STAT 359 - Assignment 5

Isabella Pelletier

12/1/2020

Question 1.

Find a model that best describes beef consumption in the United States. Complete a full analysis of the data (initial plots, model selection, residual plots etc). Discuss your results

```
beef = read.table(file="~/Desktop/R/beef.txt", header=TRUE, sep="")
attach(beef)
library(knitr)
kable(beef, caption = 'Beef consumption data', align='c')
```

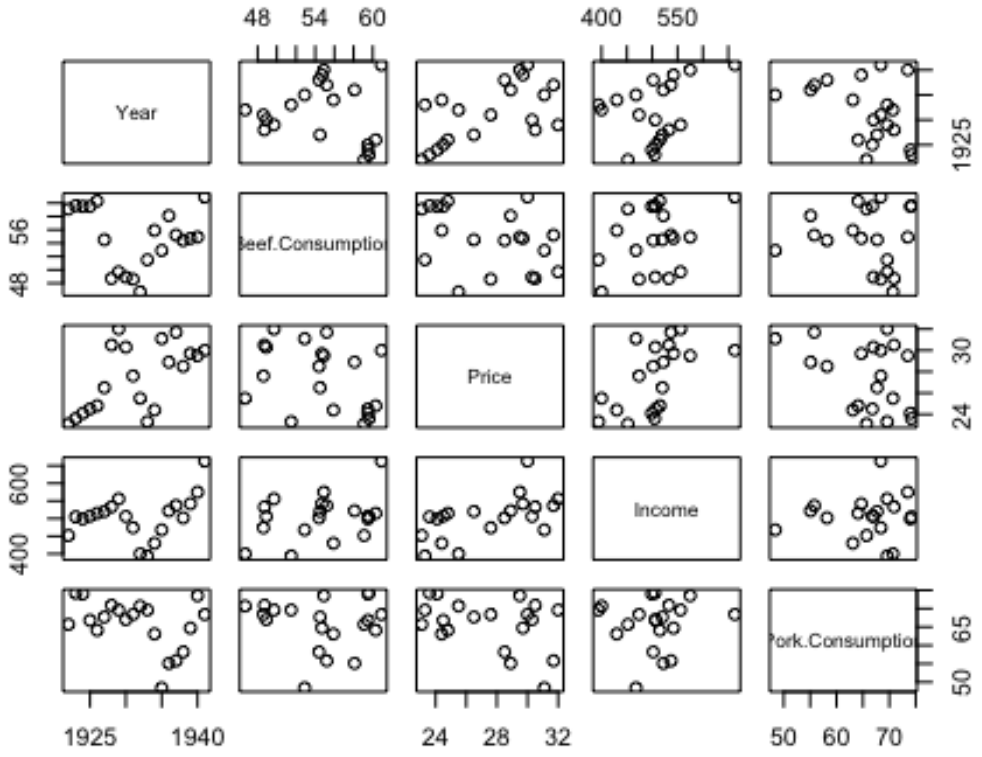
Beef consumption data

Year	Beef.Consumption	Price	Income	Pork.Consumption
1922	59.1	23.1	452	65.7
1923	59.6	23.6	505	74.2
1924	59.5	24.1	499	74.0
1925	59.5	24.5	507	66.8
1926	60.3	24.8	515	64.1
1927	54.5	26.5	520	67.7
1928	48.7	30.5	533	70.9
1929	49.7	32.0	556	69.6
1930	48.9	30.3	506	67.0
1931	48.6	27.6	474	68.4
1932	46.7	25.5	400	70.7
1933	51.5	23.3	394	69.6
1934	55.9	24.4	430	63.1
1935	52.9	31.1	468	48.4
1936	58.1	28.9	522	55.1
1937	55.2	31.7	537	55.8

1938	54.4	28.5	502	58.2
1939	54.7	29.7	542	64.7
1940	54.9	29.5	575	73.5
1941	60.9	30.0	663	68.4

- Null Hypothesis: There is no association between Beef Consumption and any of Year, Price, Income, or Pork Consumption.
- Alternative Hypothesis: There is an association with Beef Consumption and at least one of Year, Price, Income, or Pork Consumption.

```
pairs(beef)
```



- From the pairs plot we can see that Beef consumption may have a positive relationship with Income as well as a positive relationship with Pork consumption. There also may be a negative relationship between Beef consumption and Price. The relationship between Beef Consumption and Year will have to be looked into further to draw more information.

