CSCI E-106:Assignment 8 Solutions

Due Date: November 16, 2020 at 7:20 pm EST

Instructions

Students should submit their reports on Canvas. The report needs to clearly state what question is being solved, step-by-step walk-through solutions, and final answers clearly indicated. Please solve by hand where appropriate.

Please submit two files: (1) a R Markdown file (.Rmd extension) and (2) a PDF document, word, or html generated using knitr for the .Rmd file submitted in (1) where appropriate. Please, use RStudio Cloud for your solutions.

Solutions:

Problem 1

Refer to the the Efficacy of Nosocomial Infection Control (SENIC) data set. The primary objective of the Study on was to determine whether infection surveillance and control programs have reduced the rates of nosocomial (hospitalacquired) infection in United States hospitals. This data set consists of a random sample of 113 hospitals selected from the original 338 hospitals surveyed. Each line of the dataset has an identification number and provides information on 11 variables for a single hospital. The data presented here are for the 1975-76 study period. (15 points, 5 points each)

a-) Second-order regression model is to be fitted for relating number of nurses (Y) to available facilities and services (X).

$\hat{Y}$=150.079+7.066X+0.101X^2. All variables are significant and R^2 is 65%.

```
SENIC <- read.csv("/cloud/project/SENIC.csv")
Y=SENIC$Number.of.nurses
X=SENIC$Available.facilities.and.services
X1=scale(X,scale=FALSE)
X12=X1^2
f1<-lm(Y~X1+X12)
summary(f1)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X12)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -244.32  -39.42   -4.55   26.48  336.48
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
```

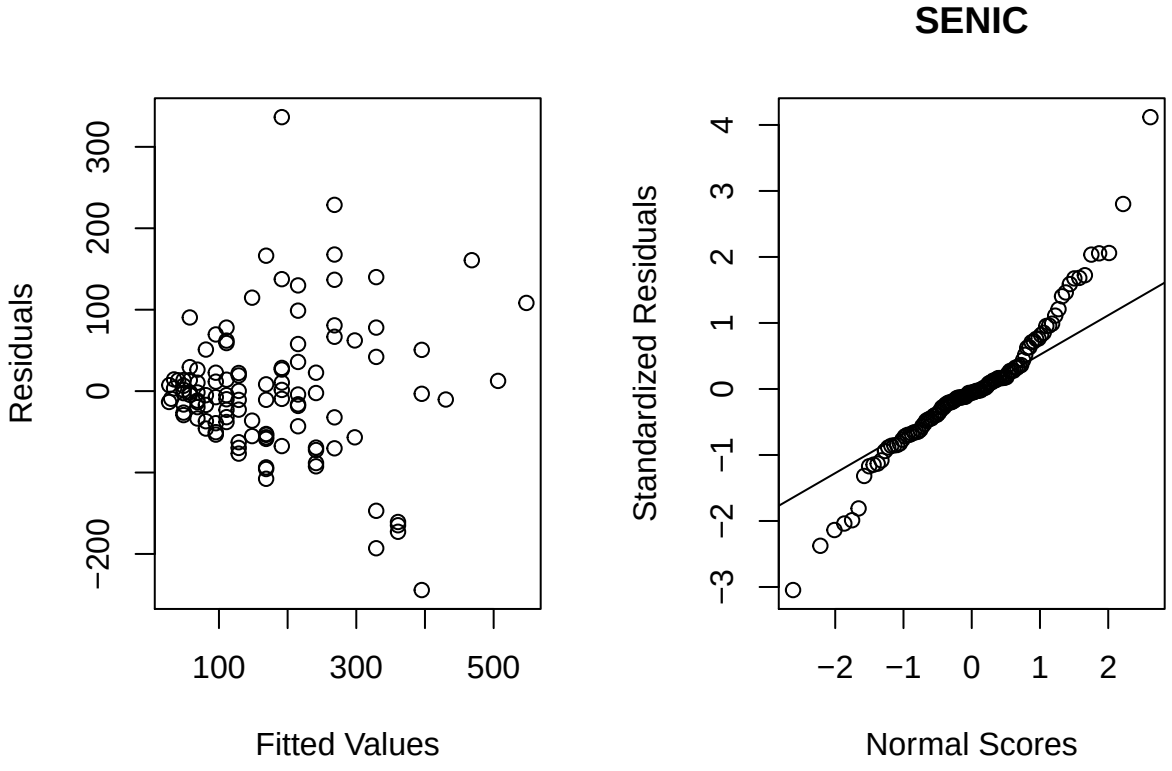
```
## (Intercept) 150.07915    9.94139  15.096  < 2e-16 ***
## X1            7.06617    0.51253  13.787  < 2e-16 ***
## X12           0.10116    0.02723   3.716  0.00032 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 82.31 on 110 degrees of freedom
## Multiple R-squared:  0.6569, Adjusted R-squared:  0.6507
## F-statistic: 105.3 on 2 and 110 DF,  p-value: < 2.2e-16
```

```
anova(f1)
```

```
## Analysis of Variance Table
##
## Response: Y
##            Df  Sum Sq Mean Sq F value    Pr(>F)
## X1          1 1333486 1333486 196.837 < 2.2e-16 ***
## X12         1   93533   93533  13.806 0.0003203 ***
## Residuals 110  745204    6775
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

b-) Plot the residuals against the fitted values. How well does the second-order model appear to fit the data? The plot indicates heteroscedasticity. I also checked the QQ plot and it indicates that the departurtures from the normal distribution.

```
par(mfrow=c(1,2))
plot(f1$fitted.values,f1$residuals,xlab="Fitted Values",ylab="Residuals")
stdres = rstandard(f1)
qqnorm(stdres,ylab="Standardized Residuals",xlab="Normal Scores",main="SENIC")
qqline(stdres)
```



c-) Test whether the quadratic term can be dropped from the regression model; use α=0.1, State the alternatives, decision rule, and conclusion.

H_0:The quadratic term can be dropped from the regression model. H_a:The quadratic term can NOT be dropped from the regression model.

P value is 0.0003, Reject H_0. The quadratic term can NOT be dropped from the regression model.

