Multiple Linear Regression

1. Start by exploring the data. Use R and create summary statistics and plots for each variable. Comment on what you see.

```
happy <- read.csv("happy(1).csv")
head(happy)
```

```
##   Happiness Age Gender Married Children Income Parents Smoke
## 1        37  16      0       0        0     52   -1.86     1
## 2        35  28      1       0        0     30   -1.47     1
## 3        57  16      0       1        1    176    0.40     0
## 4        63  23      1       0        0    146   -0.48     0
## 5        50  18      0       1        0     28    0.21     0
## 6        48  30      0       1        2     76    0.01     1
```

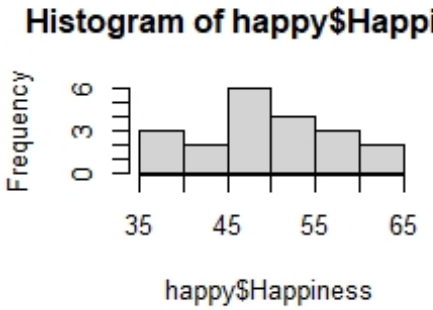
```
str(happy)
```

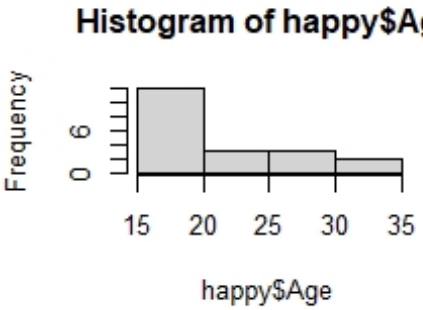
```
## 'data.frame':    20 obs. of  8 variables:
##  $ Happiness: int  37 35 57 63 50 48 48 36 41 52 ...
##  $ Age      : int  16 28 16 23 18 30 19 19 34 16 ...
##  $ Gender   : int  0 1 0 1 0 0 0 1 0 1 ...
##  $ Married  : int  0 0 1 0 1 1 1 0 0 0 ...
##  $ Children : int  0 0 1 0 0 2 0 2 2 0 ...
##  $ Income   : int  52 30 176 146 28 76 90 32 128 38 ...
##  $ Parents  : num  -1.86 -1.47 0.4 -0.48 0.21 0.01 0.5 1.59 1.98 -0.88 ...
##  $ Smoke    : int  1 1 0 0 0 1 0 0 0 0 ...
```

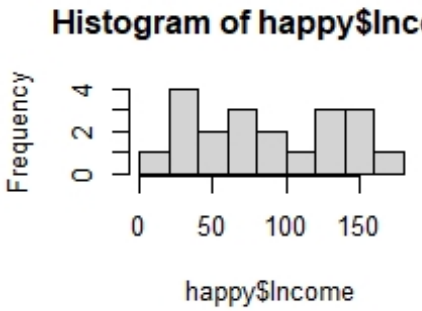
```
summary(happy)
```

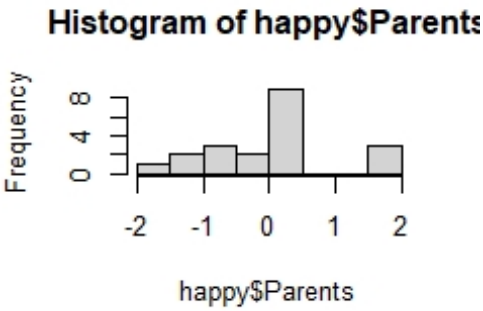
```
##    Happiness          Age            Gender        Married        Children
##  Min.   :35.00   Min.   :16.00   Min.   :0.00   Min.   :0.00   Min.   :0.0
##  1st Qu.:45.50   1st Qu.:16.00   1st Qu.:0.00   1st Qu.:0.00   1st Qu.:0.0
##  Median :50.00   Median :19.00   Median :1.00   Median :0.00   Median :0.0
##  Mean   :49.55   Mean   :21.30   Mean   :0.55   Mean   :0.45   Mean   :0.6
##  3rd Qu.:55.50   3rd Qu.:25.25   3rd Qu.:1.00   3rd Qu.:1.00   3rd Qu.:1.0
##  Max.   :63.00   Max.   :34.00   Max.   :1.00   Max.   :1.00   Max.   :2.0
##      Income         Parents            Smoke
##  Min.   : 12.0   Min.   :-1.8600   Min.   :0.0
##  1st Qu.: 47.0   1st Qu.:-0.7050   1st Qu.:0.0
##  Median : 83.0   Median : 0.0600   Median :0.0
##  Mean   : 88.9   Mean   : 0.0005   Mean   :0.3
##  3rd Qu.:135.5   3rd Qu.: 0.4250   3rd Qu.:1.0
##  Max.   :176.0   Max.   : 1.9800   Max.   :1.0
```

```
par(mfrow = c(2, 2))
hist(happy$Happiness)
hist(happy$Age)
hist(happy$Income)
hist(happy$Parents)
```