- We will start with a model including 4 main effects, 6 2-way interactions (Year x Price, Year x Income, Year x Pork, Price x Income, Price x Pork, Income x Pork), 4 3-way interactions (Year x Price x Income, Year x Price x Pork, Price x Income x Pork, Year x Income, Pork), 1 4-way interaction (Year x Price x Income x Pork), and an intercept leading to 16 parameters. This is more than 20/3 parameters so we will have to look out for over-parameterization.

```
beef.m1=lm(Beef.Consumption~Year*Price*Income*Pork.Consumption)
summary(beef.m1)

##
## Call:
## lm(formula = Beef.Consumption ~ Year * Price * Income * Pork.Consumption)
##
## Residuals:
##         1         2         3         4         5         6         7
8
## -0.007943 -0.487122  0.652720  0.418527 -0.088981 -0.772595  0.969648 -0.7
84458
##         9        10        11        12        13        14        15
16
##  0.530981 -0.717858 -0.370838  0.327339  0.046011 -0.009858  0.612594 -0.0
69940
##        17        18        19        20
## -0.529615  0.213412  0.258054 -0.190079
##
## Coefficients:
##                                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)                         -1.390e+06  9.934e+05  -1.399    0.234
## Year                                 7.188e+02  5.136e+02   1.400    0.234
## Price                                4.921e+04  3.631e+04   1.355    0.247
## Income                               2.641e+03  1.932e+03   1.367    0.243
## Pork.Consumption                     2.118e+04  1.515e+04   1.398    0.235
## Year:Price                          -2.545e+01  1.877e+01  -1.356    0.247
## Year:Income                         -1.366e+00  9.990e-01  -1.367    0.243
## Price:Income                        -9.310e+01  7.038e+01  -1.323    0.256
## Year:Pork.Consumption               -1.095e+01  7.833e+00  -1.398    0.235
## Price:Pork.Consumption              -7.505e+02  5.521e+02  -1.359    0.246
## Income:Pork.Consumption             -4.016e+01  2.955e+01  -1.359    0.246
## Year:Price:Income                    4.815e-02  3.639e-02   1.323    0.256
## Year:Price:Pork.Consumption          3.881e-01  2.854e-01   1.360    0.245
## Year:Income:Pork.Consumption         2.077e-02  1.528e-02   1.360    0.246
## Price:Income:Pork.Consumption        1.417e+00  1.073e+00   1.321    0.257
## Year:Price:Income:Pork.Consumption -7.328e-04  5.546e-04  -1.321    0.257
##
## Residual standard error: 1.101 on 4 degrees of freedom
## Multiple R-squared:  0.9873, Adjusted R-squared:  0.9397
## F-statistic: 20.73 on 15 and 4 DF,  p-value: 0.004867
```

- The initial model has an R-squared = 0.9873 and the 4-way interaction is not significant (p-value = 0.257).

```
beef.m2 = update(beef.m1,.~.-Year:Price:Income:Pork.Consumption)
summary(beef.m2)

##
## Call:
## lm(formula = Beef.Consumption ~ Year + Price + Income + Pork.Consumption +
##     Year:Price + Year:Income + Price:Income + Year:Pork.Consumption +
##     Price:Pork.Consumption + Income:Pork.Consumption + Year:Price:Income +
##     Year:Price:Pork.Consumption + Year:Income:Pork.Consumption +
##     Price:Income:Pork.Consumption)
##
## Residuals:
##       1        2        3        4        5        6        7        8
## -0.10453 -1.00229  1.40724  0.21346  0.26262 -1.00370  0.64146 -0.40309
##       9       10       11       12       13       14       15       16
##  0.21089 -0.54202 -0.14163  0.01063  0.28906  0.09799  0.62722 -0.16072
##      17       18       19       20
## -0.96792  0.66468  0.05928 -0.15863
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    -8.043e+04  7.686e+04  -1.046    0.343
## Year                            4.195e+01  3.980e+01   1.054    0.340
## Price                           1.278e+03  1.855e+03   0.689    0.522
## Income                          9.471e+01  1.535e+02   0.617    0.564
## Pork.Consumption                1.197e+03  1.019e+03   1.174    0.293
## Year:Price                     -6.723e-01  9.550e-01  -0.704    0.513
## Year:Income                    -4.951e-02  7.948e-02  -0.623    0.561
## Price:Income                   -1.435e-01  2.417e+00  -0.059    0.955
## Year:Pork.Consumption          -6.226e-01  5.285e-01  -1.178    0.292
## Price:Pork.Consumption         -2.142e+01  1.980e+01  -1.082    0.329
## Income:Pork.Consumption        -1.201e+00  2.067e+00  -0.581    0.587
## Year:Price:Income               9.417e-05  1.237e-03   0.076    0.942
## Year:Price:Pork.Consumption     1.120e-02  1.020e-02   1.099    0.322
## Year:Income:Pork.Consumption    6.274e-04  1.073e-03   0.585    0.584
## Price:Income:Pork.Consumption -4.261e-04  7.306e-04  -0.583    0.585
##
## Residual standard error: 1.181 on 5 degrees of freedom
## Multiple R-squared:  0.9818, Adjusted R-squared:  0.9307
## F-statistic: 19.22 on 14 and 5 DF,  p-value: 0.002073
```

- Year x Price x Income is the least significant 3-way interaction.

```
beef.m3 = update(beef.m2,.~.-Year:Price:Income)
summary(beef.m3)
```

```
##
## Call:
## lm(formula = Beef.Consumption ~ Year + Price + Income + Pork.Consumption +
##     Year:Price + Year:Income + Price:Income + Year:Pork.Consumption +
##     Price:Pork.Consumption + Income:Pork.Consumption + Year:Price:Pork.Con
sumption +
##     Year:Income:Pork.Consumption + Price:Income:Pork.Consumption)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.04707 -0.22057  0.01727  0.29223  1.38399
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   -7.764e+04  6.168e+04  -1.259    0.255
## Year                           4.051e+01  3.201e+01   1.265    0.253
## Price                          1.176e+03  1.176e+03   1.000    0.356
## Income                         8.983e+01  1.274e+02   0.705    0.507
## Pork.Consumption               1.189e+03  9.263e+02   1.283    0.247
## Year:Price                    -6.201e-01  6.071e-01  -1.021    0.347
## Year:Income                   -4.701e-02  6.612e-02  -0.711    0.504
## Price:Income                   4.045e-02  3.754e-02   1.078    0.323
## Year:Pork.Consumption         -6.188e-01  4.805e-01  -1.288    0.245
## Price:Pork.Consumption        -2.108e+01  1.760e+01  -1.197    0.276
## Income:Pork.Consumption       -1.207e+00  1.887e+00  -0.639    0.546
## Year:Price:Pork.Consumption    1.103e-02  9.081e-03   1.215    0.270
## Year:Income:Pork.Consumption   6.308e-04  9.788e-04   0.644    0.543
## Price:Income:Pork.Consumption -4.576e-04  5.505e-04  -0.831    0.438
##
## Residual standard error: 1.078 on 6 degrees of freedom
## Multiple R-squared:  0.9817, Adjusted R-squared:  0.9422
## F-statistic:  24.8 on 13 and 6 DF,  p-value: 0.0003903
```

