S510 HW 4

Erick Castillo

11/22/2021

Problem 1

This problem uses the **divusa** data set. The divorce column will be used as a response variable, with the remaining columns as predictors. Find the "best" model with the following methods.

A. Use stepwise regression with AIC.

```
redu.mod <- lm(divorce~1, data = divusa)
full.mod <- lm(divorce~., data = divusa)
step(redu.mod, scope = list(lower = redu.mod, upper = full.mod))
```

```
## Start:  AIC=268.19
## divorce ~ 1
##
##              Df Sum of Sq     RSS    AIC
## + femlab      1   2024.42  418.10 134.28
## + year        1   1888.22  554.31 155.99
## + birth       1   1272.98 1169.54 213.48
## + marriage    1    697.17 1745.36 244.31
## + unemployed  1    108.33 2334.19 266.69
## <none>                    2442.53 268.19
## + military    1      0.84 2441.68 270.16
##
## Step:  AIC=134.28
## divorce ~ femlab
##
##              Df Sum of Sq     RSS    AIC
## + birth       1    113.73  304.38 111.83
## + year        1     29.70  388.41 130.60
## + marriage    1     13.34  404.76 133.78
## <none>                     418.10 134.28
## + military    1      1.93  416.17 135.92
## + unemployed  1      1.48  416.62 136.00
## - femlab      1   2024.42 2442.53 268.19
##
## Step:  AIC=111.83
## divorce ~ femlab + birth
##
##              Df Sum of Sq     RSS     AIC
## + marriage    1     94.54  209.84  85.196
## + unemployed  1     44.43  259.94 101.683
```

```
## + year        1     15.54  288.84 109.798
## <none>                     304.38 111.834
## + military    1      0.87  303.50 113.613
## - birth       1    113.73  418.10 134.278
## - femlab      1    865.16 1169.54 213.483
##
## Step:  AIC=85.2
## divorce ~ femlab + birth + marriage
##
##              Df Sum of Sq     RSS     AIC
## + year        1     26.76  183.08  76.691
## + unemployed  1      6.85  202.99  84.639
## + military    1      5.66  204.18  85.089
## <none>                     209.84  85.196
## - marriage    1     94.54  304.38 111.834
## - birth       1    194.92  404.76 133.781
## - femlab      1    949.45 1159.29 214.805
##
## Step:  AIC=76.69
## divorce ~ femlab + birth + marriage + year
##
##              Df Sum of Sq    RSS     AIC
## + military    1    20.957 162.12  69.330
## <none>                    183.08  76.691
## + unemployed  1     0.651 182.43  78.417
## - year        1    26.761 209.84  85.196
## - marriage    1   105.757 288.84 109.798
## - femlab      1   137.509 320.59 117.829
## - birth       1   183.446 366.53 128.140
##
## Step:  AIC=69.33
## divorce ~ femlab + birth + marriage + year + military
##
##              Df Sum of Sq    RSS     AIC
## <none>                    162.12  69.330
## + unemployed  1     1.925 160.20  70.410
## - military    1    20.957 183.08  76.691
## - year        1    42.054 204.18  85.089
## - marriage    1   126.643 288.77 111.779
## - femlab      1   158.003 320.13 119.718
## - birth       1   172.826 334.95 123.203

##
## Call:
## lm(formula = divorce ~ femlab + birth + marriage + year + military,
##     data = divusa)
##
## Coefficients:
## (Intercept)       femlab        birth     marriage         year     military
##    405.6167       0.8548      -0.1101       0.1593      -0.2179      -0.0412
```

Answer: From the above output, it is clear that the best model, with the smallest AIC, is the one with femlab, birth, marriage, year, and military as predictor variables.