```
fr<-lm(Y~X)
anova(fr,f1)
```

```
## Analysis of Variance Table
##
## Model 1: Y ~ X
## Model 2: Y ~ X1 + X12
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    111 838737
## 2    110 745204  1     93533 13.806 0.0003203 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Problem 2

Use the fortune data under the faraway r library, data(prostate,package="faraway"). Use the prostate data with lpsa as the response and the other variables as predictors.

Implement the following variable selection methods to determine the "best" model: (40 points, 10 points each)

a-) Backward elimination

The model is $lpsa$=-0.268 + 0.552$lcavol$ + 0.509lweight +0.666*svi and R^2 is 62%. All variables are signficant at $\alpha = 0.05$.

```
library(olsrr)
```

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

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

```
library(datasets)
data(prostate,package="faraway")
k1<-lm(lpsa~.,data=prostate)
k2<-ols_step_backward_p(k1,prem = 0.05,details=TRUE)
```

```
## Backward Elimination Method
## ---------------------------
##
## Candidate Terms:
##
## 1 . lcavol
## 2 . lweight
## 3 . age
## 4 . lbph
## 5 . svi
```

```
## 6 . lcp
## 7 . gleason
## 8 . pgg45
##
## We are eliminating variables based on p value...
##
## - gleason
##
## Backward Elimination: Step 1
##
##  Variable gleason Removed
##
##                         Model Summary
## ----------------------------------------------------------------
## R                       0.809       RMSE                 0.705
## R-Squared               0.654       Coef. Var           28.436
## Adj. R-Squared          0.627       MSE                  0.497
## Pred R-Squared          0.584       MAE                  0.521
## ----------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                                ANOVA
## -------------------------------------------------------------------
##                Sum of
##               Squares        DF    Mean Square      F         Sig.
## -------------------------------------------------------------------
## Regression     83.713         7         11.959    24.078    0.0000
## Residual       44.204        89          0.497
## Total         127.918        96
## -------------------------------------------------------------------
##
##                                  Parameter Estimates
## ----------------------------------------------------------------------------------------
##       model      Beta    Std. Error    Std. Beta      t        Sig      lower    upper
## ----------------------------------------------------------------------------------------
## (Intercept)     0.954         0.829                 1.150    0.253    -0.694    2.602
##      lcavol     0.592         0.086        0.604    6.879    0.000     0.421    0.762
##     lweight     0.448         0.168        0.193    2.672    0.009     0.115    0.782
##         age    -0.019         0.011       -0.125   -1.747    0.084    -0.041    0.003
##        lbph     0.108         0.058        0.135    1.853    0.067    -0.008    0.223
##         svi     0.758         0.241        0.272    3.140    0.002     0.278    1.237
##         lcp    -0.104         0.090       -0.127   -1.155    0.251    -0.284    0.075
##       pgg45     0.005         0.003        0.130    1.549    0.125    -0.002    0.012
## ----------------------------------------------------------------------------------------
##
##
## - lcp
##
## Backward Elimination: Step 2
##
##  Variable lcp Removed
##
```

```
##                          Model Summary
## ----------------------------------------------------------------
## R                        0.806       RMSE                 0.706
## R-Squared                0.649       Coef. Var           28.489
## Adj. R-Squared           0.626       MSE                  0.499
## Pred R-Squared           0.586       MAE                  0.530
## ----------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                                ANOVA
## -------------------------------------------------------------------
##                Sum of
##               Squares        DF    Mean Square      F         Sig.
## -------------------------------------------------------------------
## Regression     83.051         6         13.842    27.766    0.0000
## Residual       44.867        90          0.499
## Total         127.918        96
## -------------------------------------------------------------------
##
##                                  Parameter Estimates
## ----------------------------------------------------------------------------------------
##       model      Beta    Std. Error    Std. Beta      t        Sig      lower    upper
## ----------------------------------------------------------------------------------------
## (Intercept)     0.980         0.831                 1.180    0.241    -0.670    2.630
##      lcavol     0.546         0.076         0.557     7.141    0.000     0.394    0.698
##     lweight     0.449         0.168         0.193     2.674    0.009     0.116    0.783
##         age    -0.017         0.011        -0.113    -1.593    0.115    -0.039    0.004
##        lbph     0.106         0.058         0.133     1.817    0.072    -0.010    0.221
##         svi     0.642         0.220         0.230     2.920    0.004     0.205    1.078
##       pgg45     0.004         0.003         0.086     1.150    0.253    -0.003    0.010
## ----------------------------------------------------------------------------------------
##
##
## - pgg45
##
## Backward Elimination: Step 3
##
##  Variable pgg45 Removed
##
##                          Model Summary
## ----------------------------------------------------------------
## R                        0.803       RMSE                 0.707
## R-Squared                0.644       Coef. Var           28.539
## Adj. R-Squared           0.625       MSE                  0.500
## Pred R-Squared           0.588       MAE                  0.534
## ----------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                                ANOVA
## -------------------------------------------------------------------
```

```
##                Sum of
##               Squares        DF    Mean Square      F         Sig.
## ----------------------------------------------------------------------
## Regression     82.392         5         16.478    32.938    0.0000
## Residual       45.526        91          0.500
## Total         127.918        96
## ----------------------------------------------------------------------
##
##                                  Parameter Estimates
## ------------------------------------------------------------------------------------------
##       model      Beta    Std. Error    Std. Beta      t        Sig      lower    upper
## ------------------------------------------------------------------------------------------
## (Intercept)     0.951         0.832                 1.143    0.256    -0.701    2.603
##      lcavol     0.566         0.075        0.578    7.583    0.000     0.417    0.714
##     lweight     0.424         0.167        0.182    2.539    0.013     0.092    0.755
##         age    -0.015         0.011       -0.096   -1.385    0.170    -0.036    0.006
##        lbph     0.112         0.058        0.141    1.927    0.057    -0.003    0.227
##         svi     0.721         0.209        0.259    3.449    0.001     0.306    1.136
## ------------------------------------------------------------------------------------------
##
##
## - age
##
## Backward Elimination: Step 4
##
##  Variable age Removed
##
##                          Model Summary
## ----------------------------------------------------------------
## R                       0.798       RMSE                 0.711
## R-Squared               0.637       Coef. Var           28.681
## Adj. R-Squared          0.621       MSE                  0.505
## Pred R-Squared          0.590       MAE                  0.549
## ----------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                                 ANOVA
## ----------------------------------------------------------------------
##                Sum of
##               Squares        DF    Mean Square      F         Sig.
## ----------------------------------------------------------------------
## Regression     81.433         4         20.358    40.292    0.0000
## Residual       46.485        92          0.505
## Total         127.918        96
## ----------------------------------------------------------------------
##
##                                  Parameter Estimates
## ------------------------------------------------------------------------------------------
##       model      Beta    Std. Error    Std. Beta      t        Sig      lower    upper
## ------------------------------------------------------------------------------------------
## (Intercept)     0.146         0.597                 0.244    0.808    -1.041    1.332
##      lcavol     0.550         0.074        0.561    7.422    0.000     0.403    0.697
```

```
##     lweight    0.391         0.166        0.168    2.355    0.021     0.061    0.721
##        lbph    0.090         0.056        0.113    1.604    0.112    -0.021    0.202
##         svi    0.712         0.210        0.255    3.390    0.001     0.295    1.129
## ----------------------------------------------------------------------------------
##
##
## - lbph
##
## Backward Elimination: Step 5
##
##  Variable lbph Removed
##
##                         Model Summary
## ----------------------------------------------------------------
## R                       0.791       RMSE                 0.717
## R-Squared               0.626       Coef. Var           28.922
## Adj. R-Squared          0.614       MSE                  0.514
## Pred R-Squared          0.587       MAE                  0.564
## ----------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                                ANOVA
## ---------------------------------------------------------------------
##                Sum of
##               Squares        DF    Mean Square      F         Sig.
## ---------------------------------------------------------------------
## Regression     80.133         3         26.711    51.985    0.0000
## Residual       47.785        93          0.514
## Total         127.918        96
## ---------------------------------------------------------------------
##
##                                 Parameter Estimates
## ----------------------------------------------------------------------------------------
##       model      Beta    Std. Error    Std. Beta      t        Sig      lower    upper
## ----------------------------------------------------------------------------------------
## (Intercept)    -0.268         0.543                -0.493    0.623    -1.347    0.811
##      lcavol     0.552         0.075        0.563     7.388    0.000     0.403    0.700
##     lweight     0.509         0.150        0.219     3.386    0.001     0.210    0.807
##         svi     0.666         0.210        0.239     3.176    0.002     0.250    1.083
## ----------------------------------------------------------------------------------------
##
##
##
## No more variables satisfy the condition of p value = 0.05
##
##
## Variables Removed:
##
## - gleason
## - lcp
## - pgg45
## - age
```

```
## - lbph
##
##
## Final Model Output
## ------------------
##
##                         Model Summary
## ----------------------------------------------------------------
## R                        0.791       RMSE                 0.717
## R-Squared                0.626       Coef. Var           28.922
## Adj. R-Squared           0.614       MSE                  0.514
## Pred R-Squared           0.587       MAE                  0.564
## ----------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                                ANOVA
## --------------------------------------------------------------------
##                Sum of
##               Squares        DF    Mean Square      F         Sig.
## --------------------------------------------------------------------
## Regression     80.133         3         26.711    51.985    0.0000
## Residual       47.785        93          0.514
## Total         127.918        96
## --------------------------------------------------------------------
##
##                                  Parameter Estimates
## ----------------------------------------------------------------------------------------------
##       model      Beta    Std. Error    Std. Beta      t        Sig      lower    upper
## ----------------------------------------------------------------------------------------------
## (Intercept)    -0.268         0.543                -0.493    0.623    -1.347    0.811
##      lcavol     0.552         0.075        0.563    7.388    0.000     0.403    0.700
##     lweight     0.509         0.150        0.219    3.386    0.001     0.210    0.807
##         svi     0.666         0.210        0.239    3.176    0.002     0.250    1.083
## ----------------------------------------------------------------------------------------------
```

```
k2
```

