# STAT512 Division 1 HW 5

*Yi Yang*

*10/22/2018*

1. In a regression analysis of on-the-job head injuries of warehouse laborers caused by falling objects, $Y$ is a measure of severity of the injury, $X_1$ is an index reflecting both the weight of the object and the distance it fell, and $X_2$ and $X_3$ are indicator variables for nature of head protection worn at the time of the accident, coded as follows:

| Type of protection | $X_2$ | $X_3$ |
|---|---|---|
| Hard hat | 1 | 0 |
| Bump cap | 0 | 1 |
| None | 0 | 0 |

The response function to be used in the study is $\mathbb{E}\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$.

a) Develop the response function for each type of pretection category.

b) For each of the following questions, specify the $H_0$ and $H_a$ for the appropriate test with the appropriate symbols. b.1) When $X_1$ is fixed, does wearing a bump cap reduce the expected severity of injury as compared with wearing no protection? b.2) When $X_1$ is fixed, is the expected severity of injury the same when wearing a hard hat as when wearing a bump cap?

**Solution:** a) Hard hat:

$$\mathbb{E}\{Y\} = \beta_0 + \beta_2 + \beta_1 X_1$$

Bump cap:

$$\mathbb{E}\{Y\} = \beta_0 + \beta_3 + \beta_1 X_1$$

None:

$$\mathbb{E}\{Y\} = \beta_0 + \beta_1 X_1$$

b) b1.

$$H_0 : \beta_3 \geq 0, \quad H_a : \beta_3 < 0$$

b2.

$$H_0 : \beta_2 = \beta_3, \quad H_a : \beta_2 \neq \beta_3$$

2. A tax consultant studied the current relation between selling price and assessed valuation of one-family residential dwelling in a large tax district by obtaining data for a random sample of 16 recent sales transactions located on corner lots and 48 transactions not located on corner lots. Data is in valuation.csv

Assume the regression model is $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$

a) Plot the sample data for the two populations (corner lots vs noncorner lots) in one scatter plot with different symbolic mark for each population. Do you think the regression relations are the same for the two population?

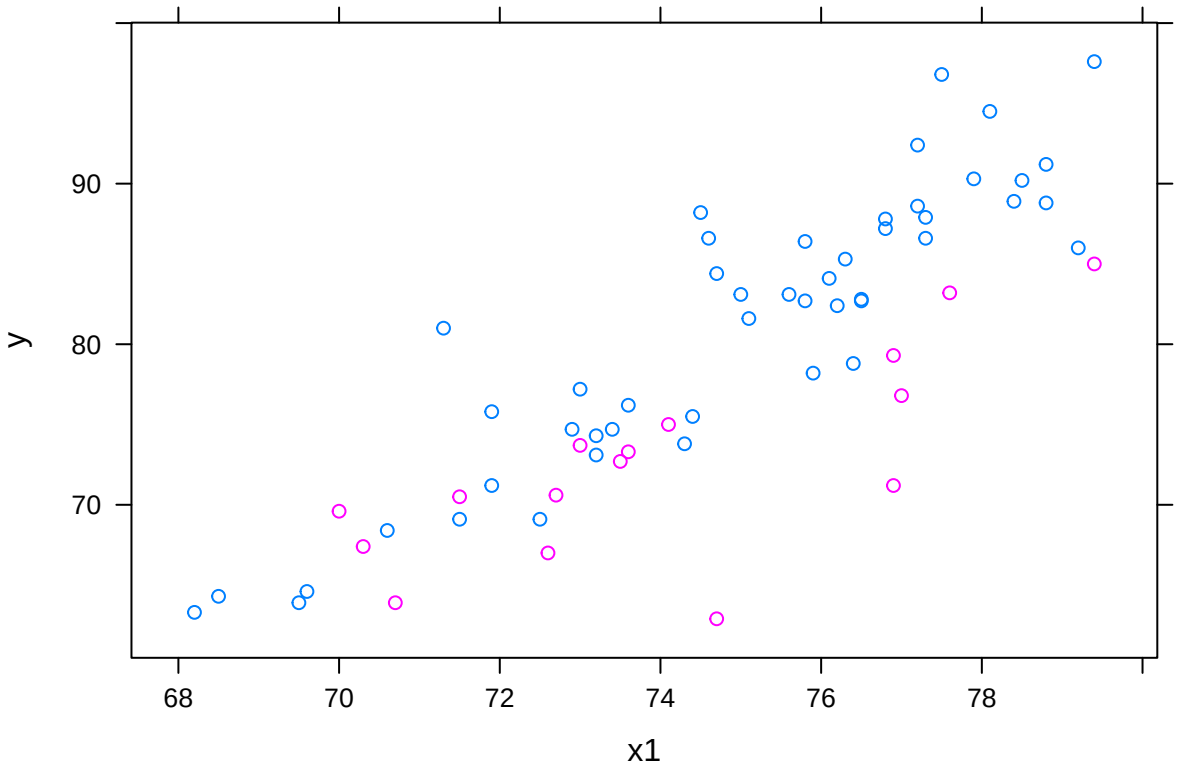
b) Test for identity of the regression functions for dwellings on corner lots and dwellings in other locations. $\alpha = 0.05$.

