Homework 5 (50 pts)
due 10/26

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, X1 is an index reflecting both the weight of the object and the distance it fell, and X2 and X3 are indicator variables for nature of head protection worn at the time of the accident, coded as follows:

Type of protection	X2	X3
Hard hat	1	0
Bump cap	0	1
None	0	0

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

a) (4) Develop the response function for each type of protection category.

Hard hat: $E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2$

Bump cap: $E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_3$

None: $E\{Y\} = \beta_0 + \beta_1 X_1$

b) (6) For each of the following questions, specify the Ho and Ha for the appropriate test with the appropriate symbols.

b.1) When X1 is fixed, does wearing a bump cap reduce the expected severity of injury as compared with wearing no protection?

$H_0: \beta_3 = 0$ (No protection) $E\{Y\} = \beta_0 + \beta_1 X_1$

$H_a: \beta_3 < 0$ (Bump cap) $E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_3$

b.2) When X1 is fixed, is the expected severity of injury the same when wearing a hard hat as when wearing a bump cap?

topic 14

$H_0: \beta_2 = \beta_3 = \beta_{new}$ (Same) $E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_{new}$

$H_a: \beta_2 \neq \beta_3$ (different)

$E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 \; for\ Hard\ Hat$

$E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_3 \; for\ Bump\ Cap$

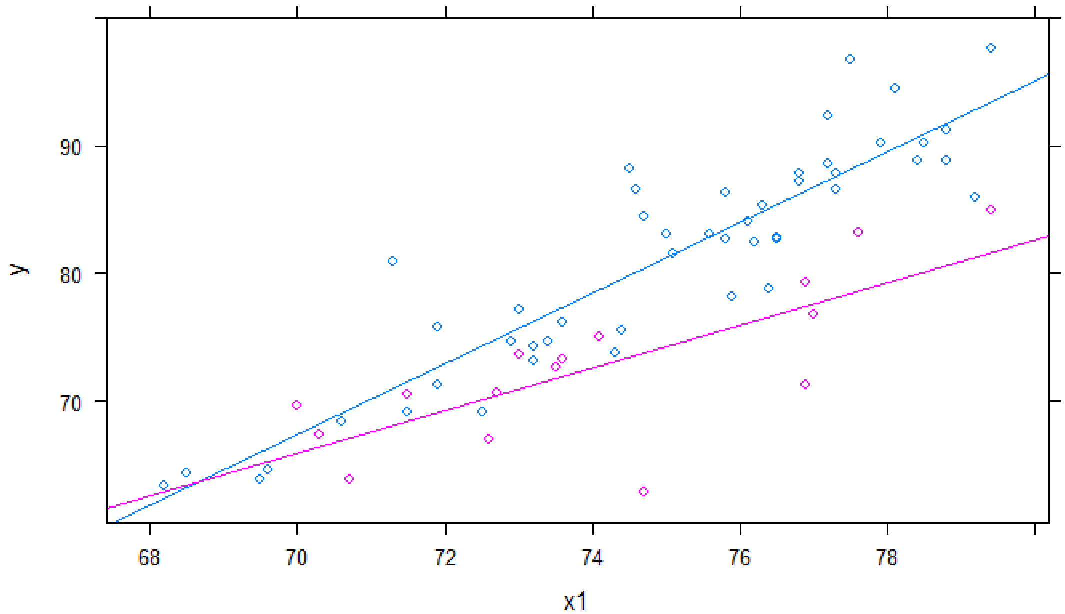
2. A tax consultant studied the current relation between selling price and assessed valuation of one-family residential dwelling in a large tax district by obtaining data for a random sample of 16 recent sales transactions located on corner lots and 48 transactions not located on corner lots. Data is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

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

R code is shown in Appendix.

a)(4) Plot the sample data for the two populations (corner lots vs non-corner 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?

Figure 1.



No, they are not the same for the two population, the two regression lines have different slope and intercept.

b)(6) Test for identity of the regression functions for dwellings on corner lots and dwellings in other locations. $\alpha = 0.05$.