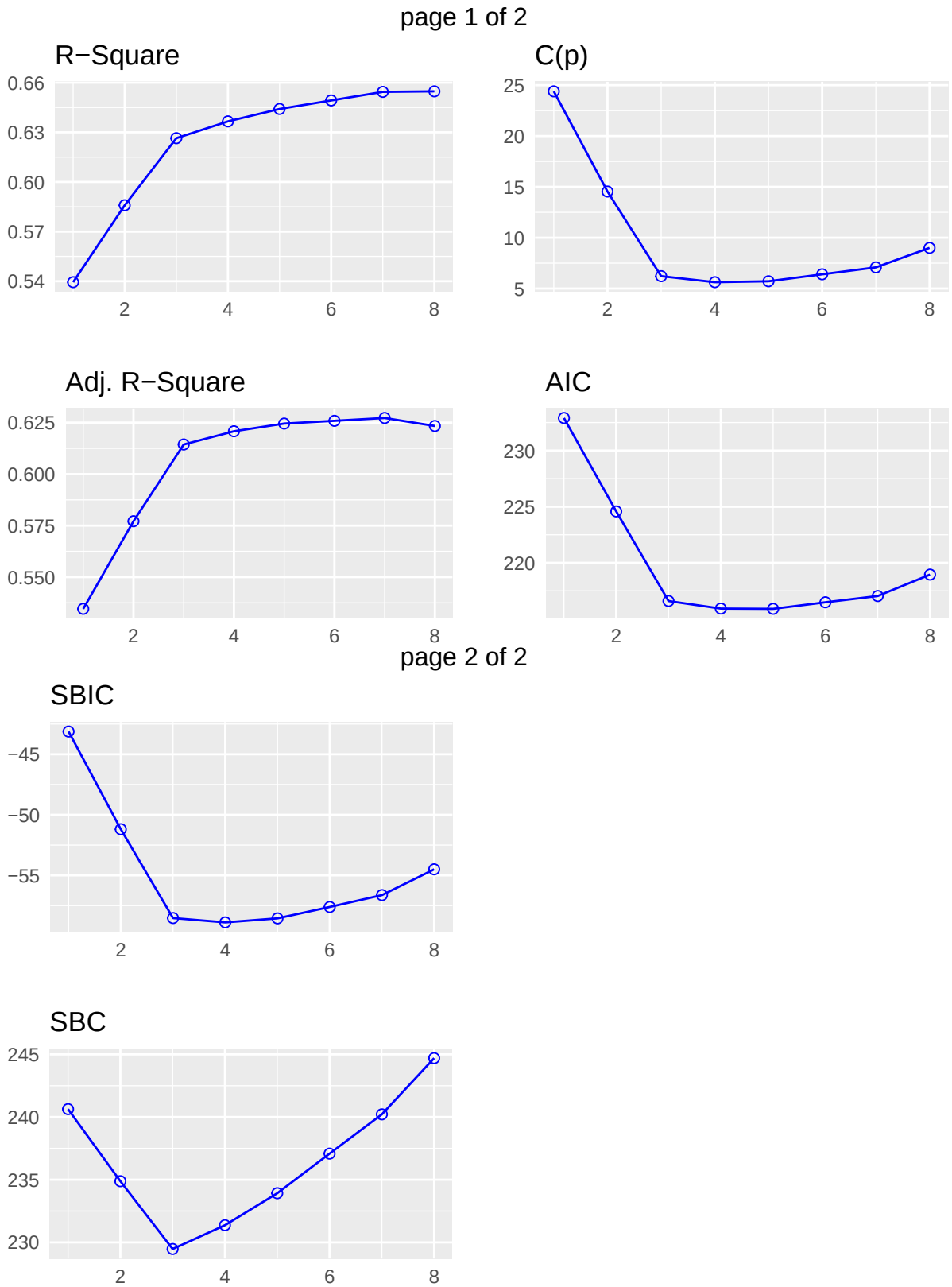
```
##
##
##                           Elimination Summary
## ------------------------------------------------------------------------
##         Variable                 Adj.
## Step    Removed     R-Square    R-Square     C(p)       AIC        RMSE
## ------------------------------------------------------------------------
##    1    gleason       0.6544      0.6273    7.0822    217.0428    0.7048
##    2    lcp           0.6493      0.6259    6.4020    216.4854    0.7061
##    3    pgg45         0.6441      0.6245    5.7150    215.8997    0.7073
##    4    age           0.6366      0.6208    5.6264    215.9223    0.7108
##    5    lbph          0.6264      0.6144    6.2169    216.5979    0.7168
## ------------------------------------------------------------------------
```

