STAT 425: HW4

Sharvi Toamr

10/22/2021

Contents

Problem 1: 1
Problem 2 7

Problem 1:

Use prostate data with lpsa as the response and the other variables as predictors.

```
library(faraway)
data("prostate")

# Fitting linear model with all variables
g = lm(lpsa ~ ., data = prostate)
summary(g)
```

```
##
## Call:
## lm(formula = lpsa ~ ., data = prostate)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7331 -0.3713 -0.0170  0.4141  1.6381
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.669337   1.296387   0.516  0.60693
## lcavol       0.587022   0.087920   6.677 2.11e-09 ***
## lweight      0.454467   0.170012   2.673  0.00896 **
## age         -0.019637   0.011173  -1.758  0.08229 .
## lbph         0.107054   0.058449   1.832  0.07040 .
## svi          0.766157   0.244309   3.136  0.00233 **
## lcp         -0.105474   0.091013  -1.159  0.24964
## gleason      0.045142   0.157465   0.287  0.77503
## pgg45        0.004525   0.004421   1.024  0.30886
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7084 on 88 degrees of freedom
```

```
## Multiple R-squared:  0.6548, Adjusted R-squared:  0.6234
## F-statistic: 20.86 on 8 and 88 DF,  p-value: < 2.2e-16
```

Implement the following variable selection methods to determine the best model:

(a) Backward elimination

```r
#removing gleason variable as it most insignificant (p-value > 0.05)
g = update(g, . ~ . - gleason)
summary(g)
```

```
##
## Call:
## lm(formula = lpsa ~ lcavol + lweight + age + lbph + svi + lcp +
##     pgg45, data = prostate)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.73117 -0.38137 -0.01728  0.43364  1.63513
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.953926   0.829439   1.150  0.25319
## lcavol       0.591615   0.086001   6.879 8.07e-10 ***
## lweight      0.448292   0.167771   2.672  0.00897 **
## age         -0.019336   0.011066  -1.747  0.08402 .
## lbph         0.107671   0.058108   1.853  0.06720 .
## svi          0.757734   0.241282   3.140  0.00229 **
## lcp         -0.104482   0.090478  -1.155  0.25127
## pgg45        0.005318   0.003433   1.549  0.12488
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7048 on 89 degrees of freedom
## Multiple R-squared:  0.6544, Adjusted R-squared:  0.6273
## F-statistic: 24.08 on 7 and 89 DF,  p-value: < 2.2e-16
```

```r
# Eliminating lcp variable as it most insignificant (p-value > 0.05)
g = update(g, . ~ . - lcp)
summary(g)
```

```
##
## Call:
## lm(formula = lpsa ~ lcavol + lweight + age + lbph + svi + pgg45,
##     data = prostate)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.77711 -0.41708  0.00002  0.40676  1.59681
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)  0.980085   0.830665   1.180  0.24116
## lcavol       0.545770   0.076431   7.141 2.31e-10 ***
## lweight      0.449450   0.168078   2.674  0.00890 **
## age         -0.017470   0.010967  -1.593  0.11469
## lbph         0.105755   0.058191   1.817  0.07249 .
## svi          0.641666   0.219757   2.920  0.00442 **
## pgg45        0.003528   0.003068   1.150  0.25331
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7061 on 90 degrees of freedom
## Multiple R-squared:  0.6493, Adjusted R-squared:  0.6259
## F-statistic: 27.77 on 6 and 90 DF,  p-value: < 2.2e-16
```

```
# Eliminating pgg45 variable as it most insignificant (p-value > 0.05)
g = update(g, . ~ . - pgg45)
summary(g)
```

```
##
## Call:
## lm(formula = lpsa ~ lcavol + lweight + age + lbph + svi, data = prostate)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.83505 -0.39396  0.00414  0.46336  1.57888
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.95100    0.83175   1.143 0.255882
## lcavol       0.56561    0.07459   7.583 2.77e-11 ***
## lweight      0.42369    0.16687   2.539 0.012814 *
## age         -0.01489    0.01075  -1.385 0.169528
## lbph         0.11184    0.05805   1.927 0.057160 .
## svi          0.72095    0.20902   3.449 0.000854 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7073 on 91 degrees of freedom
## Multiple R-squared:  0.6441, Adjusted R-squared:  0.6245
## F-statistic: 32.94 on 5 and 91 DF,  p-value: < 2.2e-16
```

```
# Eliminating age variable as it most insignificant (p-value > 0.05)
g = update(g, . ~ . - age)
summary(g)
```

```
##
## Call:
## lm(formula = lpsa ~ lcavol + lweight + lbph + svi, data = prostate)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.82653 -0.42270  0.04362  0.47041  1.48530
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.14554    0.59747   0.244  0.80809
## lcavol       0.54960    0.07406   7.422 5.64e-11 ***
## lweight      0.39088    0.16600   2.355  0.02067 *
## lbph         0.09009    0.05617   1.604  0.11213
## svi          0.71174    0.20996   3.390  0.00103 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7108 on 92 degrees of freedom
## Multiple R-squared:  0.6366, Adjusted R-squared:  0.6208
## F-statistic: 40.29 on 4 and 92 DF,  p-value: < 2.2e-16
```

```
# Eliminating lbph variable as it most insignificant (p-value > 0.05)
g = update(g, . ~ . - lbph)
summary(g)
```

```
##
## Call:
## lm(formula = lpsa ~ lcavol + lweight + svi, data = prostate)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.72964 -0.45764  0.02812  0.46403  1.57013
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.26809    0.54350  -0.493  0.62298
## lcavol       0.55164    0.07467   7.388  6.3e-11 ***
## lweight      0.50854    0.15017   3.386  0.00104 **
## svi          0.66616    0.20978   3.176  0.00203 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7168 on 93 degrees of freedom
## Multiple R-squared:  0.6264, Adjusted R-squared:  0.6144
## F-statistic: 51.99 on 3 and 93 DF,  p-value: < 2.2e-16
```

The final model using backward elimination has 3 covariates and is: lpsa ~ lcavol + lweight + svi

(b) AIC

```
# Using the best subset selection
library(leaps)
b = regsubsets(lpsa ~ ., data = prostate)
rs = summary(b)
rs$which
```