Figure 2. (Full Model)

```
> summary(fit)

Call:
lm(formula = y ~ x1 + x2, data = data)

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(fit)
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
```

$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon_i$

$H_0: \beta_2 = 0$

$H_a: \beta_2 \neq 0$

Full Model: $\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2$

Reduced Model: $\hat{Y} = \beta_0 + \beta_1 X_1$

From Figure 2, $df_R = n - (p-1) = 64 - (3-1) = 62$

SSE(F) = 1022.1, $df_F = n - p = 64 - 3 = 61$

$$F^* = \frac{\frac{SSE(R) - SSE(F)}{df_R - df_F}}{\frac{SSE(F)}{df_F}} = \frac{\frac{SSR(X2|X1)}{62-61}}{\frac{SSE(X1,X2)}{64-3}} = \frac{\frac{453.1}{62-61}}{\frac{1022.1}{61}} = 27.04148$$

Critical value: $F(1-\alpha; df_R - df_F, df_F) = F(0.95; 1{,}61) = 4.03$

Since $F^* = 27.04148 > F(1-\alpha; df_R - df_F, df_F) = F(0.95; 1{,}61) = 4.03$, reject H_0

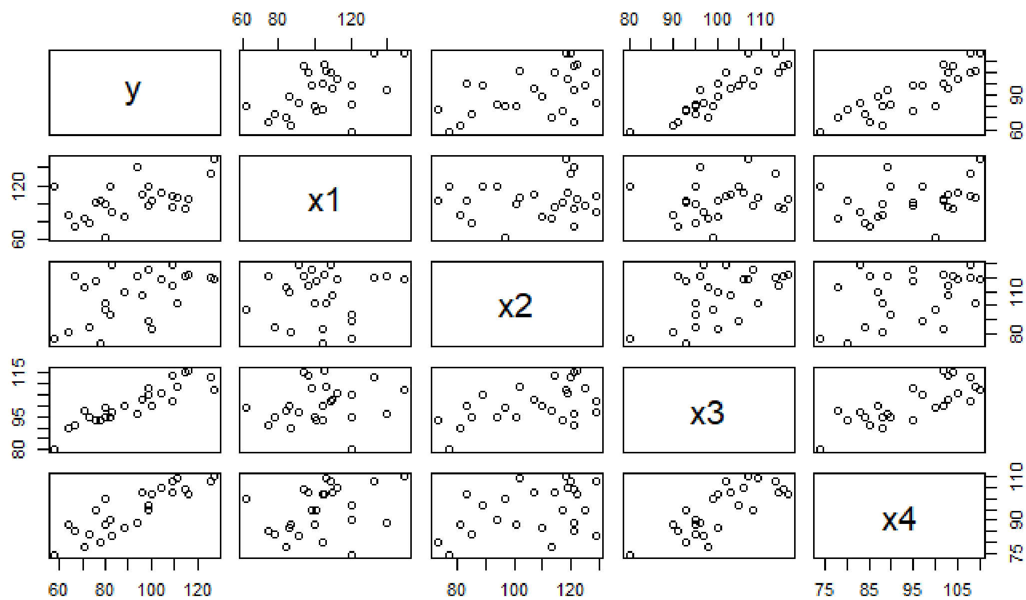
we reject $H_0: \beta_2 = 0$, so β_2 would not be equal to zero, which means β_2 has impact on Y. Hence, we are 95% confident that regression functions are different for dwellings on corner lots and dwellings in other locations.

3. (Use R for the question) A personnel officer in a governmental agency administered four newly developed aptitude 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 (X1, X2, X3, X4) and the job proficiency score (Y) for the 25 employees were recorded in proficiency.csv

R code is shown in Appendix.

a). (4) 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?

Figure 1. Scatter Plot Matrix



From Figure 1, we can see that both the relationship between y and x3 and the relationship between y and x4 are linear and highly correlated, and the relationship between y and x1 looks linear and correlated, but the relationship between y and x2 is weakly correlated. Furthermore, x3 and x4 are highly correlated.

