HW7-Solutions

Problem 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)

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

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)

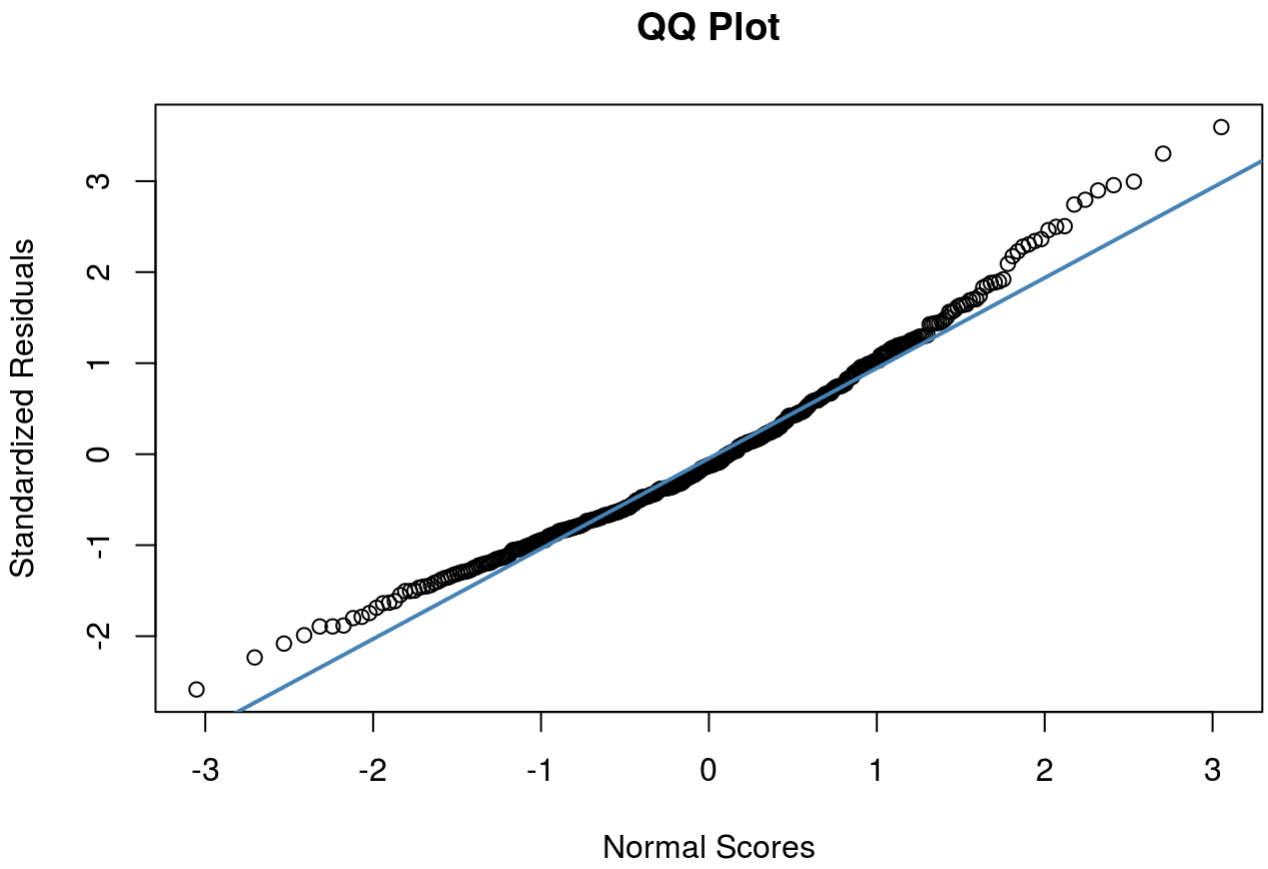
a)

_Solution: See below for the regression model coefficents, the R-Square is 24% and indicating a poor fit. However, the QQ plot does NOT indicate any problem with normality assumptions.

```
library(knitr)
CDI <- read.csv("/cloud/project/CDI.csv")
#SCR=serious crime rate
#PD=population density
#UR=unemployment rate
SCR <-CDI$Total.serious.crimes/CDI$Total.population
PD<-CDI$Total.population/CDI$Land.area
UR<-CDI$Percent.unemployment
PD1<-PD-mean(PD)
PD2<-PD1^2
UR1<-UR-mean(UR)
UR2<-UR1^2
PD.UR<-PD1*UR1
f1a<-lm(SCR~PD1+UR1+PD2+UR2+PD.UR)
summary(f1a)
```

```
##
## Call:
## lm(formula = SCR ~ PD1 + UR1 + PD2 + UR2 + PD.UR)
##
## 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)  5.629e-02  1.260e-03  44.662  < 2e-16 ***
## PD1          4.585e-06  9.841e-07   4.659 4.23e-06 ***
## UR1         -8.800e-05  6.276e-04  -0.140   0.8886
## PD2          2.698e-12  5.932e-11   0.045   0.9637
## UR2          1.629e-04  9.541e-05   1.708   0.0884 .
## PD.UR        8.334e-07  4.091e-07   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
```

```
stdei<- rstandard(f1a)
qqnorm(stdei,ylab="Standardized Residuals",xlab="Normal Scores", main="QQ Plot")
qqline(stdei,col = "steelblue", lwd = 2)
```



b)

Solution:

$$H_o : \beta_{11} = \beta_{22} = \beta_{12} = 0$$
$$Ha : At\ least\ one\ term\ is\ different\ than\ zero.$$