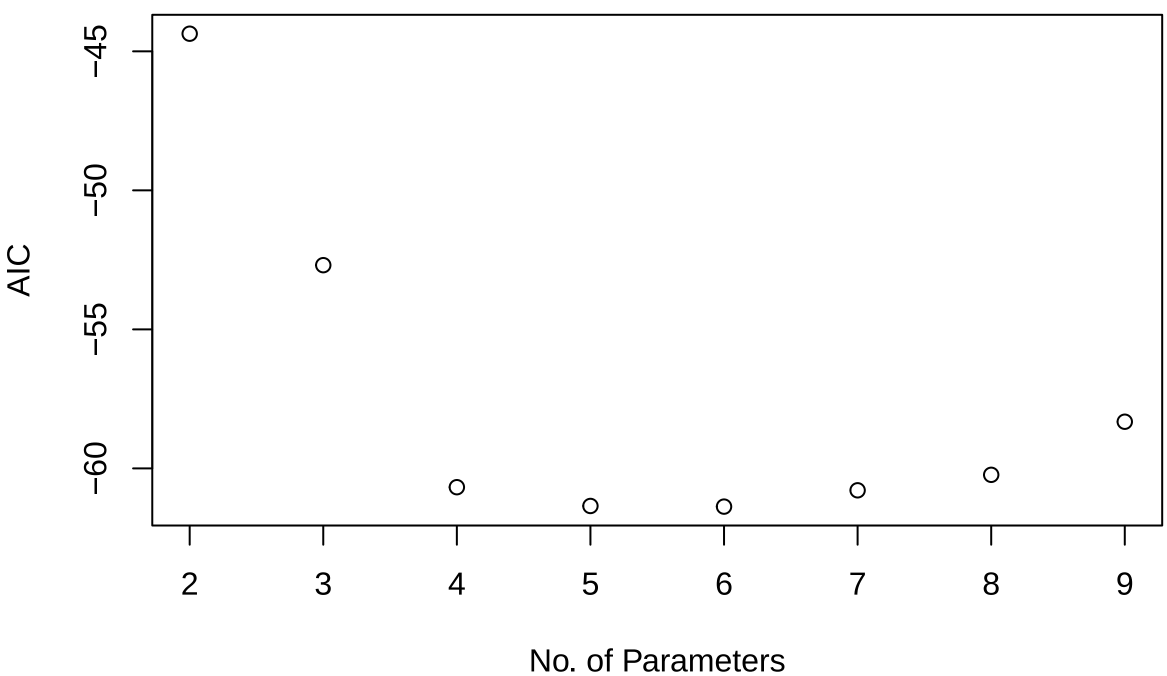
```
##   (Intercept) lcavol lweight   age  lbph   svi   lcp gleason pgg45
## 1        TRUE   TRUE   FALSE FALSE FALSE FALSE FALSE   FALSE FALSE
## 2        TRUE   TRUE    TRUE FALSE FALSE FALSE FALSE   FALSE FALSE
```

```
## 3         TRUE    TRUE    TRUE FALSE FALSE  TRUE FALSE   FALSE FALSE
## 4         TRUE    TRUE    TRUE FALSE  TRUE  TRUE FALSE   FALSE FALSE
## 5         TRUE    TRUE    TRUE  TRUE  TRUE  TRUE FALSE   FALSE FALSE
## 6         TRUE    TRUE    TRUE  TRUE  TRUE  TRUE FALSE   FALSE  TRUE
## 7         TRUE    TRUE    TRUE  TRUE  TRUE  TRUE  TRUE   FALSE  TRUE
## 8         TRUE    TRUE    TRUE  TRUE  TRUE  TRUE  TRUE    TRUE  TRUE
```

```r
# Using AIC
n = dim(prostate)[1]

msize = 2:9

Aic = n * log(rs$rss / n) + 2 * msize

AIC_size = msize[which.min(Aic)]
AIC_size
```

```
## [1] 6
```

```r
# plot
plot(msize, Aic, xlab = "No. of Parameters", ylab = "AIC")
```



Number of variables selected using AIC is 6(including the intercept).

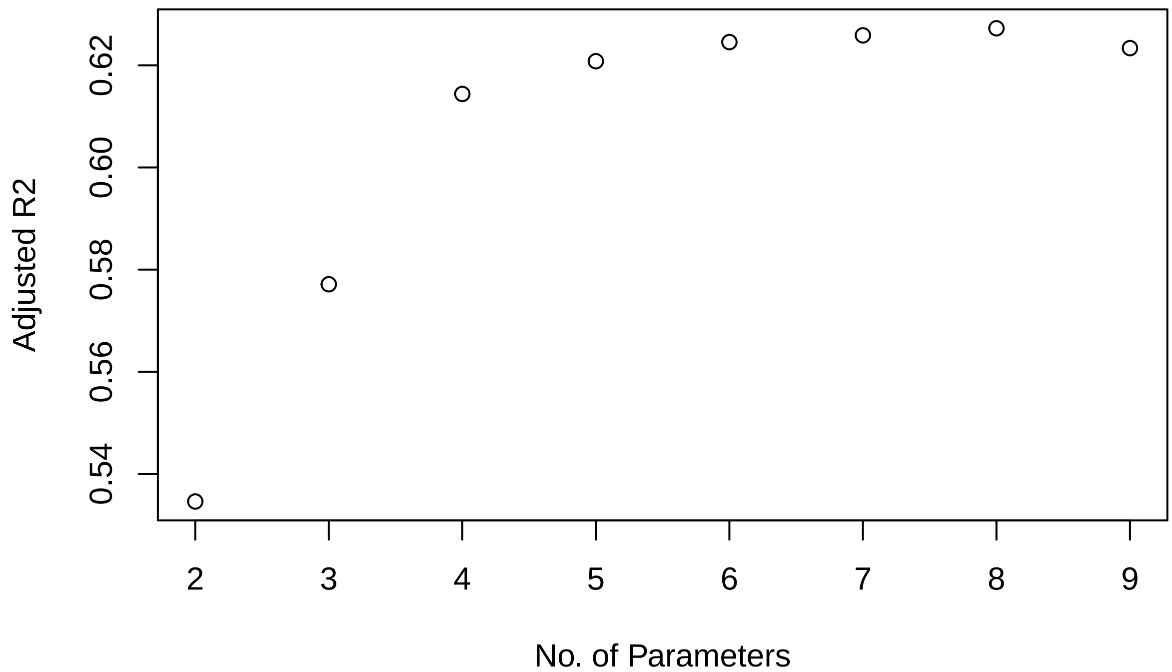
The best model using AIC : lpsa ~ lcavol + lweight + age + lbph + svi

c) Adjusted R2

```r
#Using adjusted R2

adjr2_modelsize = msize[which.max(rs$adjr2)]
adjr2_modelsize
```

```
## [1] 8
```

```r
#plot
plot(msize, rs$adjr2, xlab = "No. of Parameters", ylab = "Adjusted R2")
```



Number of variables selected using Adjusted R2 is 8(including the intercept).

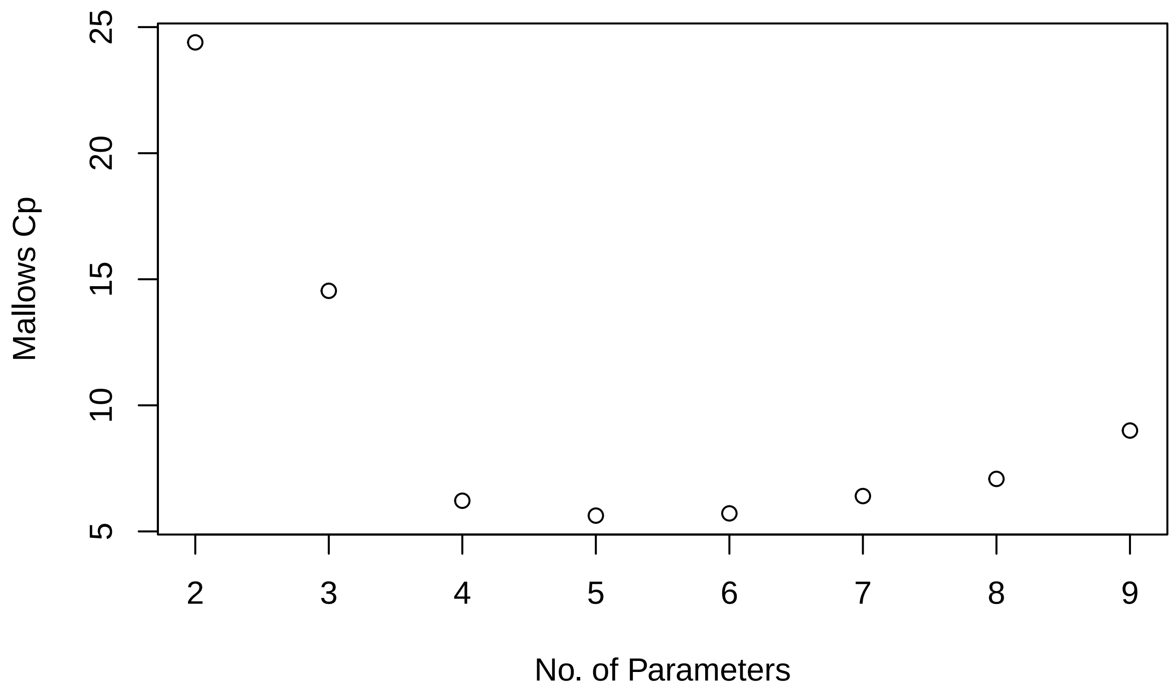
The best model using Adjusted R2 :lpsa ~ lcavol + lweight + age + lbph + svi + lcp + pgg45

(d) Mallows Cp

```r
#Using Mallow's Cp
Cp_modelsize = msize[which.min(rs$cp)]
Cp_modelsize
```

```
## [1] 5
```

```r
#plot
plot(msize, rs$cp, xlab = "No. of Parameters", ylab = "Mallows Cp")
```



