stat359_A5_wducharme

Wesley Ducharme

2024-11-23

Question 1

```
beef <- read.table(file="C:/Users/wesch/uvic/stat359 Data Analysis/data1/beef.txt",
↪  header=TRUE)
```

Plots for visualizing possible postive relationships

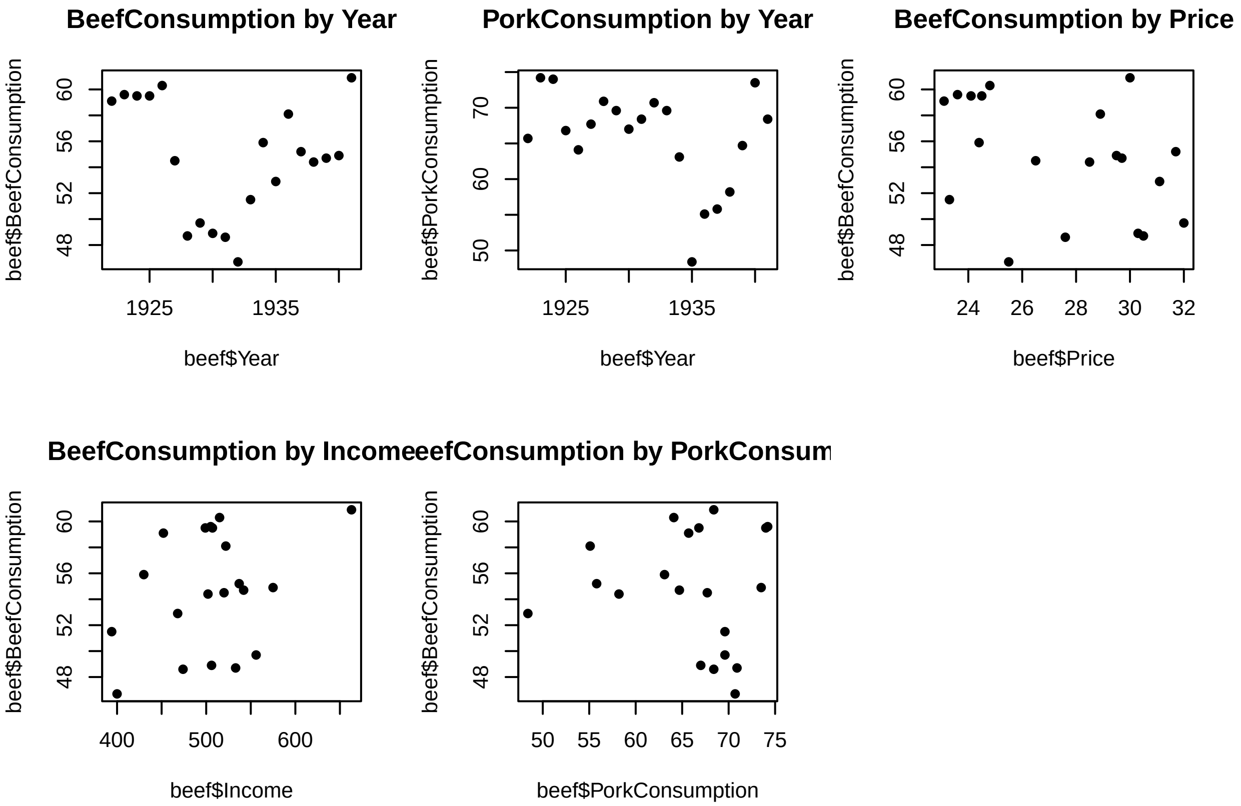
```
par(mfrow=c(2,3))
plot(beef$Year,beef$BeefConsumption,pch=16)
title("BeefConsumption by Year")

plot(beef$Year,beef$PorkConsumption,pch=16)
title("PorkConsumption by Year")

plot(beef$Price,beef$BeefConsumption,pch=16)
title("BeefConsumption by Price")

plot(beef$Income,beef$BeefConsumption,pch=16)
title("BeefConsumption by Income")

plot(beef$PorkConsumption,beef$BeefConsumption,pch=16)
title("BeefConsumption by PorkConsumption")
```



BeefConsumption by Year

PorkConsumption by Year

BeefConsumption by Price

BeefConsumption by IncomeeefConsumption by PorkConsum

The Scatterplots show a roughly positive relationship between BeefConsuption and Income all other varibles seem to have a roughly negative relationship.

