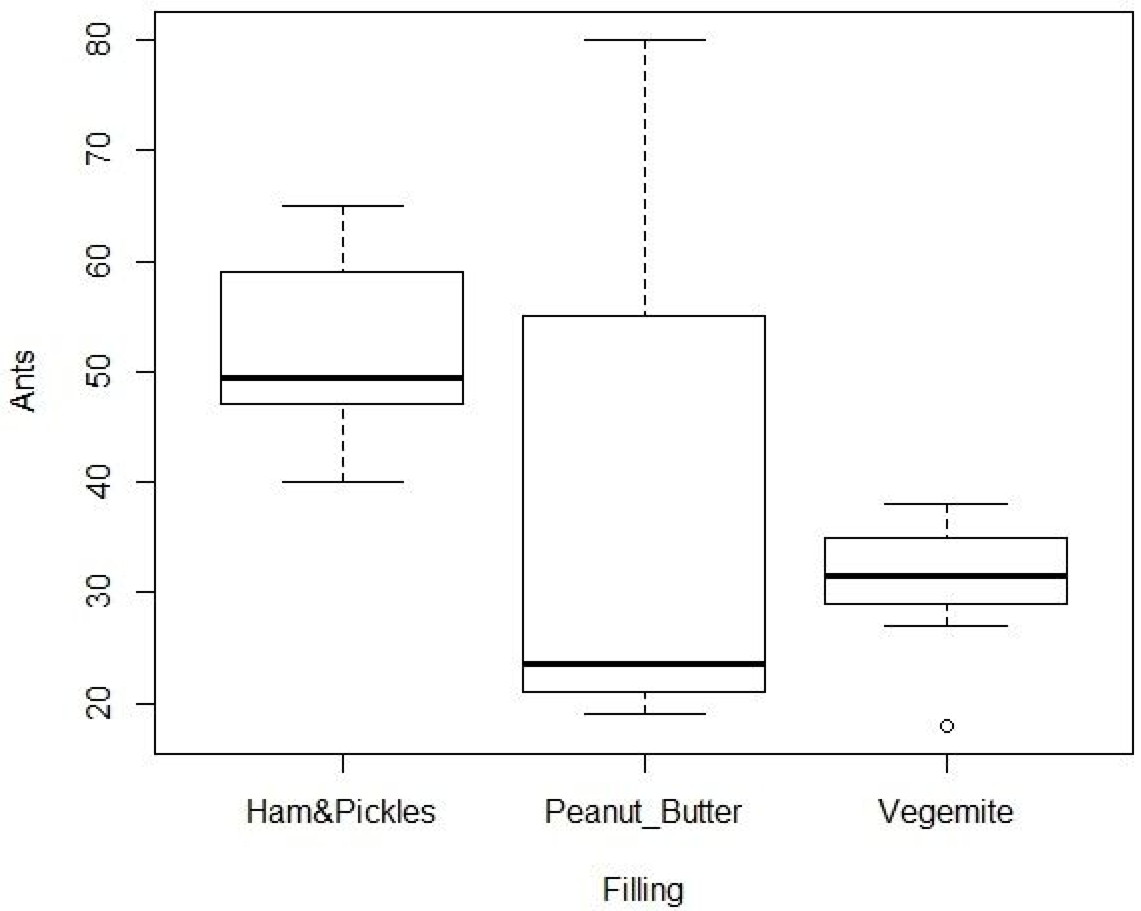
With an F-statistic of 464 on 3 and 311 DF as well as an adjusted R-squared of 0.816 (meaning 81.6% of our data is explained by our model), we conclude that calories are influenced by each of our three predictors but most significantly by Fiber at 25.3-38 calories per gram of fiber with 95% confidence. Fat appears to have about half as much influence with 15-17 calories per gram of fat at 95% confidence. Age is much less influential with a slightly negative effect of -4.8 to -0.2 calories for every year old an individual is (95% confidence).