b-) AIC. on the graph, the elbow point is for 3 variables models. The variables are lcavol, lweight, and svi. The same model in part a.

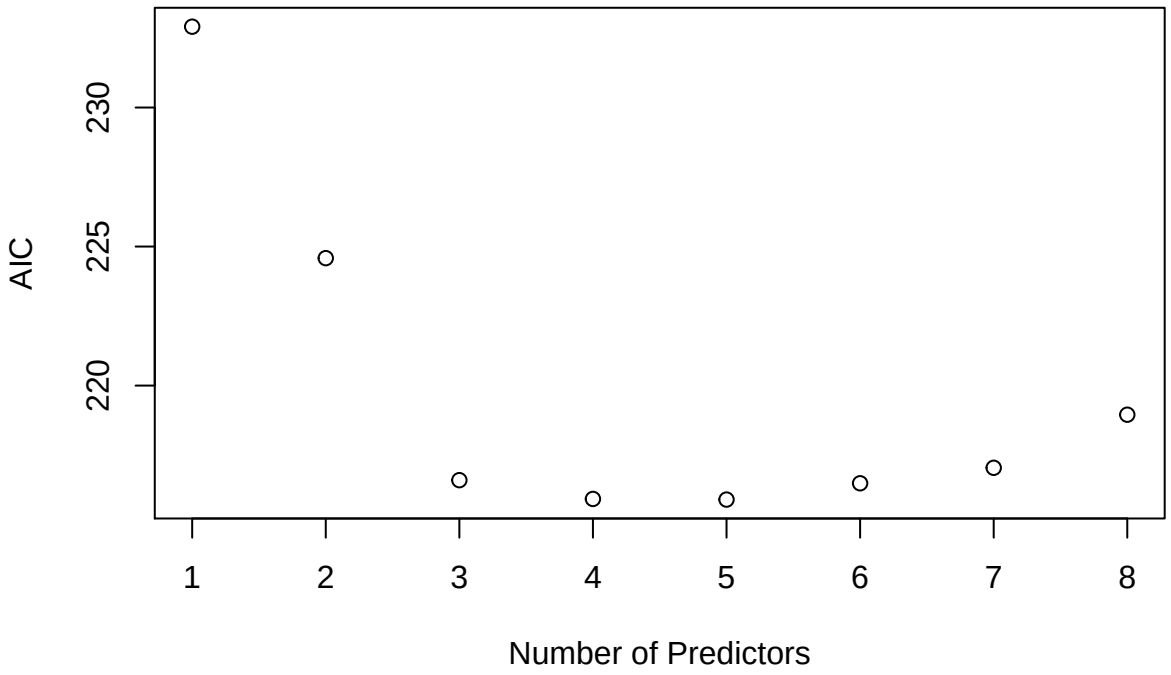
```
b1<-ols_step_best_subset(k1)
b1
```

```
##                   Best Subsets Regression
## ---------------------------------------------------------
## Model Index    Predictors
## ---------------------------------------------------------
##      1         lcavol
##      2         lcavol lweight
##      3         lcavol lweight svi
##      4         lcavol lweight lbph svi
##      5         lcavol lweight age lbph svi
##      6         lcavol lweight age lbph svi pgg45
##      7         lcavol lweight age lbph svi lcp pgg45
##      8         lcavol lweight age lbph svi lcp gleason pgg45
## ---------------------------------------------------------
##
##                                                  Subsets Regression Summary
## ----------------------------------------------------------------------------------------------------------------
##                        Adj.        Pred
## Model    R-Square    R-Square    R-Square     C(p)        AIC         SBIC        SBC        MSEP
## ----------------------------------------------------------------------------------------------------------------
##   1        0.5394      0.5346      0.5179    24.3946    232.9080    -43.1213    240.6321    60.1553
##   2        0.5859      0.5771      0.5534    14.5415    224.5837    -51.1891    234.8825    54.6631
##   3        0.6264      0.6144      0.5869     6.2169    216.5979    -58.5271    229.4714    49.8518
##   4        0.6366      0.6208      0.5898     5.6264    215.9223    -58.8841    231.3705    49.0284
##   5        0.6441      0.6245      0.5882     5.7150    215.8997    -58.5526    233.9227    48.5502
##   6        0.6493      0.6259      0.5856     6.4020    216.4854    -57.6133    237.0831    48.3850
##   7        0.6544      0.6273      0.5835     7.0822    217.0428    -56.6319    240.2152    48.2125
##   8        0.6548      0.6234       0.576     9.0000    218.9522    -54.5019    244.6993    48.7211
## ----------------------------------------------------------------------------------------------------------------
## 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
```

```
plot(b1)
```