Will now fit simple linear models for for BeefConsmption by each variable.

beef and price

```
beef_price_lm <- lm(beef$BeefConsumption ~ beef$Price)
summary(beef_price_lm)
```

```
##
## Call:
## lm(formula = beef$BeefConsumption ~ beef$Price)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.0749 -2.9155  0.7659  2.8780  7.6136
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  69.8765     8.7696   7.968 2.59e-07 ***
## beef$Price   -0.5530     0.3172  -1.743   0.0984 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.261 on 18 degrees of freedom
## Multiple R-squared:  0.1444, Adjusted R-squared:  0.0969
## F-statistic: 3.039 on 1 and 18 DF,  p-value: 0.09836
```

beef and income

```
beef_income_lm <- lm(beef$BeefConsumption ~ beef$Income)
summary(beef_income_lm)
```

```
##
## Call:
## lm(formula = beef$BeefConsumption ~ beef$Income)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.6961 -2.4996 -0.3197  3.5459  5.7756
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 41.76383    8.27790   5.045 8.42e-05 ***
## beef$Income  0.02558    0.01628   1.571    0.134
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.32 on 18 degrees of freedom
## Multiple R-squared:  0.1206, Adjusted R-squared:  0.07173
## F-statistic: 2.468 on 1 and 18 DF,  p-value: 0.1336
```

beef and pork

```
beef_pork_lm <- lm(beef$BeefConsumption ~ beef$PorkConsumption)
summary(beef_pork_lm)
```

```
##
## Call:
## lm(formula = beef$BeefConsumption ~ beef$PorkConsumption)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.7455 -3.4481  0.0049  4.5286  6.3445
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)           57.8253    10.2308   5.652 2.32e-05 ***
## beef$PorkConsumption  -0.0478     0.1547  -0.309    0.761
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.595 on 18 degrees of freedom
## Multiple R-squared:  0.005276,   Adjusted R-squared:  -0.04999
## F-statistic: 0.09548 on 1 and 18 DF,  p-value: 0.7609
```

beef and year

```
beef_year_lm <- lm(beef$BeefConsumption ~ beef$Year)
summary(beef_year_lm)
```

```
##
## Call:
## lm(formula = beef$BeefConsumption ~ beef$Year)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.9268 -3.5768  0.9611  3.9279  7.2300
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 260.0289   341.6341   0.761    0.456
## beef$Year    -0.1063     0.1769  -0.601    0.555
##
## Residual standard error: 4.561 on 18 degrees of freedom
## Multiple R-squared:  0.01968,    Adjusted R-squared:  -0.03479
## F-statistic: 0.3613 on 1 and 18 DF,  p-value: 0.5553
```

From the above simple models we can conclude that no single predictor varialbe significantly explains BeefConsumption.

Will now fit a full model involving all predictors.

```r
beef_fullmodel <- lm(beef$BeefConsumption ~ beef$Year + beef$Price + beef$Income +
↪  beef$PorkConsumption)

summary(beef_fullmodel)
```

```
##
## Call:
## lm(formula = beef$BeefConsumption ~ beef$Year + beef$Price +
##     beef$Income + beef$PorkConsumption)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.5008 -0.8400  0.1059  0.7248  2.4291
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)          119.511788 134.028204   0.892    0.387
## beef$Year             -0.014977   0.069890  -0.214    0.833
## beef$Price            -1.836727   0.162524 -11.301 9.77e-09 ***
## beef$Income            0.083311   0.007105  11.726 5.93e-09 ***
## beef$PorkConsumption  -0.418001   0.057270  -7.299 2.61e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.415 on 15 degrees of freedom
## Multiple R-squared:  0.9214, Adjusted R-squared:  0.9005
## F-statistic: 43.97 on 4 and 15 DF,  p-value: 4.105e-08
```

With the full model we have an R-squared value of 0.9214 so 92.14% of the variation in BeefConsumption is explained by the predictor variables in the model.