Q2.A) *Exploratory analysis: obtain numerical (mean & variance) and graphical summary for the number of ants against fillings. From there, what assumption(s) for analysis of variance appear(s) violated?*

Table 1: Numerical summary.

```
> tapply(sandwich.df$Ants, sandwich.df$Filling, var)
 Ham&Pickles Peanut_Butter      Vegemite
        58.5          480.0           31.6
> tapply(sandwich.df$Ants, sandwich.df$Filling, mean)
 Ham&Pickles Peanut_Butter      Vegemite
        51.5           37.2           31.0
```

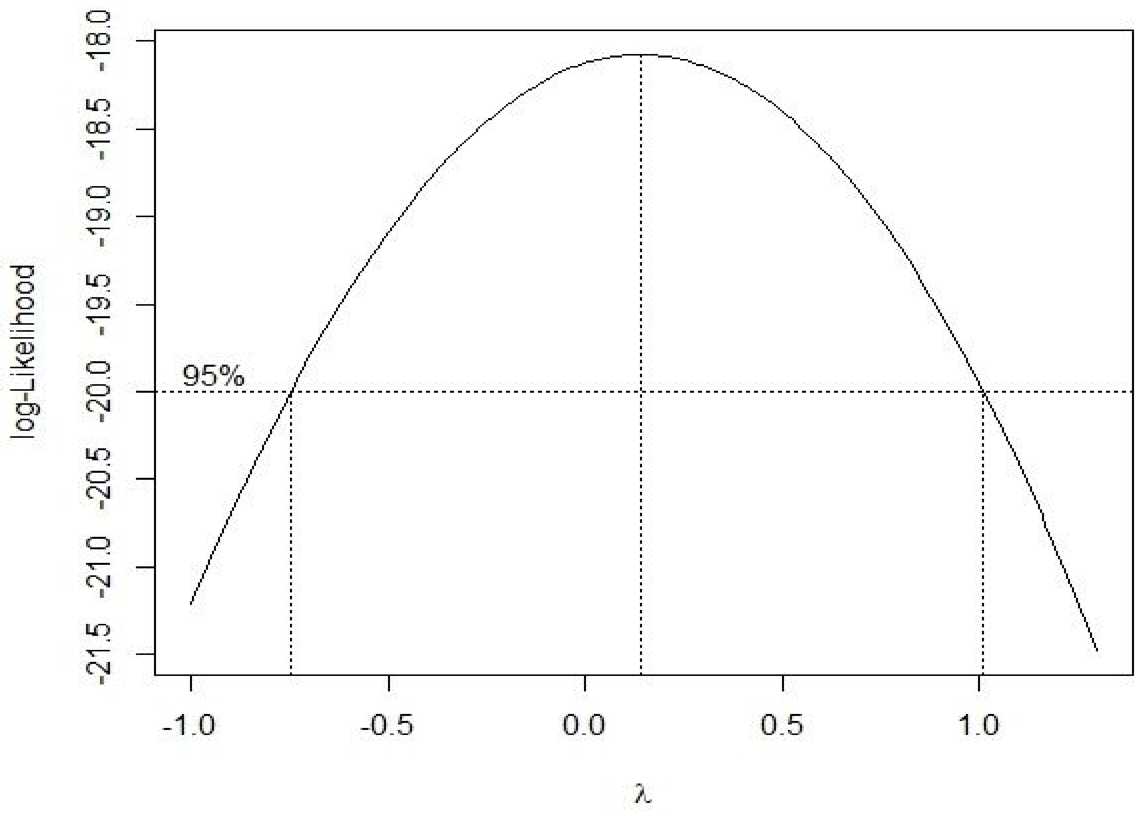
Figure 1: Boxplot graphic.