- Next to be removed is Year x Income x Pork.Consumption.

```
beef.m4 = update(beef.m3,.~.-Year:Income:Pork.Consumption)
summary(beef.m4)
```

```
##
## Call:
## lm(formula = Beef.Consumption ~ Year + Price + Income + Pork.Consumption +
##     Year:Price + Year:Income + Price:Income + Year:Pork.Consumption +
##     Price:Pork.Consumption + Income:Pork.Consumption + Year:Price:Pork.Con
sumption +
##     Price:Income:Pork.Consumption)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.16912 -0.21105  0.07984  0.26544  1.54938
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   -4.241e+04  2.736e+04  -1.550   0.1651
## Year                           2.221e+01  1.415e+01   1.570   0.1605
## Price                          1.456e+03  1.047e+03   1.391   0.2067
## Income                         7.747e+00  3.829e+00   2.023   0.0827 .
## Pork.Consumption               6.599e+02  4.114e+02   1.604   0.1527
## Year:Price                    -7.616e-01  5.419e-01  -1.406   0.2027
## Year:Income                   -4.421e-03  1.844e-03  -2.398   0.0476 *
## Price:Income                   2.923e-02  3.184e-02   0.918   0.3891
## Year:Pork.Consumption         -3.441e-01  2.126e-01  -1.619   0.1495
## Price:Pork.Consumption        -2.503e+01  1.580e+01  -1.584   0.1571
## Income:Pork.Consumption        9.429e-03  1.392e-02   0.678   0.5198
## Year:Price:Pork.Consumption    1.303e-02  8.172e-03   1.594   0.1549
## Price:Income:Pork.Consumption -3.063e-04  4.768e-04  -0.643   0.5410
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.032 on 7 degrees of freedom
## Multiple R-squared:  0.9805, Adjusted R-squared:  0.947
## F-statistic: 29.28 on 12 and 7 DF,  p-value: 8.245e-05
```

- Next to be removed is Price x Income x Pork.Consumption.

```
beef.m5 = update(beef.m4,.~.-Price:Income:Pork.Consumption)
summary(beef.m5)

##
## Call:
## lm(formula = Beef.Consumption ~ Year + Price + Income + Pork.Consumption +
##     Year:Price + Year:Income + Price:Income + Year:Pork.Consumption +
##     Price:Pork.Consumption + Income:Pork.Consumption + Year:Price:Pork.Con
sumption)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.08161 -0.34174  0.05013  0.31196  1.56990
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   -4.331e+04  2.630e+04  -1.646   0.1383
## Year                           2.253e+01  1.361e+01   1.655   0.1364
## Price                          1.461e+03  1.007e+03   1.450   0.1851
## Income                         8.839e+00  3.303e+00   2.676   0.0281 *
## Pork.Consumption               6.785e+02  3.950e+02   1.718   0.1242
```

```
## Year:Price                  -7.590e-01  5.216e-01  -1.455   0.1837
## Year:Income                 -4.680e-03  1.732e-03  -2.702   0.0270 *
## Price:Income                 8.868e-03  2.907e-03   3.050   0.0158 *
## Year:Pork.Consumption       -3.516e-01  2.043e-01  -1.721   0.1235
## Price:Pork.Consumption      -2.543e+01  1.519e+01  -1.673   0.1328
## Income:Pork.Consumption      5.298e-04  1.300e-03   0.408   0.6943
## Year:Price:Pork.Consumption  1.316e-02  7.864e-03   1.674   0.1327
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9938 on 8 degrees of freedom
## Multiple R-squared:  0.9793, Adjusted R-squared:  0.9509
## F-statistic: 34.43 on 11 and 8 DF,  p-value: 1.613e-05
```

- Next we remove the last 3-way interaction Year x Price x Pork.Consumption.

```
beef.m6 = update(beef.m5,.~.-Year:Price:Pork.Consumption)
summary(beef.m6)

##
## Call:
## lm(formula = Beef.Consumption ~ Year + Price + Income + Pork.Consumption +
##     Year:Price + Year:Income + Price:Income + Year:Pork.Consumption +
##     Price:Pork.Consumption + Income:Pork.Consumption)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.30638 -0.28445 -0.09464  0.30663  1.49017
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              6.142e+02  1.989e+03   0.309   0.7645
## Year                    -1.948e-01  1.040e+00  -0.187   0.8556
## Price                   -2.214e+02  7.703e+01  -2.874   0.0184 *
## Income                   7.875e+00  3.563e+00   2.210   0.0544 .
## Pork.Consumption         1.872e+01  2.785e+01   0.672   0.5184
## Year:Price               1.119e-01  3.920e-02   2.856   0.0189 *
## Year:Income             -4.188e-03  1.870e-03  -2.240   0.0518 .
## Price:Income             6.273e-03  2.694e-03   2.328   0.0449 *
## Year:Pork.Consumption   -1.043e-02  1.473e-02  -0.708   0.4967
## Price:Pork.Consumption   5.501e-03  2.616e-02   0.210   0.8381
## Income:Pork.Consumption  1.870e-03  1.122e-03   1.667   0.1299
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.089 on 9 degrees of freedom
## Multiple R-squared:  0.9721, Adjusted R-squared:  0.941
## F-statistic: 31.33 on 10 and 9 DF,  p-value: 8.801e-06
```