Figure 2. Correlation Matrix

```
           y        x1        x2        x3        x4
y  1.0000000 0.5144107 0.4970057 0.8970645 0.8693865
x1 0.5144107 1.0000000 0.1022689 0.1807692 0.3266632
x2 0.4970057 0.1022689 1.0000000 0.5190448 0.3967101
x3 0.8970645 0.1807692 0.5190448 1.0000000 0.7820385
x4 0.8693865 0.3266632 0.3967101 0.7820385 1.0000000
```

The correlation between y and x3 is 0.8970645 and the correlation between y and x4 is 0.8593865, which supports the findings above that both the relationship between y and x3 and the relationship between y and x4 are linear and highly correlated. Furthermore, the correlation between y and x1 is 0.5144107, which is not very bad, and we can say that they are correlated. However, the correlation between y and x2 is 0.4970057, which is a little bit lower, which means they are weakly correlated. Moreover, the correlation between x3 and x4 is 0.7820385, which is very high, so we can conclude that x3 and x4 are highly correlated and there is multicollinearity in the model.

b). (4) Fit the multiple function containing gall four predictors at first-order terms. Does it appear that all predictor variables should be retained?

Figure 3.

```
> summary(fit1)

Call:
lm(formula = y ~ x1 + x2 + x3 + x4, data = pro)

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
```

Figure 4.

```
> Anova(fit1)
Anova Table (Type II tests)

Response: y
           Sum Sq Df F value    Pr(>F)    
x1         759.83  1 45.2310 1.524e-06 ***
x2          12.22  1  0.7274   0.40383    
x3        1064.15  1 63.3465 1.262e-07 ***
x4         260.74  1 15.5215   0.00081 ***
Residuals  335.98 20                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

From Figure 3, $\beta_2 = 0.04829$ which is much lower than other parameters and it is close to zero, while the p-value for β_2 is 0.40383>$\alpha = 0.05$, so we do not reject $H_0: \beta_2 = 0$.

From Figure 4, the SSR(X2|X1, X3, X4) = 12.22 which is much lower the sum of squares of other 3 predictor variables in Type II test. Plus, F* = 0.7274 < F(0.05, 1, 25-5)=4.35, so we do not reject $H_0: \beta_2 = 0$.

Hence, we are 95% confident that X2 has no impact on Y, so we can drop X2 from the model.

c). (4) Select the best subset regression models according to the $R^2_{adj}, Cp, AIC_p, BIC_p, and\ PRESS$ and discuss your selection. Fit the model to the data in proficiency.csv

Figure 5.

