CS-E-106: Data Modeling

Assignment 7

Instructor: Hakan Gogtas
Submitted by: *Saurabh Kulkarni*

Due Date: *11/18/2019*

Question 1: Refer to the CDI data set. A regression model relating serious crime rate (Y, total serious crimes divided by total population) to population density (X1, total population divided by land area) and unemployment rate (X3) is to be constructed. (15 pts)

(a) Fit second-order regression model (equation 8.8 on the book). Plot the residuals against the fitted values. How well does the second-order model appear to fit the data? What is R2? (5pts)

Solution:

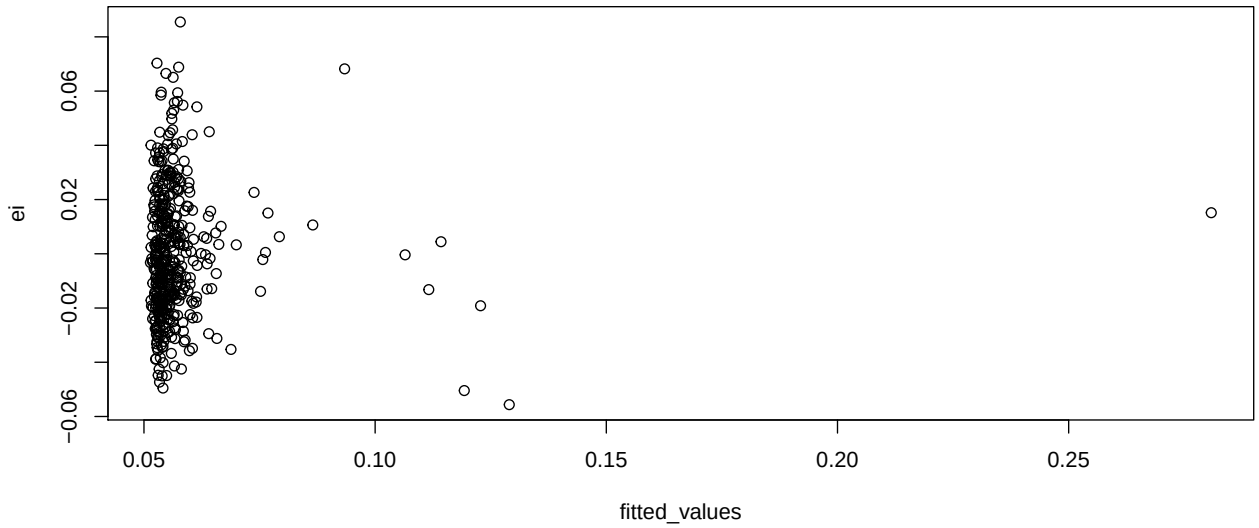
```
cdi_data = read.csv("CDI.csv")

df_1 = cdi_data

Y = df_1$Total.serious.crimes/df_1$Total.population
X1 = df_1$Total.population/df_1$Land.area
X3 = df_1$Percent.unemployment

df_1$x1 = (X1-mean(X1))/sqrt(var(X1))
df_1$x3 = (X3-mean(X3))/sqrt(var(X3))
df_1$x1sqr = df_1$x1^2
df_1$x3sqr = df_1$x3^2
df_1$x1x3 = df_1$x1*df_1$x3
```

```
lm_cdi_1a = lm(Y~x1+x3+x1sqr+x3sqr+x1x3, data=df_1)
summary(lm_cdi_1a)
```

```
##
## Call:
## lm(formula = Y ~ x1 + x3 + x1sqr + x3sqr + x1x3, data = df_1)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.055642 -0.016851 -0.002889  0.014810  0.085485
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.0562880  0.0012603  44.662  < 2e-16 ***
## x1           0.0100630  0.0021599   4.659 4.23e-06 ***
## x3          -0.0002057  0.0014672  -0.140   0.8886
## x1sqr        0.0000130  0.0002857   0.045   0.9637
## x3sqr        0.0008905  0.0005215   1.708   0.0884 .
## x1x3         0.0042761  0.0020994   2.037   0.0423 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.02383 on 434 degrees of freedom
```

```
## Multiple R-squared:  0.2485, Adjusted R-squared:  0.2398
## F-statistic:  28.7 on 5 and 434 DF,  p-value: < 2.2e-16
```

```
ei = lm_cdi_1a$residuals
fitted_values = lm_cdi_1a$fitted.values
plot(fitted_values, ei)
```




```
coeffs = summary(lm_cdi_1a)$coefficients
cat(sprintf("The regression model is Yhat = %f + %f*x1 + %f*x3 + %f*x1sqr + %f*x3sqr + %f*x1x3\n",
            coeffs[1,1], coeffs[2,1], coeffs[3,1], coeffs[4,1], coeffs[5,1], coeffs[6,1]))
```

```
## The regression model is Yhat = 0.056288 + 0.010063*x1 + -0.000206*x3 + 0.000013*x1sqr + 0.000891*x3sqr + 0
```

```
cat(sprintf("R^2: %f\n", summary(lm_cdi_1a)$r.squared))
```

```
## R^2: 0.248475
```

We can see that the regression model is not a very great fit based on the R^2 and the residual plot (we can see outliers and non-constant variance in the error terms).

(b) Test whether or not all quadratic and interaction terms can be dropped from the regression model; use = .01. State the alternatives, decision rule, and conclusion. (5pts)

Solution:

```
lm_cdi_1b = lm(Y~x1+x3, data=df_1)
summary(lm_cdi_1b)
```

```
##
## Call:
## lm(formula = Y ~ x1 + x3, data = df_1)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.053806 -0.016940 -0.003898  0.014680  0.084508
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.0572864  0.0011445  50.054   <2e-16 ***
## x1          0.0131098  0.0011461  11.439   <2e-16 ***
## x3          0.0008457  0.0011461   0.738    0.461
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.02401 on 437 degrees of freedom
## Multiple R-squared:  0.2318, Adjusted R-squared:  0.2283
## F-statistic: 65.92 on 2 and 437 DF,  p-value: < 2.2e-16
```

```
anova(lm_cdi_1b, lm_cdi_1a)
```

```
## Analysis of Variance Table
##
## Model 1: Y ~ x1 + x3
## Model 2: Y ~ x1 + x3 + x1sqr + x3sqr + x1x3
##   Res.Df     RSS Df Sum of Sq      F  Pr(>F)
## 1    437 0.25186
## 2    434 0.24638  3  0.005477 3.2159 0.02278 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
FStar = 3.2159 # from the above anova

df_diff = 3
df_E = lm_cdi_1b$df.residual
alpha = 0.01
FTest = qf(1-alpha, df_diff, df_E)
print(FTest)
```

```
## [1] 3.826715
```

Hypotheses:

$H_0 : \beta_{11} = \beta_{33} = \beta_{13} = 0$

H_a : Not all β's are equal to zero

Decision Rules:

If $F^* \leq 3.8267151$, conclude H_0

If $F^* > 3.8267151$, conclude H_a

Conclusion:

Since our test statistic, $F^* = 3.2159$, and $3.2159 \leq 3.8267151$, we conclude H_0. Also, we can see that the p-value is 0.02278 (from ANOVA) which is greater that the given $\alpha = 0.01$. Thus, we can remove all the quadratic and interaction terms.

(c) Instead of the predictor variable population density, total population (X1) and land area (X2) are to be employed as separate predictor variables, in addition to unemployment rate (X3). The regression model should contain linear and quadratic terms for total population, and linear terms only for land area and unemployment rate. (No interaction terms are to be included in this model.) Fit this regression model and obtain R2. Is this coefficient of multiple determination substantially different from the one for the regression model in part a? (5pts)

Solution:

```
X1 = cdi_data$Total.population
X2 = cdi_data$Land.area
X3 = cdi_data$Percent.unemployment

x1 = (X1-mean(X1))/sqrt(var(X1))
```

```r
x2 = (X2-mean(X2))/sqrt(var(X2))
x3 = (X3-mean(X3))/sqrt(var(X3))
x1sqr = x1^2
```

```r
lm_cdi_1c = lm(Y~x1+x2+x3+x1sqr)
summary(lm_cdi_1c)
```

```
##
## Call:
## lm(formula = Y ~ x1 + x2 + x3 + x1sqr)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.05967 -0.01704 -0.00303  0.01410  0.19106
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.0584998  0.0012291  47.595  < 2e-16 ***
## x1           0.0177099  0.0021398   8.276 1.57e-15 ***
## x2          -0.0008643  0.0012589  -0.687    0.493
## x3           0.0015955  0.0012395   1.287    0.199
## x1sqr       -0.0012161  0.0002130  -5.710 2.10e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.02539 on 435 degrees of freedom
## Multiple R-squared:  0.1444, Adjusted R-squared:  0.1365
## F-statistic: 18.35 on 4 and 435 DF,  p-value: 6.022e-14
```

```r
anova(lm_cdi_1c, lm_cdi_1a)
```

```
## Analysis of Variance Table
##
## Model 1: Y ~ x1 + x2 + x3 + x1sqr
## Model 2: Y ~ x1 + x3 + x1sqr + x3sqr + x1x3
##   Res.Df     RSS Df Sum of Sq      F   Pr(>F)
## 1    435 0.28051
## 2    434 0.24638  1  0.034121 60.103 6.46e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
cat(sprintf("R^2: %f\n", summary(lm_cdi_1c)$r.squared))
```

```
## R^2: 0.144398
```

Interpretation: We can see from the ANOVA that the model in Q1(c) is substantially different than the one in part (a), since the p-value is very low (<<0.001). Also, the R^2 is substantially different for the two models.