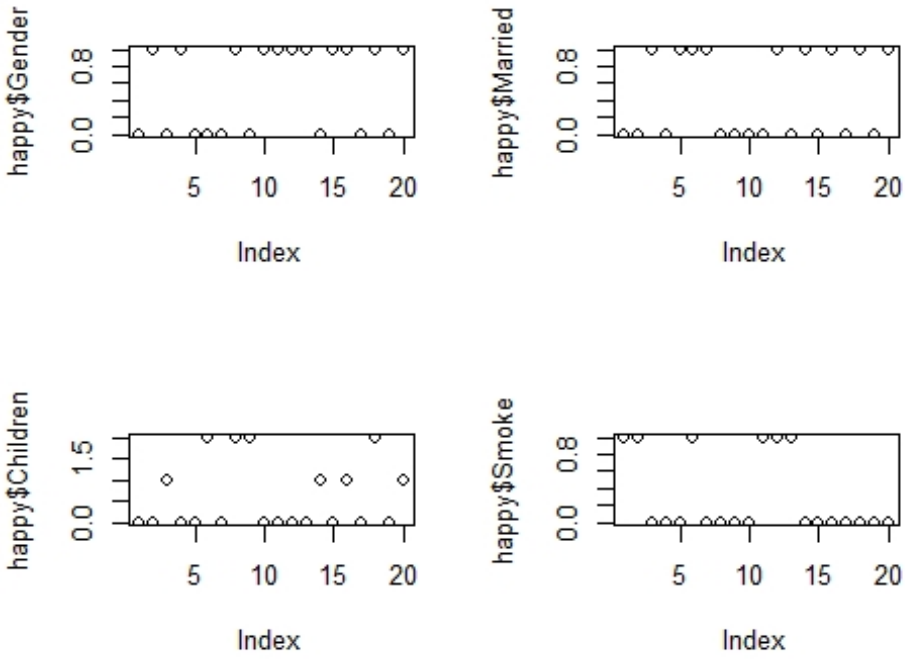









```r
plot(happy$Gender)
plot(happy$Married)
plot(happy$Children)
plot(happy$Smoke)
```



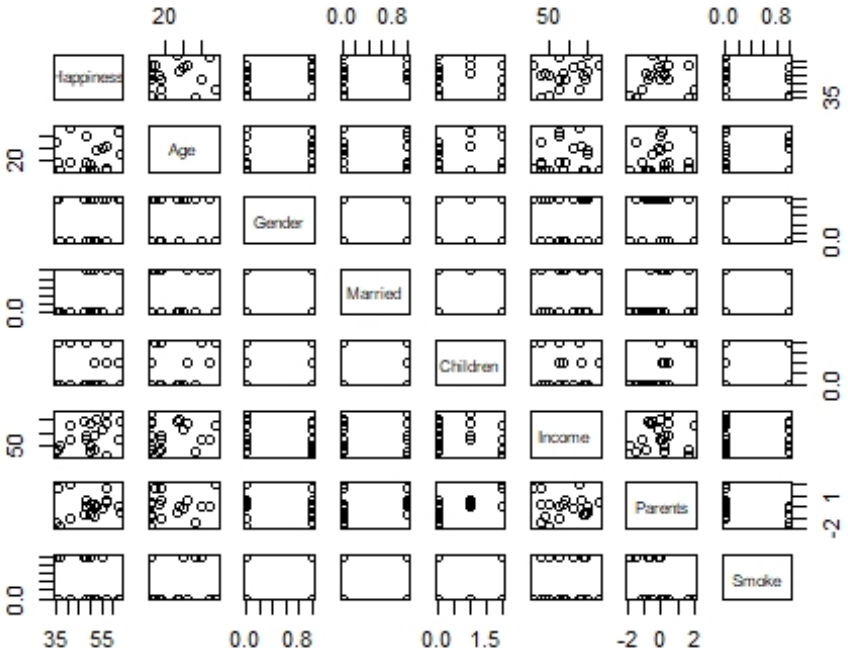
The histogram for happiness resembles a normal distribution so there is no need for log/transformation and there's no extreme values to be concerned about. All explanatory variables have no extreme values to be concerned about.