The number of variables selected using Mallows Cp is 5(including the intercept)

The best model using Mallows Cp :lpsa ~ lcavol + lweight + lbph + svi

Problem 2

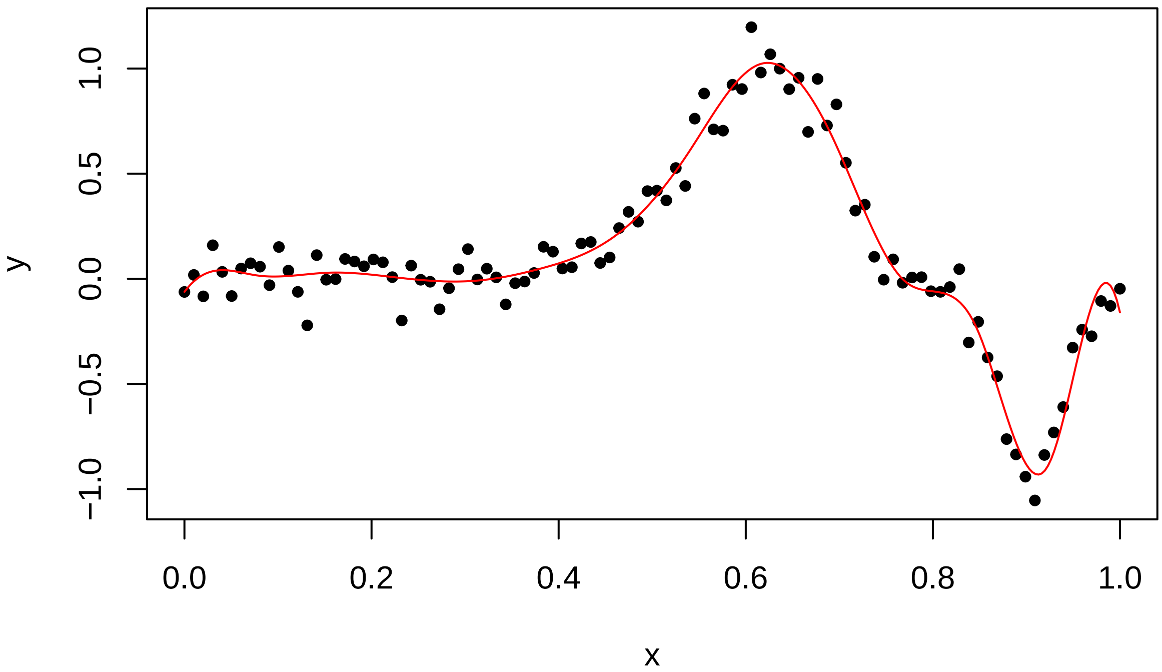
(a) Fit regression splines with 12 evenly-spaced knots using y ~ bs(x; knots = : : :). You need to load the splines package. Display the fit on top of the data.

```
set.seed(1)
library(splines)

my_func = function(x)
  sin(2 * pi * x ^ 3) ^ 3
x = seq(0, 1, length.out = 100)
y = my_func(x) + 0.1 * rnorm(100)

#knots
m = 12

#interval is 0 to 1
myknots = (1:m) / (m + 1)

#the design matrix
F = bs(x, knots = myknots, intercept = TRUE)

#fitting the spline
fit = lm(y ~ F - 1)
summary(fit)
```

```
##
```

```
## Call:
## lm(formula = y ~ F - 1)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.242916 -0.061566  0.004795  0.056278  0.197285
##
## Coefficients:
##       Estimate Std. Error t value Pr(>|t|)
## F1  -0.0620357  0.0837770  -0.740  0.46107
## F2   0.0910676  0.1104052   0.825  0.41179
## F3  -0.0250419  0.1087408  -0.230  0.81843
## F4   0.0495003  0.0866508   0.571  0.56935
## F5   0.0008568  0.0805556   0.011  0.99154
## F6  -0.0273809  0.0788082  -0.347  0.72913
## F7   0.0465322  0.0783079   0.594  0.55396
## F8   0.1608915  0.0781761   2.058  0.04268 *
## F9   0.5569560  0.0781761   7.124 3.32e-10 ***
## F10  1.1983190  0.0783079  15.303  < 2e-16 ***
## F11  0.7767364  0.0788082   9.856 1.13e-15 ***
## F12 -0.2436833  0.0805556  -3.025  0.00330 **
## F13  0.1496560  0.0866508   1.727  0.08782 .
## F14 -1.7192909  0.1087408 -15.811  < 2e-16 ***
## F15  0.3525417  0.1104052   3.193  0.00198 **
## F16 -0.1596822  0.0837770  -1.906  0.06007 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09673 on 84 degrees of freedom
## Multiple R-squared:  0.962,  Adjusted R-squared:  0.9548
## F-statistic:   133 on 16 and 84 DF,  p-value: < 2.2e-16
```

```r
#prediction line on plot
#plot
plot(x, y, type = "p", pch = 20)
lines(spline(x, predict(fit)), col = "red", lty = 1)
```



b) Compute the AIC for this model.

```
#number of observation
n=100

#number of parameters/df
p = dim(model.matrix(fit))[2]

rss = sum(fit$residuals^2)

#AIC
AIC = n*log(rss/n) + 2*p

sprintf("AIC for the model is:%f",AIC)
```

```
## [1] "AIC for the model is:-452.604139"
```

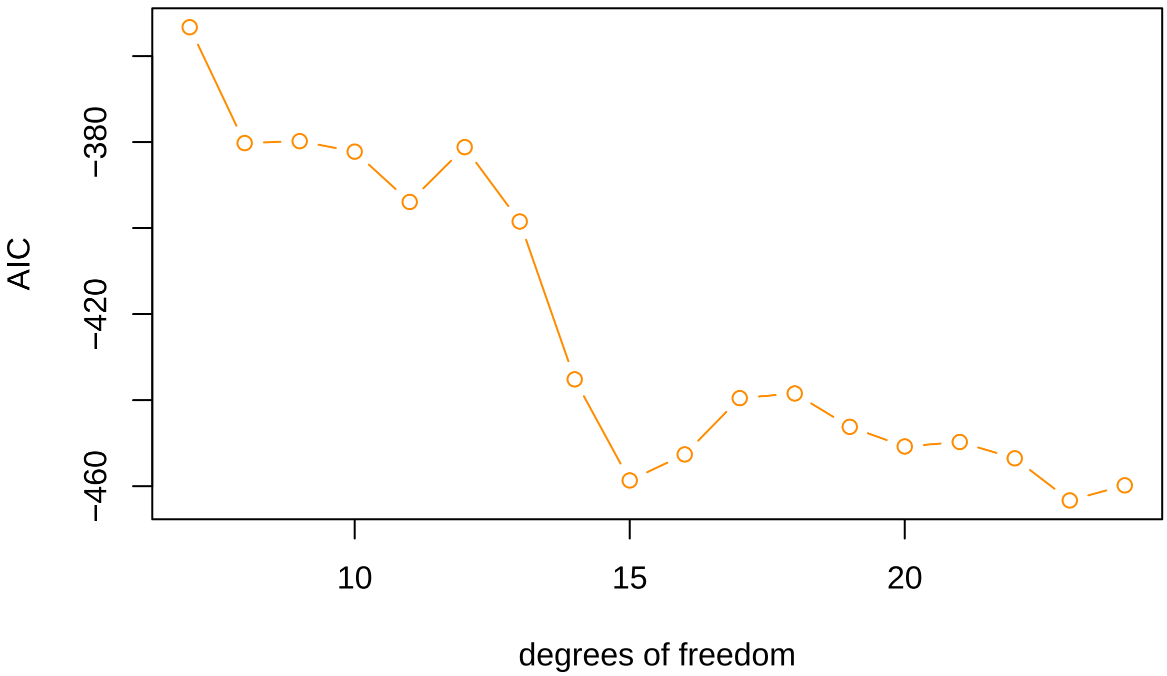
(c) Compute the Adjusted R2

```
adj_r_sq = summary(fit)$adj.r.squared
sprintf("Adjusted R2:%f", round(adj_r_sq, 4))
```

```
## [1] "Adjusted R2:0.954800"
```