From Table 1 we see a clear difference in variance between Peanut_Butter and the other two types of sandwich filling. This is also very apparent when looking at the boxplot in Figure 1. We can see a massive range of variance for the peanut butter and substantially less for the other two categories, particularly Vegemite. There is also a clear positive skewness for Ham&Pickles and more so for Peanut_Butter. Skewness and difference in variance are the two most violated assumptions here.

Q2.B) *Transformation: Use the boxcox function in R to find suitable transformation(s). Explain why the transformation(s) you have chosen is(are) deemed appropriate but the other standard power transformations that were discussed in lectures are not.*

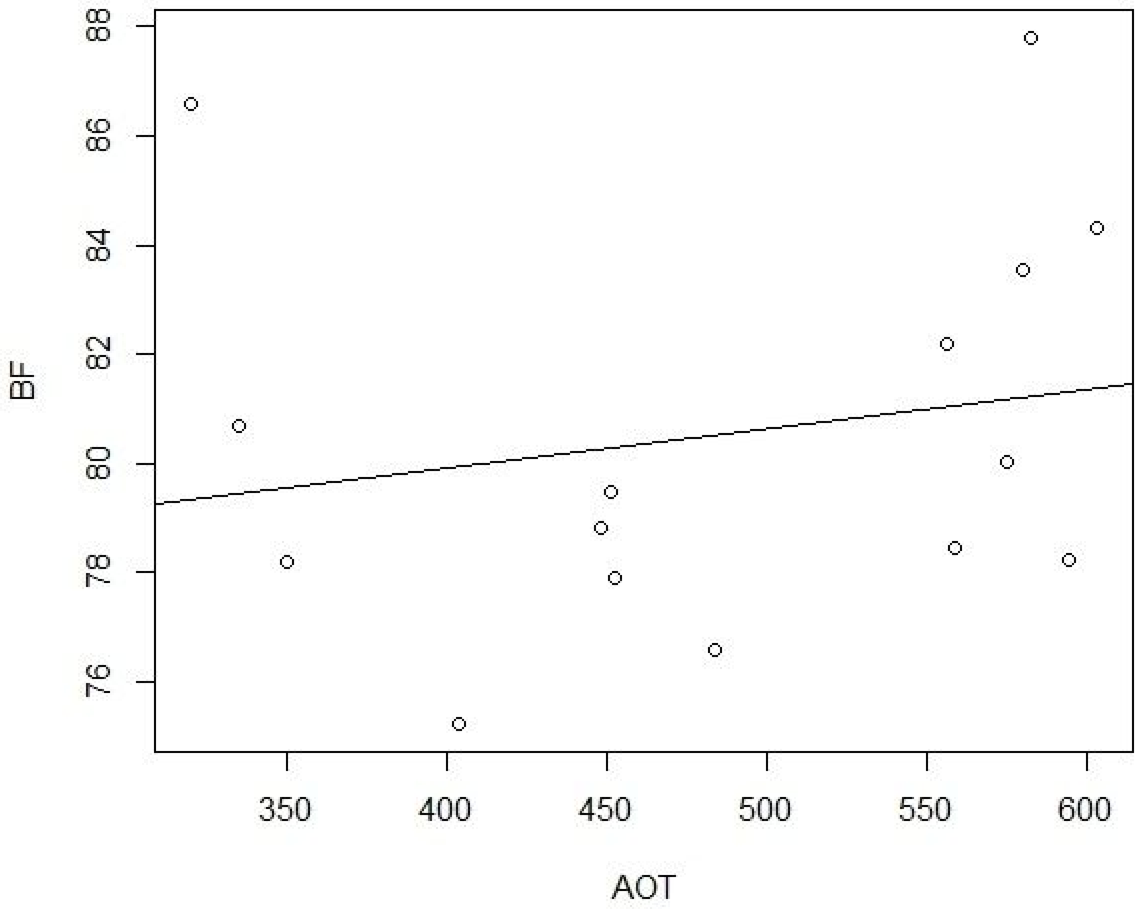
Figure 1: Boxcox output graphic.