**Solution:** a)

```r
val <- read.csv('data/valuation.csv', header = TRUE)
y <- val$y
x1 <- val$x1
x2 <- val$x2
require('lattice')
```

```
## Loading required package: lattice
```

```r
xyplot(y~x1, group=x2, data=val)
```




It seems the regression relations are not the same by oberserving the scatter plot.

b) The hypothesis test is given by

$$H_0 : \beta_2 = 0, \quad H_a : \beta_2 \neq 0$$

```r
val.mod <- lm(y~x1 + x2, val)
summary(val.mod)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2, data = val)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.4141  -2.2927  -0.1456   1.8678   9.2341
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -107.4597    13.5509   -7.93 5.80e-11 ***
## x1             2.5165     0.1806   13.93  < 2e-16 ***
## x2            -6.2057     1.1933   -5.20 2.45e-06 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.093 on 61 degrees of freedom
## Multiple R-squared:  0.8014, Adjusted R-squared:  0.7949
## F-statistic: 123.1 on 2 and 61 DF,  p-value: < 2.2e-16
```

```
anova(val.mod)
```

```
## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## x1         1 3670.9  3670.9 219.083 < 2.2e-16 ***
## x2         1  453.1   453.1  27.044 2.447e-06 ***
## Residuals 61 1022.1    16.8
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
qf(0.95,1,61)
```

```
## [1] 3.998494
```

The partial F test is given by

$$F^* = \frac{SSR(X_2|X_1)/1}{MSE} = \frac{453.1}{16.8} = 26.97$$

The critical value at significance level $\alpha = 0.05$ is 3.998494, which is less than $F^*$. Therefore, the null hypothesis can be rejected, the two regression functions are not identical.

3. (Use R for the question) A personnel officer in a governmental agency administered four newly aptitudde tests to each of the 25 applicants for entry level clerical positions in the agency. For purpose of study, all 25 applicants were accepted for positions irrespective of their test scores. After a probationary period, each applicant was rated for proficiency on the job. The scores on the four tests $(X_1, X_2, X_3, X_4)$ and the job proficiency score $(Y)$ for the 25 employees were recorded in proficiency.csv

a) Obtain the scatter plot matrix and the correlation matrix of the $X$ variables, what do the scatter plots suggest about the nature of the function relationship between the response variable and each of the predictor variables?

b) Fit the multiple function containing all four predictors at first order terms. Does it appear that all predictor variables should be retained?

c) Select the best subset regression models according to the $R^2adj$, $C_p$, $AIC_p$, $BIC_p$, and $PRESS$

and discuss your selection. Fit the model to the data in proficiency.csv

d) To assess internally the predictive ability of the regression model identified in c), compare the PRESS and SSE, what does this comparison suggest about the validity of MSE as in indicative of the predictive ability of the fitted model?

e) Run a 5 fold cross validation on the model identified in c).