d) Compute the AIC for all models with a number of knots between 3 and 20 inclusive. Plot the AIC as a function of the number of degrees of freedom. Which model is the best?

```r
n = length(x)
AIC = c()
df = c()
m = 3:20

#knots
for (i in 1:length(m))
{
  #interval is 0 to 1
  myknots = (1:m[i]) / (m[i] + 1)

  #the design matrix
  F = bs(x, knots = myknots, intercept = TRUE)

  #fitting the spline
  fit = lm(y ~ F - 1)

  #number of parameters/df
  p = dim(model.matrix(fit))[2]
  df[i] = p

  rss = sum(fit$residuals ^ 2)

  #AIC
  aic = n * log(rss / n) + 2 * p
  AIC[i] = aic

}

plot(df,
     AIC,
     type = "b",
     col = "darkorange",
     xlab = "degrees of freedom")
```




```
min_AIC = min(AIC)
best_df = df[which.min(AIC)]
best_m = m[which.min(AIC)]

sprintf(
  "For the best model: AIC= %f; degrees of freedom = %i; number of knots = %i",
  min_AIC,
  best_df,
  best_m
)
```

```
## [1] "For the best model: AIC= -463.297385; degrees of freedom = 23; number of knots = 19"
```

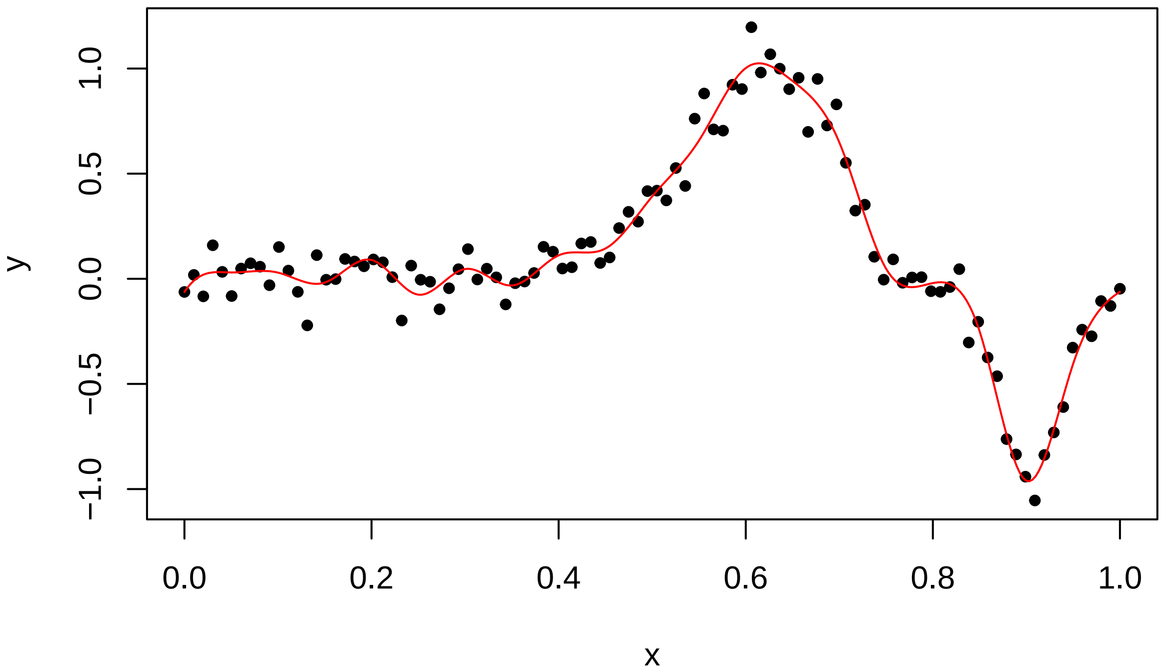
(e) Plot the fit for your selected model on top of the data.

```
#knots
m = 19

#interval is 0 to 1
myknots = (1:m) / (m + 1)

#the design matrix
F = bs(x, knots = myknots, intercept = TRUE)

#fitting the spline
best_fit = lm(y ~ F - 1)
summary(best_fit)
```

```
##
## Call:
```

```
## lm(formula = y ~ F - 1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.20315 -0.05091  0.00974  0.04643  0.18323
##
## Coefficients:
##      Estimate Std. Error t value Pr(>|t|)
## F1  -0.062621   0.084055  -0.745   0.4585
## F2   0.050443   0.122414   0.412   0.6814
## F3   0.011493   0.123277   0.093   0.9260
## F4   0.065175   0.098893   0.659   0.5118
## F5  -0.095631   0.092142  -1.038   0.3026
## F6   0.209614   0.090207   2.324   0.0228 *
## F7  -0.204366   0.089653  -2.280   0.0254 *
## F8   0.154101   0.089496   1.722   0.0891 .
## F9  -0.132422   0.089453  -1.480   0.1429
## F10  0.184422   0.089443   2.062   0.0426 *
## F11  0.064392   0.089440   0.720   0.4737
## F12  0.421085   0.089440   4.708 1.08e-05 ***
## F13  0.601223   0.089440   6.722 2.78e-09 ***
## F14  1.118232   0.089443  12.502  < 2e-16 ***
## F15  0.939373   0.089453  10.501  < 2e-16 ***
## F16  0.769532   0.089496   8.598 7.13e-13 ***
## F17 -0.122735   0.089653  -1.369   0.1750
## F18  0.008672   0.090207   0.096   0.9237
## F19 -0.034325   0.092142  -0.373   0.7105
## F20 -1.369069   0.098893 -13.844  < 2e-16 ***
## F21 -0.253911   0.123277  -2.060   0.0428 *
## F22 -0.109483   0.122414  -0.894   0.3739
## F23 -0.061554   0.084055  -0.732   0.4662
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0893 on 77 degrees of freedom
## Multiple R-squared:  0.9703, Adjusted R-squared:  0.9615
## F-statistic: 109.5 on 23 and 77 DF,  p-value: < 2.2e-16
```

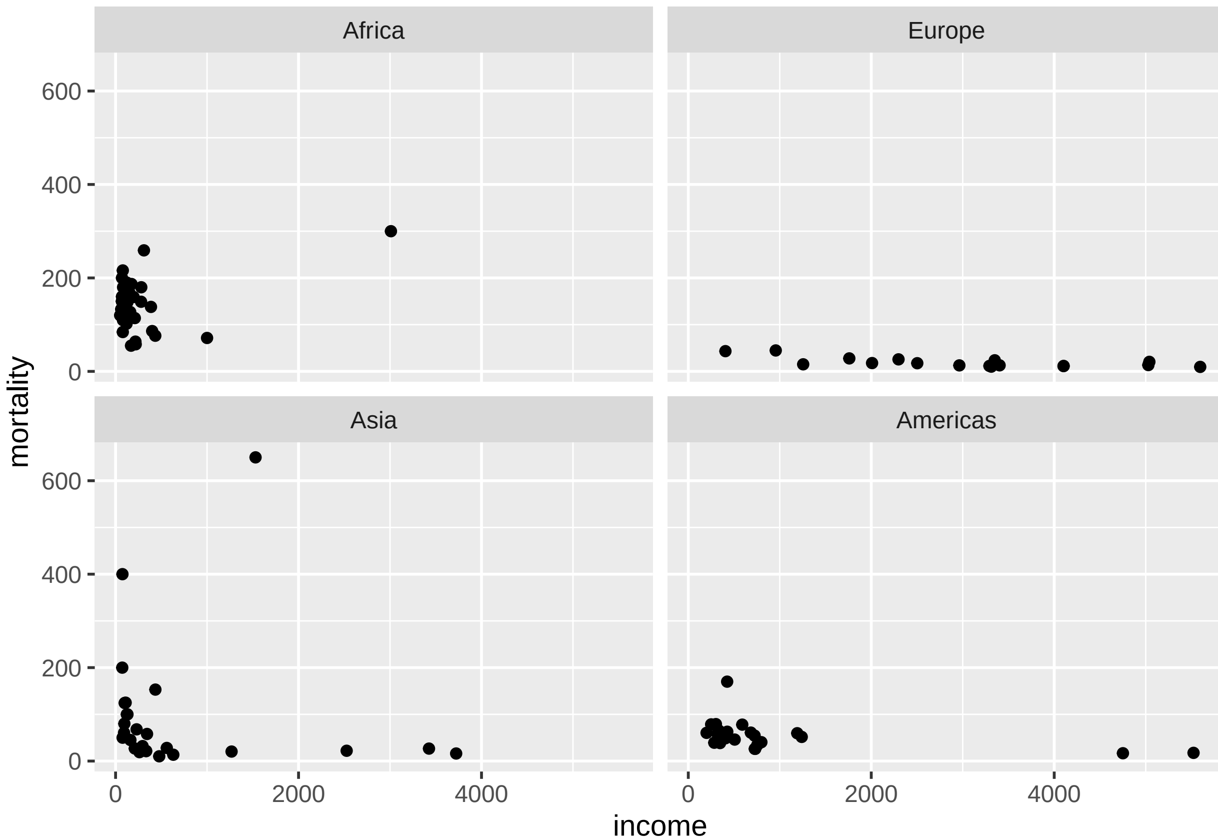
```
#plot
plot(x, y, type = "p", pch = 20)
lines(spline(x, predict(best_fit)), col = "red", lty = 1)
```