Question 2 Refer to the CDI data set. The number of active physicians (Y) is to be regressed against total population (X1), total personal income (X2), and geographic region (X3, X4, X5). (15pts)

(a) Fit a first-order regression model. Let X3 =1 if NE and 0 otherwise, X4 = 1 if NC and 0 otherwise, and X5 = 1I if S and 0 otherwise. (5pts)

Solution:

```
df = cdi_data
Y = df$Number.of.active.physicians
X1 = df$Total.population
X2 = df$Total.personal.income

X3 = ifelse(df$Geographic.region==1, 1, 0)
X4 = ifelse(df$Geographic.region==2, 1, 0)
X5 = ifelse(df$Geographic.region==3, 1, 0)
```

```
lm_cdi_2a = lm(Y~X1+X2+X3+X4+X5)
summary(lm_cdi_2a)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + X3 + X4 + X5)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1866.8  -207.7   -81.5    72.4  3721.7
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.075e+02  7.028e+01  -2.952  0.00332 **
## X1           5.515e-04  2.835e-04   1.945  0.05243 .
## X2           1.070e-01  1.325e-02   8.073  6.8e-15 ***
## X3           1.490e+02  8.683e+01   1.716  0.08685 .
## X4           1.455e+02  8.515e+01   1.709  0.08817 .
## X5           1.912e+02  8.003e+01   2.389  0.01731 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 566.1 on 434 degrees of freedom
## Multiple R-squared:  0.9011, Adjusted R-squared:  0.8999
## F-statistic: 790.7 on 5 and 434 DF,  p-value: < 2.2e-16
```

(b) Examine whether the effect for the northeastern region on number of active physicians differs from the effect for the north central region by constructing an appropriate 90 percent confidence interval. Interpret your interval estimate. (5pts)

```
alpha = 0.1
confint(lm_cdi_2a, level=1-alpha)
```

```
##                        5 %          95 %
## (Intercept) -3.233460e+02 -91.645500368
## X1           8.407549e-05   0.001018844
## X2           8.516238e-02   0.128860685
## X3           5.886209e+00 292.152934584
## X4           5.162733e+00 285.890158945
## X5           5.929211e+01 323.140497010
```

Interpretation:

The confidence interval estimates for Northeastern region (X3) are not significantly different that those for Northcentral region (X4).

(c) Test whether any geographic effects are present; use = .10. State the alternatives, decision rule, and conclusion. What is the P-value of the test? (5pts)

```r
lm_cdi_2c = lm(Y~X1+X2)
summary(lm_cdi_2c)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1849.1  -198.3   -71.4    39.7  3755.3
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -6.444e+01  3.283e+01  -1.963   0.0503 .
## X1           5.310e-04  2.775e-04   1.914   0.0563 .
## X2           1.072e-01  1.297e-02   8.269 1.64e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 568 on 437 degrees of freedom
## Multiple R-squared:  0.8998, Adjusted R-squared:  0.8993
## F-statistic:  1961 on 2 and 437 DF,  p-value: < 2.2e-16
```

```r
anova(lm_cdi_2c, lm_cdi_2a)
```

```
## Analysis of Variance Table
##
## Model 1: Y ~ X1 + X2
## Model 2: Y ~ X1 + X2 + X3 + X4 + X5
##   Res.Df       RSS Df Sum of Sq      F Pr(>F)
## 1    437 140967081
## 2    434 139093455  3   1873626 1.9487  0.121
```

```r
FStar = 1.9487 # from the above anova

df_diff = 3
df_E = lm_cdi_1b$df.residual
alpha = 0.1
FTest = qf(1-alpha, df_diff, df_E)
print(FTest)
```

```
## [1] 2.096362
```

Hypotheses:

$H_0 : \beta_3 = \beta_4 = \beta_5 = 0$

H_a : Not all β's are equal to zero

Decision Rules:

If $F^* \leq 2.0963615$, conclude H_0

If $F^* > 2.0963615$, conclude H_a

Conclusion:

Since our test statistic, $F^* = 1.9487$, and $1.9487 \leq 2.0963615$, we conclude H_0. Thus, the geographic effects are not present.

Quuestion 3 Refer to the Lung pressure Data. Increased arterial blood pressure in the lungs frequently leads to the development of heart failure in patients with chronic obstructive pulmonary disease (COPD). The standard method for determining arterial lung pressure is invasive, technically difficult, and involves some risk to the patient. Radionuclide imaging is a noninvasive, less risky method for estimating arterial pressure in the lungs. To investigate the predictive ability of this method, a cardiologist collected data on 19 mild-to-moderate COPD patients. The data includes the invasive measure of systolic pulmonary arterial pressure (Y) and three potential noninvasive predictor variables. Two were obtained by using radionuclide imaging emptying rate of blood into the pumping chamber or the heart (X1) and ejection rate of blood pumped out of the heart into the lungs (X2) and the third predictor variable measures blood gas (X3). (25pts)

(a) Fit the multiple regression function containing the three predictor variables us first-order terms. Does it appear that all predictor variables should be retained? (5pts)

Solution:

```
lung_data = read.csv("Lung Pressure.csv")
lm_lung_3a = lm(Y~X1+X2+X3, data=lung_data)
summary(lm_lung_3a)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + X3, data = lung_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -16.075 -12.064  -0.988   7.707  32.315
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 87.18750   21.55246   4.045  0.00106 **
## X1          -0.56448    0.42791  -1.319  0.20691
## X2          -0.51315    0.22449  -2.286  0.03723 *
## X3          -0.07196    0.45457  -0.158  0.87633
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.42 on 15 degrees of freedom
## Multiple R-squared:  0.6141, Adjusted R-squared:  0.5369
## F-statistic: 7.957 on 3 and 15 DF,  p-value: 0.002083
```

```
anova(lm_lung_3a)
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## X1         1 3577.1  3577.1 17.1920 0.0008615 ***
## X2         1 1384.4  1384.4  6.6535 0.0209379 *
## X3         1    5.2     5.2  0.0251 0.8763340
## Residuals 15 3121.0   208.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interpretation:

We see that the p-value for X2 and X3 show a good linear relation with Y, as they add a significant amount

of SSR (based on ANOVA above and looking at the respective p-values). But X3 does not appear to add a significant value to the model when X1 and X2 are already present.

(b) Using first-order and second-order terms for each of the three predictor variables (centered around the mean) in the pool of potential X variables (including cross products of the first order terms), find the three best hierarchical subset regression models according to the R2a,p criterion. (5pts)

```
df = lung_data

Y = df$Y
X1 = df$X1
X2 = df$X2
X3 = df$X3

df$x1 = (X1-mean(X1))/sqrt(var(X1))
df$x2 = (X2-mean(X2))/sqrt(var(X2))
df$x3 = (X3-mean(X3))/sqrt(var(X3))
df$x1sqr = df$x1^2
df$x2sqr = df$x2^2
df$x3sqr = df$x3^2
df$x1x2 = df$x1*df$x2
df$x1x3 = df$x1*df$x3
df$x2x3 = df$x2*df$x3
```

```
lm_lung_3b1 = lm(Y~x1+x2+x3+x1sqr+x2sqr+x3sqr+x1x2+x1x3+x2x3, data=df)
summary(lm_lung_3b1)
```

```
##
## Call:
## lm(formula = Y ~ x1 + x2 + x3 + x1sqr + x2sqr + x3sqr + x1x2 +
##     x1x3 + x2x3, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -10.825  -5.223  -1.236   4.804  20.784
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  41.9225     7.4385   5.636 0.000319 ***
## x1           -6.5083     4.9997  -1.302 0.225338
## x2          -16.0564     6.2420  -2.572 0.030073 *
## x3           -1.7484     3.6376  -0.481 0.642244
## x1sqr        -0.4055     8.1713  -0.050 0.961503
## x2sqr        -2.7073     8.6629  -0.313 0.761767
## x3sqr        -5.7702     5.6451  -1.022 0.333396
## x1x2          9.7756    15.4349   0.633 0.542266
## x1x3          7.3913    10.9729   0.674 0.517494
## x2x3        -10.0528    10.6616  -0.943 0.370353
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.39 on 9 degrees of freedom
## Multiple R-squared:  0.8291, Adjusted R-squared:  0.6583
## F-statistic: 4.852 on 9 and 9 DF,  p-value: 0.01383
```

```r
lm_lung_3b2 = update(lm_lung_3b1,.~.-x1sqr)
summary(lm_lung_3b2)
```

```
##
## Call:
## lm(formula = Y ~ x1 + x2 + x3 + x2sqr + x3sqr + x1x2 + x1x3 +
##     x2x3, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -10.849  -5.351  -1.374   4.931  20.884
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   41.692      5.520   7.553 1.94e-05 ***
## x1            -6.583      4.524  -1.455   0.1763
## x2           -16.046      5.919  -2.711   0.0219 *
## x3            -1.741      3.449  -0.505   0.6246
## x2sqr         -2.449      6.567  -0.373   0.7170
## x3sqr         -5.612      4.426  -1.268   0.2336
## x1x2           9.043      4.257   2.124   0.0596 .
## x1x3           7.045      8.043   0.876   0.4016
## x2x3          -9.778      8.642  -1.131   0.2843
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.76 on 10 degrees of freedom
## Multiple R-squared:  0.8291, Adjusted R-squared:  0.6923
## F-statistic: 6.063 on 8 and 10 DF,  p-value: 0.005164
```