- The least significant 2-way interaction is Price x Pork.Consumption so this will be removed first.

```
beef.m7 = update(beef.m6,.~.-Price:Pork.Consumption)
summary(beef.m7)
```

```
##
## Call:
## lm(formula = Beef.Consumption ~ Year + Price + Income + Pork.Consumption +
##     Year:Price + Year:Income + Price:Income + Year:Pork.Consumption +
##     Income:Pork.Consumption)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.35768 -0.33818 -0.07641  0.28363  1.46755
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              6.706e+02  1.874e+03   0.358  0.72791
## Year                    -2.299e-01  9.763e-01  -0.236  0.81857
## Price                   -2.131e+02  6.295e+01  -3.385  0.00695 **
## Income                   7.841e+00  3.385e+00   2.317  0.04302 *
## Pork.Consumption         1.498e+01  2.040e+01   0.734  0.47953
## Year:Price               1.078e-01  3.233e-02   3.336  0.00754 **
## Year:Income             -4.169e-03  1.776e-03  -2.348  0.04081 *
## Price:Income             6.267e-03  2.562e-03   2.446  0.03450 *
## Year:Pork.Consumption   -8.410e-03  1.061e-02  -0.793  0.44640
## Income:Pork.Consumption  1.835e-03  1.055e-03   1.739  0.11265
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.035 on 10 degrees of freedom
## Multiple R-squared:  0.9719, Adjusted R-squared:  0.9467
## F-statistic: 38.48 on 9 and 10 DF,  p-value: 1.37e-06
```

- Next we remove Year x Pork.Consumption.

```
beef.m8 = update(beef.m7,.~.-Year:Pork.Consumption)
summary(beef.m8)
```

```
##
## Call:
## lm(formula = Beef.Consumption ~ Year + Price + Income + Pork.Consumption +
##     Year:Price + Year:Income + Price:Income + Income:Pork.Consumption)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -1.43734 -0.28697 -0.11836  0.00022  1.79589
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              1.750e+03  1.265e+03   1.384  0.19392
## Year                    -7.915e-01  6.603e-01  -1.199  0.25584
## Price                   -2.338e+02  5.630e+01  -4.152  0.00161 **
## Income                   8.964e+00  3.022e+00   2.966  0.01283 *
## Pork.Consumption        -1.181e+00  5.084e-01  -2.323  0.04033 *
## Year:Price               1.185e-01  2.887e-02   4.106  0.00174 **
## Year:Income             -4.745e-03  1.593e-03  -2.978  0.01255 *
## Price:Income             6.364e-03  2.516e-03   2.530  0.02799 *
## Income:Pork.Consumption  1.653e-03  1.012e-03   1.633  0.13084
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.018 on 11 degrees of freedom
## Multiple R-squared:  0.9702, Adjusted R-squared:  0.9485
## F-statistic: 44.72 on 8 and 11 DF,  p-value: 2.608e-07
```

- Next we remove Income x Pork.Consumption.

```
beef.m9 = update(beef.m8,.~.-Income:Pork.Consumption)
summary(beef.m9)

##
## Call:
## lm(formula = Beef.Consumption ~ Year + Price + Income + Pork.Consumption +
##     Year:Price + Year:Income + Price:Income)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.5416 -0.4934 -0.1553  0.7153  1.7881
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         1.893e+03  1.347e+03   1.405  0.18528
## Year               -9.020e-01  7.009e-01  -1.287  0.22239
## Price              -2.167e+02  5.903e+01  -3.670  0.00320 **
## Income              7.815e+00  3.136e+00   2.492  0.02833 *
## Pork.Consumption -3.553e-01  5.399e-02  -6.582 2.61e-05 ***
## Year:Price          1.100e-01  3.030e-02   3.630  0.00345 **
## Year:Income        -4.077e-03  1.643e-03  -2.481  0.02889 *
## Price:Income        5.215e-03  2.577e-03   2.023  0.06589 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.086 on 12 degrees of freedom
```

```
## Multiple R-squared:  0.9629, Adjusted R-squared:  0.9413
## F-statistic: 44.55 on 7 and 12 DF,  p-value: 1.19e-07
```

- Next we remove Price x Income.

```
beef.m10 = update(beef.m9,.~.-Price:Income)
summary(beef.m10)

##
## Call:
## lm(formula = Beef.Consumption ~ Year + Price + Income + Pork.Consumption +
##     Year:Price + Year:Income)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.9306 -0.4066  0.1341  0.4486  1.7845
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       2.353e+03  1.477e+03   1.593   0.1352
## Year             -1.176e+00  7.652e-01  -1.537   0.1484
## Price            -1.630e+02  5.868e+01  -2.778   0.0157 *
## Income            4.073e+00  2.818e+00   1.445   0.1720
## Pork.Consumption -3.309e-01  5.855e-02  -5.652 7.91e-05 ***
## Year:Price        8.356e-02  3.042e-02   2.747   0.0166 *
## Year:Income      -2.069e-03  1.457e-03  -1.420   0.1791
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.208 on 13 degrees of freedom
## Multiple R-squared:  0.9503, Adjusted R-squared:  0.9274
## F-statistic: 41.43 on 6 and 13 DF,  p-value: 9.802e-08
```