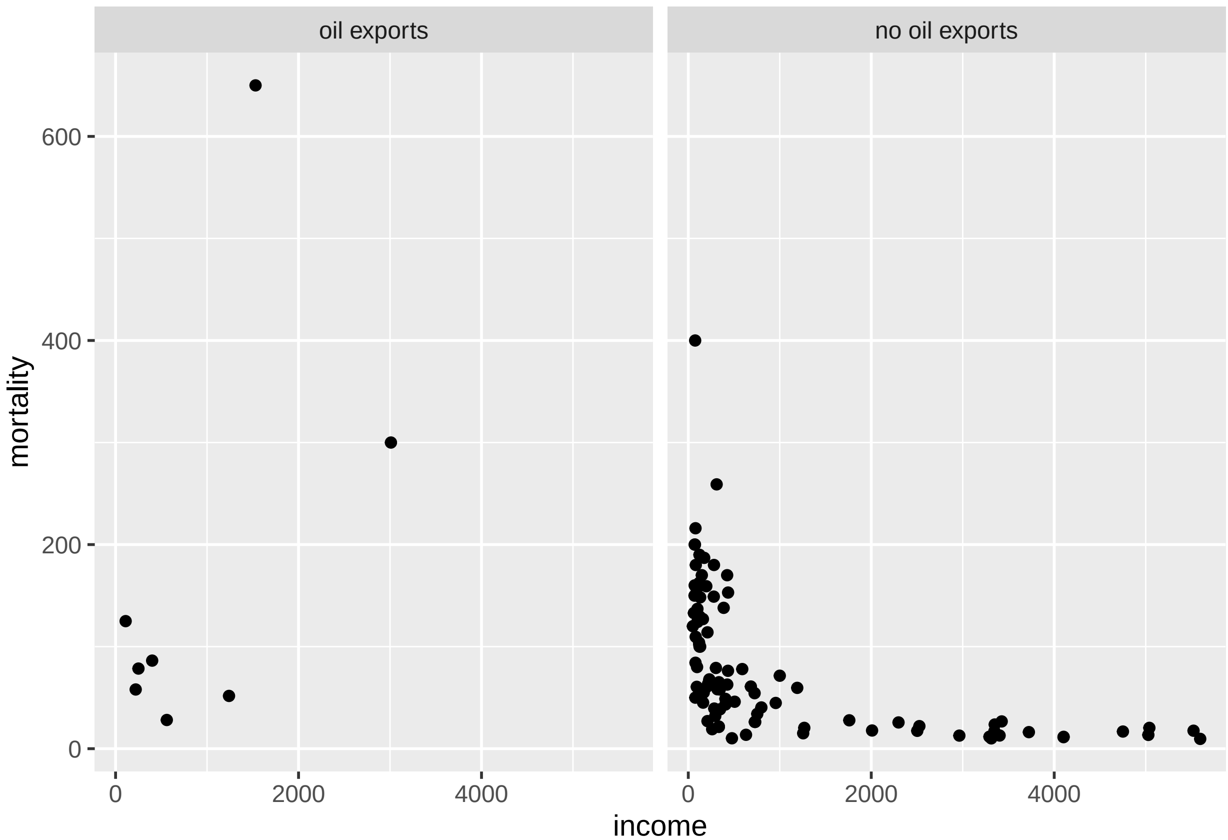
4. Problem 3GR: Using the infmort data, find a model for the infant mortality in terms of the other variables. Consider region and oil as categorical predictors.

(a) Plot the data and comment on your results.

```
library(faraway)
data("infmort")

library(ggplot2)
#by region
ggplot(aes(x = income, y = mortality), data = infmort) + geom_point() +
  facet_wrap( ~ region)
```




```r
#by oil
ggplot(aes(x = income, y = mortality), data = infmort) + geom_point() +
  facet_wrap( ~ oil)
```



From the plot we can make the following observations:

1. Most of the countries in Africa fall under low income compared to other regions and mortality is generally high
2. Low income countries of Asia, on average, have high mortality as compared to high income countries.
3. In Europe, mortality rates are almost the same in all the countries irrespective of their PCI per capita income
4. In Americas, most of the countries are towards the low income. Mortality,on average, is slightly high for those countries with respect to the 2 other countries which have high per capita income.
5. For no oil exports countries the mortality rates are high when the per capita income is low and the mortality rates are low if the income is high.
6. For the countries which export oil, not enough to make an observation.**

(b) Fit a full model considering all potential interactions between the continuous and cate- gorical predictors. Comment on your results.

```
#checking the class of region and oil predictors
class(infmort$oil)
```

```
## [1] "factor"
```

```r
class(infmort$region)
```

```
## [1] "factor"
```

```r
#fitting the full model
full_model = lm(mortality ~  income*oil*region, data = infmort)
summary(full_model)
```

```
##
## Call:
## lm(formula = mortality ~ income * oil * region, data = infmort)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -168.006  -21.666   -2.137   11.719  310.277
##
## Coefficients: (2 not defined because of singularities)
##                                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)                              45.69666   44.57686   1.025 0.308148
## income                                    0.08463    0.02536   3.337 0.001247
## oilno oil exports                       108.73976   46.86716   2.320 0.022672
## regionEurope                           -119.78192   34.73945  -3.448 0.000872
## regionAsia                              -80.89596   68.07225  -1.188 0.237916
## regionAmericas                           39.57102   84.30492   0.469 0.639973
## income:oilno oil exports                 -0.15423    0.06238  -2.473 0.015361
## income:regionEurope                       0.06453    0.05776   1.117 0.266992
## income:regionAsia                         0.32841    0.06017   5.458 4.48e-07
## income:regionAmericas                    -0.11170    0.08392  -1.331 0.186666
## oilno oil exports:regionEurope                 NA         NA      NA       NA
## oilno oil exports:regionAsia             18.14860   70.89640   0.256 0.798565
## oilno oil exports:regionAmericas       -130.48532   86.91040  -1.501 0.136879
## income:oilno oil exports:regionEurope          NA         NA      NA       NA
## income:oilno oil exports:regionAsia      -0.28503    0.08363  -3.408 0.000992
## income:oilno oil exports:regionAmericas   0.17192    0.10184   1.688 0.094959
##
## (Intercept)
## income                                 **
## oilno oil exports                      *
## regionEurope                           ***
## regionAsia
## regionAmericas
## income:oilno oil exports               *
## income:regionEurope
## income:regionAsia                      ***
## income:regionAmericas
## oilno oil exports:regionEurope
## oilno oil exports:regionAsia
## oilno oil exports:regionAmericas
## income:oilno oil exports:regionEurope
## income:oilno oil exports:regionAsia    ***
## income:oilno oil exports:regionAmericas .
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 56 on 87 degrees of freedom
##   (4 observations deleted due to missingness)
## Multiple R-squared:  0.6691, Adjusted R-squared:  0.6196
## F-statistic: 13.53 on 13 and 87 DF,  p-value: 7.929e-16
```

From the model summary results, we can infer the following:

-The above predictors explains 67% of the variance in infant mortality rate

-Income is statistically significant in predicting the infant mortality rate (p-value < 0.05)

-Change in effect of income on mortality for Asia wrt Africa is significant but not for Europe or Americas for countries with oil exports

-Change in effect of income on mortality for no oil export countries wrt oil export countries is significant in Africa

-The difference between mean mortality rate of europe and africa is signficant, assuming income and oil status remain same

-The difference between mean mortality rate of no oil export countries w.r.t oil export countries is significant, assuming income and region stays same.