As we can see here, our 95% line allows for a range of Lambda choices, the ideal Lambda (λ) appears to lie somewhere near 0.14. But since our range of choices is so large, we can choose either the Log transformation ($\lambda = 0$) or the Square root transformation ($\lambda = 0.5$). However, we prefer the transformation with the highest log-Likelihood which in this case is the Log transformation.

Q3.A) *Plot BF against AOT. What does the plot suggest?*

Figure 1: Scatter plot of BF against AOT.



The points are unevenly scattered across the plot and show a curvilinear shape. With the limited data available to us and the non-linear scattering of the data, we cannot use a linear model. We will use a polynomial regression model.

Q3.B) *Determine the order of the polynomial model required to fit the data. With reference to relevant output, justify each step in the process.*

We will start by modeling polynomials of order one, two, and three, to then analyse their differences using a Partial F-test and see which has the best fit for our dataset.

Tully McDonald
220196038

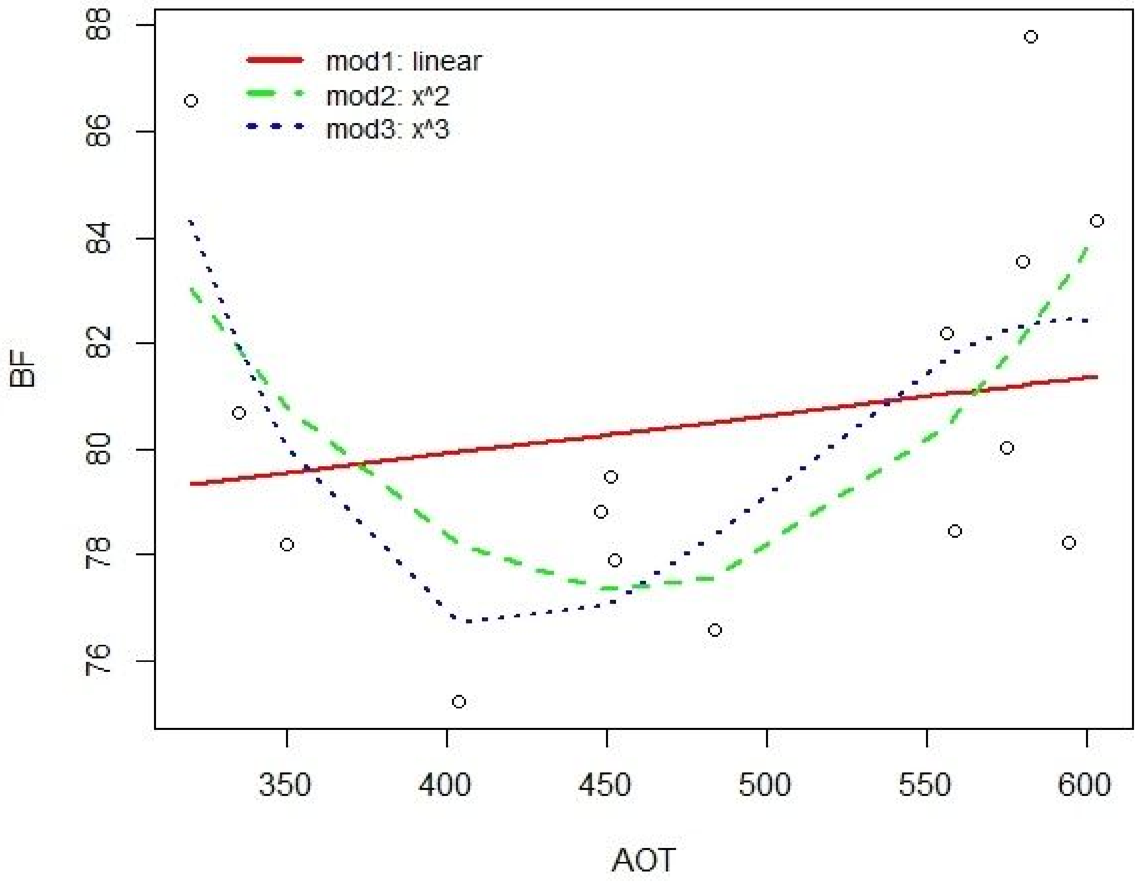
Table 1: Partial F-test using ANOVA of mod1 (linear) against mod2 (order 2).

```
Model 1: BF ~ AOT
Model 2: BF ~ AOT + I(AOT^2)
   Res.Df    RSS    Df    Sum of Sq      F     Pr(>F)
1    13        179
2    12        106     1        72.4        8.16    0.014
```