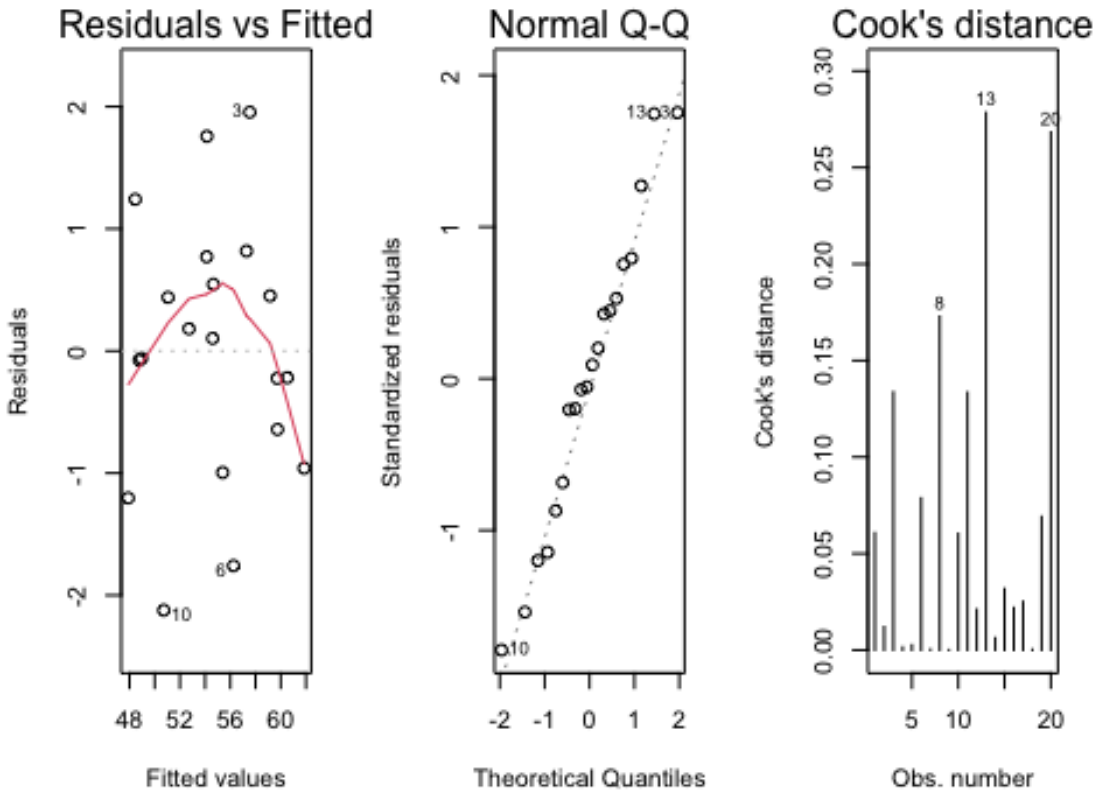
- Next we remove Year x Income.

```
beef.m11 = update(beef.m10,.~.-Year:Income)
summary(beef.m11)

##
## Call:
## lm(formula = Beef.Consumption ~ Year + Price + Income + Pork.Consumption +
##     Year:Price)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.12258 -0.72192  0.02075  0.60224  1.95468
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)       3.250e+03  1.383e+03   2.351   0.0339 *
## Year             -1.639e+00  7.170e-01  -2.285   0.0384 *
## Price            -1.170e+02  5.069e+01  -2.309   0.0367 *
## Income            7.133e-02  8.203e-03   8.696 5.12e-07 ***
## Pork.Consumption -3.608e-01  5.656e-02  -6.379 1.71e-05 ***
## Year:Price        5.977e-02  2.630e-02   2.273   0.0393 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.252 on 14 degrees of freedom
## Multiple R-squared:  0.9426, Adjusted R-squared:  0.9221
## F-statistic: 45.98 on 5 and 14 DF,  p-value: 3.385e-08
```

- All remaining terms are significant at the 0.05 level. The remaining 6 terms is less than 20/3 so we do not have to worry about over-paramaterization.
- The initial model had an R-squared value of 0.9873 while the best fit model has an R-squared value of 0.9426 which is still quite acceptable. The p-value of 3.385e-08 allows us to reject the null hypothesis of no interaction.
- Now we examine some model diagnostics.

```
par(mfrow=c(1,3))
plot(beef.m11, which=c(1,2,4))
```



- The variance of the residuals is quite constant and the distribution of the residuals in the Q-Q plot does not appear to deviate from normality. The slight deviations and variability shown can be explained by the very small sample size of the data. Observations 13 and 20 in Cook's distance look as though they appear influential but they do not appear to be extreme in any of the other plots so we will leave them in.
- In summary, all 4 original factors as well as the 2-way interaction of Year x Price appear to be related to Beef Consumption.
- Year, Price, and Pork Consumption are negatively related to Beef Consumption, while Income and Year x Price are Positively related to Beef Consumption.
- Final model: Beef Consumption = 3.250e+03 - 1.639e+00xYear - 1.170e+02xPrice + 7.133e-02xIncome - 3.608e-01xPork.Consumption + 5.977e-02xYearxPrice
- This model explains roughly 94% of the variability in Beef Consumption in this data set.