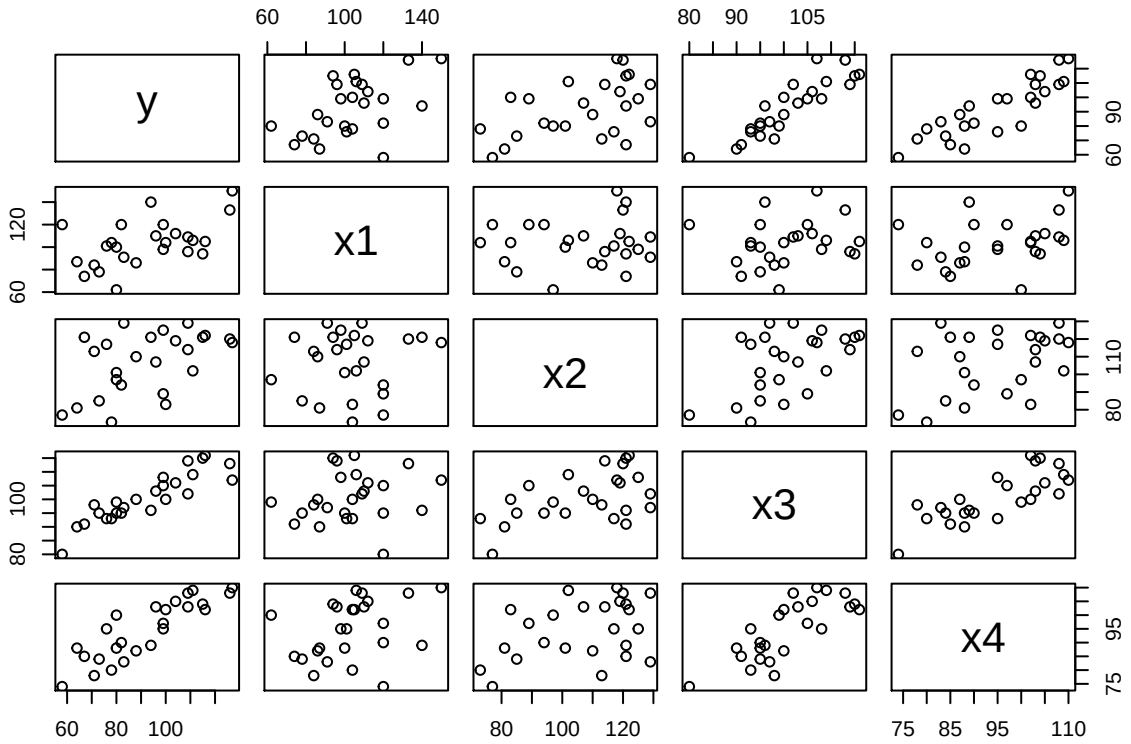
f) To assess externally the validity of the regression model identified in c), 25 additional applicants for entry-level clerical positions in the agency were similarly tested and hired irrespective of their test scores. The data is in proficiencyTest.csv

Fit the model identified in c) to the validation data set. Compare the regression coefficients and their estimated standard deviation to the results in c). Do the estimates for the validation data set appear to be reasonably similar to those obtained for the model-building data set (proficiency.csv)?

**Solution:**

a)

```
prof <- read.csv('data/proficiency.csv',header = TRUE)
plot(prof)
```




From scatter plot, the response variable has a strong linear relationship with $X_3$ and $X_4$, but has the least relationship with $X_2$.

b.

```
prof.mod <- lm(y~x1+x2+x3+x4, prof)
summary(prof.mod)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3 + x4, data = prof)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.9779 -3.4506  0.0941  2.4749  5.9959
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -124.38182    9.94106 -12.512 6.48e-11 ***
## x1             0.29573    0.04397   6.725 1.52e-06 ***
## x2             0.04829    0.05662   0.853  0.40383
## x3             1.30601    0.16409   7.959 1.26e-07 ***
## x4             0.51982    0.13194   3.940  0.00081 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.099 on 20 degrees of freedom
## Multiple R-squared:  0.9629, Adjusted R-squared:  0.9555
```

```
## F-statistic: 129.7 on 4 and 20 DF,  p-value: 5.262e-14
```