```
> BestSub(pro[,2:5], pro$y, num=4)
  p 1 2 3 4      SSEp        r2    r2.adj         Cp      AICp      SBCp    PRESSp
1 2 0 0 1 0 1768.0228 0.8047247 0.7962344  84.246496 110.46853 112.90629 2064.5976
1 2 0 0 0 1 2210.6887 0.7558329 0.7452170 110.597414 116.05459 118.49234 2548.6349
1 2 1 0 0 0 6658.1453 0.2646184 0.2326452 375.344689 143.61801 146.05576 7791.5994
1 2 0 1 0 0 6817.5291 0.2470147 0.2142762 384.832454 144.20941 146.64717 7991.0964
2 3 1 0 1 0  606.6574 0.9329956 0.9269043  17.112978  85.72721  89.38384  760.9744
2 3 0 0 1 1 1111.3126 0.8772573 0.8660988  47.153985 100.86053 104.51716 1449.6001
2 3 1 0 0 1 1672.5853 0.8152656 0.7984716  80.565307 111.08125 114.73788 2109.8967
2 3 0 1 1 0 1755.8127 0.8060733 0.7884436  85.519650 112.29528 115.95191 2206.6460
3 4 1 0 1 1  348.1970 0.9615422 0.9560482   3.727399  73.84732  78.72282  471.4520
3 4 1 1 1 0  596.7207 0.9340931 0.9246779  18.521465  87.31433  92.18984  831.1521
3 4 0 1 1 1 1095.8078 0.8789698 0.8616797  48.231020 102.50928 107.38479 1570.5610
3 4 1 1 0 1 1400.1275 0.8453581 0.8232664  66.346500 108.63607 113.51157 1885.8454
4 5 1 1 1 1  335.9775 0.9628918 0.9554702   5.000000  74.95421  81.04859  518.9885
```

For SSEp, I choose the smallest value 335.9775. For R^2, I choose the largest value 0.9628918. For R^2_{adj}, I choose the largest value 0.9560482. For Cp, I choose the closest value to the number of predictor variables 3.727399. For AICp, I choose the smallest value 73.84732. For SBCp, I choose the smallest value 78.72282. For PRESSp, I choose the smallest value 471.452. I marked all of them in the Figure 5.

From above table, we can see that the SSEp and R^2 both show that the Full model containing X1, X2, X3, X4 is a good model, and other measurements($R^2_{adj}, Cp, AIC_p, BIC_p, and$ PRESSp) suggest that the model with X1, X3, X4 containing in it (without X2) is the best subset regression model.

The regression line is

$Y_i = \beta_0 + \beta_1 X_1 + \beta_3 X_3 + \beta_4 X_4 + \epsilon_i$

Figure 6.

```
> summary(best)

Call:
lm(formula = y ~ x1 + x3 + x4, data = pro)

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
```

From R output, the $\beta_0 = -124.20002,\ \beta_1 = 0.29633,\ \beta_3 = 1.35697,\ \beta_4$=0.51742

so the regression line is:

$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_3 X_3 + \beta_4 X_4 = -124.20002 + 0.29633X_1 + 1.35697X_3 + 0.51742X_4$

d). (5) 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 indicator of the predictive ability of the fitted model?

$PRESS_p = \sum (Y_i - \hat{Y}_{i(i)})^2$

SSE = $\sum(Y_i - \hat{Y}_i)^2$

$PRESS_p$ is different from SSE because, in $PRESS_p$, each fitted value $\hat{Y}_i$ is obtained by deleting the ith case from the data set, and the model is estimated by the remaining n-1 cases, and then use the fitted regression function to obtain the predicted value $\hat{Y}_{i(i)}$ for ith case. For the MSE, just like the SSE, it needs to be contained in a range and it will change as the range changes. However, PRESS will not be influenced by the change of range because it is estimated by deleting the ith case from the data set, so it won't be affected. When we want to make prediction, it is always better to use PRESS rather than SSE, because it has the highest predictive ability

Therefore, PRESS is the best method as the tool of predictive ability.

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

Figure 7.

	nvmax <dbl>	RMSE <dbl>	Rsquared <dbl>	MAE <dbl>	RMSESD <dbl>	RsquaredSD <dbl>	MAESD <dbl>
1	4	4.195349	0.9586264	3.769349	1.587165	0.03628518	1.774073

The model with X1, X3, X4 containing in it (without X2) has four parameters $\beta_0,\ \beta_1,\ \beta_3,\ \beta_4$, so we set nvmax = 4.

f). (8) 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)?

Figure 8.