B. Use best subsets regression with R^2_{adj}.

```r
attach(divusa)
r2.mod <- regsubsets(cbind(year, unemployed, femlab, marriage, birth, military), divorce)
summary.mod <- summary(r2.mod)

summary.mod$which
```

```
##   (Intercept)  year unemployed femlab marriage birth military
## 1        TRUE FALSE      FALSE   TRUE    FALSE FALSE    FALSE
## 2        TRUE FALSE      FALSE   TRUE    FALSE  TRUE    FALSE
## 3        TRUE FALSE      FALSE   TRUE     TRUE  TRUE    FALSE
## 4        TRUE  TRUE      FALSE   TRUE     TRUE  TRUE    FALSE
## 5        TRUE  TRUE      FALSE   TRUE     TRUE  TRUE     TRUE
## 6        TRUE  TRUE       TRUE   TRUE     TRUE  TRUE     TRUE
```

```r
summary.mod$rsq
```

```
## [1] 0.8288227 0.8753838 0.9140885 0.9250448 0.9336249 0.9344132
```

Answer: The model with the best R^2_{adj} is the one with all the predictors present in the model. That is, year, unemployed, femlab, marriage, birth, and military are all in.

C. Use best subsets regression with adjusted Mallow's C_p.

```r
summary.mod$which
```

```
##   (Intercept)  year unemployed femlab marriage birth military
## 1        TRUE FALSE      FALSE   TRUE    FALSE FALSE    FALSE
## 2        TRUE FALSE      FALSE   TRUE    FALSE  TRUE    FALSE
## 3        TRUE FALSE      FALSE   TRUE     TRUE  TRUE    FALSE
## 4        TRUE  TRUE      FALSE   TRUE     TRUE  TRUE    FALSE
## 5        TRUE  TRUE      FALSE   TRUE     TRUE  TRUE     TRUE
## 6        TRUE  TRUE       TRUE   TRUE     TRUE  TRUE     TRUE
```

```r
summary.mod$cp
```

```
## [1] 109.695444  62.001274  22.692257  12.998703   5.841314   7.000000
```

```r
detach(divusa)
```

Answer: We can conclude that the "best" model, when using Mallow's C_p is the one that has year, femlab, marriage, birth, and military as predictors in the model. Notice that for this model, $C_p \approx 5.84 < p = 6$, meaning that this model has the least amount of bias.