Table 2: Partial F-test using ANOVA of mod2 (order 2) against mod3 (order 3).

```
Model 1: BF ~ AOT + I(AOT^2)
Model 2: BF ~ AOT + I(AOT^2) + I(AOT^3)
   Res.Df    RSS    Df    Sum of Sq      F     Pr(>F)
1    12       106.5
2    11        94.5     1        12            1.4    0.26
```

Figure 2: Scatter plot with the different models plotted as lines.



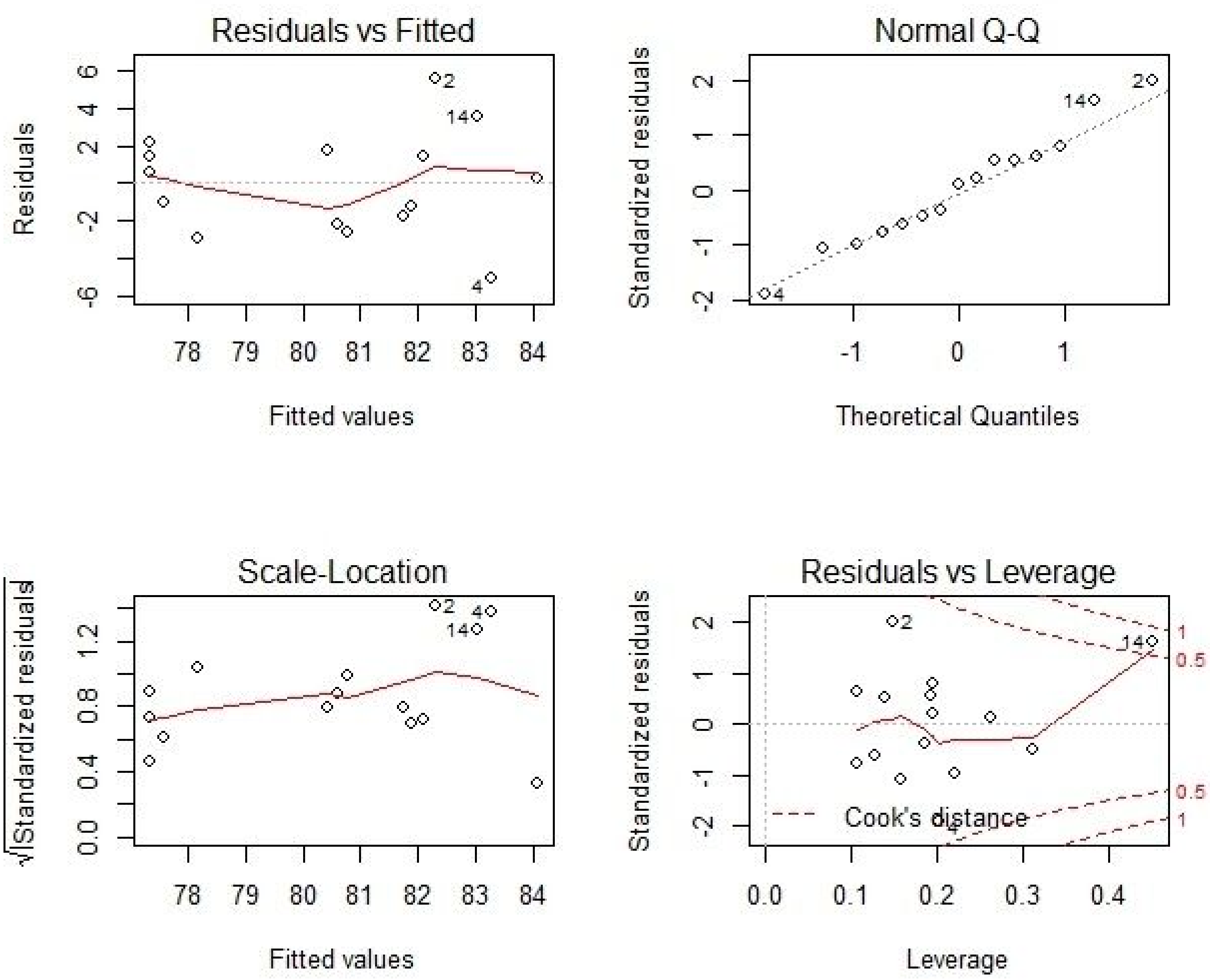
The Partial F-test provides a quick way to check the significance of the difference between our models.