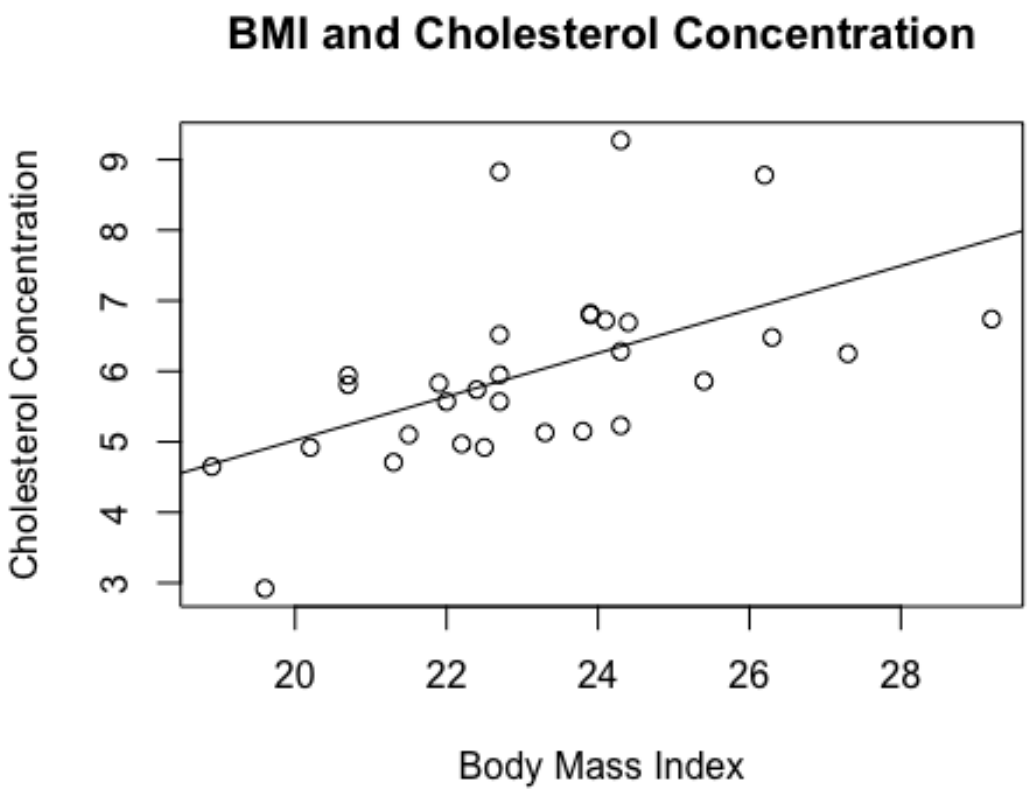
Question 2.

Use multiple regression to test whether serum cholesterol is associated with body mass index when age is included in the model. Consider carefully how both variables should be included, include initial plots, residual plots etc. Discuss your results carefully.

```
chol = read.table(file="~/Desktop/R/chol.txt", header=TRUE, sep="")
attach(chol)
```

- Null Hypothesis: There is no association between Cholesterol concentration and BMI when age is included in the model.
- Alternative Hypothesis: There is an association between Cholesterol concentration and BMI when age is included in the model.
- We will first fit a simple linear regression model with just Cholesterol concentration and BMI.

```
plot(BMI, CHOL, ylab="Cholesterol Concentration", xlab="Body Mass Index")
title('BMI and Cholesterol Concentration' )
model.simple = lm(CHOL~BMI)
abline(model.simple)
```



- From the plot of BMI and Cholesterol concentration it appears that there is a positive relationship between the two.

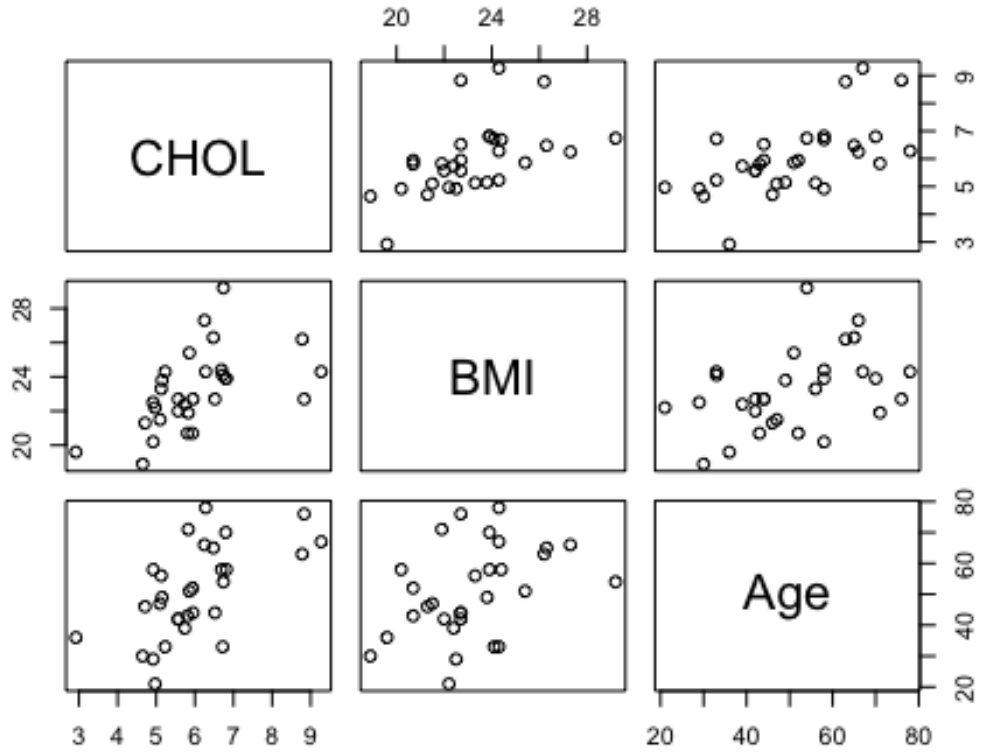
```
summary(model.simple)

##
## 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
```

- The model on its own does not appear to be a good fit with only 28.65% of the variability in Cholesterol concentration explained by BMI.
- For testing the relationship between Cholesterol concentration and BMI we obtain a p-value = 0.002303 so we reject the null hypothesis that they are not related.

```
2*(1-pt(3.353,length(BMI)-2))

## [1] 0.002305173

coef(model.simple)[2]

##       BMI
## 0.3089659

confint(model.simple)

##                 2.5 %    97.5 %
## (Intercept) -5.551855 3.2381970
## BMI          0.120233 0.4976987
```