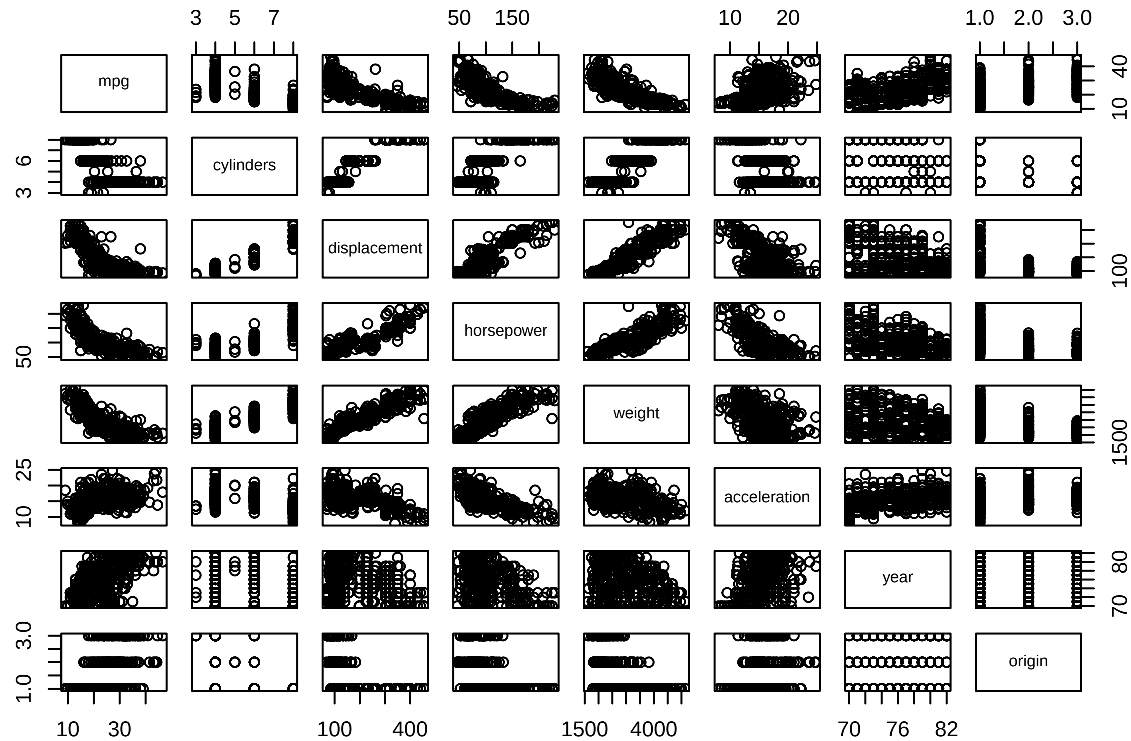
Problem 2

This problem uses the **Auto** data set.

A. Produce a scatterplot matrix which includes all the variables in the data set.

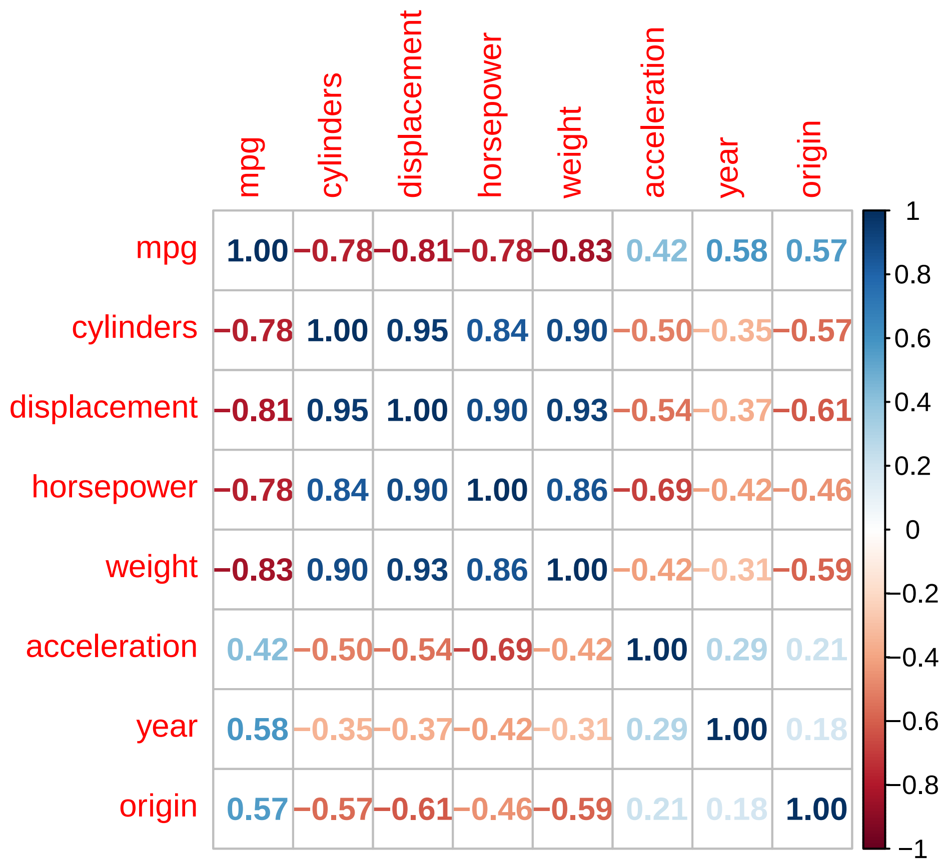
In this case, I will omit the name column. It has a total of 301 unique strings, meaning that these will eat up the degrees of freedom if they are included as categorical variables.

```r
Auto1 <- subset(Auto, select = -c(name))
pairs(Auto1)
```



B. Compute and visualize the matrix of correlations between the above variables.

```r
corrplot(cor(Auto1), method = 'number')
```