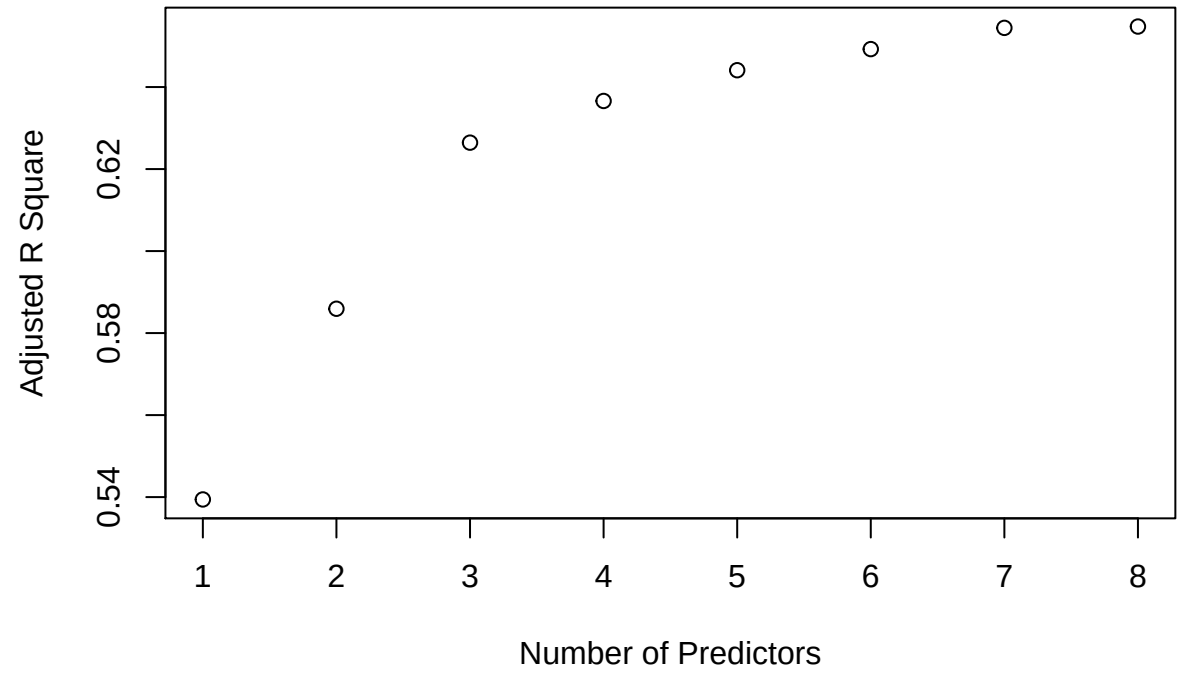



```
plot(1:8,b1$aic,xlab="Number of Predictors",ylab="AIC")
```



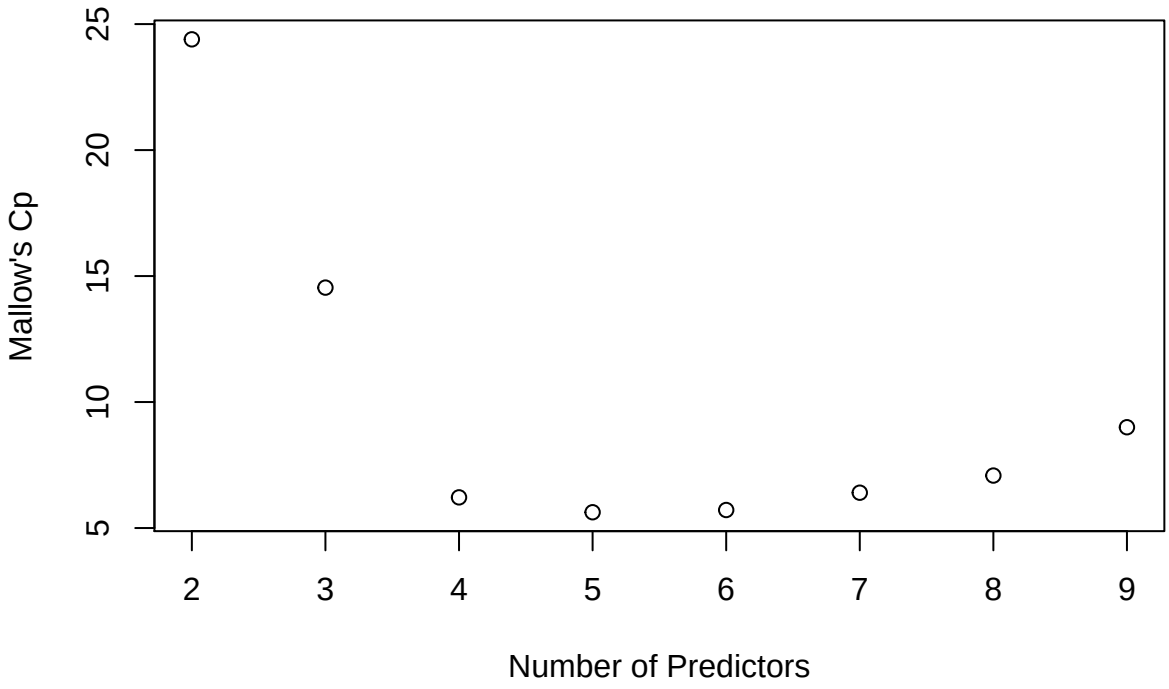
c-) Adjusted R_a^2. on the graph, the elbow point is for 3 variables models. The variables are lcavol, lweight, and svi. The same model in part a and b.

```
plot(1:8,b1$rsquare,xlab="Number of Predictors",ylab="Adjusted R Square")
```