- A unit increase in BMI will increase Cholesterol concentration by 0.0023 with 95% confidence interval (0.120,0.498).
- This analysis concludes that there is a positive association between Cholesterol concentration and BMI.
- Now we will add the age variable into the equation.

```
chol.data = data.frame(CHOL=CHOL,BMI=BMI,Age=Age)
pairs(chol.data)
```



- From the pairs plot we can see that Cholesterol concentration is also positively associated with age as we knew. It also appears that BMI and Age are positively correlated. There looks to be a slight curvature in the graph of CHOl and BMI so we will attempt to fix this with a quadratic model.

```
chol.multiple = lm(CHOL~BMI*Age+I(BMI^2))
summary(chol.multiple)
```

```
##
## Call:
## lm(formula = CHOL ~ BMI * Age + I(BMI^2))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.36116 -0.69790 -0.08948  0.48762  2.21358
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.083e+01  1.431e+01  -1.455    0.158
## BMI          1.950e+00  1.239e+00   1.574    0.128
## Age          2.168e-02  1.986e-01   0.109    0.914
## I(BMI^2)    -3.748e-02  2.953e-02  -1.269    0.216
## BMI:Age      7.306e-04  8.589e-03   0.085    0.933
```

```
##
## Residual standard error: 0.9909 on 25 degrees of freedom
## Multiple R-squared:  0.5062, Adjusted R-squared:  0.4272
## F-statistic: 6.406 on 4 and 25 DF,  p-value: 0.001083
```

- In the multiple regression model including age we have increased the R-squared value with R-squared = 0.5062
- The 2-way interaction of BMI x Age is not significant so we will remove it

```
chol.multiple2 = update(chol.multiple,.~.-BMI:Age)
summary(chol.multiple2)
```

```
##
## Call:
## lm(formula = CHOL ~ BMI + Age + I(BMI^2))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.34670 -0.69412 -0.08862  0.46843  2.22448
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -20.95735   13.95259  -1.502  0.14514
## BMI           1.92523    1.18229   1.628  0.11550
## Age           0.03854    0.01345   2.864  0.00816 **
## I(BMI^2)     -0.03618    0.02474  -1.462  0.15572
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9718 on 26 degrees of freedom
## Multiple R-squared:  0.506,  Adjusted R-squared:  0.449
## F-statistic: 8.878 on 3 and 26 DF,  p-value: 0.0003209
```

- Remove the quadratic term for BMI.

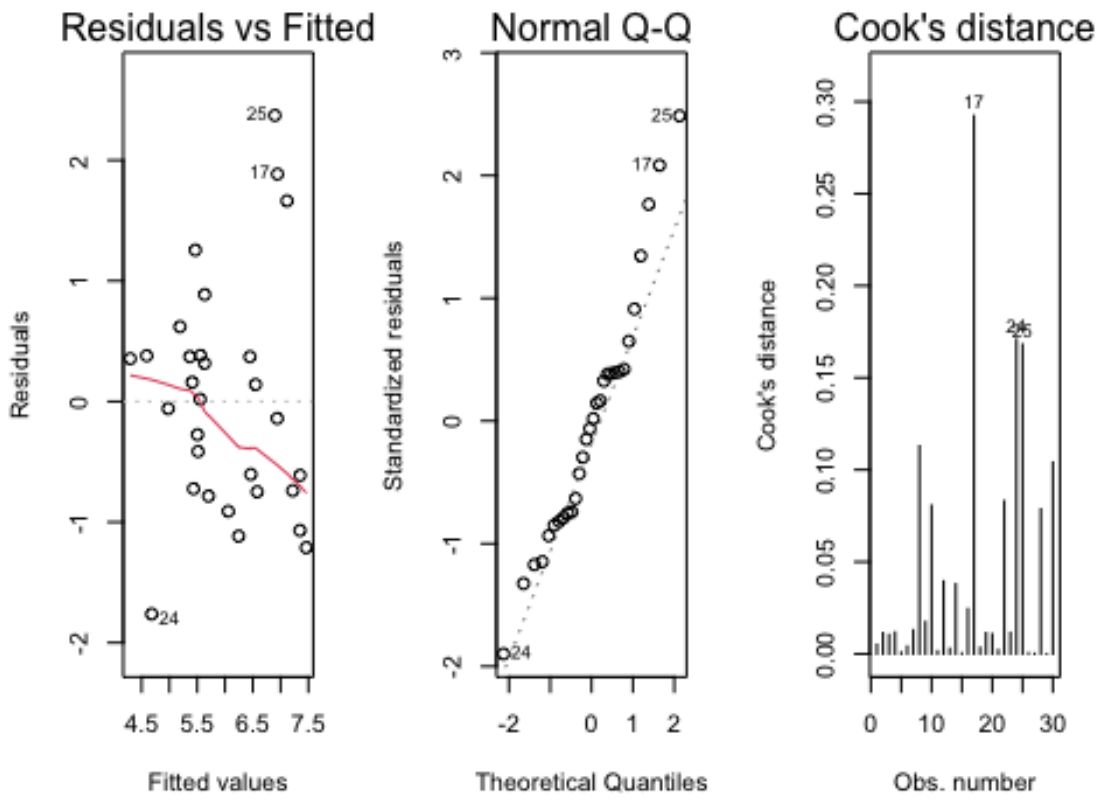
```
chol.multiple3 = update(chol.multiple2,.~.-I(BMI^2))
summary(chol.multiple3)
```

```
##
## Call:
## lm(formula = CHOL ~ BMI + Age)
##
## 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
## BMI          0.20137    0.08876   2.269  0.03149 *
## Age          0.04097    0.01363   3.006  0.00567 **
## ---
## 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
```