C. Perform multiple linear regression with **mpg** as the response variable, with all other variables as predictors.

```
mod1 <- lm(mpg~., data = Auto1)
summary(mod1)
```

```
##
## Call:
## lm(formula = mpg ~ ., data = Auto1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.5903 -2.1565 -0.1169  1.8690 13.0604
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -17.218435   4.644294  -3.707  0.00024 ***
## cylinders     -0.493376   0.323282  -1.526  0.12780
## displacement   0.019896   0.007515   2.647  0.00844 **
## horsepower    -0.016951   0.013787  -1.230  0.21963
## weight        -0.006474   0.000652  -9.929  < 2e-16 ***
## acceleration   0.080576   0.098845   0.815  0.41548
## year           0.750773   0.050973  14.729  < 2e-16 ***
## origin         1.426141   0.278136   5.127 4.67e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 3.328 on 384 degrees of freedom
## Multiple R-squared:  0.8215, Adjusted R-squared:  0.8182
## F-statistic: 252.4 on 7 and 384 DF,  p-value: < 2.2e-16
```

i. Is there a relationship between the predictors and the response?

Answer: There appears to be a strong relationship between mpg and the predictors present in the model. Notice that $R^2_{adj} = 0.8182$, which is high. There are a few predictors that do not appear to be significant given the presence of the other predictors. These include **cylinders, horsepower, and acceleration.**

ii. Which predictors appear to have a statistically significant relationship with the response?

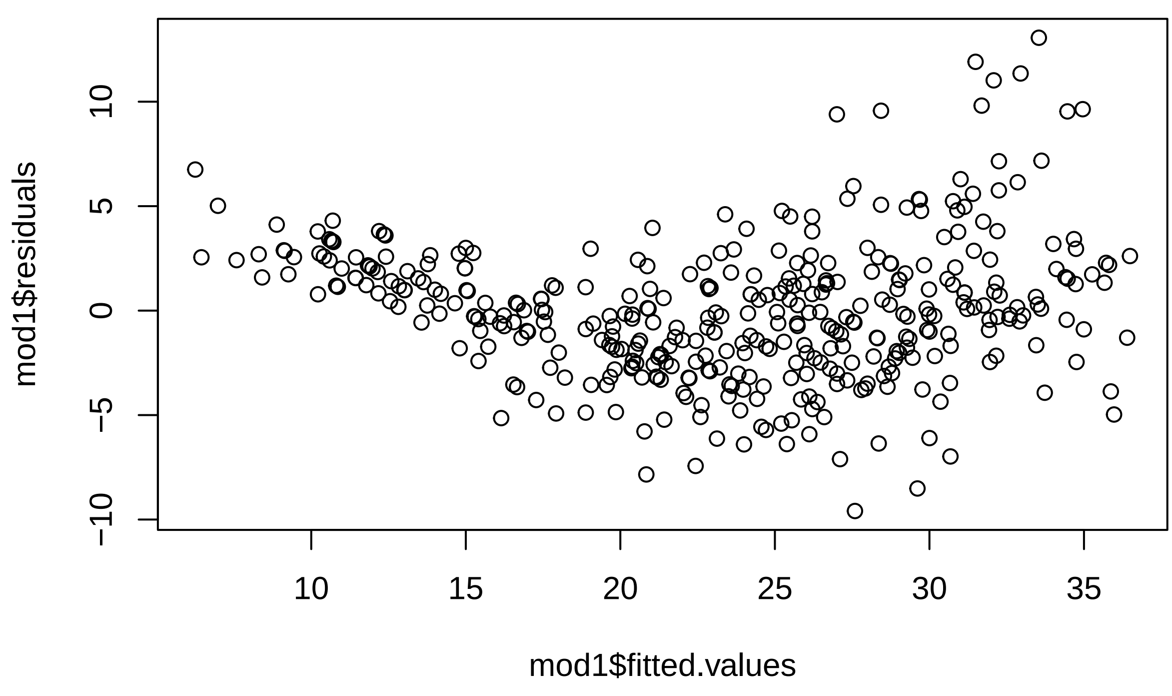
Answer: It appears that **displacement, weight, year, and origin** all have a statistically significant relationship with mpg, given the presence of the other predictors in the model.

iii. What does the coefficient for the **year** variable suggest?