P value is 0.02278>0.01, Accept Null, the terms can be dropped. See below for the Rcode

```
f1aR<-lm(SCR~PD1+UR1)
anova(f1aR,f1a)
```

```
## Analysis of Variance Table
##
## Model 1: SCR ~ PD1 + UR1
## Model 2: SCR ~ PD1 + UR1 + PD2 + UR2 + PD.UR
##   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
```

c)

Solution: See below for the regression model coefficents, the R-Square is 14% compared to 24% in part a. The model performance is weaker than part a.

```
X1<-CDI$Total.population
X2<-CDI$Land.area
X3<-CDI$Percent.unemployment
X11<-X1-mean(X1)
X12<-X11^2
f1c<-lm(SCR~X11+X12+X2+X3)
summary(f1c)
```

```
##
## Call:
## lm(formula = SCR ~ X11 + X12 + X2 + X3)
##
## 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)  5.458e-02  3.631e-03  15.031  < 2e-16 ***
## X11          2.942e-08  3.555e-09   8.276 1.57e-15 ***
## X12         -3.356e-15  5.878e-16  -5.710 2.10e-08 ***
## X2          -5.576e-07  8.123e-07  -0.687    0.493
## X3           6.824e-04  5.302e-04   1.287    0.199
## ---
## 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
```

Problem 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)

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)

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

a)

Solution: See below for the rcode. R square is 90%. X2 and X5 are significnat at 95% confidence level (alpha=0.05).

```
CDI <- read.csv("/cloud/project/CDI.csv")
Y<-CDI$Number.of.active.physicians
X1<-CDI$Total.population
X2<-CDI$Total.personal.income
X3 <- as.numeric(CDI$Geographic.region == 1)
X4 <- as.numeric(CDI$Geographic.region == 2)
X5 <- as.numeric(CDI$Geographic.region == 3)
f2a<-lm(Y~X1+X2+X3+X4+X5)
summary(f2a)
```

```
##
## 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)

Solution: X3 and X4 are not significant. Confindence for all betas also show that X3 and X4 are not signficiant. I also calcualte the confidence interval for b3-b4, which covers zero, indicating that there is no difference.

```
confint(f2a)
```

```
##                     2.5 %        97.5 %
## (Intercept) -3.456304e+02 -69.361106971
## X1          -5.828310e-06   0.001108748
## X2           8.095958e-02   0.133063482
## X3          -2.164623e+01 319.685375829
## X4          -2.183695e+01 312.889843723
## X5           3.391581e+01 348.516796471
```

```
X<-model.matrix(f2a)
XXInv<-solve(t(X)%*%X)
at<-anova(f2a)
MSE<-at$`Mean Sq`[6]
Var<-MSE*(XXInv)
Varb34=Var[4,4]+Var[5,5]-2*Var[4,5]
cbind((1.490e+02-1.455e+02)-sqrt(Varb34)*qt(.95,434),(1.490e+02-1.455e+02)+sqrt(Varb34)*qt(.95,434))
```

```
##           [,1]     [,2]
## [1,] -126.4089 133.4089
```

c)

Solution:

$$H_o : \beta_3 = \beta_4 = \beta_5 = 0$$
$$Ha : At\ least\ one\ term\ is\ different\ than\ zero.$$

#P value is 0.121>0.10, Accept Null, the terms can be dropped. No geographic effects. See below for the Rcode

```
f2aR<-lm(Y~X1+X2)
anova(f2aR,f2a)
```

```
## 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
```

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

a)

Solution: Only X2 is significnat and the R square is 61%.