```r
lm_lung_3b3 = update(lm_lung_3b2,.~.-x2sqr)
summary(lm_lung_3b3)
```

```
##
## Call:
## lm(formula = Y ~ x1 + x2 + x3 + x3sqr + x1x2 + x1x3 + x2x3, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.7381  -5.6479  -0.8885   4.9743  21.6337
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   40.557      4.420   9.175 1.74e-06 ***
## x1            -5.967      4.043  -1.476  0.16803
## x2           -17.190      4.859  -3.537  0.00465 **
## x3            -1.405      3.196  -0.440  0.66869
## x3sqr         -5.067      4.011  -1.263  0.23261
## x1x2           7.850      2.696   2.912  0.01415 *
## x1x3           5.565      6.716   0.829  0.42492
## x2x3          -7.216      5.034  -1.433  0.17955
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 11.29 on 11 degrees of freedom
## Multiple R-squared:  0.8267, Adjusted R-squared:  0.7164
## F-statistic: 7.496 on 7 and 11 DF,  p-value: 0.001847
```

```
lm_lung_3b4 = update(lm_lung_3b3,.~.-x1x3)
summary(lm_lung_3b4)
```

```
##
## Call:
## lm(formula = Y ~ x1 + x2 + x3 + x3sqr + x1x2 + x2x3, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -13.257  -5.658  -1.406   3.468  25.410
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  40.0233     4.3158   9.274 8.04e-07 ***
## x1           -5.5849     3.9640  -1.409  0.18424
## x2          -16.9858     4.7894  -3.547  0.00402 **
## x3           -0.9495     3.1069  -0.306  0.76514
## x3sqr        -3.6895     3.6025  -1.024  0.32597
## x1x2          8.0186     2.6530   3.022  0.01061 *
## x2x3         -4.4262     3.6935  -1.198  0.25391
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.14 on 12 degrees of freedom
## Multiple R-squared:  0.8159, Adjusted R-squared:  0.7238
## F-statistic: 8.863 on 6 and 12 DF,  p-value: 0.0007742
```

```
lm_lung_3b5 = update(lm_lung_3b4,.~.-x3sqr)
summary(lm_lung_3b5)
```

```
##
## Call:
## lm(formula = Y ~ x1 + x2 + x3 + x1x2 + x2x3, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -13.463  -4.918  -2.334   3.967  24.931
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  37.0841     3.2294  11.483 3.54e-08 ***
## x1           -6.6514     3.8319  -1.736  0.10622
## x2          -15.9137     4.6824  -3.399  0.00475 **
## x3           -0.6321     3.0972  -0.204  0.84144
## x1x2          8.7699     2.5544   3.433  0.00445 **
## x2x3         -1.8817     2.7382  -0.687  0.50402
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.16 on 13 degrees of freedom
## Multiple R-squared:  0.7998, Adjusted R-squared:  0.7228
```

```
## F-statistic: 10.39 on 5 and 13 DF,  p-value: 0.0003511
```

```
lm_lung_3b6 = update(lm_lung_3b5,.~.-x3)
summary(lm_lung_3b6)
```

```
##
## Call:
## lm(formula = Y ~ x1 + x2 + x1x2 + x2x3, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -12.553  -5.250  -2.092   4.397  25.213
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   37.101      3.116  11.907 1.03e-08 ***
## x1            -6.923      3.468  -1.997  0.06570 .
## x2           -15.444      3.936  -3.924  0.00153 **
## x1x2           8.777      2.465   3.561  0.00313 **
## x2x3          -1.829      2.631  -0.695  0.49840
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.77 on 14 degrees of freedom
## Multiple R-squared:  0.7991, Adjusted R-squared:  0.7418
## F-statistic: 13.93 on 4 and 14 DF,  p-value: 8.695e-05
```

```
lm_lung_3b7 = update(lm_lung_3b6,.~.-x2x3)
summary(lm_lung_3b7)
```

```
##
## Call:
## lm(formula = Y ~ x1 + x2 + x1x2, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.3075  -6.6602  -0.5824   4.6284  24.0398
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   37.888      2.852  13.284 1.06e-09 ***
## x1            -7.511      3.305  -2.273 0.038173 *
## x2           -14.098      3.367  -4.187 0.000794 ***
## x1x2           8.690      2.419   3.592 0.002667 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.58 on 15 degrees of freedom
## Multiple R-squared:  0.7922, Adjusted R-squared:  0.7507
## F-statistic: 19.06 on 3 and 15 DF,  p-value: 2.233e-05
```

We can see that `lm_lung_3b5`, `lm_lung_3b6` and `lm_lung_3b7` are the three best hierarchical subset regression models based on Adjusted R-squared.

```
model_hlm_3b = regsubsets(Y~x1+x2+x3+x1sqr+x2sqr+x3sqr+x1x2+x1x3+x2x3, data=df)
regs = summary(model_hlm_3b)
```

```
results_df = data.frame(regs$which)
results_df$adjr2 = regs$adjr2
results_df[order(results_df$adjr2),]
```

```
##   X.Intercept.    x1   x2    x3 x1sqr x2sqr x3sqr  x1x2  x1x3  x2x3
## 1         TRUE FALSE TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## 2         TRUE FALSE TRUE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE
## 8         TRUE  TRUE TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE
## 7         TRUE  TRUE TRUE  TRUE FALSE FALSE  TRUE  TRUE  TRUE  TRUE
## 6         TRUE  TRUE TRUE FALSE FALSE FALSE  TRUE  TRUE  TRUE  TRUE
## 5         TRUE  TRUE TRUE FALSE FALSE FALSE  TRUE  TRUE FALSE  TRUE
## 3         TRUE  TRUE TRUE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE
## 4         TRUE  TRUE TRUE FALSE  TRUE  TRUE FALSE FALSE FALSE FALSE
##       adjr2
## 1 0.5329124
## 2 0.6857448
## 8 0.6923417
## 7 0.7164219
## 6 0.7354852
## 5 0.7430869
## 3 0.7506631
## 4 0.7506701
```

(c) Is there much difference in R2a,p for the three best subset models? (5pts)

Solution:

`lm_lung_3b5`, `lm_lung_3b6` and `lm_lung_3b7` have an adjusted R^2 of 0.7228, 0.7418, 0.7507 respectively. Thus, they have more or less the same adjusted R^2.

(d) Calculate the PRESS statistic and compare it to SSE. What does this comparison suggest about the validity of MSE as an indicator of the predictive ability of the fitted model? (5pts)

Solution:

We will be using the function PRESS() from `qpcR` library to calculate the PRESS statistic

```
help("PRESS")
```

```
PRESS(lm_lung_3b7)
```

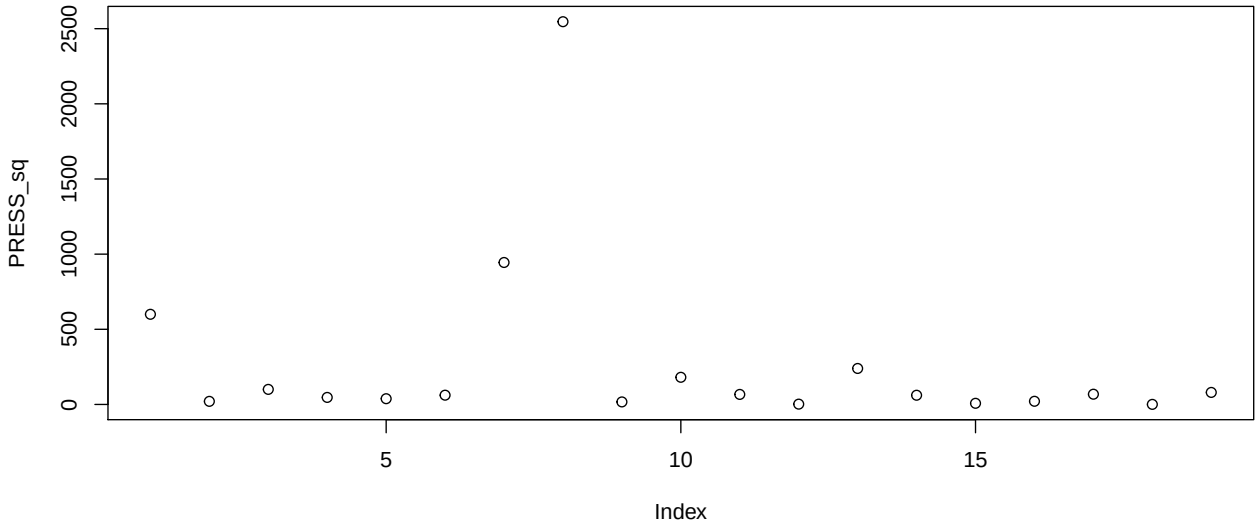
```
## .........10.........
```

```
## $stat
## [1] 5102.494
##
## $residuals
##  [1]  24.4919932   4.5388932 -10.0110553  -6.8391583   6.1822953
##  [6]   7.8544580  30.7316691 -50.4624853   4.1036637 -13.4401289
## [11]  -8.1659387   1.4535580 -15.4726134   7.8289481   2.7623672
## [16]  -4.5912117  -8.2534854  -0.6765019  -8.9428342
##
## $P.square
## [1] 0.3691032
```

```
PRESS_sq = (PRESS(lm_lung_3b7)$residuals)^2
```

```
## .........10.........
```

```r
plot(PRESS_sq)
```




```r
anova(lm_lung_3b7)
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## x1         1 3577.1  3577.1  31.929 4.611e-05 ***
## x2         1 1384.4  1384.4  12.357  0.003124 **
## x1x2       1 1445.8  1445.8  12.905  0.002667 **
## Residuals 15 1680.5   112.0
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

SSE is 1680.5 for the best model chosen above and the PRESS stat is 5102.5. This means that there are a few observations in the data set are significantly driving the model's coefficients.