(c) Use a sequential ANOVA to determine the best model.

```
# Sequential ANOVA model
anova(full_model)
```

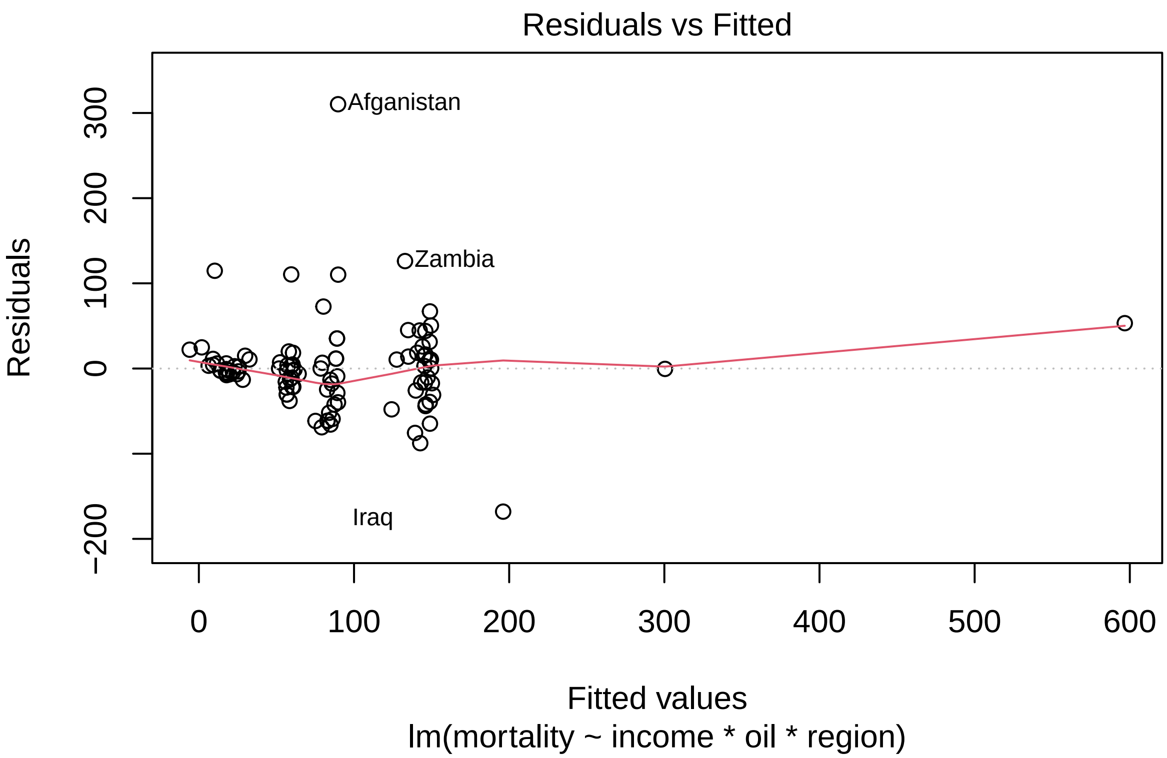
```
## Analysis of Variance Table
##
## Response: mortality
##                   Df Sum Sq Mean Sq F value    Pr(>F)
## income             1  90086   90086 28.7241 6.760e-07 ***
## oil                1  56902   56902 18.1432 5.166e-05 ***
## region             3 109016   36339 11.5866 1.839e-06 ***
## income:oil         1  96435   96435 30.7484 3.106e-07 ***
## income:region      3  74531   24844  7.9214 9.903e-05 ***
## oil:region         2  46491   23245  7.4118  0.001066 **
## income:oil:region  2  78180   39090 12.4638 1.739e-05 ***
## Residuals         87 272855    3136
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

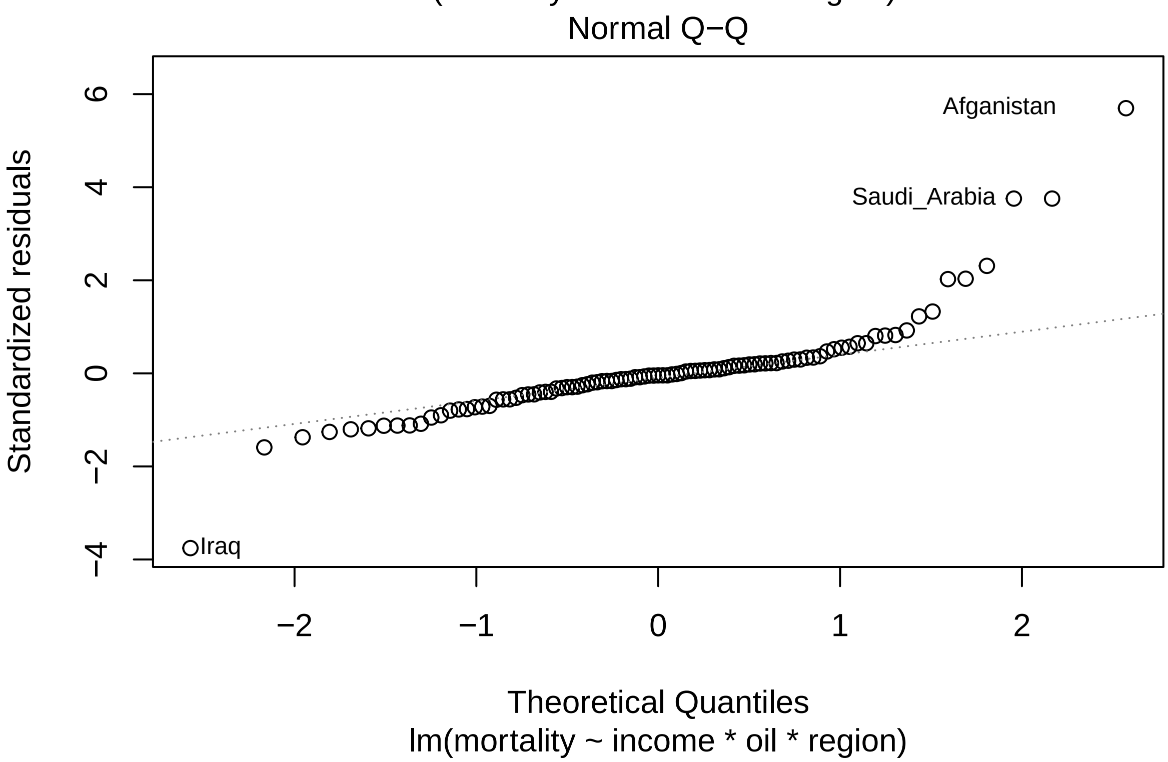
From sequential ANOVA, we see that all terms are significant. So our full model is the best model

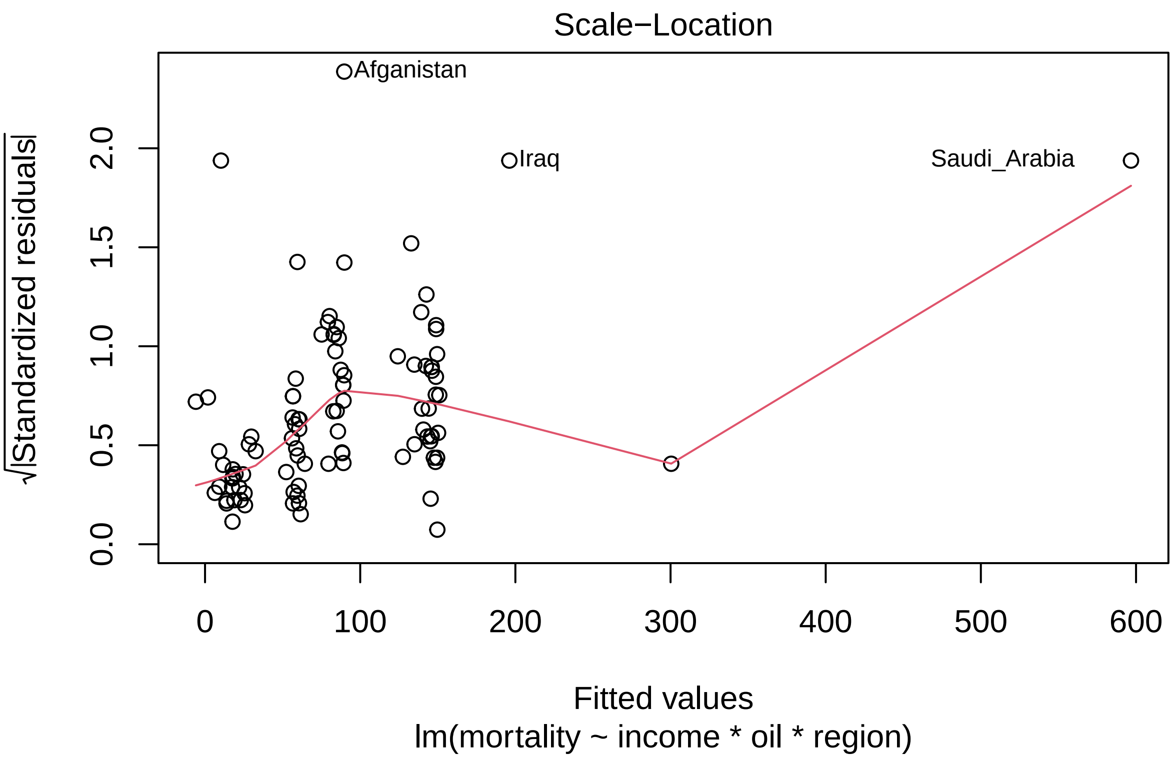
(d) Check the assumptions of your model using appropriate diagnostics. Be alert for trans- formations and/or unusual points and make adjustments if necessary.