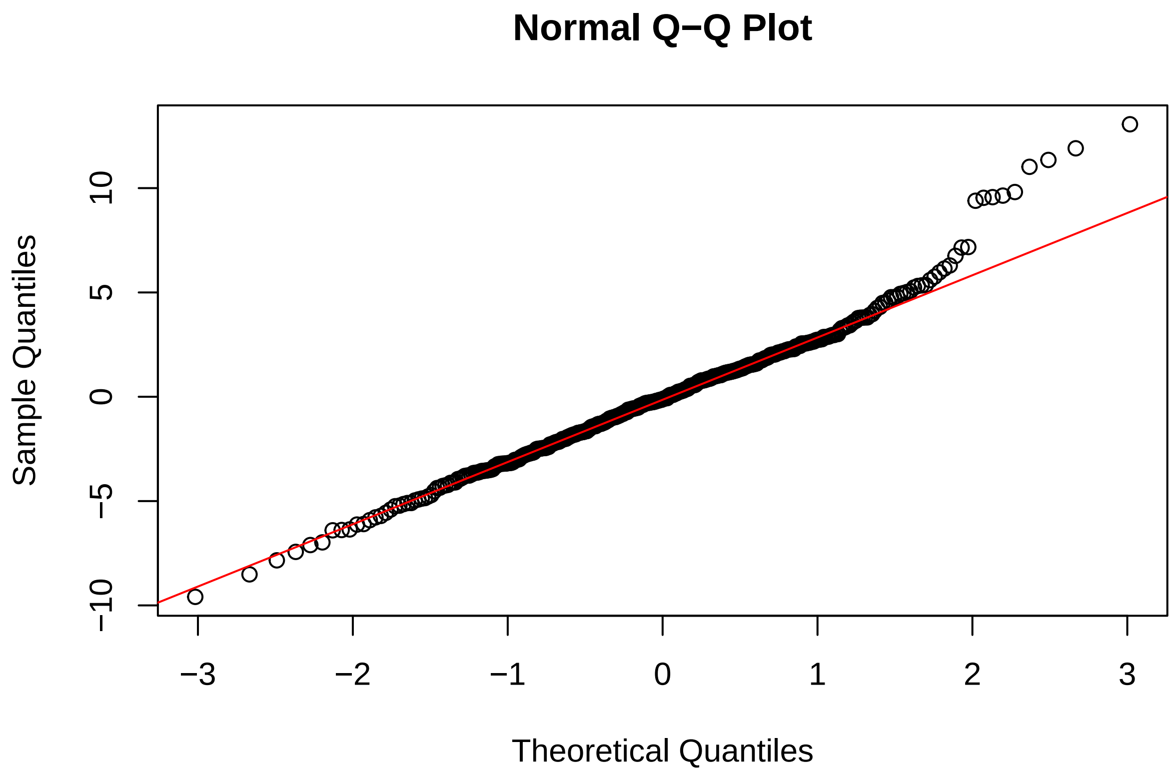
Answer: $\hat{\beta}_{year} \approx 0.751$. This means that for every year that passes, we would expect the average mpg for cars to increase by 0.751, holding all the other variables in the model constant.

D. Perform a residual analysis, are there any problems with this fit?

```
plot(mod1$fitted.values, mod1$residuals)
```




```
qqnorm(mod1$residuals)
qqline(mod1$residual, col = 'red')
```




```r
shapiro.test(mod1$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  mod1$residuals
## W = 0.97659, p-value = 5.768e-06
```

Answer: The fit vs. residuals plot shows that there is a lack of linearity. There is also a fanning pattern with the residuals, meaning that there is non-constant variance. Finally, the Normal QQ-plot indicates that there is an issue with the normality of the residuals. This claim is further supported by the Shaprio-Wilk test output, which has a very small p-value. This means that there is very strong evidence to suggest that the residuals are not normally distributed.