- Null hypothesis: There is no significant difference between the two models.
- Alternative hypothesis: The second model is significantly better than the first.

Table 1 shows a significant p-value of less than 0.05 which means the order 2 polynomial is better than order 1. Now checking Table 2, we can see the order 3 polynomial is not significantly better than order 2. Figure 2 visualises the models, we can visually see that mod2 appears to be our best choice of polynomial order. So, we will use the order 2 polynomial model.

Q3.C) *For the final model in (b), check the model assumptions, and also identify any potential outliers or influential points.*

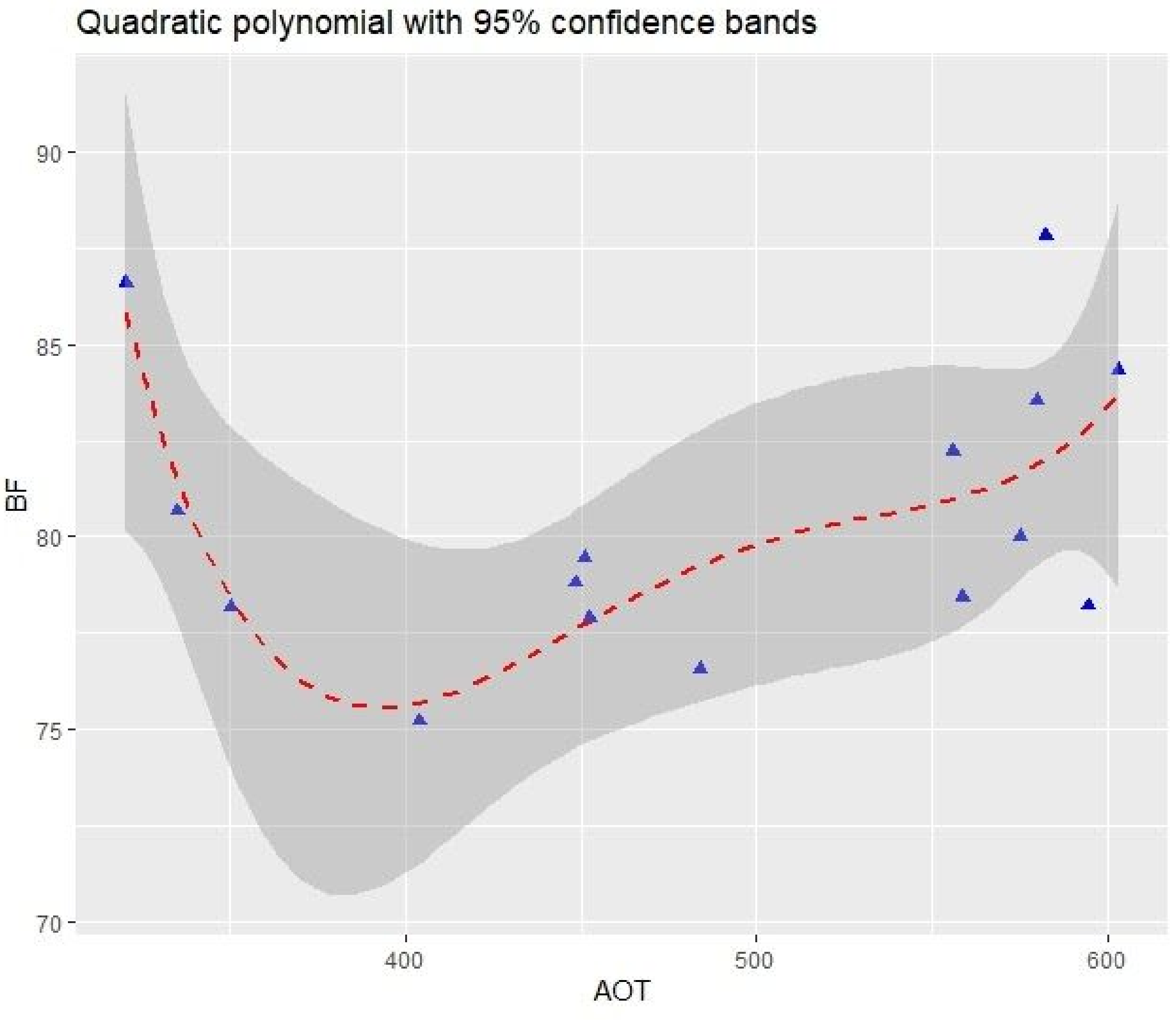
Figure 3: Model assumptions analysis plots.