```
Lung.Pressure <- read.csv("/cloud/project/Lung Pressure.csv")
f3a<-lm(Y~X1+X2+X3,data=Lung.Pressure)
summary(f3a)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + X3, data = Lung.Pressure)
##
## 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
```

b)

Solution: There are different ways to solve this problem. You can use the best subset algorithms. Alternatively, get the all possible models and then slect the models with the highest adjusted R square. I used the latter approach, please see the code below.

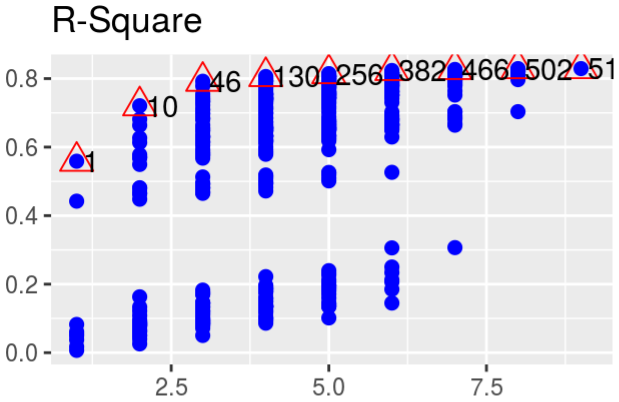
```
Y<-Lung.Pressure$Y
X1<-Lung.Pressure$X1
X2<-Lung.Pressure$X2
X3<-Lung.Pressure$X3
Xi1<-X1-mean(X1)
Xi2<-X2-mean(X2)
Xi3<-X3-mean(X3)
X11<-Xi1^2
X22<-Xi2^2
X33<-Xi3^2
X12<-Xi1*Xi2
X13<-Xi1*Xi3
X23<-Xi2*Xi3
library(olsrr)
```

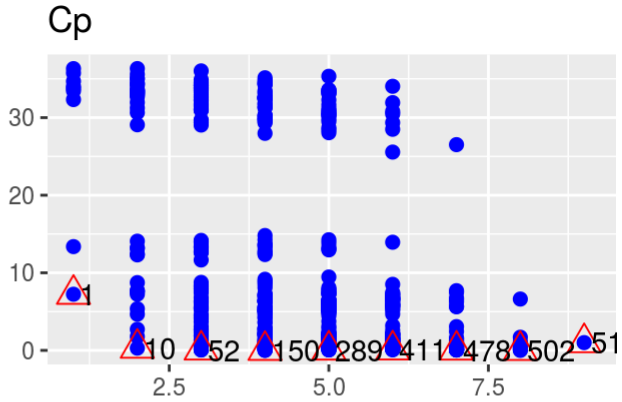
```
##
## Attaching package: 'olsrr'
```

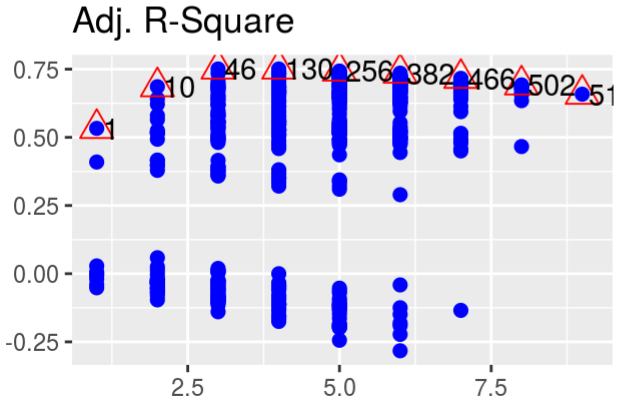
```
## The following object is masked from 'package:datasets':
##
##     rivers
```

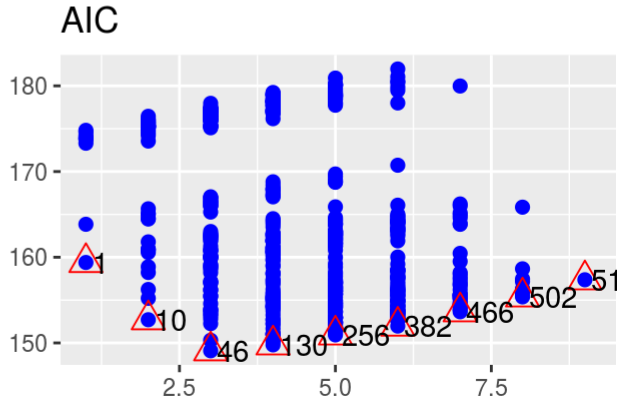
```
q3<-data.frame(cbind(Y,Xi1,Xi2,Xi3,X11,X22,X33,X12,X13,X23))
f3b<-lm(Y ~Xi1+Xi2+Xi3+X11+X22+X33+X12+X13+X23,data=q3)
res<- ols_step_all_possible(f3b)
plot(res)
```

page 1 of 2

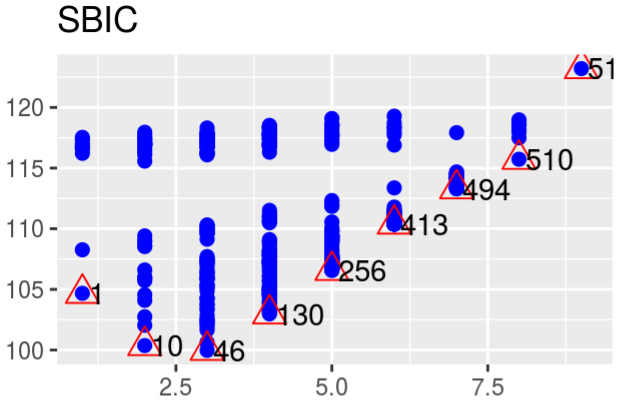


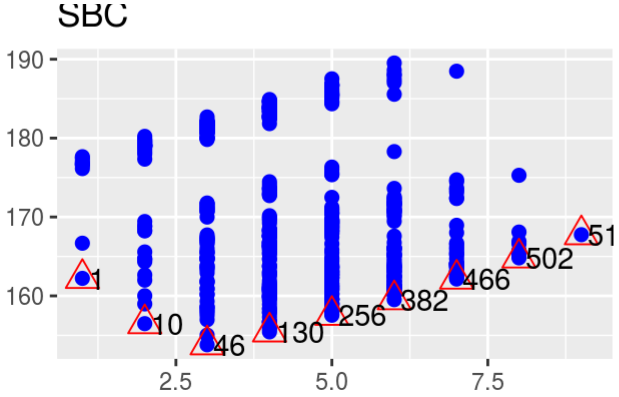






page 2 of 2






```
res1<-data.frame(cbind(res$adjr,1:length(res$adjr)))
res2<-res1[order(res1[,1],decreasing = TRUE),]
res[res2[1:3,2],]
```

```
## # A tibble: 3 x 6
##   Index     N Predictors      `R-Square` `Adj. R-Square` `Mallow's Cp`
##   <int> <int> <chr>                <dbl>           <dbl>         <dbl>
## 1   130     4 Xi1 Xi2 X11 X22      0.806           0.751        1.21
## 2    46     3 Xi1 Xi2 X12          0.792           0.751       -0.0561
## 3   131     4 Xi1 Xi2 X11 X23      0.802           0.746        1.41
```

c)

Solution: They are very close to each other.

d)

Solution: MSE=153.6, PRESS=15908.91, press[8]=1587. MSE is not reliable, PRESS indicates that there are influential or outlier observations.