E. Are there any outliers? Are there any high-leverage points?

```r
# the following code identifies high leverage points.
hv1 <- hatvalues(mod1)
which(hv1 > 3*(mod1$rank/dim(Auto1)[1]))
```

```
##  9 14 27 28 29
##  9 14 27 28 29
```

```r
# the following code identifies outliers.
rstan1 <- rstandard(mod1)
which(rstan1 > 3 | rstan1 < -3)
```

```
## 245 323 326 327
## 243 321 324 325
```

Answer: From the above output, I can tell that there are 5 high-leverage points, including the 9th, 14th, 27th, 28th, and 29th obsevations in the Auto1 data set.

There are also 4 outliers including the 243rd, 321st, 324th, and 325th entries in the Auto1 data set.

F. Use the add1 function to find at least one significant interaction term. Update the model in part **C**.

```r
add1(mod1, ~.+displacement*horsepower+horsepower*weight+
       weight*acceleration+acceleration*displacement+
       horsepower*acceleration , data = Auto1, test ='F')
```

```
## Single term additions
##
## Model:
## mpg ~ cylinders + displacement + horsepower + weight + acceleration +
##     year + origin
##                           Df Sum of Sq    RSS    AIC F value    Pr(>F)
## <none>                                 4252.2 950.50
## displacement:horsepower    1   1003.62 3248.6 846.97 118.325 < 2.2e-16 ***
## horsepower:weight          1    961.33 3290.9 852.04 111.881 < 2.2e-16 ***
## weight:acceleration        1    473.56 3778.7 906.22  47.999 1.813e-11 ***
## displacement:acceleration  1    489.38 3762.8 904.57  49.812 7.991e-12 ***
## horsepower:acceleration    1    464.24 3788.0 907.18  46.940 2.933e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
mod2 <- update(mod1, .~.+displacement*horsepower)
summary(mod2)
```

```
##
## Call:
## lm(formula = mpg ~ cylinders + displacement + horsepower + weight +
##     acceleration + year + origin + displacement:horsepower, data = Auto1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.7010 -1.6009 -0.0967  1.4119 12.6734
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            -1.894e+00  4.302e+00  -0.440  0.66007
## cylinders               6.466e-01  3.017e-01   2.143  0.03275 *
## displacement           -7.487e-02  1.092e-02  -6.859 2.80e-11 ***
## horsepower             -1.975e-01  2.052e-02  -9.624  < 2e-16 ***
## weight                 -3.147e-03  6.475e-04  -4.861 1.71e-06 ***
## acceleration           -2.131e-01  9.062e-02  -2.351  0.01921 *
```

```
## year                      7.379e-01  4.463e-02  16.534  < 2e-16 ***
## origin                    6.891e-01  2.527e-01   2.727  0.00668 **
## displacement:horsepower   5.236e-04  4.813e-05  10.878  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.912 on 383 degrees of freedom
## Multiple R-squared:  0.8636, Adjusted R-squared:  0.8608
## F-statistic: 303.1 on 8 and 383 DF,  p-value: < 2.2e-16
```

Answer: The significant interaction term that I added to my model was displacement*horsepower.