2. Use suitable graphs and plots to explore the relationship between variables. Comment on your graphs.

```
cor(happy)
```

```
##              Happiness          Age       Gender     Married    Children
## Happiness  1.00000000  0.034047716  0.239291282  0.44259417 -0.02041684
## Age        0.03404772  1.000000000 -0.005154982 -0.02921156  0.36758851
## Gender     0.23929128 -0.005154982  1.000000000 -0.19191919 -0.07537784
## Married    0.44259417 -0.029211562 -0.191919192  1.00000000  0.32663729
## Children  -0.02041684  0.367588506 -0.075377836  0.32663729  1.00000000
## Income     0.38364569  0.134700517 -0.137572229  0.04154306 -0.12566616
## Parents    0.18088505  0.219582207 -0.039765063  0.30383396  0.75398133
## Smoke     -0.34682959  0.059694428  0.153522062 -0.15352206 -0.21821789
##                 Income      Parents       Smoke
## Happiness  0.383645690  0.180885054 -0.34682959
## Age        0.134700517  0.219582207  0.05969443
## Gender    -0.137572229 -0.039765063  0.15352206
## Married    0.041543056  0.303833961 -0.15352206
## Children  -0.125666156  0.753981325 -0.21821789
## Income     1.000000000 -0.003676385  0.04215375
## Parents   -0.003676385  1.000000000 -0.55689549
## Smoke      0.042153754 -0.556895491  1.00000000
```

```
pairs(happy)
```



The matrix plot is quite successful in capturing the relationship of the predictors and the response. There is a positive linear relationship between happiness and parents, and income and happiness. It seems that females are happier than male, married people are happier than non-married people, people with one child are happier those with no or two children, and non-smokers are happier than smokers.

In terms of the relationship between the predictors, parents and income show a fairly positive linear relationship indicating that we may only need in the final model. In addition, it is hard to tell the relationship of other explanatory variables with gender, married, children and smoke as their values are between 0 and 2.

Therefore we'll use the correlation summary. Age and children has a positive correlation of 0.367588506, married and children has 0.326,63729, married and parent has 0.30383396, children and parents has 0.75398133 whereas parents and smoke has a negative correlation of -0.556895491. The highest correlation is between children and parents, so we may only need one in our best model.

So far, parent is correlated with many predictors. We may not need parent in our best model, but we will have to run tests to see if that is the case.

Overall, the correlation summary tells us that the predictors married, income and smoke have a significant correlation with happiness, with income being the highest correlation. We can expect that they will be included in the best model.

3. Next, fit a linear model starting with all the main effects. Reduce your model to the most parsimonious final model. Make sure you look at the residuals. Don't forget to check that the variables that are "factors" are being recognized in R, for example the variable married has only two levels, yes or no.

```
factor_Gender <- as.factor(happy$Gender)
factor_Married <- as.factor(happy$Married)
factor_Smoke <- as.factor(happy$Smoke)

model1 <- lm(Happiness ~ Age + factor_Gender + factor_Married +
               Children + Income + Parents + factor_Smoke, data = happy)
summary(model1)

##
## Call:
## lm(formula = Happiness ~ Age + factor_Gender + factor_Married +
##     Children + Income + Parents + factor_Smoke, data = happy)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -7.807 -4.267  1.117  3.118  8.421
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     35.88628    6.26606   5.727  9.5e-05 ***
## Age              0.14873    0.26399   0.563   0.5835
## factor_Gender1   7.24347    2.90245   2.496   0.0281 *
## factor_Married1  8.53957    3.02062   2.827   0.0153 *
## Children        -1.54372    3.11432  -0.496   0.6291
## Income           0.06754    0.03042   2.220   0.0464 *
## Parents         -1.03135    2.80250  -0.368   0.7193
## factor_Smoke1   -8.02949    4.01477  -2.000   0.0687 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.116 on 12 degrees of freedom
## Multiple R-squared:  0.6457, Adjusted R-squared:  0.4391
## F-statistic: 3.125 on 7 and 12 DF,  p-value: 0.04018
```