```
library(qpcR)
```

```
## Loading required package: MASS
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:olsrr':
##
##     cement
```

```
## Loading required package: minpack.lm
```

```
## Loading required package: rgl
```

```
## Warning in rgl.init(initValue, onlyNULL): RGL: unable to open X11 display
```

```
## Warning: 'rgl_init' failed, running with rgl.useNULL = TRUE
```

```
## Loading required package: robustbase
```

```
## Loading required package: Matrix
```

```
p<-PRESS(f3b)
```

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

```
p1<-p$residuals^2
p1
```

```
##  [1] 1156.0969911    0.1984776  225.4453320  230.5551592 2240.4964663
##  [6]  414.9324620 1587.0379921 4856.1047714   84.5753684  322.6878954
## [11]   23.4878176    0.4086422  508.0897845  671.8597983 3318.1509096
## [16]    3.1415396   71.8941418    0.8954438  192.8555491
```

```
sum(p1)
```

```
## [1] 15908.91
```

```
anova(f3b)
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## Xi1        1 3577.1  3577.1 23.2954 0.0009384 ***
## Xi2        1 1384.4  1384.4  9.0156 0.0148934 *
## Xi3        1    5.2     5.2  0.0340 0.8578897
## X11        1 1338.9  1338.9  8.7196 0.0161447 *
## X22        1  221.6   221.6  1.4429 0.2603337
## X33        1   34.7    34.7  0.2262 0.6457225
## X12        1    2.4     2.4  0.0159 0.9024954
## X13        1    4.9     4.9  0.0317 0.8626981
## X23        1  136.5   136.5  0.8891 0.3703534
## Residuals  9 1382.0   153.6
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Problem 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)

a)

Solution: Please see the best model below with, I used the regsubset library, you could use olssr library as well. The model has the adjusted squre of 54%.

```
Website.Developer <- read.csv("/cloud/project/Website Developer.csv")
f4<-lm(Websites.delivered~Process.change+Backlog.of.orders+Team.experience+factor(Team.number),data=Website.Developer )
#library(olsrr)
#Best Subset Regression
#k4<-ols_step_best_subset(f4, details = FALSE)
#plot(k4)
library(leaps)
b <- regsubsets(Websites.delivered~Process.change+Backlog.of.orders+Team.experience+factor(Team.number),data=Website.Develop
er )
rs <- summary(b)
rs$adjr2
```

```
## [1] 0.4644287 0.4902534 0.5148451 0.5333350 0.5485836 0.5519415 0.5498679
## [8] 0.5473631
```

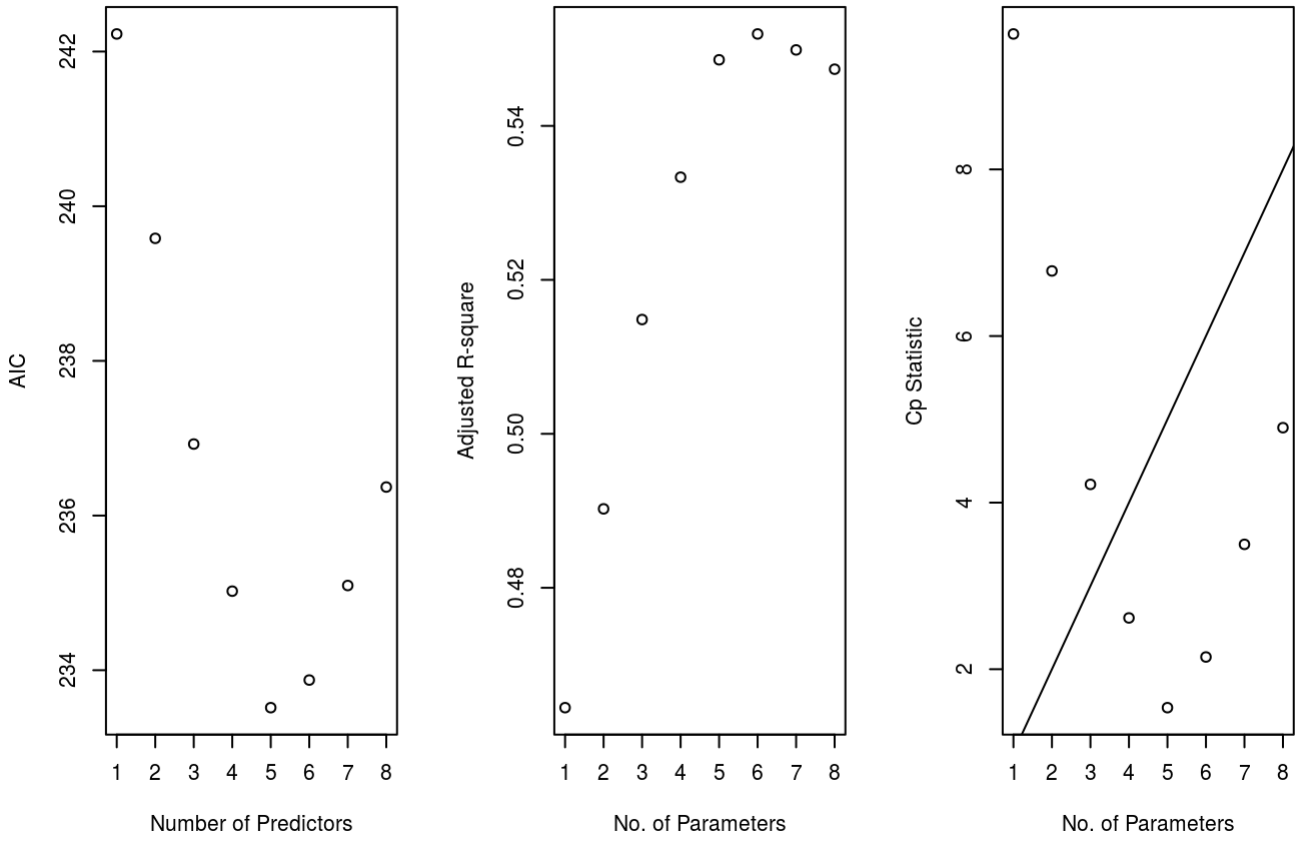
```
rs$which[6,]
```

```
##           (Intercept)        Process.change     Backlog.of.orders
##                  TRUE                  TRUE                  TRUE
##       Team.experience  factor(Team.number)2  factor(Team.number)3
##                 FALSE                 FALSE                 FALSE
##  factor(Team.number)4  factor(Team.number)5  factor(Team.number)6
##                 FALSE                  TRUE                 FALSE
##  factor(Team.number)7  factor(Team.number)8  factor(Team.number)9
##                  TRUE                  TRUE                  TRUE
## factor(Team.number)10 factor(Team.number)11 factor(Team.number)12
##                 FALSE                 FALSE                 FALSE
## factor(Team.number)13
##                 FALSE
```

```
AIC <- 73*log(rs$rss/73) + (2:9)*2
par(mfrow=c(1,3))
plot(AIC ~ I(1:8), ylab="AIC", xlab="Number of Predictors")
plot(1:8,rs$adjr2,xlab="No. of Parameters",ylab="Adjusted R-square")
which.max(rs$adjr2)
```