(e) Case 8 alone accounts for approximately one-half of the entire PRESS statistic. Would you recommend modification of the model because of the strong impact of this case? What are some corrective action options that would lessen the effect of case 8? (5pts)

Solution:

- PRESS statistic for case 8 is ≈ 2500. This clearly indicates that case 8 is an outlier.
- Thus, case 8 should be taken out from the model building data set and the same model can be refitted.

```r
summary(lm(formula = Y ~ x1 + x2 + x1x2, data = df[-8,]))
```

```
##
## Call:
## lm(formula = Y ~ x1 + x2 + x1x2, data = df[-8, ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.2577  -6.0590  -0.7347   3.0510  24.9438
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  35.94847    2.88041  12.480 5.64e-09 ***
```

```
## x1           -0.07454    5.19684  -0.014 0.988758
## x2          -17.30750    3.62782  -4.771 0.000299 ***
## x1x2         16.12087    4.74878   3.395 0.004358 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.894 on 14 degrees of freedom
## Multiple R-squared:  0.8237, Adjusted R-squared:  0.7859
## F-statistic:  21.8 on 3 and 14 DF,  p-value: 1.53e-05
```

We can see a significant improvement in Adjusted R^2.

Question 4 Refer to the Website developer data set. Management is interested in determining what variables have the greatest impact on production output in the release of new customer websites. Data on 13 three-person website developed teams consisting of a project manager, a designer. and a developer are provided in the data set. Production data from January 2001 through August 2002 include four potential predictors; (1) the change in the website development process. (2) the size of the backlog of orders, (3) the team effect, and (4) the number of months experience of each team. (10 pts)

(a) Develop a best subset model for predicting production output. Justify your choice of model. Assess your model's ability to predict and discuss its use as a tool for management decisions. (10 pts)

Solution:

```
website_data = read.csv("Website Developer.csv")
website_data$Process.change = as.factor(website_data$Process.change)
website_data$Team.number = as.factor(website_data$Team.number)
summary(website_data)
```

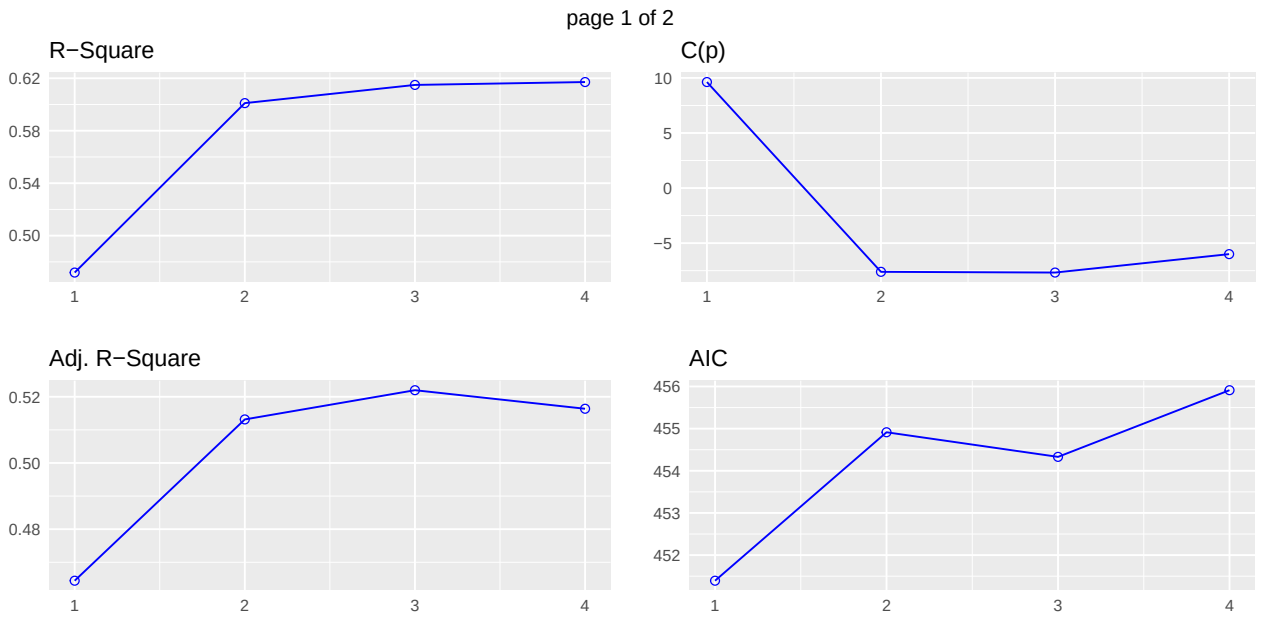
```
##  Websites.delivered Backlog.of.orders  Team.number Team.experience
##  Min.   : 0.000     Min.   : 3.00     1      : 7   Min.   : 2.00
##  1st Qu.: 3.000     1st Qu.:23.00     2      : 7   1st Qu.: 6.00
##  Median : 7.000     Median :28.00     3      : 7   Median :11.00
##  Mean   : 9.041     Mean   :27.82     4      : 7   Mean   :10.85
##  3rd Qu.:13.000     3rd Qu.:34.00     5      : 7   3rd Qu.:15.00
##  Max.   :30.000     Max.   :45.00     6      : 7   Max.   :21.00
##                                       (Other):31
##  Process.change      Year         Quarter
##  0:47           Min.   :2001   Min.   :1.000
##  1:26           1st Qu.:2001   1st Qu.:1.000
##                 Median :2002   Median :2.000
##                 Mean   :2002   Mean   :2.342
##                 3rd Qu.:2002   3rd Qu.:3.000
##                 Max.   :2002   Max.   :4.000
##
```

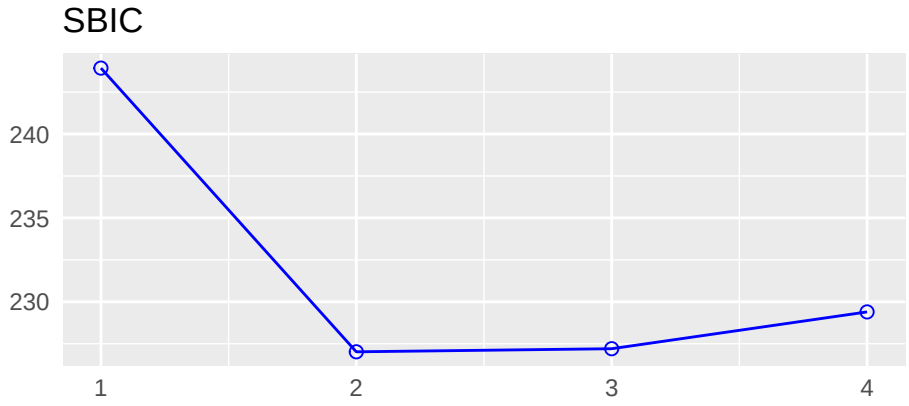
```
model = lm(Websites.delivered~Process.change+Backlog.of.orders+
             Team.number+Team.experience, data=website_data)
summary(model)
```