Correlation between BeefConsumption and year does not appear too be statistically significant so we will try a reduced model with the Year variable removed and compare it to the full model.

```
beef_reducedmodel <- lm (beef$BeefConsumption ~ beef$Price + beef$Income +
↪  beef$PorkConsumption)
summary(beef_reducedmodel)
```

```
##
## Call:
## lm(formula = beef$BeefConsumption ~ beef$Price + beef$Income +
##     beef$PorkConsumption)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.44996 -0.87212  0.04715  0.73206  2.47242
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          90.813646   5.266047  17.245 9.28e-12 ***
## beef$Price           -1.849850   0.145990 -12.671 9.32e-10 ***
## beef$Income           0.083190   0.006868  12.113 1.80e-09 ***
## beef$PorkConsumption -0.415085   0.053945  -7.695 9.15e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.372 on 16 degrees of freedom
## Multiple R-squared:  0.9212, Adjusted R-squared:  0.9064
## F-statistic: 62.33 on 3 and 16 DF,  p-value: 4.799e-09
```

```
anova(beef_reducedmodel, beef_fullmodel)
```

```
## Analysis of Variance Table
##
## Model 1: beef$BeefConsumption ~ beef$Price + beef$Income + beef$PorkConsumption
## Model 2: beef$BeefConsumption ~ beef$Year + beef$Price + beef$Income +
##     beef$PorkConsumption
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     16 30.111
## 2     15 30.019  1  0.091904 0.0459 0.8332
```

Comparing the full model including The Year and the reduced model without the year we can conclude that Year does not significantly improve our model so the reduced model will be the model we will proceed with.

Will now check for possible interaction that may improve the fit of our model.

Checking for a significant interaction between Price of beef and PorkConsumption

```
beef_reducedmodel_int1 <- lm (beef$BeefConsumption ~ beef$Price + beef$Income +
↪  beef$PorkConsumption + beef$Price*beef$PorkConsumption)
summary(beef_reducedmodel_int1)
```