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

```
plot(1:8,rs$cp,xlab="No. of Parameters",ylab="Cp Statistic")
abline(0,1)
```



Problem 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) ###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) ###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)

a)

Solution: Based on the code below, model with 2 predictor variables is the best model. PSA=Cancer.volume+Capsular.penetration

```
Prostate.Cancer <- read.csv("/cloud/project/Prostate Cancer.csv")
set.seed(567)
sample.ind <- sample(1:nrow(Prostate.Cancer), size = 65)
devq5 <- Prostate.Cancer[sample.ind,]
holdoutq5 <- Prostate.Cancer[-sample.ind,]
f5<-lm(PSA.level~Cancer.volume+Weight+Age+Benign.prostatic.hyperplasia+Seminal.vesicle.invasion+Capsular.penetration+
 Gleason.score,data=devq5)
library(olsrr)
ols_step_both_p(f5,prem=0.05,details=TRUE)
```

```
## Stepwise Selection Method
## -------------------------
##
## Candidate Terms:
##
## 1. Cancer.volume
## 2. Weight
## 3. Age
## 4. Benign.prostatic.hyperplasia
## 5. Seminal.vesicle.invasion
## 6. Capsular.penetration
## 7. Gleason.score
##
## We are selecting variables based on p value...
##
##
## Stepwise Selection: Step 1
##
## - Capsular.penetration added
##
##                          Model Summary
## ----------------------------------------------------------------
## R                       0.665       RMSE                30.176
## R-Squared               0.442       Coef. Var          133.878
## Adj. R-Squared          0.433       MSE                910.572
## Pred R-Squared          0.235       MAE                 16.877
## ----------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                                ANOVA
## -------------------------------------------------------------------------
##                   Sum of
##                  Squares        DF    Mean Square      F         Sig.
## -------------------------------------------------------------------------
## Regression     45363.742         1      45363.742    49.819    0.0000
## Residual       57366.062        63        910.572
## Total         102729.803        64
## -------------------------------------------------------------------------
```

```
##
##                                  Parameter Estimates
## ----------------------------------------------------------------------------------------
##                model     Beta    Std. Error    Std. Beta      t       Sig      lower     upper
## ----------------------------------------------------------------------------------------
##          (Intercept)    6.700         4.364                1.535    0.130    -2.021    15.421
## Capsular.penetration    6.360         0.901        0.665    7.058    0.000     4.559     8.161
## ----------------------------------------------------------------------------------------
##
##
##
## Stepwise Selection: Step 2
##
## - Cancer.volume added
##
##                          Model Summary
## ---------------------------------------------------------------
## R                       0.710       RMSE                28.650
## R-Squared               0.505       Coef. Var          127.109
## Adj. R-Squared          0.489       MSE                820.818
## Pred R-Squared          0.295       MAE                 14.967
## ---------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                                ANOVA
## --------------------------------------------------------------------
##                   Sum of
##                  Squares        DF    Mean Square      F         Sig.
## --------------------------------------------------------------------
## Regression     51839.108         2      25919.554    31.578    0.0000
## Residual       50890.696        62        820.818
## Total         102729.803        64
## --------------------------------------------------------------------
##
##                                  Parameter Estimates
## ----------------------------------------------------------------------------------------
##                model     Beta    Std. Error    Std. Beta      t       Sig      lower    upper
## ----------------------------------------------------------------------------------------
##          (Intercept)    0.452         4.703                0.096    0.924    -8.949    9.853
```

```
## Capsular.penetration    3.738         1.266        0.391    2.952    0.004     1.207    6.269
##        Cancer.volume    1.720         0.612        0.372    2.809    0.007     0.496    2.944
## ----------------------------------------------------------------------------------------------
##
##
##
##                          Model Summary
## ----------------------------------------------------------------
## R                       0.710       RMSE                 28.650
## R-Squared               0.505       Coef. Var           127.109
## Adj. R-Squared          0.489       MSE                 820.818
## Pred R-Squared          0.295       MAE                  14.967
## ----------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                                ANOVA
## --------------------------------------------------------------------
##                   Sum of
##                  Squares        DF    Mean Square      F         Sig.
## --------------------------------------------------------------------
## Regression     51839.108         2      25919.554    31.578    0.0000
## Residual       50890.696        62        820.818
## Total         102729.803        64
## --------------------------------------------------------------------
##
##                                    Parameter Estimates
## ----------------------------------------------------------------------------------------------
##                model     Beta    Std. Error    Std. Beta      t       Sig      lower    upper
## ----------------------------------------------------------------------------------------------
##          (Intercept)    0.452         4.703                0.096    0.924    -8.949    9.853
## Capsular.penetration    3.738         1.266        0.391    2.952    0.004     1.207    6.269
##        Cancer.volume    1.720         0.612        0.372    2.809    0.007     0.496    2.944
## ----------------------------------------------------------------------------------------------
##
##
##
## No more variables to be added/removed.
##
##
```

```
## Final Model Output
## ------------------
##
##                          Model Summary
## ----------------------------------------------------------------
## R                       0.710       RMSE                 28.650
## R-Squared               0.505       Coef. Var           127.109
## Adj. R-Squared          0.489       MSE                 820.818
## Pred R-Squared          0.295       MAE                  14.967
## ----------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                                ANOVA
## --------------------------------------------------------------------
##                   Sum of
##                  Squares        DF    Mean Square      F         Sig.
## --------------------------------------------------------------------
## Regression     51839.108         2      25919.554    31.578    0.0000
## Residual       50890.696        62        820.818
## Total         102729.803        64
## --------------------------------------------------------------------
##
##                                    Parameter Estimates
## ----------------------------------------------------------------------------------------------
##                model     Beta    Std. Error    Std. Beta      t       Sig      lower    upper
## ----------------------------------------------------------------------------------------------
##          (Intercept)    0.452         4.703                0.096    0.924    -8.949    9.853
## Capsular.penetration    3.738         1.266        0.391    2.952    0.004     1.207    6.269
##        Cancer.volume    1.720         0.612        0.372    2.809    0.007     0.496    2.944
## ----------------------------------------------------------------------------------------------
```

```
##
##                                 Stepwise Selection Summary
## ----------------------------------------------------------------------------------------------
##                                Added/                   Adj.
## Step          Variable         Removed     R-Square    R-Square     C(p)        AIC        RMSE
## ----------------------------------------------------------------------------------------------
##    1    Capsular.penetration    addition       0.442       0.433    10.6000    631.3454    30.1757
##    2       Cancer.volume        addition       0.505       0.489     4.5180    625.5601    28.6499
## ----------------------------------------------------------------------------------------------
```

```
f4f<-lm(PSA.level~Cancer.volume+Capsular.penetration,data=devq5)
summary(f4f)
```

```
##
## Call:
## lm(formula = PSA.level ~ Cancer.volume + Capsular.penetration,
##     data = devq5)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -67.191  -4.595   1.055   5.135 141.423
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)           0.4517     4.7028   0.096  0.92379
## Cancer.volume         1.7197     0.6123   2.809  0.00664 **
## Capsular.penetration  3.7378     1.2663   2.952  0.00446 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 28.65 on 62 degrees of freedom
## Multiple R-squared:  0.5046, Adjusted R-squared:  0.4886
## F-statistic: 31.58 on 2 and 62 DF,  p-value: 3.493e-10
```