The multiple linear regression function is given by

$$Y = -124.38 + 0.286X_1 + 0.048X_2 + 1.306X_3 + 0.520X_4$$

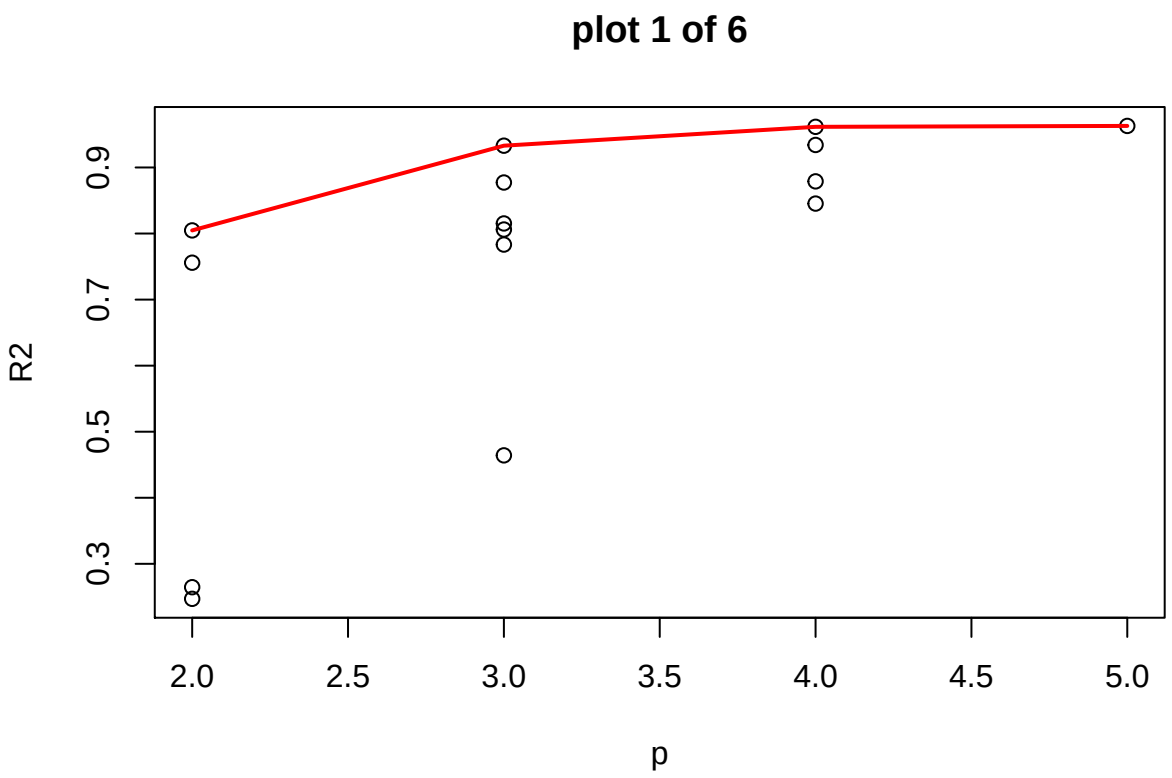
From the T test results in the summary table, $X_1$, $X_3$ and $X_4$ should be retained in the model.

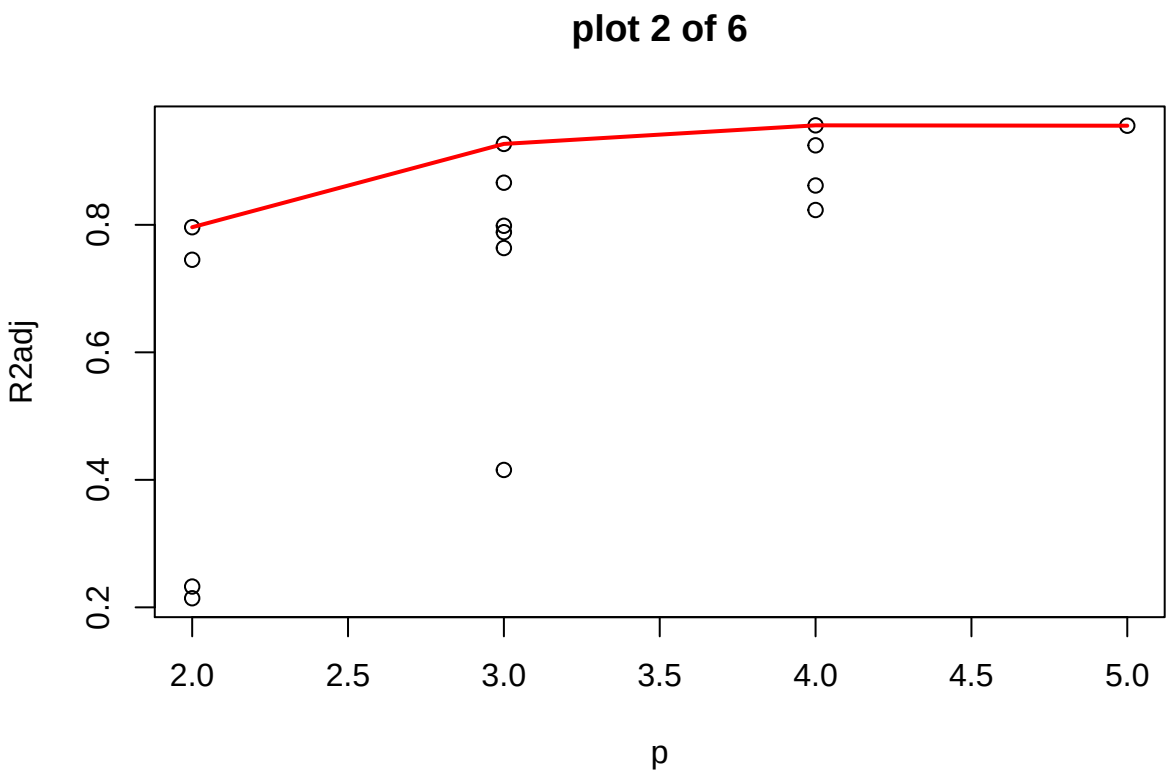
c)