To check the model assumptions we will use the Residuals vs Fitted, Normal Q-Q, Scale-Location, and Residuals vs Leverage plots. First, looking at the Residuals vs Fitted plot, we see acceptable random scattering of the points above and below the 0 line with one (maybe two) outliers. These outliers seem to have a slight effect on the model which you can more easily see when looking at the Scale-Location plot with a miniscule positive increase. Point 14 of these outliers has a particular influence on the model as seen in the Residuals vs Leverage plot. The Normal Q-Q plot features a moderately straight line with a few points (previously identified as outliers) showing potential curvature in the tails, this has potential to violate our model assumptions. However, with the limited data we have, it's difficult to decide.

Q3.D) *Plot the fitted values on a scatter plot of the data. Include a plot of the 95% confidence bands on your graph.*

Figure 4: 95% confidence bands.



Q3.E) *Predict the mean BF for people with AOT of 300mm of Hg. Show your calculations (do not use Rstudio).*

Table 3: Summary of mod2.

```
Coefficients:
              Estimate    Std. Error   t value    Pr(>|t|)
(Intercept)   1.42e+02    2.30e+01     6.17       4.8e-05
AOT          -2.82e-01    1.02e-01    -2.78       0.017
I(AOT^2)      3.09e-04    1.08e-04     2.86       0.014

Residual standard error: 2.98 on 12 degrees of freedom
Multiple R-squared:  0.428,    Adjusted R-squared:  0.333
F-statistic: 4.49 on 2 and 12 DF,  p-value: 0.0349
```

$$E(BF) \;=\; 142 \;-0.282(300) \;+\; 0.000309(300^2)$$
$$E(BF) \;=\; 85.21$$

Q3.F) *Taking into account the context of the problem, summarise your findings (one paragraph).*

Based on the data available to us we were able to find a suitable model, a quadratic polynomial regression model. Using our analysis tools we could confidently say our model is useful for predicting within the scope of our data points. However, our R-Squared of 0.428 and Adjusted R-Squared of 0.33 show a slightly low percentage of data explanation, this may be due to various things such as the minimal amount of data points to work with. Nonetheless, with our p-value of less than 0.05 we are happy with our conclusion that cerebral blood flow (BF) in humans can be predicted using arterial oxygen tension (AOT).