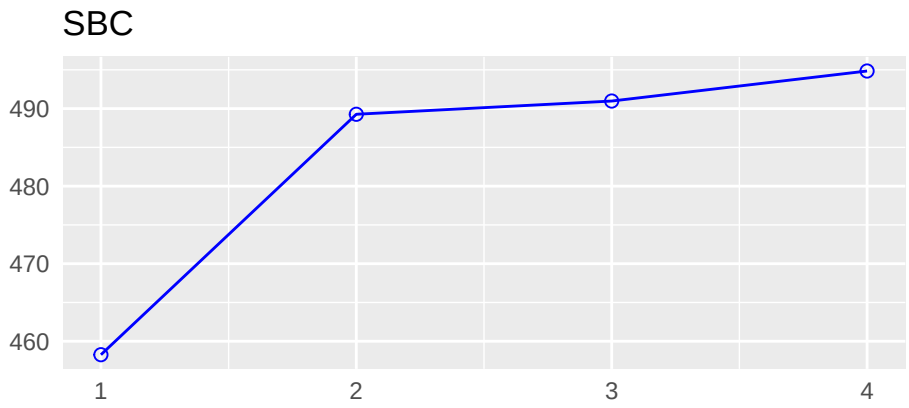
```
##
## Call:
## lm(formula = Websites.delivered ~ Process.change + Backlog.of.orders +
##     Team.number + Team.experience, data = website_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
```

```
## -9.860 -2.906 -0.474  2.951  9.944
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -1.55674    3.08312  -0.505 0.615560
## Process.change1    7.62008    2.05296   3.712 0.000469 ***
## Backlog.of.orders  0.08832    0.12247   0.721 0.473758
## Team.number2       2.61335    2.63689   0.991 0.325840
## Team.number3       1.15547    2.63323   0.439 0.662462
## Team.number4       3.39072    2.63370   1.287 0.203145
## Team.number5       5.85308    2.64020   2.217 0.030634 *
## Team.number6       3.53315    2.64510   1.336 0.186946
## Team.number7       9.53068    3.25722   2.926 0.004923 **
## Team.number8       9.61382    3.24649   2.961 0.004460 **
## Team.number9       7.23480    2.76643   2.615 0.011391 *
## Team.number10      2.19056    2.94036   0.745 0.459334
## Team.number11      5.23963    3.73260   1.404 0.165819
## Team.number12      6.61600    4.75645   1.391 0.169646
## Team.number13      2.51647    2.63883   0.954 0.344297
## Team.experience    0.12982    0.22677   0.572 0.569260
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.926 on 57 degrees of freedom
## Multiple R-squared:  0.6171, Adjusted R-squared:  0.5164
## F-statistic: 6.125 on 15 and 57 DF,  p-value: 2.213e-07
```

```
#Best Subset Regression
k1<-ols_step_best_subset(model, details = FALSE)
plot(k1)
```








```
k1
```

```
##                           Best Subsets Regression
## ---------------------------------------------------------------------------
## Model Index    Predictors
## ---------------------------------------------------------------------------
##      1         Process.change
##      2         Process.change Team.number
##      3         Process.change Backlog.of.orders Team.number
##      4         Process.change Backlog.of.orders Team.number Team.experience
## ---------------------------------------------------------------------------
##
##                                                    Subsets Regression Summary
## ------------------------------------------------------------------------------------------------------------
##                         Adj.        Pred
## Model   R-Square    R-Square   R-Square    C(p)      AIC      SBIC      SBC      MSEP      FPE     HSP
## ------------------------------------------------------------------------------------------------------------
##  1      0.4719     0.4644     0.4358    9.6264   451.3927   243.9312   458.2641   27.6317   27.6106   0.38
##  2      0.6011     0.5132     0.3049   -7.6069   454.9140   227.0186   489.2709   25.4818   25.4333   0.35
##  3      0.6149     0.5220     0.3151   -7.6723   454.3303   227.2028   490.9776   25.3877   25.3007   0.35
##  4      0.6171     0.5164     0.2873   -6.0000   455.9118   229.3962   494.8496   26.0687   25.9298   0.36
## ------------------------------------------------------------------------------------------------------------
## AIC: Akaike Information Criteria
##  SBIC: Sawa's Bayesian Information Criteria
##  SBC: Schwarz Bayesian Criteria
```

```
##  MSEP: Estimated error of prediction, assuming multivariate normality
##  FPE: Final Prediction Error
##  HSP: Hocking's Sp
##  APC: Amemiya Prediction Criteria
```

Interpretation

We can see above that all the model selection criteria point us to model #2 - looking at the plots, they all have the "elbow" at model #2.

```
model_4 = lm(Websites.delivered~Process.change+Team.number, data=website_data)
summary(model_4)
```

```
##
## Call:
## lm(formula = Websites.delivered ~ Process.change + Team.number,
##     data = website_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.7386 -3.3903 -0.3903  3.4536  9.8955
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)         1.962      1.903   1.031  0.30680
## Process.change1     9.634      1.268   7.599 2.61e-10 ***
## Team.number2        2.714      2.642   1.027  0.30843
## Team.number3        1.143      2.642   0.433  0.66689
## Team.number4        3.429      2.642   1.298  0.19942
## Team.number5        5.714      2.642   2.163  0.03461 *
## Team.number6        3.429      2.642   1.298  0.19942
## Team.number7        8.471      3.110   2.724  0.00847 **
## Team.number8        8.971      3.110   2.885  0.00546 **
## Team.number9        6.660      2.750   2.422  0.01855 *
## Team.number10       1.585      2.898   0.547  0.58653
## Team.number11       3.949      3.445   1.146  0.25627
## Team.number12       2.904      4.065   0.714  0.47777
## Team.number13       2.286      2.642   0.865  0.39045
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.943 on 59 degrees of freedom
## Multiple R-squared:  0.6011, Adjusted R-squared:  0.5132
## F-statistic: 6.838 on 13 and 59 DF,  p-value: 8.638e-08
```

```
anova(model_4)
```

```
## Analysis of Variance Table
##
## Response: Websites.delivered
##                 Df  Sum Sq Mean Sq F value    Pr(>F)
## Process.change   1 1704.80 1704.80 69.7848 1.384e-11 ***
## Team.number     12  466.75   38.90  1.5922    0.1189
## Residuals       59 1441.33   24.43
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
calc_mspr <- function(model, df, Y_str){
  yhat = predict(model, df)
  yi = df[,Y_str]

  MSPR = sum((yi-yhat)^2)/nrow(df)
  MSPR
}
```

```r
MSPR = calc_mspr(model=model_4, df=website_data, Y_str=c("Websites.delivered"))
print(MSPR)
```

```
## [1] 19.74427
```

Interpretation:

60% of the variation in `Websites.delivered` is explained by our model (with variables `Process.change` and `Team.number`). We can also see that the MSE and the MSPR are not significantly different. Thus, the model is certainly not a perfect fit for the give data set. However, it does a good enough job of pointing the management in the right direction as to where it should focus its efforts to drive more efficiency.

- First recommendation to the management would be that a change in website development process can significantly improve the production output
- Second, teams 5,7, 8, 9 are doing a good job and are surely the high performers (higher β's). While, team number 3 and 10 are laggind behind.

Question 5 Refer to the Prostate cancer data set. Serum prostate-specific antigen (PSA) was determined in 97 men with advanced prostate cancer. PSA is a well-established screening test for prostate cancer and the oncologists wanted to examine the correlation between level of PSA and a number of clinical measures for men who were about to undergo radical prostatectomy. The measures are cancer volume, prostate weight, patient age, the amount of benign prostatic hyperplasia, seminal vesicle invasion, capsular penetration, and Gleason score. (15 Pts)

(a) Select a random sample of 65 observations to use as the model-building data set. Develop a best subset model for predicting PSA. Justify your choice of model. Assess your model's ability to predict and discuss its usefulness to the oncologists. (5pts)

Solution:

```r
prostate_data = read.csv("Prostate Cancer.csv")
prostate_data$Seminal.vesicle.invasion = as.factor(prostate_data$Seminal.vesicle.invasion)
prostate_data$Gleason.score = as.factor(prostate_data$Gleason.score)
summary(prostate_data)
```

```
##    PSA.level       Cancer.volume       Weight           Age
##  Min.   :  0.651   Min.   : 0.2592   Min.   : 10.70   Min.   :41.00
##  1st Qu.:  5.641   1st Qu.: 1.6653   1st Qu.: 29.37   1st Qu.:60.00
##  Median : 13.330   Median : 4.2631   Median : 37.34   Median :65.00
##  Mean   : 23.730   Mean   : 6.9987   Mean   : 45.49   Mean   :63.87
##  3rd Qu.: 21.328   3rd Qu.: 8.4149   3rd Qu.: 48.42   3rd Qu.:68.00
##  Max.   :265.072   Max.   :45.6042   Max.   :450.34   Max.   :79.00
##  Benign.prostatic.hyperplasia Seminal.vesicle.invasion
##  Min.   : 0.000               0:76
##  1st Qu.: 0.000               1:21
##  Median : 1.350
##  Mean   : 2.535
##  3rd Qu.: 4.759
##  Max.   :10.278
##  Capsular.penetration Gleason.score
```

```
##  Min.   : 0.0000      6:33
##  1st Qu.: 0.0000      7:43
##  Median : 0.4493      8:21
##  Mean   : 2.2454
##  3rd Qu.: 3.2544
##  Max.   :18.1741
```

```
set.seed(1234)
train_ind = sample(1:nrow(prostate_data), 65)
test_ind = setdiff(1:nrow(prostate_data), train_ind)
train_df = prostate_data[train_ind,]
test_df = prostate_data[test_ind,]
```

```
exc_cols = c("Seminal.vesicle.invasion","Gleason.score")
cor(prostate_data[,-which(names(prostate_data) %in% exc_cols)])
```

```
##                               PSA.level Cancer.volume      Weight
## PSA.level                     1.00000000   0.624150588 0.026213430
## Cancer.volume                 0.62415059   1.000000000 0.005107148
## Weight                        0.02621343   0.005107148 1.000000000
## Age                           0.01719938   0.039094423 0.164323714
## Benign.prostatic.hyperplasia -0.01648649  -0.133209431 0.321848748
## Capsular.penetration          0.55079252   0.692896688 0.001578905
##                                     Age Benign.prostatic.hyperplasia
## PSA.level                     0.01719938                   -0.01648649
## Cancer.volume                 0.03909442                   -0.13320943
## Weight                        0.16432371                    0.32184875
## Age                           1.00000000                    0.36634121
## Benign.prostatic.hyperplasia 0.36634121                    1.00000000
## Capsular.penetration          0.09955535                   -0.08300865
##                               Capsular.penetration
## PSA.level                            0.550792517
## Cancer.volume                        0.692896688
## Weight                               0.001578905
## Age                                  0.099555351
## Benign.prostatic.hyperplasia        -0.083008649
## Capsular.penetration                 1.000000000
```