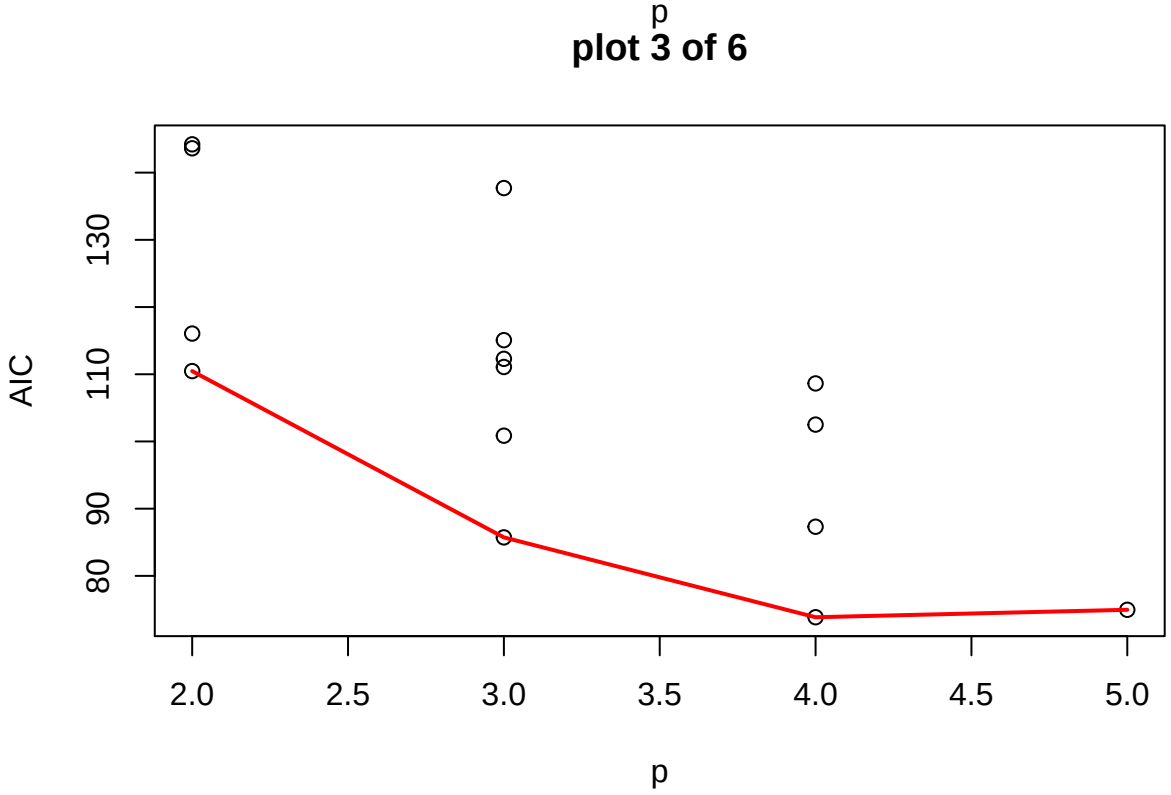
```
library(ALSM)
```

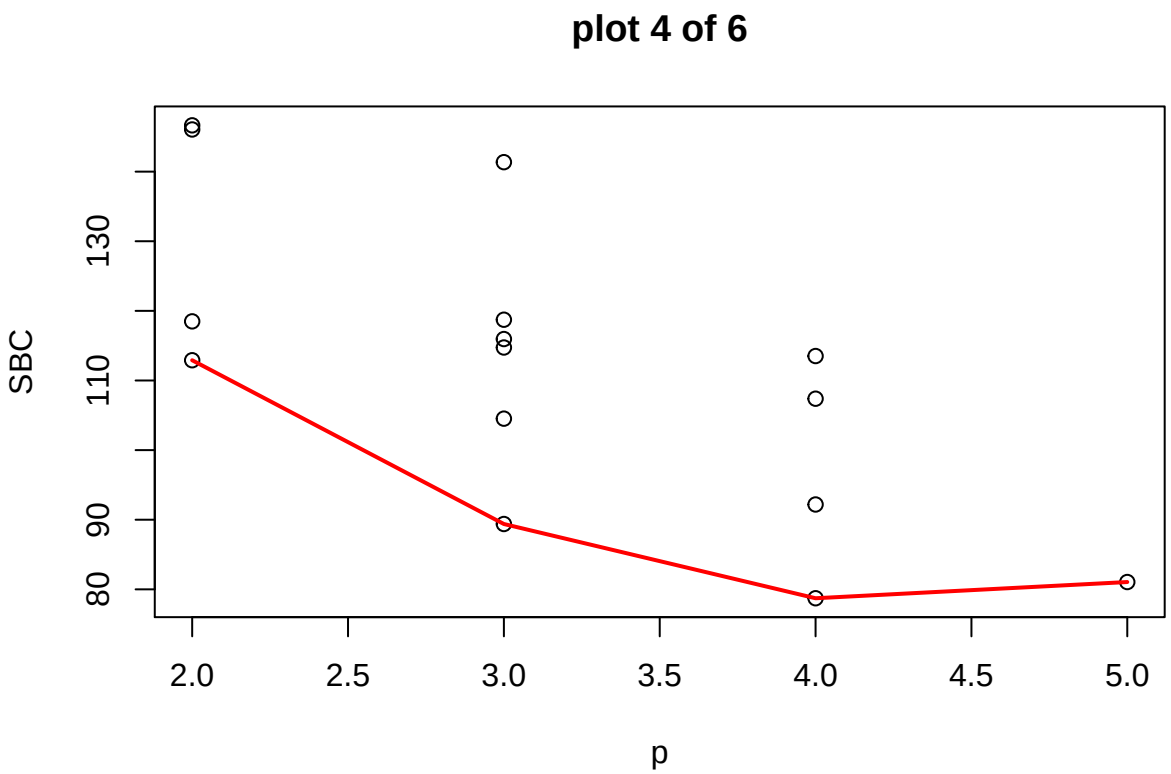
```
## Loading required package: leaps
## Loading required package: SuppDists
## Loading required package: car
## Loading required package: carData
##
## Attaching package: 'ALSM'
## The following object is masked from 'package:lattice':
##
##     oneway
```