a)

Solution: Based on the code below, model with 2 predictor variables is the best model. PSA=Cancer.volume+Capsular.penetration

```
Prostate.Cancer <- read.csv("/cloud/project/Prostate Cancer.csv")
set.seed(567)
sample.ind <- sample(1:nrow(Prostate.Cancer), size = 65)
devq5 <- Prostate.Cancer[sample.ind,]
holdoutq5 <- Prostate.Cancer[-sample.ind,]
f5<-lm(PSA.level~Cancer.volume+Weight+Age+Benign.prostatic.hyperplasia+Seminal.vesicle.invasion+Capsular.penetration+
 Gleason.score,data=devq5)
library(olsrr)
ols_step_both_p(f5,prem=0.05,details=TRUE)
```

```
## Stepwise Selection Method
## -------------------------
##
## Candidate Terms:
##
## 1. Cancer.volume
## 2. Weight
## 3. Age
## 4. Benign.prostatic.hyperplasia
## 5. Seminal.vesicle.invasion
## 6. Capsular.penetration
## 7. Gleason.score
##
## We are selecting variables based on p value...
##
##
## Stepwise Selection: Step 1
##
## - Capsular.penetration added
##
##                          Model Summary
## ----------------------------------------------------------------
## R                       0.665       RMSE                30.176
## R-Squared               0.442       Coef. Var          133.878
## Adj. R-Squared          0.433       MSE                910.572
## Pred R-Squared          0.235       MAE                 16.877
## ----------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                                ANOVA
## --------------------------------------------------------------------
##                   Sum of
##                  Squares        DF    Mean Square      F         Sig.
## --------------------------------------------------------------------
## Regression     45363.742         1      45363.742    49.819    0.0000
## Residual       57366.062        63        910.572
## Total         102729.803        64
## --------------------------------------------------------------------
```

```
##
##                                  Parameter Estimates
## ----------------------------------------------------------------------------------------
##                model     Beta    Std. Error    Std. Beta      t       Sig      lower     upper
## ----------------------------------------------------------------------------------------
##          (Intercept)    6.700         4.364                1.535    0.130    -2.021    15.421
## Capsular.penetration    6.360         0.901        0.665    7.058    0.000     4.559     8.161
## ----------------------------------------------------------------------------------------
##
##
##
## Stepwise Selection: Step 2
##
## - Cancer.volume added
##
##                          Model Summary
## ---------------------------------------------------------------
## R                       0.710       RMSE                28.650
## R-Squared               0.505       Coef. Var          127.109
## Adj. R-Squared          0.489       MSE                820.818
## Pred R-Squared          0.295       MAE                 14.967
## ---------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                                ANOVA
## -------------------------------------------------------------------
##                   Sum of
##                  Squares        DF    Mean Square      F         Sig.
## -------------------------------------------------------------------
## Regression     51839.108         2      25919.554    31.578    0.0000
## Residual       50890.696        62        820.818
## Total         102729.803        64
## -------------------------------------------------------------------
##
##                                  Parameter Estimates
## ----------------------------------------------------------------------------------------
##                model     Beta    Std. Error    Std. Beta      t       Sig      lower    upper
## ----------------------------------------------------------------------------------------
##          (Intercept)    0.452         4.703                0.096    0.924    -8.949    9.853
```

```
## Capsular.penetration    3.738         1.266        0.391    2.952    0.004     1.207    6.269
##        Cancer.volume    1.720         0.612        0.372    2.809    0.007     0.496    2.944
## ----------------------------------------------------------------------------------------------
##
##
##
##                          Model Summary
## ----------------------------------------------------------------
## R                       0.710       RMSE                 28.650
## R-Squared               0.505       Coef. Var           127.109
## Adj. R-Squared          0.489       MSE                 820.818
## Pred R-Squared          0.295       MAE                  14.967
## ----------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                                ANOVA
## --------------------------------------------------------------------
##                   Sum of
##                  Squares        DF    Mean Square      F         Sig.
## --------------------------------------------------------------------
## Regression     51839.108         2      25919.554    31.578    0.0000
## Residual       50890.696        62        820.818
## Total         102729.803        64
## --------------------------------------------------------------------
##
##                                    Parameter Estimates
## ----------------------------------------------------------------------------------------------
##                model     Beta    Std. Error    Std. Beta      t       Sig      lower    upper
## ----------------------------------------------------------------------------------------------
##          (Intercept)    0.452         4.703                0.096    0.924    -8.949    9.853
## Capsular.penetration    3.738         1.266        0.391    2.952    0.004     1.207    6.269
##        Cancer.volume    1.720         0.612        0.372    2.809    0.007     0.496    2.944
## ----------------------------------------------------------------------------------------------
##
##
##
## No more variables to be added/removed.
##
##
```