```
> BestSub(pro[,2:5], pro$y, num=1)
  p 1 2 3 4      SSEp        r2    r2.adj        Cp      AICp      SBCp    PRESSp
1 2 0 0 1 0 1768.0228 0.8047247 0.7962344 84.246496 110.46853 112.90629 2064.5976
2 3 1 0 1 0  606.6574 0.9329956 0.9269043 17.112978  85.72721  89.38384  760.9744
3 4 1 0 1 1  348.1970 0.9615422 0.9560482  3.727399  73.84732  78.72282  471.4520
4 5 1 1 1 1  335.9775 0.9628918 0.9554702  5.000000  74.95421  81.04859  518.9885
>
> BestSub(prof[,2:5], prof$y, num=1)
  p 1 2 3 4      SSEp        r2    r2.adj        Cp      AICp      SBCp    PRESSp
1 2 0 0 1 0 1593.9706 0.7886362 0.7794465 74.116237 107.87769 110.31544 2001.7963
2 3 1 0 1 0  471.8126 0.9374367 0.9317491  9.154243  79.44265  83.09928  677.2045
3 4 1 0 1 1  385.4536 0.9488880 0.9415863  6.000988  76.38863  81.26413  638.7037
4 5 1 1 1 1  335.1627 0.9555567 0.9466681  5.000000  74.89350  80.98788  604.2496
```

Figure 9.

```
> summary(fit2)

Call:
lm(formula = y ~ x1 + x2 + x3 + x4, data = prof)

Residuals:
    Min      1Q  Median      3Q     Max 
-7.3312 -2.6117 -0.1671  3.0793  6.3339 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -123.69739   11.33348 -10.914 7.12e-10 ***
x1             0.31204    0.04519   6.905 1.05e-06 ***
x2             0.07924    0.04574   1.732   0.0986 .  
x3             1.28105    0.23381   5.479 2.31e-05 ***
x4             0.48220    0.19125   2.521   0.0203 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.094 on 20 degrees of freedom
Multiple R-squared:  0.9556,    Adjusted R-squared:  0.9467 
F-statistic: 107.5 on 4 and 20 DF,  p-value: 3.174e-13
```

Figure 10.

```
> Anova(fit2)
Anova Table (Type II tests)

Response: y
          Sum Sq Df F value    Pr(>F)    
x1        799.03  1 47.6802 1.047e-06 ***
x2         50.29  1  3.0010   0.09861 .  
x3        503.06  1 30.0191 2.309e-05 ***
x4        106.53  1  6.3571   0.02029 *  
Residuals 335.16 20                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

From Figure 8, we can easily notice that the

Comparing Figure 3&4 with Figure 9&10,

From Figure 9, $\beta_2 = 0.07924$ which is bigger than that in Figure 3, while the p-value for β_2 is 0.0983 which is much smaller than that in Figure 3, this time if we compare the p-value to significance level = 0.1, p-value would be lower than α, then we may reject H_0: $\beta_2 = 0$.

From Figure 10, the SSR(X2|X1, X3, X4) = 50.29 which is bigger than that in Figure 4. Plus, F* = 3.0010 >F(0.1, 1, 25-5)=2.97 with significance level = 0.1, so we may reject H_0: $\beta_2 = 0$.

Hence, we are 90% confident that X2 has impact on Y, so we cannot drop X2 from the model now.

In (c), the SSEp and R^2 both show that the Full model containing X1, X2, X3, X4 is a good model, and other measurements($R^2_{adj}, Cp, AIC_p, BIC_p, and$ PRESSp) suggest that the model with X1, X3, X4 containing in it (without X2) is the best subset regression model. This time, all of the indictors show that the full model with X1, X2, X3, X4 in it is the best model to fit validation data set. Hence, we can choose the full model.

Appendix

HW5 Q2

```
data<-read.csv("C:/Users/candi/Desktop/STAT 512/valuation.csv",header=TRUE,sep = ",
")
data
```