```
#checking model dianostic
plot(full_model)
```

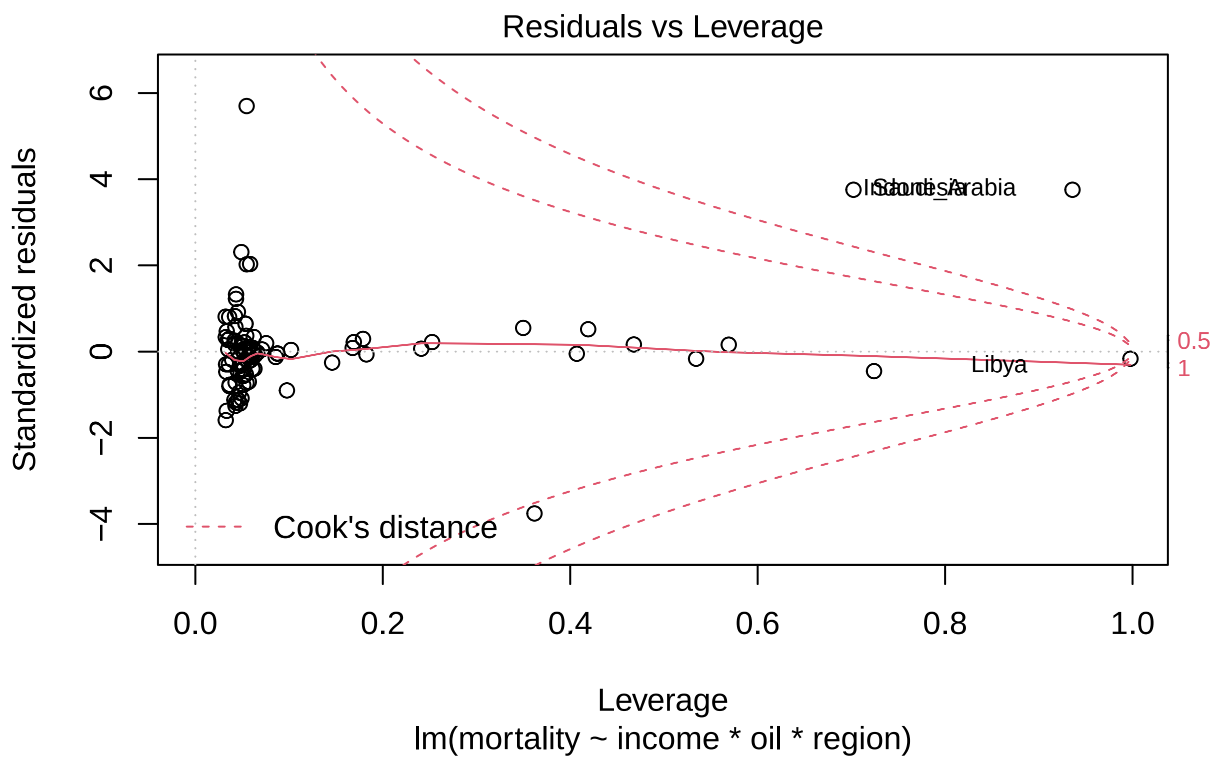
```
## Warning: not plotting observations with leverage one:
##   22, 28
```


Residuals vs Fitted
Afganistan
Zambia
Iraq
Residuals
300
200
100
0
−200
0
100
200
300
400
500
600
Fitted values
lm(mortality ~ income * oil * region)


Normal Q−Q
Afganistan
Saudi_Arabia
Iraq
Standardized residuals
6
4
2
0
−2
−4
−2
−1
0
1
2
Theoretical Quantiles
lm(mortality ~ income * oil * region)




```
## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced

## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced
```



From the above plots, we can make the following observations:

Plot-1: From residual vs fitted plot, the mean line is around zero and points seem randomly distributed. Thus, it is safe to assume that assumption of linearity is not being violated

Plot-2: From the Q-Q plot, it seems like the assumption of normality of errors is not being violated. However, we will do a Shapiro Wills test to confirm this

Plot-3: It seems like the assumption of equal variance for errors is being violated (fan shaped). However, we will do B-P test to confirm this

Plot-4: As per the cook's distance plot,'Indonesia', 'Libya' and 'Saudi Arabia' are highly influential observations