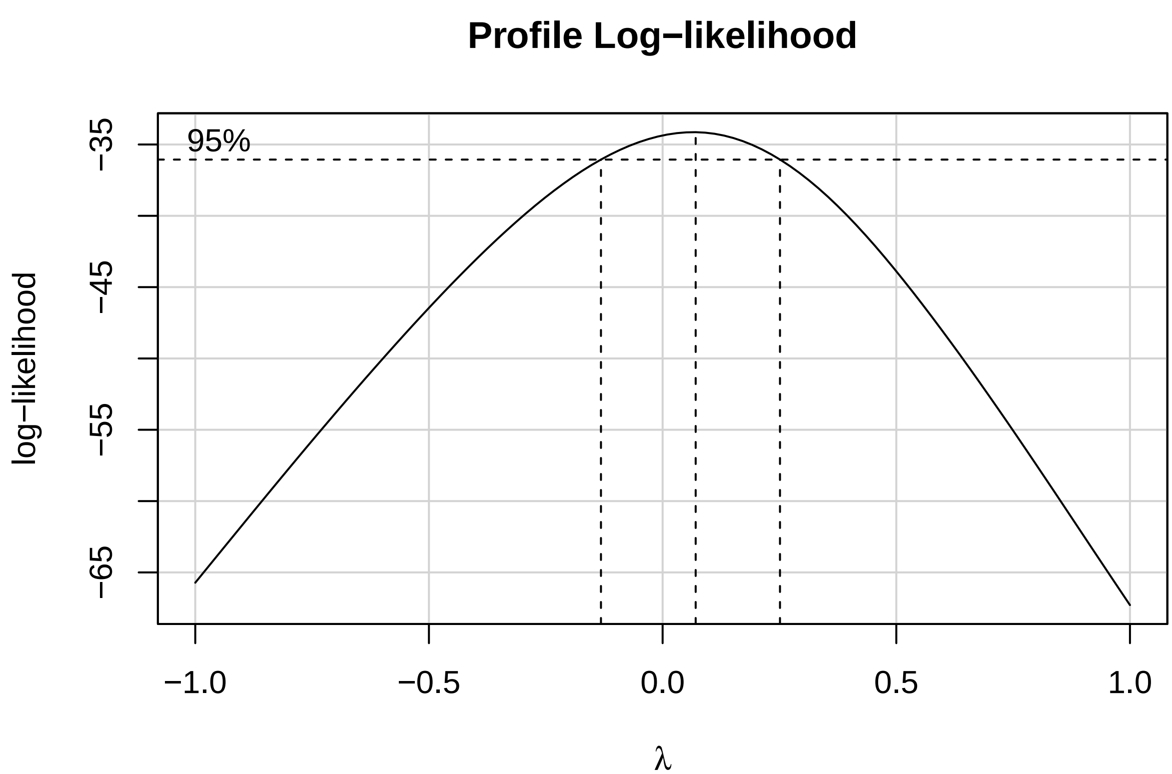
Problem 3

This problem uses the **lathe1** data set.

A. Starting with the second-order model specified in the problem, use the Box-Cox method to show that the response requires a logarithmic transformation.

```
mod3 <- lm(Life~Speed+Feed+I(Speed^2)+I(Feed^2)+Speed*Feed, data = lathe1)

mod.boxcox = boxCox(mod3, lambda = seq(-1, 1, length = 10))
```



Answer: Because $0 \in C.I$ in the above plot, we can conclude that a log() transformation of the response variable would be useful.

B. State the null and alternative hypotheses for the global F-test for the model with log(Life). Perform the test and summarize the results.

Answer: The null and alternative hypotheses are as follows:

$H_0 : \hat{\beta}_{Speed} = \hat{\beta}_{Feed} = \hat{\beta}_{Speed^2} = \hat{\beta}_{Feed^2} = \hat{\beta}_{Speed\times Feed} = 0$

H_A : at least one $B_i \neq 0$, where $i \in \{$Speed, Feed, $Speed^2$, $Feed^2$, $Speed \times Feed\}$.

The test is performed in the following chunk:

```
logLife <- log(lathe1$Life)
mod4.full <- lm(logLife~Speed+Feed+I(Speed^2)+I(Feed^2)+Speed*Feed, data = lathe1)
mod4.redu <- lm(logLife~1, data = lathe1)
anova(mod4.redu, mod4.full)
```

```
## Analysis of Variance Table
##
## Model 1: logLife ~ 1
## Model 2: logLife ~ Speed + Feed + I(Speed^2) + I(Feed^2) + Speed * Feed
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     19 41.533
## 2     14  1.237  5    40.296 91.236 3.551e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the above output, we can see that the p-value is very close to 0. This implies that there is strong evidence to suggest that at least one of the $\hat{\beta}_i$'s $\neq 0$. We would reject H_0.

C. Explain the practical meaning of the hypothesis $H_0 : \beta_1 = \beta_{11} = \beta_{12} = 0$ in the context of the above model.

Answer: This is a partial F-test that examines whether Speed, $Speed^2$, and $Speed \times Feed$ are significant in the model.

D. Perform the test in **C.** and summarize results.