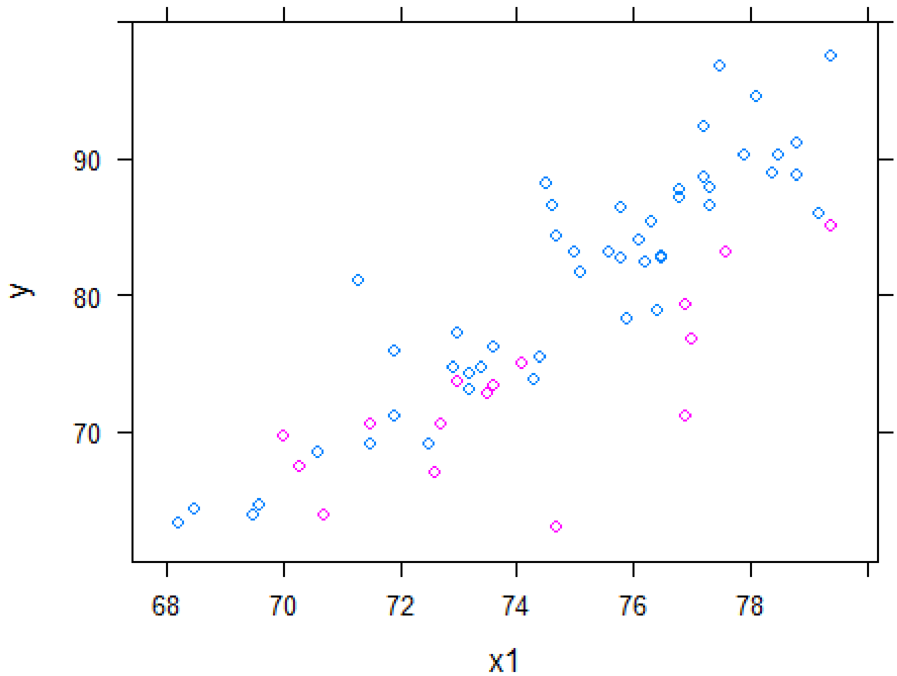
```
##       y   x1 x2
## 1  78.8 76.4  0
## 2  73.8 74.3  0
## 3  64.6 69.6  0
## 4  76.2 73.6  0
## 5  87.2 76.8  0
## 6  70.6 72.7  1
## 7  86.0 79.2  0
## 8  83.1 75.6  0
## 9  94.5 78.1  0
## 10 71.2 76.9  1
## 11 64.3 68.5  0
## 12 73.1 73.2  0
## 13 96.8 77.5  0
## 14 82.4 76.2  0
## 15 81.6 75.1  0
## 16 76.8 77.0  1
## 17 77.2 73.0  0
## 18 73.7 73.0  1
## 19 88.6 77.2  0
## 20 74.7 73.4  0
## 21 91.2 78.8  0
## 22 86.6 77.3  0
## 23 82.7 76.5  0
## 24 87.8 76.8  0
## 25 85.0 79.4  1
## 26 69.1 71.5  0
## 27 69.6 70.0  1
## 28 71.2 71.9  0
## 29 62.9 74.7  1
## 30 84.1 76.1  0
## 31 67.0 72.6  1
## 32 83.2 77.6  1
```

```
## 33 63.9 70.7  1
## 34 85.3 76.3  0
## 35 92.4 77.2  0
## 36 90.3 77.9  0
## 37 74.7 72.9  0
## 38 73.3 73.6  1
## 39 83.1 75.0  0
## 40 69.1 72.5  0
## 41 75.0 74.1  1
## 42 67.4 70.3  1
## 43 68.4 70.6  0
## 44 79.3 76.9  1
## 45 86.4 75.8  0
## 46 75.8 71.9  0
## 47 88.8 78.8  0
## 48 72.7 73.5  1
## 49 88.9 78.4  0
## 50 82.7 75.8  0
## 51 86.6 74.6  0
## 52 82.8 76.5  0
## 53 87.9 77.3  0
## 54 75.5 74.4  0
## 55 81.0 71.3  0
## 56 88.2 74.5  0
## 57 63.9 69.5  0
## 58 78.2 75.9  0
## 59 63.3 68.2  0
## 60 90.2 78.5  0
## 61 74.3 73.2  0
## 62 97.6 79.4  0
## 63 84.4 74.7  0
## 64 70.5 71.5  1
```