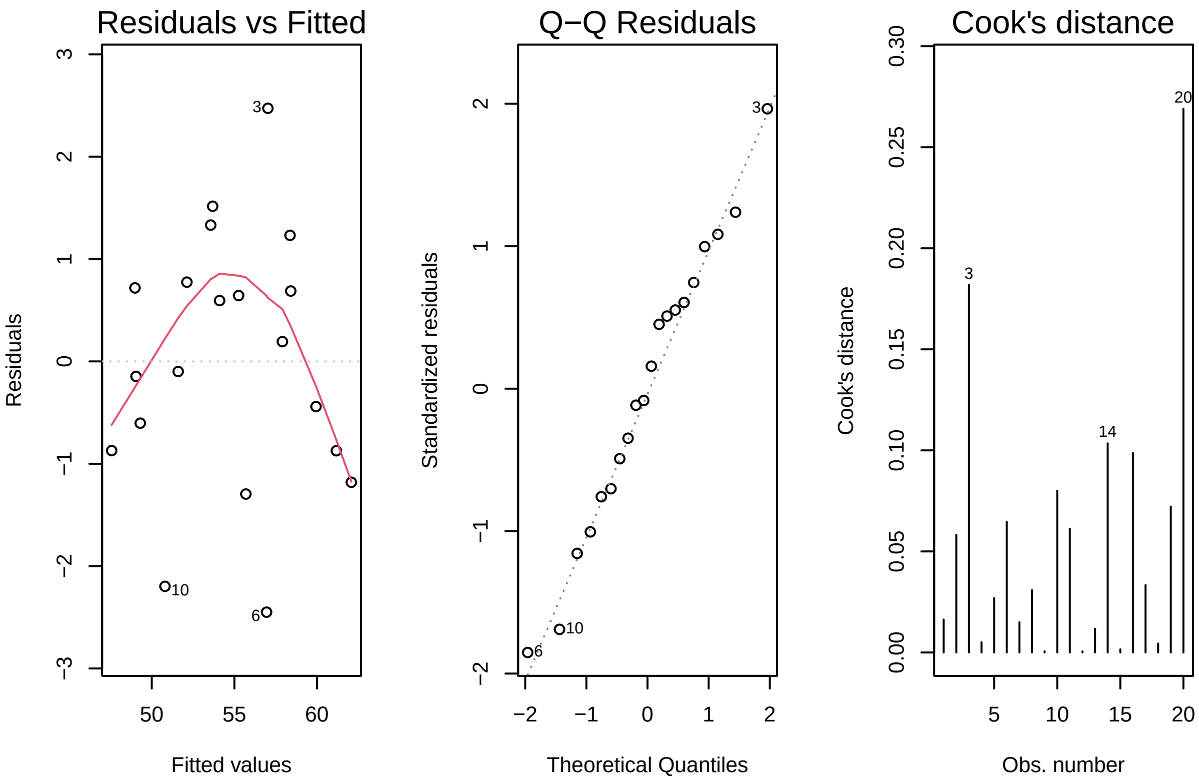
```
##
## Call:
## lm(formula = beef$BeefConsumption ~ beef$Price + beef$Income +
##     beef$PorkConsumption + beef$Price * beef$PorkConsumption)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.36261 -0.44530 -0.07958  0.81633  1.56448
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     23.213104  34.392252   0.675   0.5100
## beef$Price                       0.431027   1.156673   0.373   0.7146
## beef$Income                      0.084899   0.006371  13.327 1.02e-09 ***
## beef$PorkConsumption             0.582059   0.504697   1.153   0.2668
## beef$Price:beef$PorkConsumption -0.034183   0.017218  -1.985   0.0657 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.261 on 15 degrees of freedom
## Multiple R-squared:  0.9376, Adjusted R-squared:  0.9209
## F-statistic: 56.32 on 4 and 15 DF,  p-value: 7.412e-09
```

R-squared increases to 0.9376 but there is minimal evidence supporting this interaction with a p-value of 0.0657. Due to such low evidence we will proceed with the other reduced model.

Model diagnostics

```
par(mfrow=c(1,3))
plot(beef_reducedmodel,which=c(1,2,4))
```



The qq plot shows that the residuals follow an approximately normal distribution and a constant variance. Though it seem observation 20 may be an outltlier and have influence on the model. Will remove the observation and refit.

```
reduced_model_no20 <-
↪  update(beef_reducedmodel,.~.,subset(1:length(Beef.Consumption)!=20))
summary(reduced_model_no20)
```

```
##
## Call:
## lm(formula = beef$BeefConsumption ~ beef$Price + beef$Income +
##     beef$PorkConsumption)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.44996 -0.87212  0.04715  0.73206  2.47242
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          90.813646   5.266047  17.245 9.28e-12 ***
## beef$Price           -1.849850   0.145990 -12.671 9.32e-10 ***
## beef$Income           0.083190   0.006868  12.113 1.80e-09 ***
## beef$PorkConsumption -0.415085   0.053945  -7.695 9.15e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.372 on 16 degrees of freedom
## Multiple R-squared:  0.9212, Adjusted R-squared:  0.9064
## F-statistic: 62.33 on 3 and 16 DF,  p-value: 4.799e-09
```

In conclusion the 3 factors of Price, Income, Price, and PorkConsumption are related to BeefConsumption. These are additive effects and do not have any significant interactions between themselves.