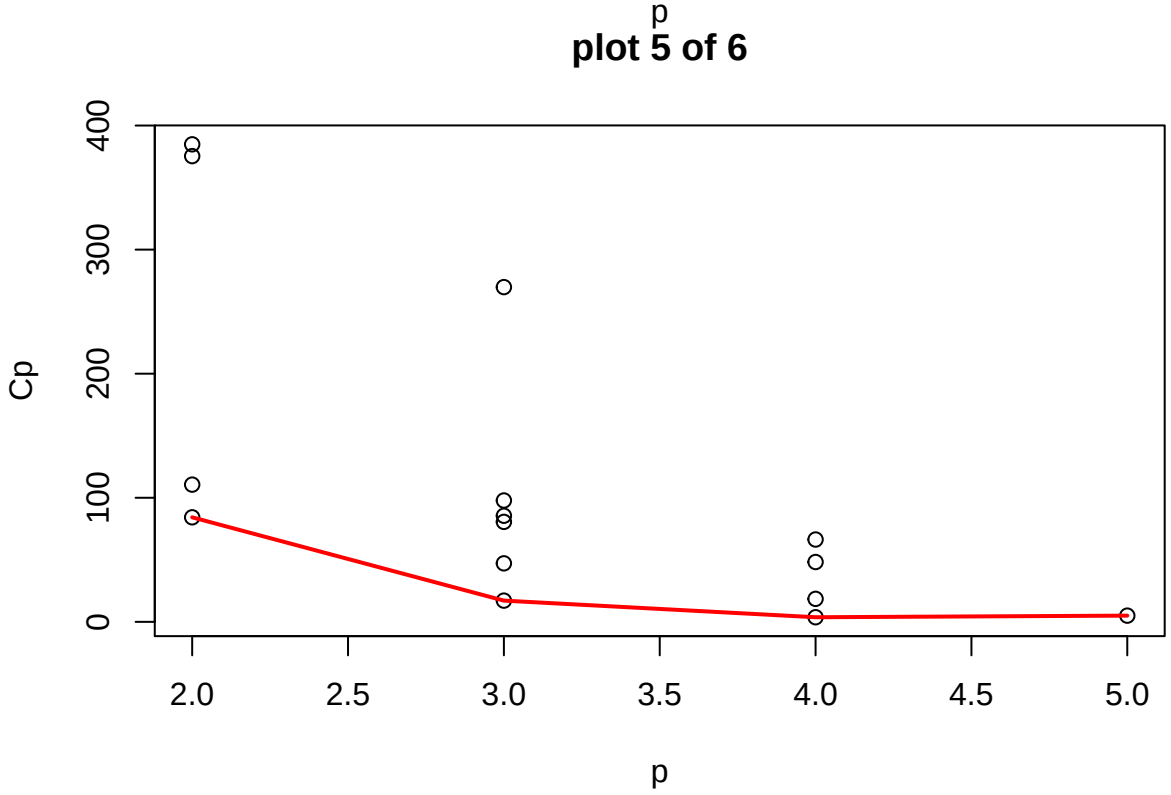
```
plotmodel.s(prof[,2:5],prof[,1])
```