Model1 is the full model with all the main effects and it is not a good model. The variables children, parents, smoke and age have p-values greater than 0.05. In addition, the residual standard error is so high meaning that there is a large difference between the observed values and predicted values.

Therefore, we need to reduce the model.

```
drop1(model1)

## Single term deletions
##
## Model:
```

```
## Happiness ~ Age + factor_Gender + factor_Married + Children +
##     Income + Parents + factor_Smoke
##                Df Sum of Sq    RSS    AIC
## <none>                      448.85 78.219
## Age             1    11.873 460.72 76.741
## factor_Gender   1   232.961 681.81 84.580
## factor_Married  1   298.950 747.80 86.428
## Children        1     9.190 458.04 76.624
## Income          1   184.416 633.27 83.103
## Parents         1     5.066 453.91 76.444
## factor_Smoke    1   149.614 598.46 81.973
```

So the full model has an AIC of 78.219. The goal is to decrease that AIC and by taking out the variable age, we will reduce the AIC of our model to 76.741. This concept is the same for children and parents.

The opposite concept applies to smoke as removing it increases the AIC, although its p-value is greater than 0.05 from previous summary. However, we'll follow the results of the AIC, so we'll keep smoke in the model.

Overall, drop the variables age, children and parents.

```
model2 <- lm(Happiness ~ factor_Gender + factor_Married + Income +
               factor_Smoke, data = happy)
summary(model2)
```

```
##
## Call:
## lm(formula = Happiness ~ factor_Gender + factor_Married + Income +
##     factor_Smoke, data = happy)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -11.220  -3.942   0.997   4.053   8.265
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     37.80236    3.61528  10.456 2.77e-08 ***
## factor_Gender1   7.08473    2.70378   2.620   0.0193 *
## factor_Married1  7.25383    2.67801   2.709   0.0162 *
## Income           0.07292    0.02725   2.676   0.0173 *
## factor_Smoke1   -6.31855    2.89319  -2.184   0.0453 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.798 on 15 degrees of freedom
## Multiple R-squared:  0.602,  Adjusted R-squared:  0.4959
## F-statistic: 5.673 on 4 and 15 DF,  p-value: 0.005501
```

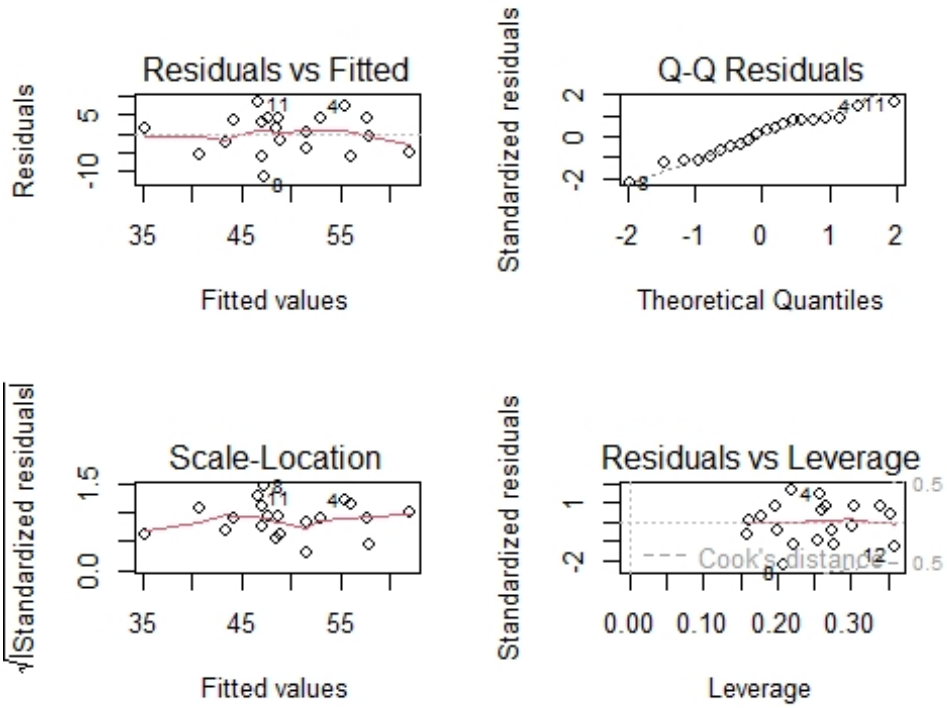
All the variables in model2 have a statistically significant p-values. The residual standard error has decreased implying that the distance between the observed values and predicted values has also decreased. The adjusted r-squared has increased to roughly 50% due to the removal of insignificant predictors. Overall, model2 is a better fit than model1.

```
anova(model1, model2)

## Analysis of Variance Table
##
## Model 1: Happiness ~ Age + factor_Gender + factor_Married + Children +
##     Income + Parents + factor_Smoke
## Model 2: Happiness ~ factor_Gender + factor_Married + Income +
factor_Smoke
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     12 448.85
## 2     15 504.20 -3   -55.354 0.4933 0.6936
```