- All remaining terms are significant

```
par(mfrow=c(1,3))
plot(chol.multiple3, which=c(1,2,4))
```



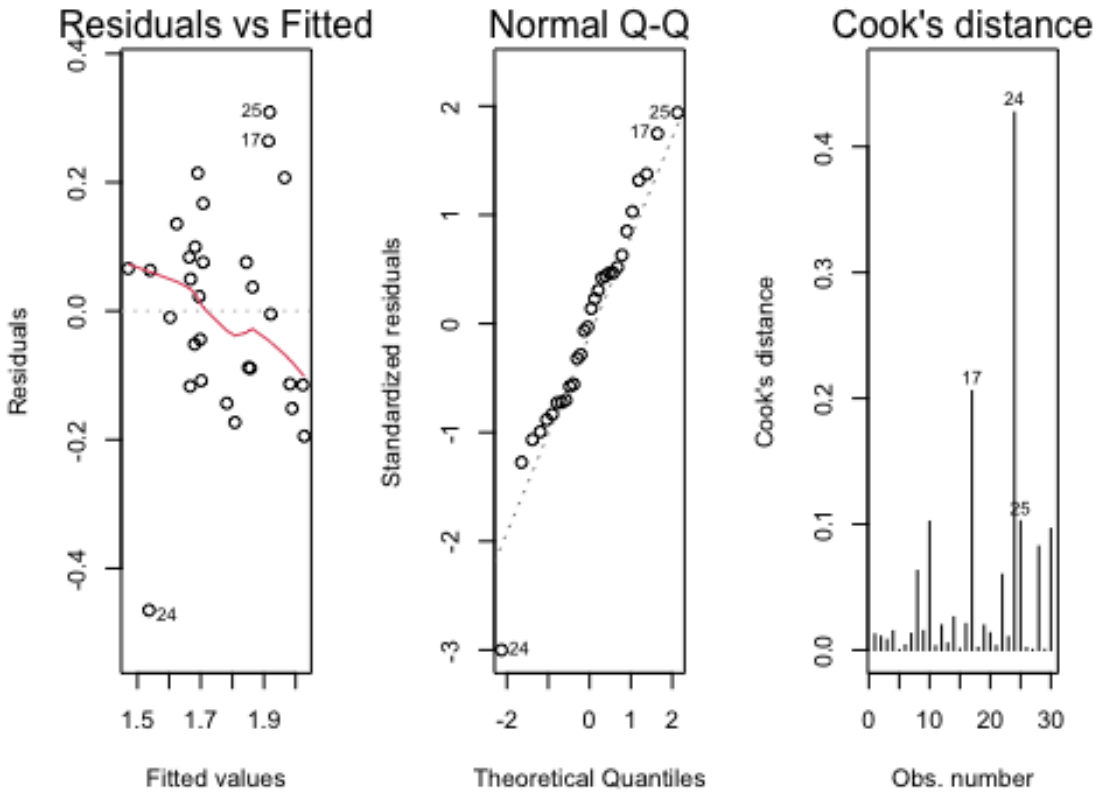
- It appears as though the variance of the residuals is not constant and increases with fitted values. The distribution of the residuals may be right skewed as well as potentially a bit left skewed.
- We will apply a log transformation to the response.

```
chol.multiple4 = update(chol.multiple3,log(.)~.)
summary(chol.multiple4)
```

```
##
## Call:
## lm(formula = log(CHOL) ~ BMI + Age)
##
## 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 .
## BMI         0.038581   0.014819   2.604   0.0148 *
## Age         0.006449   0.002276   2.834   0.0086 **
## ---
## 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
```

- Remaining terms are still significant with the p-value for BMI being 0.0148 and the p-value for Age being 0.0086. The P-value of 0.0001516 tells us that we can reject the null hypothesis and there is an interaction between BMI, Age and Cholesterol concentration.

```
par(mfrow=c(1,3))
plot(chol.multiple4,which=c(1,2,4))
```



- The variance of the residuals appears constant and the distribution is much better. The slight deviations can probably be explained by the small sample size. Observation 24 looks as though it may be potentially influential. This observation appears extreme in all of the diagnostic plots.

```
chol.multiple5 = update(chol.multiple4,.~.,subset=(1:length(CHOL)!=17))
summary(chol.multiple5)

##
## Call:
## lm(formula = log(CHOL) ~ BMI + Age, subset = (1:length(CHOL) !=
##     17))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.45922 -0.09628  0.02173  0.07950  0.33963
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.497821   0.305113   1.632  0.11482
## BMI         0.043781   0.014505   3.018  0.00563 **
```

```
## Age         0.004857   0.002352   2.065  0.04902 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1589 on 26 degrees of freedom
## Multiple R-squared:  0.4737, Adjusted R-squared:  0.4332
## F-statistic:  11.7 on 2 and 26 DF,  p-value: 0.0002379
```

- The coefficient estimates appear stable and the R-squared is also stable to remove observation 24.

```
confint(chol.multiple4)

##                    2.5 %     97.5 %
## (Intercept) -0.101383385 1.19790770
## BMI          0.008176087 0.06898639
## Age          0.001779404 0.01111773
```

- The 95% confidence interval for the coefficient of cholesterol includes 0
- In summary Cholesterol concentration is associated with BMI when Age is included in the model. The effects are additive and these factors do not appear to interact in their relationship with cholesterol concentration.
- On the log-scale, BMI and Age are positively related to cholesterol concentration.
- Final model: CHOL = exp{0.548262 + 0.038581xBMI + 0.006449xAge}
- This model explains roughly 48% of the variability of Cholesterol concentration in the data set.