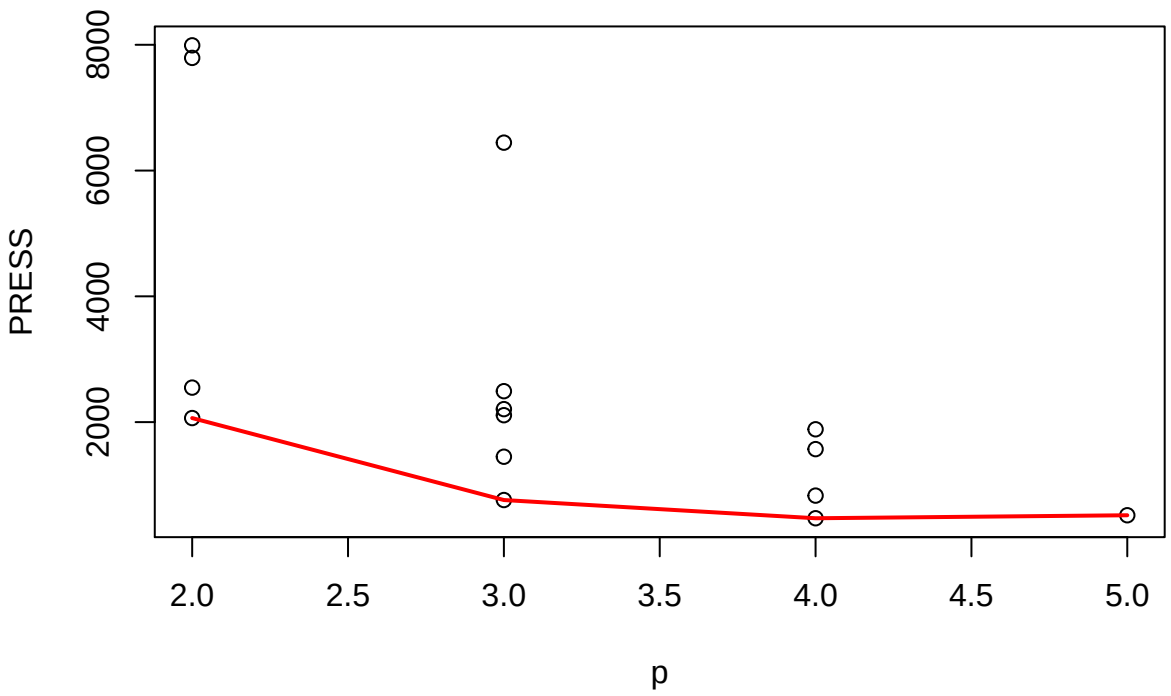



plot 2 of 6
R2adj
0.2
0.4
0.6
0.8
2.0
2.5
3.0
3.5
4.0
4.5
5.0
p


plot 3 of 6
AIC
80
90
110
130
2.0
2.5
3.0
3.5
4.0
4.5
5.0
p


plot 4 of 6
SBC
80
90
110
130
2.0
2.5
3.0
3.5
4.0
4.5
5.0
p


plot 5 of 6
Cp
0
100
200
300
400
2.0
2.5
3.0
3.5
4.0
4.5
5.0
p




```
library('leaps')
BestSub(prof[,2:5],prof[,1],num=1)
```

```
##   p 1 2 3 4      SSEp        r2    r2.adj        Cp      AICp      SBCp
## 1 2 0 0 1 0 1768.0228 0.8047247 0.7962344 84.246496 110.46853 112.90629
## 2 3 1 0 1 0  606.6574 0.9329956 0.9269043 17.112978  85.72721  89.38384
## 3 4 1 0 1 1  348.1970 0.9615422 0.9560482  3.727399  73.84732  78.72282
## 4 5 1 1 1 1  335.9775 0.9628918 0.9554702  5.000000  74.95421  81.04859
##      PRESSp
## 1 2064.5976
## 2  760.9744
## 3  471.4520
## 4  518.9885
```

From $R^2_{adj}$, it reveals that subset $X_1, X_3, X_4$ has $R^2_{adj} = 0.9560$ which attains the lowest increasing rate. Other selection critia also gives the same result.

d) Based on the comparison between $PRESS = 471.4520$ and $SSE = 348.1970$, it implies there is a large bias exixsts if we rely on the SSE criteria. Therefore, MSE is not a good indicator of the predictive ability of the fitted model.

e)

```
library(MASS)
library(leaps)
library(caret)
```

```
## Loading required package: ggplot2
```

```
set.seed(123)
train.control <- trainControl(method = 'cv',number = 5)
step.model1 <- train(y~x1+x3+x4,data=prof,method='leapBackward',tuneGrid=data.frame(nvmax=3),
```

```
                    trControl=train.control)
step.model1
```

```
## Linear Regression with Backwards Selection
##
## 25 samples
##  3 predictor
##
## No pre-processing
## Resampling: Cross-Validated (5 fold)
## Summary of sample sizes: 20, 20, 20, 20, 20
## Resampling results:
##
##   RMSE      Rsquared   MAE
##   4.195349  0.9586264  3.769349
##
## Tuning parameter 'nvmax' was held constant at a value of 3
```