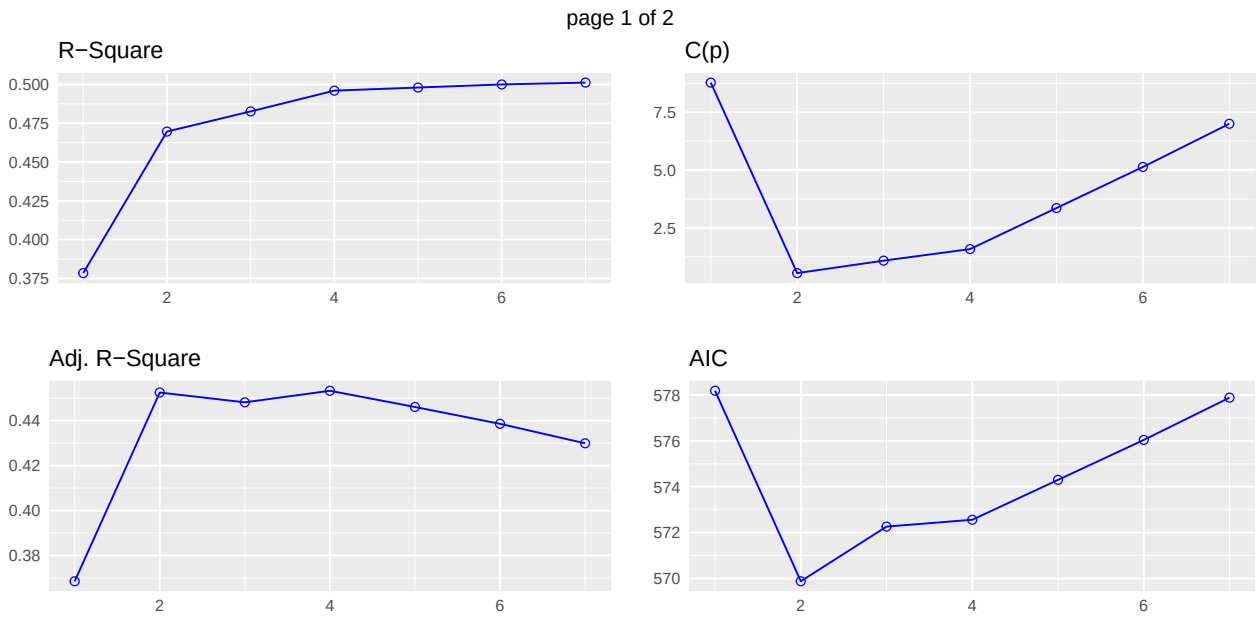
We can see that x6 and x1 are positively correlation with $r = 0.69$. However, Y has a higher correlation with x1. So we will take out x6.

```
model = lm(PSA.level~., data=train_df)
summary(model)
```

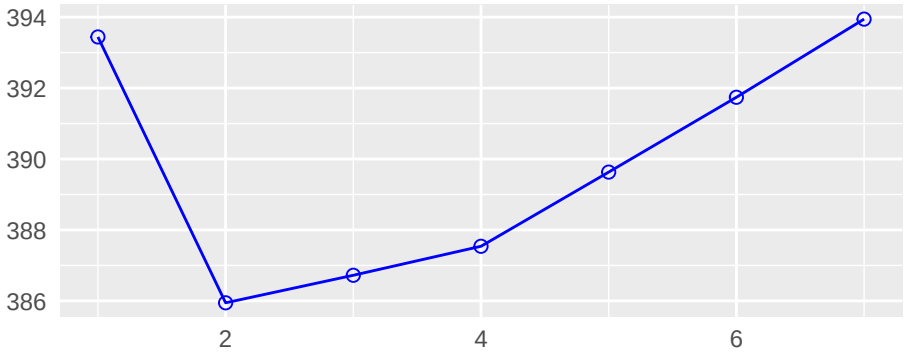
```
##
## Call:
## lm(formula = PSA.level ~ ., data = train_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -38.660  -5.811   0.158   4.343 105.314
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   33.37306   22.10793   1.510  0.13678
## Cancer.volume                  1.63231    0.51239   3.186  0.00236 **
```

```
## Weight                         0.01718    0.04703   0.365  0.71628
## Age                           -0.49894    0.36839  -1.354  0.18105
## Benign.prostatic.hyperplasia   0.35149    0.94596   0.372  0.71161
## Seminal.vesicle.invasion1     22.29067    8.33682   2.674  0.00981 **
## Capsular.penetration          -0.48942    0.99744  -0.491  0.62557
## Gleason.score7                 1.36416    6.09161   0.224  0.82362
## Gleason.score8                10.93274    8.45208   1.293  0.20115
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.05 on 56 degrees of freedom
## Multiple R-squared:  0.5012, Adjusted R-squared:  0.4299
## F-statistic: 7.033 on 8 and 56 DF,  p-value: 2.191e-06
```

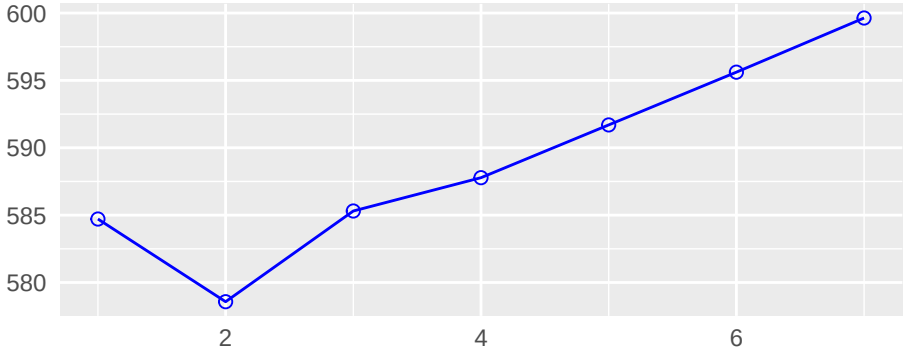
```
k1<-ols_step_best_subset(model, details = FALSE)
plot(k1)
```



SBIC



SBC




```
k1
```

```
##                                                 Best Subsets Regression
## ---------------------------------------------------------------------------------------------------------
## Model Index    Predictors
## ---------------------------------------------------------------------------------------------------------
##      1         Cancer.volume
##      2         Cancer.volume Seminal.vesicle.invasion
##      3         Cancer.volume Seminal.vesicle.invasion Gleason.score
##      4         Cancer.volume Age Seminal.vesicle.invasion Gleason.score
##      5         Cancer.volume Weight Age Seminal.vesicle.invasion Gleason.score
##    6       Cancer.volume Age Benign.prostatic.hyperplasia Seminal.vesicle.invasion Capsular.penetration
##    7       Cancer.volume Weight Age Benign.prostatic.hyperplasia Seminal.vesicle.invasion Capsular.pene
## ---------------------------------------------------------------------------------------------------------
##
##                                                 Subsets Regression Summary
## ---------------------------------------------------------------------------------------------------------
##                         Adj.        Pred
## Model   R-Square   R-Square   R-Square    C(p)     AIC      SBIC      SBC      MSEP      FPE      HSP
## ---------------------------------------------------------------------------------------------------------
##  1      0.3784     0.3686     0.2799   8.7804   578.1870   393.4416   584.7102   414.6855   414.2866   6.4
##  2      0.4696     0.4525      0.347   0.5439   569.8754   385.9464   578.5730   365.4545   364.5755   5.7
##  3      0.4826     0.4481     0.3346   1.0829   572.2607   386.7230   585.3070   374.5100   372.8888   5.8
##  4      0.4960     0.4533      0.285   1.5817   572.5587   387.5402   587.7794   377.3026   374.7619   5.8
##  5      0.4980     0.4461     0.2196   3.3569   574.2999   389.6312   591.6950   388.8740   385.1330   6.0
```

```
##  6     0.5000    0.4386    0.2073   5.1334   576.0416   391.7411   595.6111   401.0417   395.8333   6.1
##  7     0.5012    0.4299    0.1349   7.0000   577.8869   393.9447   599.6308   414.5049   407.5267   6.4
## ----------------------------------------------------------------------------------------------------------
## AIC: Akaike Information Criteria
##  SBIC: Sawa's Bayesian Information Criteria
##  SBC: Schwarz Bayesian Criteria
##  MSEP: Estimated error of prediction, assuming multivariate normality
##  FPE: Final Prediction Error
##  HSP: Hocking's Sp
##  APC: Amemiya Prediction Criteria
```

```
model_train = lm(PSA.level~Cancer.volume + Seminal.vesicle.invasion, data=train_df)
summary(model_train)
```

```
##
## Call:
## lm(formula = PSA.level ~ Cancer.volume + Seminal.vesicle.invasion,
##     data = train_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -40.270  -4.870  -0.913   4.822 111.982
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 4.5789     3.2473   1.410  0.16352
## Cancer.volume               1.7853     0.4103   4.351 5.16e-05 ***
## Seminal.vesicle.invasion1  21.3832     6.5497   3.265  0.00179 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.67 on 62 degrees of freedom
## Multiple R-squared:  0.4696, Adjusted R-squared:  0.4525
## F-statistic: 27.45 on 2 and 62 DF,  p-value: 2.901e-09
```

```
anova(model_train)
```

```
## Analysis of Variance Table
##
## Response: PSA.level
##                           Df  Sum Sq Mean Sq F value   Pr(>F)
## Cancer.volume              1 15415.5 15415.5  44.235 8.61e-09 ***
## Seminal.vesicle.invasion   1  3714.5  3714.5  10.659 0.001786 **
## Residuals                 62 21606.5   348.5
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
MSPR = calc_mspr(model=model_train, df=train_df, Y_str="PSA.level")
print(MSPR)
```

```
## [1] 332.4071
```

Interpretation:

We can see that, based on the plots and the above summary, model #2 is where the "elbow" occurs for various measures.