Price is negatively related to beef consumption so as the price gets higher beef consumption can be expected to decrease. Income has a positive relation with beef consumption so as people's incomes increase we can expect beef consumption to increase. Finally Pork consumption is negatively related to beef consumption so we can see that less beef consumption corresponds with higher pork consumption.

The final model being: BeefConsumption = 90.813646 + (-1.849850*Price) + (0.083190*Income) + (-0.415085*PorkConsumption)

This model explains approximatly 92.12% of the variability in BeefConsumption in this dataset.

Question 2

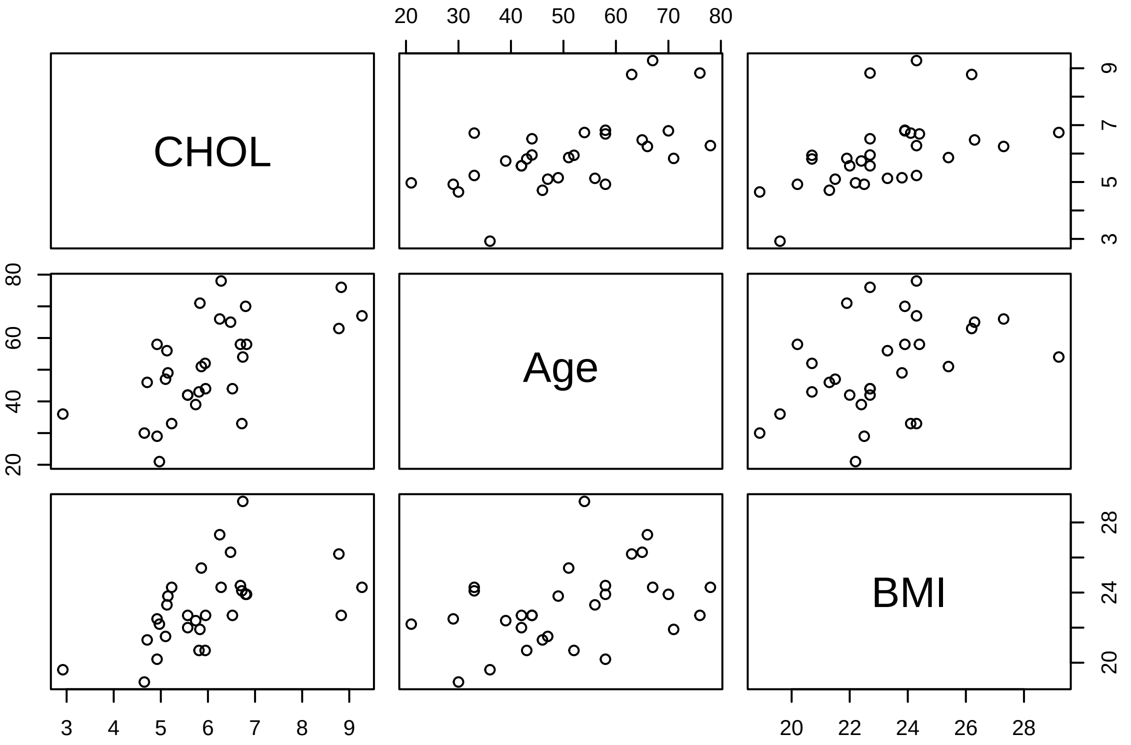
```
chol_data <- read.table(file="C:/Users/wesch/uvic/stat359 Data Analysis/data1/chol.txt",
↪  header=TRUE)
attach(chol_data)
library(knitr)

summary(chol_data)
```

```
##       CHOL            Age             BMI
##  Min.   :2.920   Min.   :21.00   Min.   :18.90
##  1st Qu.:5.135   1st Qu.:42.00   1st Qu.:21.93
##  Median :5.845   Median :50.00   Median :22.70
##  Mean   :6.005   Mean   :50.70   Mean   :23.18
##  3rd Qu.:6.647   3rd Qu.:61.75   3rd Qu.:24.30
##  Max.   :9.270   Max.   :78.00   Max.   :29.20
```

Initial scatter plots

```
pairs(chol_data)
```



From the initial plots it can be observed that CHOL has a postive relationship with Age and BMI. It also seems as there is a positive relationship between BMI and Age aswell. Not all of these relationships look extremely linear.