f)

```
prof.test <- read.csv('data/proficiencyTest.csv', header = TRUE)
colnames(prof.test) <- c('ytest','x1test','x2test','x3test','x4test')
prof.test.mod <- lm(ytest~x1test+x3test+x4test,prof.test)
summary(prof.test.mod)
```

```
##
## Call:
## lm(formula = ytest ~ x1test + x3test + x4test, data = prof.test)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.4619 -2.3836  0.6834  2.1123  7.2394
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -122.76705   11.84783 -10.362 1.04e-09 ***
## x1test         0.31238    0.04729   6.605 1.54e-06 ***
## x3test         1.40676    0.23262   6.048 5.31e-06 ***
## x4test         0.42838    0.19749   2.169   0.0417 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.284 on 21 degrees of freedom
## Multiple R-squared:  0.9489, Adjusted R-squared:  0.9416
## F-statistic:   130 on 3 and 21 DF,  p-value: 1.017e-13
```

```
summary(lm(y~x1+x3+x4,prof))
```

```
##
## Call:
## lm(formula = y ~ x1 + x3 + x4, data = prof)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.4579 -3.1563 -0.2057  1.8070  6.6083
##
```

```
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -124.20002    9.87406 -12.578 3.04e-11 ***
## x1             0.29633    0.04368   6.784 1.04e-06 ***
## x3             1.35697    0.15183   8.937 1.33e-08 ***
## x4             0.51742    0.13105   3.948 0.000735 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.072 on 21 degrees of freedom
## Multiple R-squared:  0.9615, Adjusted R-squared:  0.956
## F-statistic:   175 on 3 and 21 DF,  p-value: 5.16e-15
```

By comparing the coefficients and their standard deviation of the fitting model over the training data set and the test data set, it is concluded that the two regression functions are reasonably similar to each other.