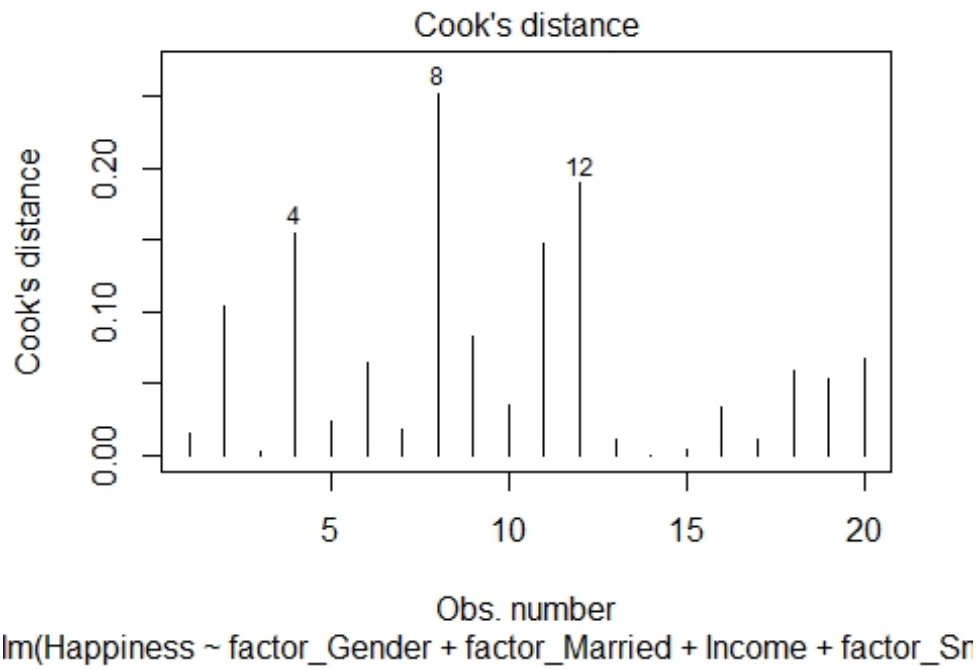
The p-value is not statistically significant meaning there is no difference between the two models. Therefore, we'll choose model2 as it has a better residual standard error and all variables are statistically significant.

```
par(mfrow = c(2, 2))
plot(model2)
```



The residuals and standardised residuals are close to zero and they show a random scatter implying constant variance. They also follow normality as they are pretty close to the line.

```
plot(model2, which=4)
```



There are no influential points as well as the points are less than 0.4.