43% of the variation in PSA.level is explained by the variables "Cancer Volume" and "Seminal vesicle

invasion". This could be significant considering the nature of the problem we are trying to solve since any correlation can lead the oncologists early detection of the ailment.

- The oncologists would be able predict the PSA based on the volume of cancer and if there was a seminal vescicle invasion.
- We can see that both these variables are postively correlated to the PSA level, which means the oncologists can determine early if a patient is in the "high-risk" zone and possibly start early treatment.

(b) Fit the regression model identified in part a to the validation data set. Compare the estimated regression coefficients and their estimated standard errors with those obtained in part a. Also compare the error mean square and coefficients of multiple determination. Does the model fitted to the validation data set yield similar estimates as the model fitted to the model-building data set? (5pts)

```
summary(model_train)
```

```
##
## Call:
## lm(formula = PSA.level ~ Cancer.volume + Seminal.vesicle.invasion,
##     data = train_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -40.270  -4.870  -0.913   4.822 111.982
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  4.5789     3.2473   1.410  0.16352
## Cancer.volume                1.7853     0.4103   4.351 5.16e-05 ***
## Seminal.vesicle.invasion1   21.3832     6.5497   3.265  0.00179 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.67 on 62 degrees of freedom
## Multiple R-squared:  0.4696, Adjusted R-squared:  0.4525
## F-statistic: 27.45 on 2 and 62 DF,  p-value: 2.901e-09
```

```
model_test = lm(PSA.level~Cancer.volume + Seminal.vesicle.invasion, data=test_df)
summary(model_test)
```

```
##
## Call:
## lm(formula = PSA.level ~ Cancer.volume + Seminal.vesicle.invasion,
##     data = test_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -68.641 -20.885  -0.658   4.821 157.758
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 -0.7378    11.0479  -0.067   0.9472
## Cancer.volume                2.7925     1.2432   2.246   0.0325 *
## Seminal.vesicle.invasion1   33.0803    28.7479   1.151   0.2593
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 47.64 on 29 degrees of freedom
## Multiple R-squared:  0.434,  Adjusted R-squared:  0.395
## F-statistic: 11.12 on 2 and 29 DF,  p-value: 0.0002604
```

```
confint(model_train)
```

```
##                                2.5 %    97.5 %
## (Intercept)               -1.9123620 11.070220
## Cancer.volume              0.9650299  2.605527
## Seminal.vesicle.invasion1  8.2906122 34.475831
```

```
confint(model_test)
```

```
##                                 2.5 %    97.5 %
## (Intercept)               -23.3333645 21.857723
## Cancer.volume               0.2498822  5.335095
## Seminal.vesicle.invasion1 -25.7156727 91.876331
```

```
anova(model_train)
```

```
## Analysis of Variance Table
##
## Response: PSA.level
##                           Df  Sum Sq Mean Sq F value   Pr(>F)
## Cancer.volume              1 15415.5 15415.5  44.235 8.61e-09 ***
## Seminal.vesicle.invasion  1  3714.5  3714.5  10.659 0.001786 **
## Residuals                 62 21606.5   348.5
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(model_test)
```

```
## Analysis of Variance Table
##
## Response: PSA.level
##                           Df Sum Sq Mean Sq F value    Pr(>F)
## Cancer.volume              1  47459   47459 20.9147 8.291e-05 ***
## Seminal.vesicle.invasion  1   3005    3005  1.3241    0.2593
## Residuals                 29  65805    2269
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interpretation

Comparing Coeficients and their s(b):

- We can see that all the three β s for the validation model have significantly higher standard errors compared to those for the ones in training model. We can also see this in the 95% confidence interval.
- Also, there is some material differences in the values for the Intercept and the β for `Seminal.vesicle.invasion`.

Comparing MSE and R-squared:

- MSE for the validation model is higher that the one for training model.
- Also, multipleR^2 for validation model is lower than that for the training model.
- This means that the training model is a better fit to the training data compared to validation model is to the validation data.
- The above effect is mostly due to the fact that there are more data points in training data set compare to the validation data set.

(c) Calculate the mean squared prediction error (equation 9.20 on the book) and compare it to MSE obtained from the model-building data set. Is there evidence of a substantial bias problem in MSE here? (5pts)

```
MSPR = calc_mspr(model=model_train, df=train_df, Y_str="PSA.level")
print(MSPR)
```

```
## [1] 332.4071
```

Interpretation:

We can see that the MSPR obtained in part(c) is not significantly different than MSE in part(a). Thus, we can say that there is no evidence of substantial bias problem here.

Question 6 Refer to Market share data set. Company executives want to be able to predict market share of their product (Y) based on merchandise price (X1), the gross Nielsen rating points (X2, an index of the amount of advertising exposure that the product received); the presence or absence of a wholesale pricing discount (X3 = 1 if discount present: otherwise X3 = 0); the presence or absence of a package promotion during the period (X4 = 1 if promotion present: otherwise X4 = 0): and year (X5). Code year as a nominal level variable and use 2000 as the referent year. (20 pts)

(a) Using only first-order terms for predictor variables, find the three best subset regression models according to the SBCp criterion. (7 pts)

Solution:

```
market_data = read.csv("Market Share.csv")
market_data$Discount.Price = as.factor(market_data$Discount.Price)
market_data$Package.Promotion = as.factor(market_data$Package.Promotion)
market_data$Year = as.factor(market_data$Year)
summary(market_data)
```

```
##   Market.Share       Price       Gross.Nielsen.Rating.Points
##  Min.   :2.230   Min.   :2.124   Min.   : 72.0
##  1st Qu.:2.473   1st Qu.:2.200   1st Qu.:268.0
##  Median :2.640   Median :2.280   Median :412.0
##  Mean   :2.664   Mean   :2.324   Mean   :388.1
##  3rd Qu.:2.880   3rd Qu.:2.420   3rd Qu.:499.5
##  Max.   :3.160   Max.   :2.781   Max.   :858.0
##
##  Discount.Price Package.Promotion     Month      Year
##  0:15           0:16               Apr    : 3   1999: 4
##  1:21           1:20               Aug    : 3   2000:12
##                                    Dec    : 3   2001:12
##                                    Feb    : 3   2002: 8
##                                    Jan    : 3
##                                    Jul    : 3
##                                    (Other):18
```

```
reg1 = regsubsets(Market.Share~Price+Gross.Nielsen.Rating.Points+Discount.Price
                  +Package.Promotion+Year, data=market_data)
res.sum = summary(reg1)
res.sum$which
```

```
##   (Intercept) Price Gross.Nielsen.Rating.Points Discount.Price1
## 1        TRUE FALSE                       FALSE            TRUE
## 2        TRUE FALSE                       FALSE            TRUE
## 3        TRUE  TRUE                       FALSE            TRUE
## 4        TRUE  TRUE                       FALSE            TRUE
## 5        TRUE  TRUE                       FALSE            TRUE
```

```
## 6        TRUE  TRUE                        TRUE            TRUE
## 7        TRUE  TRUE                        TRUE            TRUE
##   Package.Promotion1 Year2000 Year2001 Year2002
## 1              FALSE    FALSE    FALSE    FALSE
## 2               TRUE    FALSE    FALSE    FALSE
## 3               TRUE    FALSE    FALSE    FALSE
## 4               TRUE    FALSE     TRUE    FALSE
## 5               TRUE    FALSE     TRUE     TRUE
## 6               TRUE    FALSE     TRUE     TRUE
## 7               TRUE     TRUE     TRUE     TRUE
```

```
res.sum$bic
```

```
## [1] -28.16723 -28.13423 -29.79864 -28.16656 -25.65112 -22.36766 -18.81016
```

```
order(res.sum$bic)
```

```
## [1] 3 1 4 2 5 6 7
```

Price, Discount and Promotion have the lowest SBC value.

(b) Using forward stepwise regression, find the best subset of predictor variables to predict market share of their product. Use limits of 0.10 and .15 for adding or deleting a predictor, respectively. (7pts)

Solution:

```
model = lm(Market.Share~Price+Gross.Nielsen.Rating.Points+Discount.Price
                  +Package.Promotion+Year, data=market_data)
summary(model)
```

```
##
## Call:
## lm(formula = Market.Share ~ Price + Gross.Nielsen.Rating.Points +
##     Discount.Price + Package.Promotion + Year, data = market_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.33558 -0.11872  0.02459  0.08020  0.21952
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  3.034e+00  4.921e-01   6.166 1.17e-06 ***
## Price                       -2.470e-01  1.982e-01  -1.246   0.2229
## Gross.Nielsen.Rating.Points -9.653e-05  1.914e-04  -0.504   0.6181
## Discount.Price1              4.093e-01  5.385e-02   7.601 2.80e-08 ***
## Package.Promotion1           1.240e-01  5.484e-02   2.261   0.0317 *
## Year2000                    -1.324e-02  9.304e-02  -0.142   0.8879
## Year2001                    -1.220e-01  9.950e-02  -1.226   0.2303
## Year2002                    -9.630e-02  1.100e-01  -0.876   0.3887
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1529 on 28 degrees of freedom
## Multiple R-squared:  0.7326, Adjusted R-squared:  0.6657
## F-statistic: 10.96 on 7 and 28 DF,  p-value: 1.382e-06
```