Simple linear models:

```
simple_age <- lm(CHOL ~ Age)
summary(simple_age)
```

```
##
## Call:
## lm(formula = CHOL ~ Age)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.29944 -0.67361  0.02992  0.40873  2.39393
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.29561    0.70480   4.676 6.72e-05 ***
## Age          0.05344    0.01336   3.999 0.000422 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.063 on 28 degrees of freedom
## Multiple R-squared:  0.3635, Adjusted R-squared:  0.3408
## F-statistic: 15.99 on 1 and 28 DF,  p-value: 0.0004216
```

```r
simple_bmi <- lm(CHOL ~ BMI)
summary(simple_bmi)
```

```
##
## Call:
## lm(formula = CHOL ~ BMI)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.97890 -0.80623 -0.07073  0.53611  2.97330
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.15683    2.14558  -0.539   0.5940
## BMI          0.30897    0.09214   3.353   0.0023 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.125 on 28 degrees of freedom
## Multiple R-squared:  0.2865, Adjusted R-squared:  0.2611
## F-statistic: 11.24 on 1 and 28 DF,  p-value: 0.002303
```

Will now fit a full model including interactions and quadratic variables.

```r
chol_fullmodel <- lm(CHOL ~ Age + BMI + (BMI*Age))
summary(chol_fullmodel)
```

```
##
## Call:
## lm(formula = CHOL ~ Age + BMI + (BMI * Age))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.56312 -0.72399 -0.05217  0.40839  2.40946
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -6.546427   8.947853  -0.732    0.471
## Age          0.154186   0.170975   0.902    0.375
## BMI          0.457127   0.395281   1.156    0.258
## Age:BMI     -0.004933   0.007426  -0.664    0.512
##
## Residual standard error: 1.002 on 26 degrees of freedom
## Multiple R-squared:  0.4743, Adjusted R-squared:  0.4137
## F-statistic:  7.82 on 3 and 26 DF,  p-value: 0.0007002
```

Interaction between Age and BMI does not seem statistically significant with a p-value of 0.906 there is almost no evidence supporting it.

Remvoing the interaction.