```r
library(lmtest)
```

```
## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```r
#checking constant variance assumption using BP test
bptest(full_model)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  full_model
## BP = 12.714, df = 13, p-value = 0.4701
```

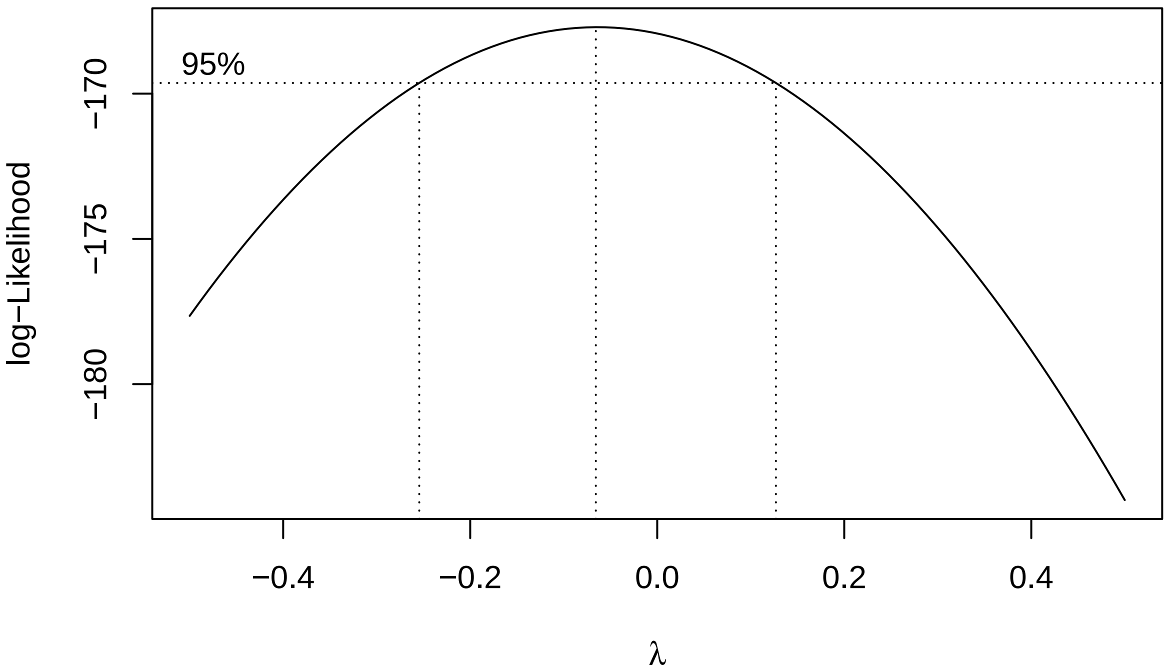
```r
#checking normality assumption using Shapiro-Wilks test
shapiro.test(residuals(full_model))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(full_model)
## W = 0.81402, p-value = 5.968e-10
```

From the studentized Breusch-Pagan test result, we can infer that the assumption of constant variance is valid.

From the Shapiro-Wilk normality test, we can infer that the assumption of normality is not valid. Thus, we might have to check for box-cox transformation

```r
#checking for boxcox
library("MASS")

boxcox(full_model,plotit = TRUE,lambda = seq(-0.5, 0.5, by = 0.1))
```



From Box-Cox transformation, we can see that lambda = 1 is not included in confidence region and thus, there is a need to do transformation for the response. We can conclude a log transformation can be done on the response variable(lamda_max ~ 0)

(e) Interpret your final model by explaining what the regression parameter estimates mean.

```r
full_model_log = lm(log(mortality) ~  income*oil*region, data = infmort)
summary(full_model_log)
```

```
##
## Call:
## lm(formula = log(mortality) ~ income * oil * region, data = infmort)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.60539 -0.24595  0.03789  0.23433  1.90154
##
## Coefficients: (2 not defined because of singularities)
##                                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)                        4.0865456  0.4500144   9.081 3.07e-14
## income                             0.0005406  0.0002560   2.112   0.0376
## oilno oil exports                  0.9322285  0.4731355   1.970   0.0520
## regionEurope                      -1.4927017  0.3507033  -4.256 5.23e-05
## regionAsia                        -0.3007901  0.6872062  -0.438   0.6627
## regionAmericas                     0.3820179  0.8510790   0.449   0.6546
## income:oilno oil exports          -0.0012175  0.0006297  -1.933   0.0564
## income:regionEurope                0.0004533  0.0005831   0.777   0.4391
## income:regionAsia                  0.0009520  0.0006075   1.567   0.1207
## income:regionAmericas             -0.0009625  0.0008472  -1.136   0.2591
## oilno oil exports:regionEurope            NA         NA      NA       NA
## oilno oil exports:regionAsia      -0.5978126  0.7157166  -0.835   0.4059
## oilno oil exports:regionAmericas  -1.3016049  0.8773819  -1.484   0.1416
```

```
## income:oilno oil exports:regionEurope           NA         NA      NA       NA
## income:oilno oil exports:regionAsia     -0.0006785  0.0008442  -0.804   0.4238
## income:oilno oil exports:regionAmericas  0.0013919  0.0010281   1.354   0.1793
##
## (Intercept)                             ***
## income                                  *
## oilno oil exports                       .
## regionEurope                            ***
## regionAsia
## regionAmericas
## income:oilno oil exports                .
## income:regionEurope
## income:regionAsia
## income:regionAmericas
## oilno oil exports:regionEurope
## oilno oil exports:regionAsia
## oilno oil exports:regionAmericas
## income:oilno oil exports:regionEurope
## income:oilno oil exports:regionAsia
## income:oilno oil exports:regionAmericas
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5654 on 87 degrees of freedom
##   (4 observations deleted due to missingness)
## Multiple R-squared:  0.7034, Adjusted R-squared:  0.6591
## F-statistic: 15.87 on 13 and 87 DF,  p-value: < 2.2e-16
```

```r
anova(full_model_log)
```

```
## Analysis of Variance Table
##
## Response: log(mortality)
##                   Df Sum Sq Mean Sq  F value    Pr(>F)
## income             1 38.172  38.172 119.4266 < 2.2e-16 ***
## oil                1  2.561   2.561   8.0129 0.0057676 **
## region             3 16.858   5.619  17.5812 5.234e-09 ***
## income:oil         1  3.961   3.961  12.3924 0.0006885 ***
## income:region      3  1.571   0.524   1.6382 0.1864055
## oil:region         2  1.434   0.717   2.2440 0.1121393
## income:oil:region  2  1.403   0.701   2.1946 0.1175318
## Residuals         87 27.808   0.320
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
#final model

final_model = lm(log(mortality) ~  income + oil + region +income*oil, data = infmort)
summary(final_model)
```

```
##
## Call:
## lm(formula = log(mortality) ~ income + oil + region + income *
```

```
##     oil, data = infmort)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.66297 -0.31903  0.03669  0.30312  1.89006
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)              4.6675511  0.3080863  15.150  < 2e-16 ***
## income                   0.0005035  0.0002246   2.242 0.027314 *
## oilno oil exports        0.2284596  0.3019431   0.757 0.451163
## regionEurope            -1.1722446  0.2377605  -4.930 3.52e-06 ***
## regionAsia              -0.7729598  0.1533491  -5.041 2.25e-06 ***
## regionAmericas          -0.7676816  0.1662493  -4.618 1.23e-05 ***
## income:oilno oil exports -0.0007921  0.0002330  -3.400 0.000992 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5854 on 94 degrees of freedom
##   (4 observations deleted due to missingness)
## Multiple R-squared:  0.6564, Adjusted R-squared:  0.6345
## F-statistic: 29.93 on 6 and 94 DF,  p-value: < 2.2e-16
```

```
anova(final_model)
```

```
## Analysis of Variance Table
##
## Response: log(mortality)
##            Df Sum Sq Mean Sq F value    Pr(>F)
## income      1 38.172  38.172 111.379 < 2.2e-16 ***
## oil         1  2.561   2.561   7.473 0.0074844 **
## region      3 16.858   5.619  16.396 1.195e-08 ***
## income:oil  1  3.961   3.961  11.557 0.0009917 ***
## Residuals  94 32.216   0.343
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Meaning of regression parameter estimates:

1. The parameter estimate of `income` is the effect of income on log of mortality for African countries with oil exports
2. The parameter estimate of `oilno oil exports` represents change in expected log (mortality) on no oil export countries wrt to oil export countries in Africa
3. The parameter estimate of `regionEurope` represents change in effect of income on log(mortality) for Europe wrt to Africa in countries with oil exports
4. The parameter estimate of `regionAsia` represents change in effect of income on log(mortality) for Asia wrt to Africa in countries with oil exports
5. The parameter estimate of `regionAmericas` represents change in effect of income on log(mortality) for Americas wrt to Africa in countries with oil exports
6. The parameter estimate of interactive term `income:oilno oil exports` represents change in effect of income on log(mortality) for no oil countries wrt to oil export countries in Africa