```
help("ols_step_forward_p")
k2<-ols_step_forward_p(model, pent=0.1, prem=0.15, details = TRUE)
```

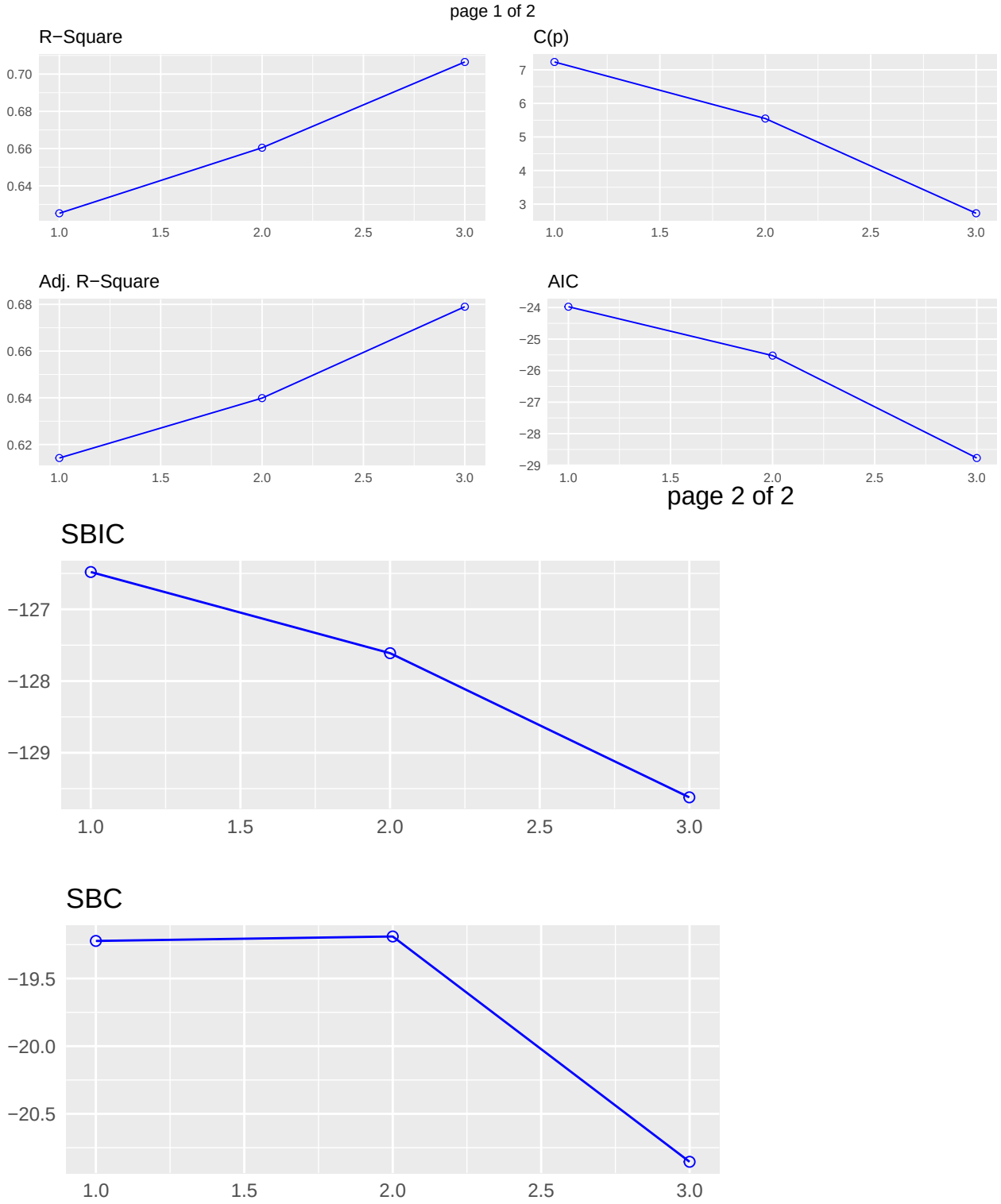
```
## Forward Selection Method
## --------------------------
##
## Candidate Terms:
##
## 1. Price
## 2. Gross.Nielsen.Rating.Points
## 3. Discount.Price
## 4. Package.Promotion
## 5. Year
##
## We are selecting variables based on p value...
##
##
## Forward Selection: Step 1
##
## - Discount.Price
##
##                         Model Summary
## ----------------------------------------------------------------
## R                       0.791       RMSE                0.164
## R-Squared               0.625       Coef. Var           6.164
## Adj. R-Squared          0.614       MSE                 0.027
## Pred R-Squared          0.584       MAE                 0.134
## ----------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                               ANOVA
## -------------------------------------------------------------------
##                Sum of
##               Squares        DF    Mean Square      F         Sig.
## -------------------------------------------------------------------
## Regression      1.530         1          1.530    56.728    0.0000
## Residual        0.917        34          0.027
## Total           2.446        35
## -------------------------------------------------------------------
##
##                                  Parameter Estimates
## ----------------------------------------------------------------------------------------
##           model     Beta    Std. Error    Std. Beta      t        Sig     lower    upper
## ----------------------------------------------------------------------------------------
##     (Intercept)    2.420         0.042                57.080    0.000    2.334    2.506
## Discount.Price1    0.418         0.056        0.791    7.532    0.000    0.305    0.531
## ----------------------------------------------------------------------------------------
##
##
##
## Forward Selection: Step 2
##
## - Package.Promotion
##
##                         Model Summary
```

```
## ----------------------------------------------------------------
## R                       0.813       RMSE                 0.159
## R-Squared               0.660       Coef. Var            5.956
## Adj. R-Squared          0.640       MSE                  0.025
## Pred R-Squared          0.600       MAE                  0.123
## ----------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                                ANOVA
## --------------------------------------------------------------------
##                Sum of
##               Squares        DF    Mean Square      F         Sig.
## --------------------------------------------------------------------
## Regression      1.616         2          0.808    32.094    0.0000
## Residual        0.831        33          0.025
## Total           2.446        35
## --------------------------------------------------------------------
##
##                                    Parameter Estimates
## ----------------------------------------------------------------------------------------------
##             model     Beta    Std. Error    Std. Beta      t        Sig      lower    upper
## ----------------------------------------------------------------------------------------------
##       (Intercept)    2.374        0.048                 49.394    0.000     2.276    2.471
##    Discount.Price1    0.403        0.054        0.762     7.426    0.000     0.293    0.513
## Package.Promotion1    0.100        0.054        0.190     1.849    0.073    -0.010    0.209
## ----------------------------------------------------------------------------------------------
##
##
##
## Forward Selection: Step 3
##
## - Price
##
##                          Model Summary
## ----------------------------------------------------------------
## R                       0.841       RMSE                 0.150
## R-Squared               0.707       Coef. Var            5.623
## Adj. R-Squared          0.679       MSE                  0.022
## Pred R-Squared          0.637       MAE                  0.118
## ----------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                                ANOVA
## --------------------------------------------------------------------
##                Sum of
##               Squares        DF    Mean Square      F         Sig.
## --------------------------------------------------------------------
## Regression      1.728         3          0.576    25.677    0.0000
## Residual        0.718        32          0.022
## Total           2.446        35
```

```
## ----------------------------------------------------------------
##
##                                 Parameter Estimates
## ----------------------------------------------------------------------------------------
##             model      Beta    Std. Error    Std. Beta      t       Sig      lower     upper
## ----------------------------------------------------------------------------------------
##       (Intercept)     3.185         0.365                8.726    0.000     2.442     3.929
##    Discount.Price1     0.399         0.051        0.755    7.787    0.000     0.295     0.504
## Package.Promotion1     0.118         0.051        0.225    2.292    0.029     0.013     0.223
##             Price    -0.353         0.157       -0.217   -2.241    0.032    -0.673    -0.032
## ----------------------------------------------------------------------------------------
##
##
##
## No more variables to be added.
##
## Variables Entered:
##
## + Discount.Price
## + Package.Promotion
## + Price
##
##
## Final Model Output
## ------------------
##
##                          Model Summary
## ---------------------------------------------------------------
## R                         0.841       RMSE                0.150
## R-Squared                 0.707       Coef. Var           5.623
## Adj. R-Squared            0.679       MSE                 0.022
## Pred R-Squared            0.637       MAE                 0.118
## ---------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                                ANOVA
## ----------------------------------------------------------------
##                Sum of
##               Squares        DF    Mean Square      F         Sig.
## ----------------------------------------------------------------
## Regression      1.728         3          0.576    25.677    0.0000
## Residual        0.718        32          0.022
## Total           2.446        35
## ----------------------------------------------------------------
##
##                                 Parameter Estimates
## ----------------------------------------------------------------------------------------
##             model      Beta    Std. Error    Std. Beta      t       Sig      lower     upper
## ----------------------------------------------------------------------------------------
##       (Intercept)     3.185         0.365                8.726    0.000     2.442     3.929
##    Discount.Price1     0.399         0.051        0.755    7.787    0.000     0.295     0.504
## Package.Promotion1     0.118         0.051        0.225    2.292    0.029     0.013     0.223
```

```
##             Price    -0.353        0.157       -0.217    -2.241    0.032    -0.673    -0.032
## ------------------------------------------------------------------------------------------------
```

```
plot(k2)
```



(c) How does the best subset according to forward stepwise regression compare with the best subset according to the SBCp criterion used in part a? (6pts)

```
k2$predictors
```

```
## [1] "Discount.Price"    "Package.Promotion" "Price"
```

We see that the best subset according to forward stepwise regression is same as the best subset according to the SBCp criterion.