Appendix: R Studio script code

```
options(digits=3, show.signif.stars = F)

source("Rfunctions.R")

nutrition <- read.table("NutritionStudy.txt", header=TRUE)

pairs(nutrition[,c(2,4,5,3)], lower.panel=panel.smooth, upper.panel=panel.cor)

nutrition_no_id <- nutrition[c(2,3,4,5)]

# Multicollinearity tests

cor.prob(nutrition_no_id) # Correlation significance

multi_test2 <- lm(Calories~Fat+Fiber+Age, data=nutrition_no_id)

summary(multi_test2)
```

```
# VIF printout
library(car)
vif(multi_test2)
# Stepwise table
formL <- formula(~ 1)
formU <- formula(~ Age+ Fat + Fiber)
no.model <- lm(Calories ~ 1, data=nutrition_no_id)
fstep.model <- step(no.model, direction = "forward",
              scope=list(lower=formL, upper=formU))
summary(fstep.model)
betaCI(fstep.model)

# Question 2
sandwich.df <- read.table("sandwich.txt", header=T)
# Declare Filling as our factor with:
sandwich.df$Filling <- factor(sandwich.df$Filling)
# Print the numerical summary (i.e mean and variance)
tapply(sandwich.df$Ants, sandwich.df$Filling, mean)
tapply(sandwich.df$Ants, sandwich.df$Filling, var)
# Now visualise the data using boxplot.
plot(Ants~Filling, data=sandwich.df)
# Pt2
library(MASS)
boxcox(Ants~Filling, lambda=seq(from=-1, to=1.3, by=0.01), data=sandwich.df)

# Question 3
hbf.df <- read.table("bloodflow.txt", header=T)
attach(hbf.df)
# Scatter plot
plot(BF~AOT, data=hbf.df)
# Polynomial models
mod1 <- lm(BF~AOT)
```

```
mod2 <- lm(BF~AOT + I(AOT^2))
mod3 <- lm(BF~AOT + I(AOT^2) + I(AOT^3))
summary(mod1)
summary(mod2)
summary(mod3)
lines(smooth.spline(AOT, predict(mod1)), col="red", lwd=2, lty=1)
lines(smooth.spline(AOT, predict(mod2)), col="green", lwd=2, lty=2)
lines(smooth.spline(AOT, predict(mod3)), col="blue", lwd=2, lty=3)
legend(330, 88, legend = c("mod1: linear","mod2: x^2","mod3: x^3"), col=c("red","green","blue"),
lty=c(1,2,3), lwd=3, bty="n", cex=0.9)
# Check the differences between them.
anova(mod1, mod2)
anova(mod2, mod3)
anova(mod3)
# Alternative method. Keep adding terms until a non-significant result (this happens after adding mod3)
#mod.1 <- lm(BF~AOT)
#anova(mod.1)
#mod.2 <- update(mod.1, .~. + I(AOT^2))
#anova(mod.2)
#mod.3 <- update(mod.2, .~. + I(AOT^3))
#anova(mod.3)
#mod.4 <- update(mod.3, .~. + I(AOT^4))
#anova(mod.4)
# Residuals vs fitted and normal qq
par(mfrow=c(2,2))
plot(mod2, which=c(1,2,3,5))
# 95% confidence plot
par(mfrow=c(1,1))
library(ggplot2)
ggplot(data=mod2, aes(x=AOT, y=BF)) + geom_point(pch=17, col="blue", size=2)+
  geom_smooth(method = "lm", formula = y ~ poly(x, 4), col="red", linetype=2)+
  labs(title="Quadratic polynomial with 95% confidence bands")
```