d-) Mallows C_p. Based on Mallow C_p, the model was selected in previous part, p=4 and C_p=6.2. For the model with lcavol,lweight, lbph, and svi; p=5 and C_p=5.6. It is a judgment call, we want to C_p to be close to p. In this model, lbph is not significant.

```
#p includes the intercept, adding one.
plot(2:9,b1$cp,xlab="Number of Predictors",ylab="Mallow's Cp")
```




```
f2<-lm(lpsa~lcavol+lweight+lbph+svi,data=prostate)
summary(f2)
```

```
##
## Call:
## lm(formula = lpsa ~ lcavol + lweight + lbph + svi, data = prostate)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.82653 -0.42270  0.04362  0.47041  1.48530
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.14554    0.59747   0.244  0.80809
## lcavol       0.54960    0.07406   7.422 5.64e-11 ***
## lweight      0.39088    0.16600   2.355  0.02067 *
## lbph         0.09009    0.05617   1.604  0.11213
## svi          0.71174    0.20996   3.390  0.00103 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7108 on 92 degrees of freedom
## Multiple R-squared:  0.6366, Adjusted R-squared:  0.6208
## F-statistic: 40.29 on 4 and 92 DF,  p-value: < 2.2e-16
```

Problem 3

Refer to the SENIC data set in problem 1. Length of stay (Y) is to be predicted, and the pool of potential predictor variables includes all other variables in the data set except medical school affiliation and region. It is believed that a model with $log(Y)$ as the response variable and the predictor variables in first-order terms with no interaction terms will be appropriate. Consider cases 57-113 to constitute the model-building data

set to be used for the following analyses.(45 points, 9 points each)

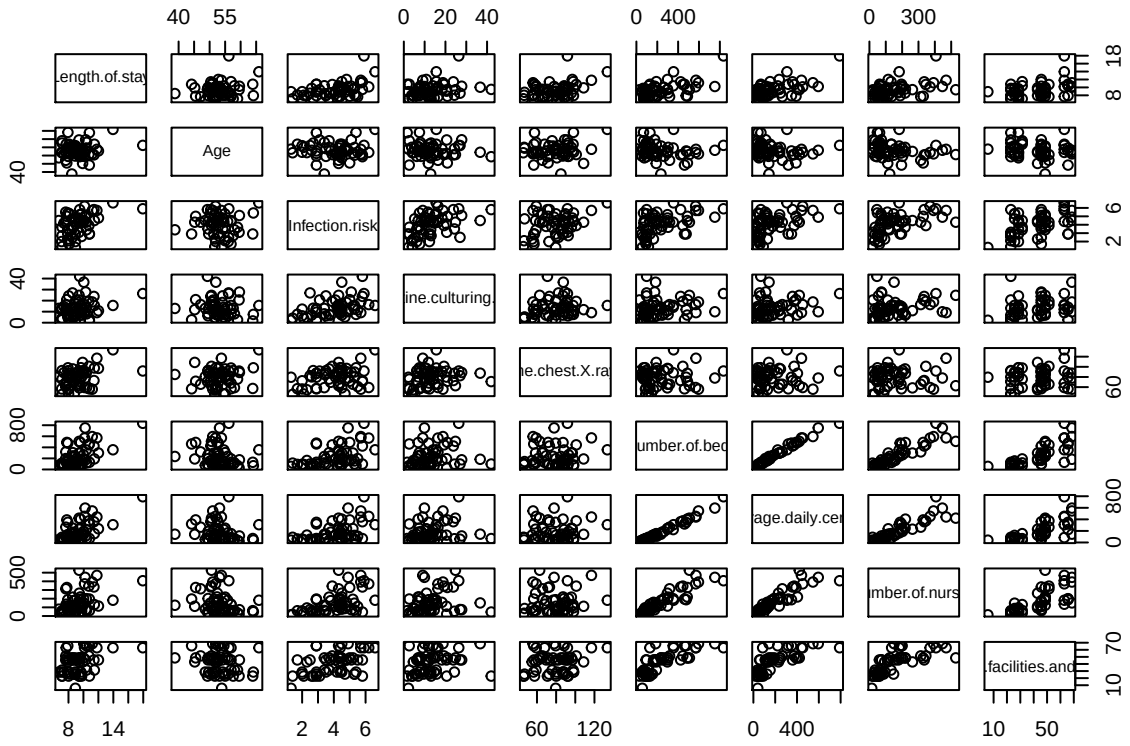
a-) Prepare separate dot plots for each of the predictor variables. Are there any noteworthy features in these plots? Comment.