4. Try some different model fitting methods such as starting with a minimal model and adding terms to it (forward selection), or using the step-both-ways option in R. Compare your final models with your backwards selection model. You can also try some different ways to make decisions, such as compare your final, best model from using AIC, BIC, or, if you have done STAT202, Mallows Cp.

```
minmod <- lm(Happiness ~ 1, data=happy)
step(minmod, direction = "forward",
 scope = list(lower = ~ 1,
 upper = ~ Age + factor_Gender + factor_Married + Children +
 Income + Parents + factor_Smoke))
```

```
## Start:  AIC=84.97
## Happiness ~ 1
##
##                   Df Sum of Sq    RSS    AIC
## + factor_Married   1   248.182 1018.8 82.612
## + Income           1   186.475 1080.5 83.788
## + factor_Smoke     1   152.402 1114.5 84.409
## <none>                         1267.0 84.973
## + factor_Gender    1    72.546 1194.4 85.793
## + Parents          1    41.454 1225.5 86.307
## + Age              1     1.469 1265.5 86.950
## + Children         1     0.528 1266.4 86.964
##
```

```
## Step:  AIC=82.61
## Happiness ~ factor_Married
##
##                 Df Sum of Sq     RSS    AIC
## + Income         1   169.321  849.45 80.977
## + factor_Gender  1   138.285  880.48 81.695
## + factor_Smoke   1   100.915  917.85 82.526
## <none>                       1018.77 82.612
## + Children       1    38.605  980.16 83.840
## + Parents        1     3.006 1015.76 84.553
## + Age            1     2.798 1015.97 84.557
##
## Step:  AIC=80.98
## Happiness ~ factor_Married + Income
##
##                 Df Sum of Sq    RSS    AIC
## + factor_Gender  1   184.921 664.53 78.067
## + factor_Smoke   1   114.452 734.99 80.083
## <none>                       849.45 80.977
## + Children       1    18.855 830.59 82.528
## + Parents        1     3.830 845.62 82.887
## + Age            1     0.010 849.44 82.977
##
## Step:  AIC=78.07
## Happiness ~ factor_Married + Income + factor_Gender
##
##                Df Sum of Sq    RSS    AIC
## + factor_Smoke  1   160.322 504.20 74.545
## <none>                      664.53 78.067
## + Children      1    15.088 649.44 79.607
## + Parents       1     2.945 661.58 79.978
## + Age           1     0.038 664.49 80.066
##
## Step:  AIC=74.54
## Happiness ~ factor_Married + Income + factor_Gender + factor_Smoke
##
##            Df Sum of Sq    RSS    AIC
## <none>                  504.20 74.545
## + Parents   1    39.350 464.85 74.920
## + Children  1    37.858 466.35 74.984
## + Age       1     0.183 504.02 76.538

##
## Call:
## lm(formula = Happiness ~ factor_Married + Income + factor_Gender +
##     factor_Smoke, data = happy)
##
## Coefficients:
##     (Intercept)  factor_Married1           Income   factor_Gender1
##        37.80236          7.25383          0.07292          7.08473
```

```
##   factor_Smoke1
##         -6.31855
```

The forward selection method shows that married, income, gender and smoke are part of the best model similar to the previous question.

```
maxmod <- lm(Happiness ~ Age + factor_Gender + factor_Married +
               Children + Income + Parents + factor_Smoke, data = happy)
step(maxmod, direction = "backward")

## Start:  AIC=78.22
## Happiness ~ Age + factor_Gender + factor_Married + Children +
##     Income + Parents + factor_Smoke
##
##                   Df Sum of Sq    RSS    AIC
## - Parents          1     5.066 453.91 76.444
## - Children         1     9.190 458.04 76.624
## - Age              1    11.873 460.72 76.741
## <none>                         448.85 78.219
## - factor_Smoke     1   149.614 598.46 81.973
## - Income           1   184.416 633.27 83.103
## - factor_Gender    1   232.961 681.81 84.580
## - factor_Married   1   298.950 747.80 86.428
##
## Step:  AIC=76.44
## Happiness ~ Age + factor_Gender + factor_Married + Children +
##     Income + factor_Smoke
##
##                   Df Sum of Sq    RSS    AIC
## - Age              1    12.431 466.35 74.984
## <none>                         453.91 76.444
## - Children         1    50.106 504.02 76.538
## - Income           1   179.793 633.71 81.117
## - factor_Smoke     1   192.884 646.80 81.526
## - factor_Gender    1   228.235 682.15 82.590
## - factor_Married   1   295.601 749.52 84.474
##
## Step:  AIC=74.98
## Happiness ~ factor_Gender + factor_Married + Children + Income +
##     factor_Smoke
##
##                   Df Sum of Sq    RSS    AIC
## - Children         1    37.858 504.20 74.545
## <none>                         466.35 74.984
## - factor_Smoke     1   183.092 649.44 79.607
## - Income           1   209.949 676.29 80.418
## - factor_Gender    1   228.836 695.18 80.969
## - factor_Married   1   283.402 749.75 82.480
##
## Step:  AIC=74.54
```

```
## Happiness ~ factor_Gender + factor_Married + Income + factor_Smoke
##
##                  Df Sum of Sq    RSS    AIC
## <none>                        504.20 74.545
## - factor_Smoke    1    160.32 664.53 78.067
## - factor_Gender   1    230.79 734.99 80.083
## - Income          1    240.66 744.86 80.349
## - factor_Married  1    246.62 750.82 80.509

##
## Call:
## lm(formula = Happiness ~ factor_Gender + factor_Married + Income +
##     factor_Smoke, data = happy)
##
## Coefficients:
##     (Intercept)   factor_Gender1  factor_Married1           Income
##        37.80236          7.08473          7.25383         0.07292
##   factor_Smoke1
##        -6.31855
```