```
## Final Model Output
## ------------------
##
##                          Model Summary
## ----------------------------------------------------------------
## R                       0.710       RMSE                 28.650
## R-Squared               0.505       Coef. Var           127.109
## Adj. R-Squared          0.489       MSE                 820.818
## Pred R-Squared          0.295       MAE                  14.967
## ----------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                                ANOVA
## --------------------------------------------------------------------
##                   Sum of
##                  Squares        DF    Mean Square      F         Sig.
## --------------------------------------------------------------------
## Regression     51839.108         2      25919.554    31.578    0.0000
## Residual       50890.696        62        820.818
## Total         102729.803        64
## --------------------------------------------------------------------
##
##                                    Parameter Estimates
## ----------------------------------------------------------------------------------------------
##                model     Beta    Std. Error    Std. Beta      t       Sig      lower    upper
## ----------------------------------------------------------------------------------------------
##          (Intercept)    0.452         4.703                0.096    0.924    -8.949    9.853
## Capsular.penetration    3.738         1.266        0.391    2.952    0.004     1.207    6.269
##        Cancer.volume    1.720         0.612        0.372    2.809    0.007     0.496    2.944
## ----------------------------------------------------------------------------------------------
```

```
##
##                                 Stepwise Selection Summary
## ----------------------------------------------------------------------------------------------
##                                Added/                   Adj.
## Step          Variable         Removed     R-Square    R-Square     C(p)        AIC        RMSE
## ----------------------------------------------------------------------------------------------
##    1    Capsular.penetration    addition       0.442       0.433    10.6000    631.3454    30.1757
##    2       Cancer.volume        addition       0.505       0.489     4.5180    625.5601    28.6499
## ----------------------------------------------------------------------------------------------
```

```
f4f<-lm(PSA.level~Cancer.volume+Capsular.penetration,data=devq5)
summary(f4f)
```

```
##
## Call:
## lm(formula = PSA.level ~ Cancer.volume + Capsular.penetration,
##     data = devq5)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -67.191  -4.595   1.055   5.135 141.423
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)           0.4517     4.7028   0.096  0.92379
## Cancer.volume         1.7197     0.6123   2.809  0.00664 **
## Capsular.penetration  3.7378     1.2663   2.952  0.00446 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 28.65 on 62 degrees of freedom
## Multiple R-squared:  0.5046, Adjusted R-squared:  0.4886
## F-statistic: 31.58 on 2 and 62 DF,  p-value: 3.493e-10
```

b)

Solution: Capsular.penetration becomes insignificant and Rsquare decreases. MSE increased from 776 to 1149.8.Indicating problem with the model stability.

```
f4f2<-lm(PSA.level~Cancer.volume+Capsular.penetration,data=holdoutq5)
summary(f4f2)
```

```
##
## Call:
## lm(formula = PSA.level ~ Cancer.volume + Capsular.penetration,
##     data = holdoutq5)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -51.179 -14.221  -0.684   6.802 164.753
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)            0.5022     9.1138   0.055   0.9564
## Cancer.volume          4.1465     1.2140   3.415   0.0019 **
## Capsular.penetration   0.1525     2.6287   0.058   0.9541
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 35.69 on 29 degrees of freedom
## Multiple R-squared:  0.3482, Adjusted R-squared:  0.3033
## F-statistic: 7.746 on 2 and 29 DF,  p-value: 0.002017
```

```
anova(f4f2)
```

```
## Analysis of Variance Table
##
## Response: PSA.level
##                      Df Sum Sq Mean Sq F value    Pr(>F)
## Cancer.volume         1  19726 19725.7 15.4889 0.0004764 ***
## Capsular.penetration  1      4     4.3  0.0034 0.9541288
## Residuals            29  36933  1273.5
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(f4f)
```

```
## Analysis of Variance Table
##
## Response: PSA.level
##                      Df Sum Sq Mean Sq F value    Pr(>F)
## Cancer.volume         1  44687   44687 54.4420 4.747e-10 ***
## Capsular.penetration  1   7152    7152  8.7135  0.004455 **
## Residuals            62  50891     821
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
MSPR<-sum(f4f2$residuals^2)/length(f4f2$residuals)
```

c)