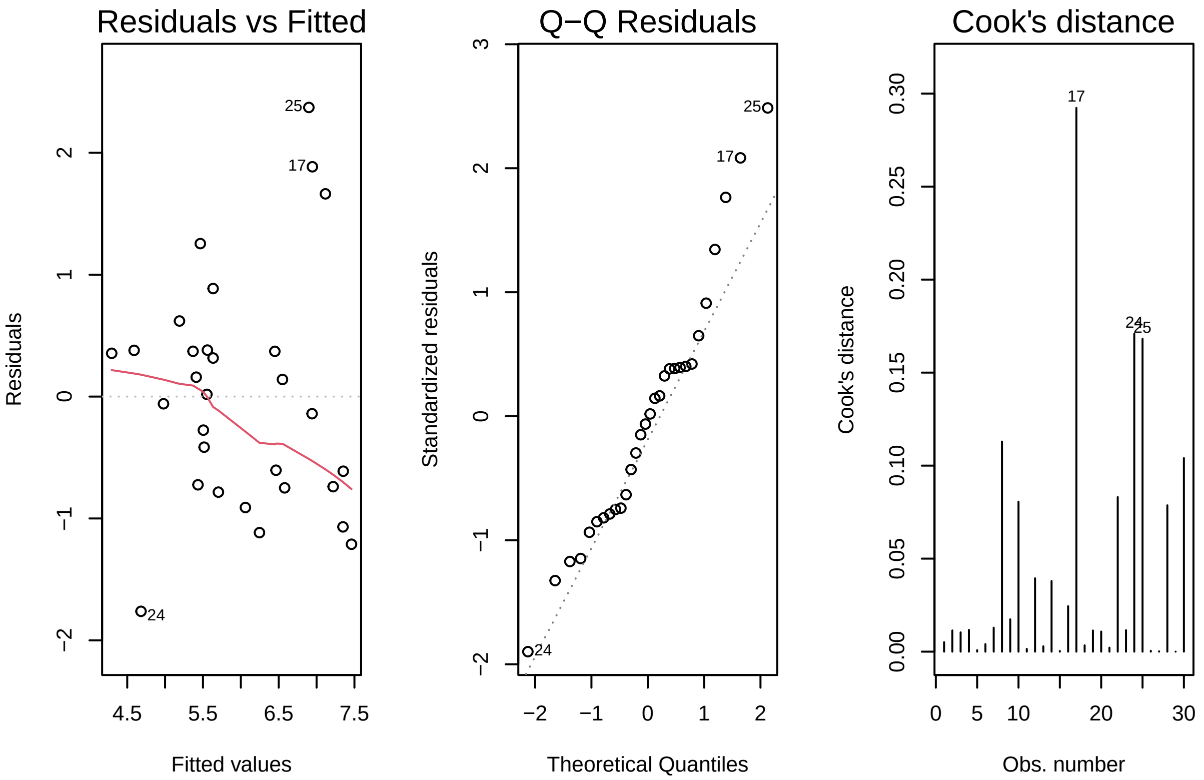
```
chol_reducedmodel <- update(chol_fullmodel,.~.- BMI:Age)
summary(chol_reducedmodel)
```

```
##
## Call:
## lm(formula = CHOL ~ Age + BMI)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7619 -0.7353 -0.0205  0.3772  2.3717
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.73983    1.89641  -0.390  0.69951
## Age          0.04097    0.01363   3.006  0.00567 **
## BMI          0.20137    0.08876   2.269  0.03149 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.992 on 27 degrees of freedom
## Multiple R-squared:  0.4654, Adjusted R-squared:  0.4258
## F-statistic: 11.75 on 2 and 27 DF,  p-value: 0.000213
```

Will proceed with chol_reducedmodel as all terms are significant.

Model diagnostics

```
par(mfrow=c(1,3))
plot(chol_reducedmodel,which=c(1,2,4))
```



Is safe to assume model ahs equal variance but residuals show a right skew and has a significant outlier at observation 17.

Will performance log transformation of the response.