Length of stay is highly correlated with Number of beds and average daily census. It is moderately correlated with Infection.risk,Number.of.nurses.

```
dim(SENIC)
```

```
## [1] 113  11
```

```
dev=SENIC[57:113,-c(7:8)]
hold=SENIC[1:56,]
plot(dev)
```




```
round(cor(dev),2)
```

```
##                                      Length.of.stay   Age Infection.risk
## Length.of.stay                                 1.00  0.19           0.47
## Age                                            0.19  1.00           0.03
## Infection.risk                                 0.47  0.03           1.00
## Routine.culturing.ratio                        0.26 -0.10           0.45
## Routine.chest.X.ray.ratio                      0.36  0.16           0.33
## Number.of.beds                                 0.59 -0.20           0.49
## Average.daily.census                           0.63 -0.17           0.50
## Number.of.nurses                               0.47 -0.24           0.53
## Available.facilities.and.services              0.40 -0.16           0.45
##                                      Routine.culturing.ratio
## Length.of.stay                                          0.26
## Age                                                    -0.10
## Infection.risk                                          0.45
## Routine.culturing.ratio                                 1.00
```

```
## Routine.chest.X.ray.ratio                          0.19
## Number.of.beds                                     0.17
## Average.daily.census                               0.20
## Number.of.nurses                                   0.24
## Available.facilities.and.services                  0.24
##                                  Routine.chest.X.ray.ratio Number.of.beds
## Length.of.stay                                        0.36           0.59
## Age                                                   0.16          -0.20
## Infection.risk                                        0.33           0.49
## Routine.culturing.ratio                               0.19           0.17
## Routine.chest.X.ray.ratio                             1.00           0.07
## Number.of.beds                                        0.07           1.00
## Average.daily.census                                  0.09           0.99
## Number.of.nurses                                      0.06           0.91
## Available.facilities.and.services                     0.13           0.76
##                                  Average.daily.census Number.of.nurses
## Length.of.stay                                   0.63             0.47
## Age                                             -0.17            -0.24
## Infection.risk                                   0.50             0.53
## Routine.culturing.ratio                          0.20             0.24
## Routine.chest.X.ray.ratio                        0.09             0.06
## Number.of.beds                                   0.99             0.91
## Average.daily.census                             1.00             0.90
## Number.of.nurses                                 0.90             1.00
## Available.facilities.and.services                0.73             0.71
##                                  Available.facilities.and.services
## Length.of.stay                                                0.40
## Age                                                          -0.16
## Infection.risk                                                0.45
## Routine.culturing.ratio                                       0.24
## Routine.chest.X.ray.ratio                                     0.13
## Number.of.beds                                                0.76
## Average.daily.census                                          0.73
## Number.of.nurses                                              0.71
## Available.facilities.and.services                             1.00
```

b-) Obtain the scatter plot matrix. Also obtain the correlation matrix of the X variables. Is there evidence of strong linear pairwise associations among the predictor variables here?

See above. The number of beds is highly correlated with average daily census, Number.of.nurses, and Available.facilities.and.services.

c-) Obtain the three best subsets according to the C_p criterion, Which of these subset models appears to have the smallest bias?

See below, model #3 has the smallest bias, the variables are

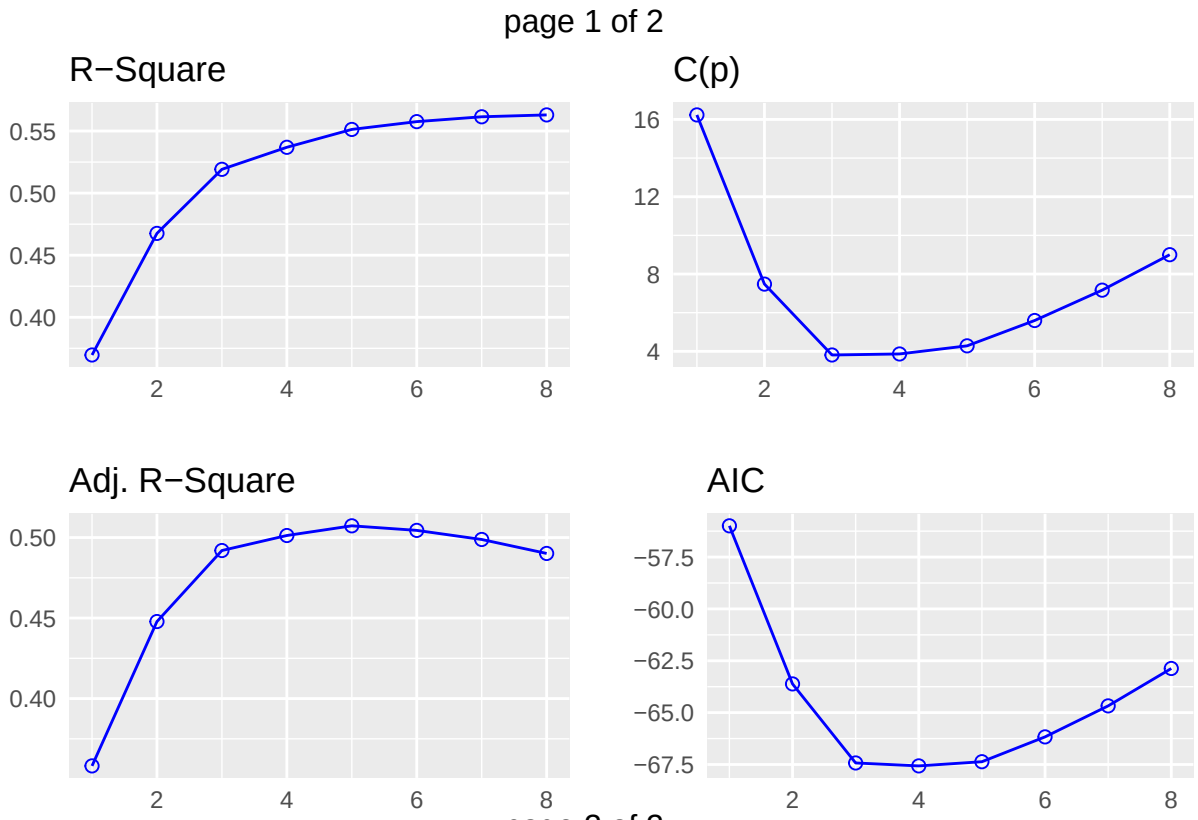
Age,Routine.chest.X.ray.ratio, and Average.daily.census. The R^2 is 51%. All variables are significant.