The result for backward selection is the same as that of forward selection. This proves that gender, married, income and smoke are all significant variables in predicting happiness.

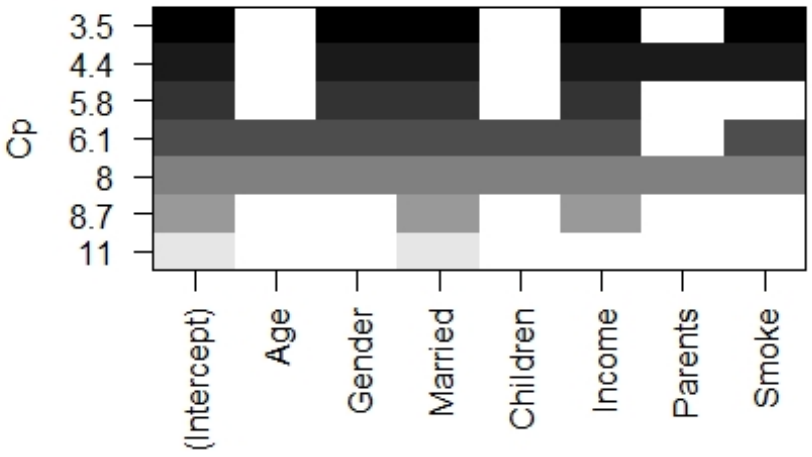
```
library(ISLR)

## Warning: package 'ISLR' was built under R version 4.4.1

library(leaps)

## Warning: package 'leaps' was built under R version 4.4.1

regfit.full <- regsubsets(Happiness~., happy, nvmax = 7)
plot(regfit.full, scale = "Cp")
```



We can use mallows Cp to select the best model. According to the results, gender, married, income and smoke are in the best model as they have the lowest Cp.

5. Discuss what your final, best model means – what effects happiness? How can we be happy?

```
model2 <- lm(Happiness ~ factor_Gender + factor_Married + Income +
               factor_Smoke, data = happy)
summary(model2)

##
## Call:
## lm(formula = Happiness ~ factor_Gender + factor_Married + Income +
##     factor_Smoke, data = happy)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -11.220  -3.942   0.997   4.053   8.265
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     37.80236    3.61528  10.456 2.77e-08 ***
## factor_Gender1   7.08473    2.70378   2.620   0.0193 *
## factor_Married1  7.25383    2.67801   2.709   0.0162 *
## Income           0.07292    0.02725   2.676   0.0173 *
## factor_Smoke1   -6.31855    2.89319  -2.184   0.0453 *
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.798 on 15 degrees of freedom
## Multiple R-squared:  0.602,  Adjusted R-squared:  0.4959
## F-statistic: 5.673 on 4 and 15 DF,  p-value: 0.005501
```

With one unit of increase in income, happiness goes up by 0.07292. Happiness increases when you are a female by 7.08473. Happiness increases when you are married by 7.25383. Happiness decreases when you are a smoker by 6.31855.

So according to this data, if you want to be happy, you should have a high income, become a female, be married and don't smoke.