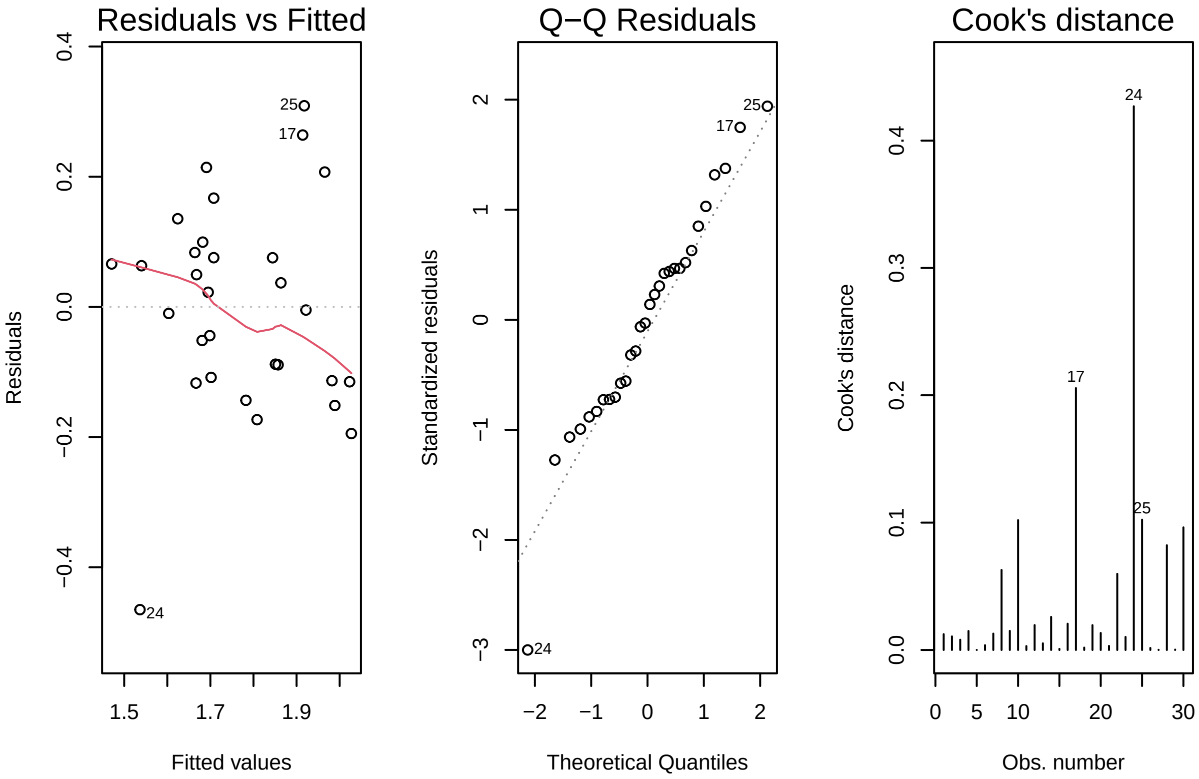
```
chollog_reducedmodel<-update(chol_reducedmodel,log(.)~.)
summary(chollog_reducedmodel)
```

```
##
## Call:
## lm(formula = log(CHOL) ~ Age + BMI)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.46502 -0.11212  0.00883  0.08151  0.30894
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.548262   0.316618   1.732   0.0948 .
## Age         0.006449   0.002276   2.834   0.0086 **
## BMI         0.038581   0.014819   2.604   0.0148 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1656 on 27 degrees of freedom
## Multiple R-squared:  0.4787, Adjusted R-squared:  0.4401
## F-statistic:  12.4 on 2 and 27 DF,  p-value: 0.0001516
```

R-squared went up a small amount.

Will check diagnostics

```
par(mfrow=c(1,3))
plot(chollog_reducedmodel,which=c(1,2,4))
```



The qq plot shows an approximately normal distribution and continued equal variance. With observation 24 being an extreme outlier.

Will remove observation 24.

```
chollog_reducedmodel2<-update(chollog_reducedmodel,.~.,subset=(1:length(CHOL)!=24))
summary(chollog_reducedmodel2)
```

```
##
## Call:
## lm(formula = log(CHOL) ~ Age + BMI, subset = (1:length(CHOL) !=
##     24))
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.186747 -0.090342 -0.005277  0.054324  0.312609
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.856019   0.276883   3.092  0.00471 **
## Age         0.005917   0.001899   3.116  0.00443 **
## BMI         0.027230   0.012724   2.140  0.04190 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1378 on 26 degrees of freedom
## Multiple R-squared:  0.4616, Adjusted R-squared:  0.4202
## F-statistic: 11.15 on 2 and 26 DF,  p-value: 0.0003195
```

Removing this point affects BMI's p-value and reduces the R-squared value a little bit.

In conclusion log of the serum cholesterol is positively associated with BMI when Age is included in the model.

Final model being: CHOL = exp{0.856019 + 0.005917*Age + 0.027230*BMI}

This model explains 46.16% of the variation of CHOL in this data set. Which is a little less than half of the variability.