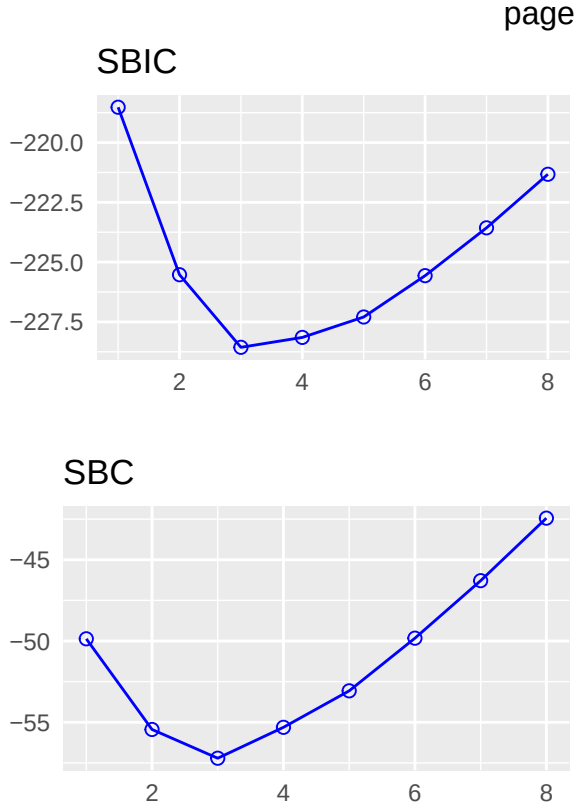
```
k1<-lm(log(Length.of.stay)~.,data=dev)
b1<-ols_step_best_subset(k1)
b1
```

```
##                                                                                                  Best Subsets Regression
## ---------------------------------------------------------------------------------------------------------------------------
## Model Index    Predictors
## ---------------------------------------------------------------------------------------------------------------------------
```

```
##      1         Average.daily.census
##      2         Routine.chest.X.ray.ratio Average.daily.census
##      3         Age Routine.chest.X.ray.ratio Average.daily.census
##      4         Age Routine.chest.X.ray.ratio Average.daily.census Number.of.nurses
##      5         Age Routine.culturing.ratio Routine.chest.X.ray.ratio Average.daily.census Number.of.
##      6         Age Infection.risk Routine.chest.X.ray.ratio Number.of.beds Average.daily.census Numbe
##      7         Age Infection.risk Routine.culturing.ratio Routine.chest.X.ray.ratio Average.daily.cen
##      8         Age Infection.risk Routine.culturing.ratio Routine.chest.X.ray.ratio Number.of.beds Av
## ----------------------------------------------------------------------------------------------------
##
##                                                Subsets Regression Summary
## ----------------------------------------------------------------------------------------------------
##                        Adj.        Pred
## Model    R-Square    R-Square    R-Square     C(p)        AIC         SBIC         SBC        MSEP
## ----------------------------------------------------------------------------------------------------
##   1        0.3696      0.3582      0.2968    16.2329    -55.9935    -218.5217    -49.8643    1.1657
##   2        0.4676      0.4478      0.3849     7.4790    -63.6155    -225.5275    -55.4433    1.0032
##   3        0.5192      0.4919      0.4244     3.8112    -67.4264    -228.5634    -57.2111    0.9234
##   4        0.5369      0.5013      0.4324     3.8638    -67.5679    -228.1496    -55.3096    0.9068
##   5        0.5513      0.5073      0.4368     4.2839    -67.3665    -227.2958    -53.0652    0.8962
##   6        0.5576      0.5045      0.4033     5.5946    -66.1694    -225.5666    -49.8250    0.9017
##   7        0.5615      0.4988       0.407     7.1658    -64.6746    -223.5663    -46.2871    0.9123
##   8        0.5630      0.4901      0.3747     9.0000    -62.8712    -221.3255    -42.4406    0.9286
## ----------------------------------------------------------------------------------------------------
## 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
```

```
plot(b1)
```






```
f<-lm(log(Length.of.stay)~Age+Routine.chest.X.ray.ratio+Average.daily.census,data=dev)
summary(f)
```

```
##
## Call:
## lm(formula = log(Length.of.stay) ~ Age + Routine.chest.X.ray.ratio +
##     Average.daily.census, data = dev)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.25937 -0.08657  0.02955  0.07747  0.21522
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                1.4055759  0.2045023   6.873 7.22e-09 ***
## Age                        0.0089343  0.0037461   2.385  0.02069 *
## Routine.chest.X.ray.ratio 0.0027050  0.0009643   2.805  0.00702 **
## Average.daily.census      0.0006738  0.0001049   6.420 3.86e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1272 on 53 degrees of freedom
## Multiple R-squared:  0.5192, Adjusted R-squared:  0.4919
## F-statistic: 19.07 on 3 and 53 DF,  p-value: 1.614e-08
```

d-) The regression model identified as best in part c is to be validated by means of the validation data set consisting of cases 1-56. Fit the regression model identified in part c as best to the validation data set. Compare the estimated regression coefficients and their estimated standard deviations with those obtained in Part C.

The coeeficents are similar but R^2 is reduced to 30%.

```
f1<-lm(log(Length.of.stay)~Age+Routine.chest.X.ray.ratio+Average.daily.census,data=hold)
summary(f1)
```

```
##
## Call:
## lm(formula = log(Length.of.stay) ~ Age + Routine.chest.X.ray.ratio +
##     Average.daily.census, data = hold)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.31188 -0.10566 -0.00886  0.09251  0.50058
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                1.4249909  0.2872850   4.960 7.92e-06 ***
## Age                        0.0091956  0.0048558   1.894  0.06383 .
## Routine.chest.X.ray.ratio 0.0035052  0.0010068   3.482  0.00102 **
## Average.daily.census      0.0003610  0.0001431   2.522  0.01476 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1497 on 52 degrees of freedom
## Multiple R-squared:  0.2934, Adjusted R-squared:  0.2526
## F-statistic: 7.196 on 3 and 52 DF,  p-value: 0.0003955
```

```
rbind(f$coefficients,f1$coefficients)
```

```
##      (Intercept)         Age Routine.chest.X.ray.ratio Average.daily.census
## [1,]    1.405576 0.008934253               0.002705048         0.0006737651
## [2,]    1.424991 0.009195614               0.003505178         0.0003610419
```

e-) Also compare the error mean squares 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?

Residual standard error: 0.1272 vs 0.1497 Multiple R-squared: 0.5192 vs 0.2934

The model coefficients and Residual standard error are very close to each other. However, R^2 is different indicating that there could be outliers in the hold out sample. The further investigation is needed.