Solution: MSPR=1042.05 and MSE=776.Indicating problem with the model stability.

```
anova(f4f2)
```

```
## Analysis of Variance Table
##
## Response: PSA.level
##                      Df Sum Sq Mean Sq F value    Pr(>F)
## Cancer.volume         1  19726 19725.7 15.4889 0.0004764 ***
## Capsular.penetration  1      4     4.3  0.0034 0.9541288
## Residuals            29  36933  1273.5
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
MSPR<-sum(f4f2$residuals^2)/length(f4f2$residuals)
MSPR
```

```
## [1] 1154.145
```

Problem 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 SECp criterion. (7 pts)

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

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

a)

Solution: The best model with SBC is X1,X3 X4 with -135.390. see below for the details.

```
Market.Share <- read.csv("/cloud/project/Market Share.csv")
Y<-Market.Share$Market.Share
X1<-Market.Share$Price
X2<-Market.Share$Gross.Nielsen.Rating.Points
X3<-Market.Share$Discount.Price
X4<-Market.Share$Package.Promotion
X5<- as.numeric(Market.Share$Year == 1999)
X6<- as.numeric(Market.Share$Year == 2001)
X7<- as.numeric(Market.Share$Year == 2002)
q6<-data.frame(cbind(Y,X1,X2,X3,X4,X5,X6,X7))
f6<-lm(Y~X1+X2+X3+X4+X5+X6+X7,data=q6)

library(leaps)
b6 <- regsubsets(Y~X1+X2+X3+X4+X5+X6+X7,data=q6)
rs6 <- summary(b6)
rs6$which
```

```
##   (Intercept)    X1    X2   X3    X4    X5    X6    X7
## 1        TRUE FALSE FALSE TRUE FALSE FALSE FALSE FALSE
## 2        TRUE FALSE FALSE TRUE  TRUE FALSE FALSE FALSE
## 3        TRUE  TRUE FALSE TRUE  TRUE FALSE FALSE FALSE
## 4        TRUE  TRUE FALSE TRUE  TRUE FALSE  TRUE FALSE
## 5        TRUE  TRUE FALSE TRUE  TRUE FALSE  TRUE  TRUE
## 6        TRUE  TRUE  TRUE TRUE  TRUE FALSE  TRUE  TRUE
## 7        TRUE  TRUE  TRUE TRUE  TRUE  TRUE  TRUE  TRUE
```

```
which.max(rs6$adjr2)
```

```
## [1] 4
```

```
par(mfrow=c(1,3))
SBC <- 36*log(rs6$rss/36) + (2:8)*log(2:8)
SBC
```

```
## [1] -130.7506 -132.3915 -135.3901 -134.8395 -133.2043 -130.6335 -127.6454
```

```
rs6$which[3:5,]
```

```
##   (Intercept)   X1    X2   X3   X4    X5    X6    X7
## 3        TRUE TRUE FALSE TRUE TRUE FALSE FALSE FALSE
## 4        TRUE TRUE FALSE TRUE TRUE FALSE  TRUE FALSE
## 5        TRUE TRUE FALSE TRUE TRUE FALSE  TRUE  TRUE
```

b)

Solution: The best model is X1,X3 X4. see below for the details.

```
library(olsrr)
k6b<-ols_step_forward_p(f6,pent=0.15,prem=0.10,details=TRUE)
```

```
## Forward Selection Method
## --------------------------
##
## Candidate Terms:
##
## 1. X1
## 2. X2
## 3. X3
## 4. X4
## 5. X5
## 6. X6
## 7. X7
##
## We are selecting variables based on p value...
##
##
## Forward Selection: Step 1
##
## - X3
##
##                        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
##          X3    0.418         0.056        0.791     7.532    0.000    0.305    0.531
## ----------------------------------------------------------------------------------
##
##
##
## Forward Selection: Step 2
##
## - X4
##
##                         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
```

```
##          X3    0.403         0.054        0.762     7.426    0.000     0.293    0.513
##          X4    0.100         0.054        0.190     1.849    0.073    -0.010    0.209
## ----------------------------------------------------------------------------------
##
##
##
## Forward Selection: Step 3
##
## - X1
##
##                         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
##          X3     0.399         0.051        0.755     7.787    0.000     0.295     0.504
##          X4     0.118         0.051        0.225     2.292    0.029     0.013     0.223
##          X1    -0.353         0.157       -0.217    -2.241    0.032    -0.673    -0.032
## ----------------------------------------------------------------------------------------
##
```

```
##
##
## No more variables to be added.
##
## Variables Entered:
##
## + X3
## + X4
## + X1
##
##
## 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
##          X3     0.399         0.051        0.755     7.787    0.000     0.295     0.504
```

```
##          X4     0.118         0.051         0.225     2.292    0.029     0.013     0.223
##          X1    -0.353         0.157        -0.217    -2.241    0.032    -0.673    -0.032
## ------------------------------------------------------------------------------------------
```

c)

Solution: The same result.