```
mod4.part <- lm(logLife~Feed+I(Feed^2), data = lathe1)
anova(mod4.part, mod4.full)
```

```
## Analysis of Variance Table
##
## Model 1: logLife ~ Feed + I(Feed^2)
## Model 2: logLife ~ Speed + Feed + I(Speed^2) + I(Feed^2) + Speed * Feed
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     17 32.300
## 2     14  1.237  3    31.063 117.22 3.726e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Answer: Because the above output's p-value is so low, we would conclude that at least one of the tested predictors is useful in the model. We would reject H_0.

E. Using Cook's distance, find the two most influential observations when using the fit of the quadratic mean function for log(Life). Explain why these observations are influential. Delete these points, and refit the model. Compare the fit with all the data.

```r
cooks.distance(mod4.full)
```

```
##            1            2            3            4            5            6 
## 0.0745581876 0.0002358999 0.1611290980 0.0293444172 0.4172638143 0.0089104068 
##            7            8            9           10           11           12 
## 0.2024479551 0.0333705363 0.7611370235 0.7088115474 0.0755462115 0.0932562838 
##           13           14           15           16           17           18 
## 0.0066483194 0.0491977930 0.0001916341 0.0121013330 0.0077362334 0.0001916341 
##           19           20 
## 0.0121013330 0.0012883357
```

Answer: From the above output, we observe that the 9th and 10th observations are both greater than 0.5. This indicates these points may be influential. The following chunk of code drops these rows and refits the model:

```r
lathe2 <- lathe1[-c(9,10), ]
logLife2 <- log(lathe2$Life)

mod5 <- lm(logLife2~Speed+Feed+I(Speed^2)+I(Feed^2)+Speed*Feed, data = lathe2)

summary(mod4.full) # original model
```

```
## 
## Call:
## lm(formula = logLife ~ Speed + Feed + I(Speed^2) + I(Feed^2) + 
##     Speed * Feed, data = lathe1)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -0.43349 -0.14576 -0.02494  0.16748  0.47992 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  1.18809    0.10508  11.307 2.00e-08 ***
## Speed       -1.58902    0.08580 -18.520 3.04e-11 ***
## Feed        -0.79023    0.08580  -9.210 2.56e-07 ***
## I(Speed^2)   0.28808    0.10063   2.863 0.012529 *  
## I(Feed^2)    0.41851    0.10063   4.159 0.000964 ***
## Speed:Feed  -0.07286    0.10508  -0.693 0.499426    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.2972 on 14 degrees of freedom
## Multiple R-squared:  0.9702, Adjusted R-squared:  0.9596 
## F-statistic: 91.24 on 5 and 14 DF,  p-value: 3.551e-10
```

```r
summary(mod5) # model with two points dropped
```

```
## 
## Call:
## lm(formula = logLife2 ~ Speed + Feed + I(Speed^2) + I(Feed^2) + 
##     Speed * Feed, data = lathe2)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.39963 -0.14660  0.00387  0.14917  0.32783
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.18809    0.08241  14.417 6.11e-09 ***
## Speed       -1.43300    0.08241 -17.388 7.10e-10 ***
## Feed        -0.79023    0.06729 -11.743 6.15e-08 ***
## I(Speed^2)   0.28022    0.12363   2.267 0.042700 *
## I(Feed^2)    0.42244    0.09217   4.583 0.000629 ***
## Speed:Feed  -0.07286    0.08241  -0.884 0.394025
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2331 on 12 degrees of freedom
## Multiple R-squared:  0.9759, Adjusted R-squared:  0.9658
## F-statistic: 97.07 on 5 and 12 DF,  p-value: 2.804e-09
```

Answer: Dropping the two most influential points slightly increased the R^2_{adj} of the model from 0.9596 to 0.9658.