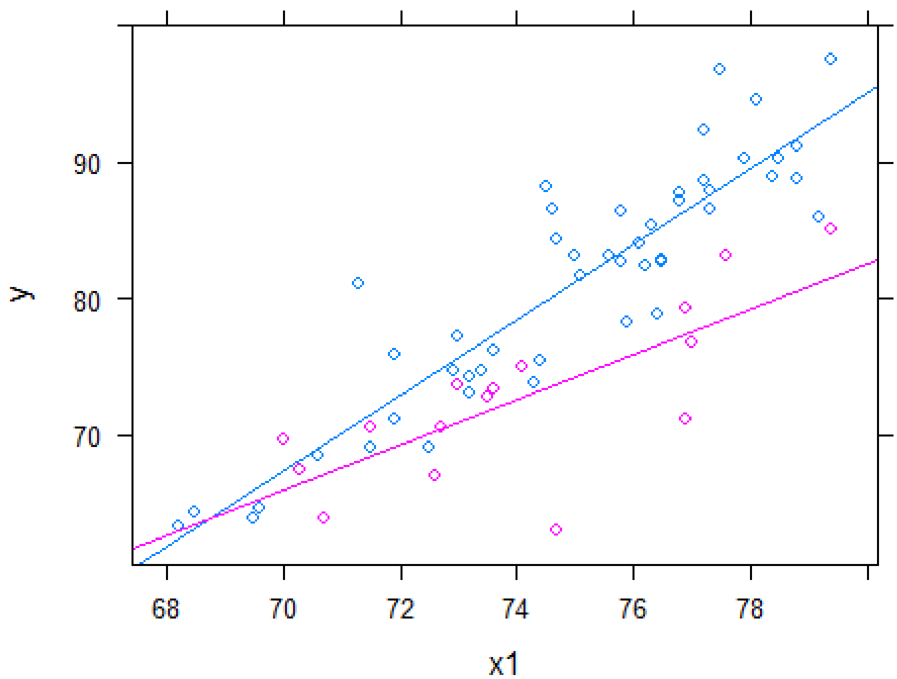
```
colnames(data)<-c("y","x1","x2")
y<-data$y
x1<-data$x1
x2<-data$x2
require("lattice")
```

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

```
xyplot(y~x1, groups=x2, data=data)
```




```r
xyplot(y~x1, groups=x2, type=c("p","r"), data=data)
```




```r
fit<-lm(y~x1+x2, data=data)
summary(fit)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2, data = data)
##
## 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(fit)
```

```
## 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
```

HW5 Q3

```
pro<-read.csv("C:/Users/candi/Desktop/STAT 512/proficiency.csv",header=TRUE,sep = "
,")
pro
```

```
##      y  x1  x2  x3  x4
## 1   88  86 110 100  87
## 2   80  62  97  99 100
## 3   96 110 107 103 103
## 4   76 101 117  93  95
## 5   80 100 101  95  88
## 6   73  78  85  95  84
## 7   58 120  77  80  74
## 8  116 105 122 116 102
```

```
## 9  104 112 119 106 105
## 10  99 120  89 105  97
## 11  64  87  81  90  88
## 12 126 133 120 113 108
## 13  94 140 121  96  89
## 14  71  84 113  98  78
## 15 111 106 102 109 109
## 16 109 109 129 102 108
## 17 100 104  83 100 102
## 18 127 150 118 107 110
## 19  99  98 125 108  95
## 20  82 120  94  95  90
## 21  67  74 121  91  85
## 22 109  96 114 114 103
## 23  78 104  73  93  80
## 24 115  94 121 115 104
## 25  83  91 129  97  83
```

```
colnames(pro)<-c("y","x1","x2","x3","x4")
y<-pro$y
x1<-pro$x1
x2<-pro$x2
x3<-pro$x3
x4<-pro$x4

#(a)
plot(pro)
cor(pro)
```

```
##            y        x1        x2        x3        x4
## y  1.0000000 0.5144107 0.4970057 0.8970645 0.8693865
## x1 0.5144107 1.0000000 0.1022689 0.1807692 0.3266632
## x2 0.4970057 0.1022689 1.0000000 0.5190448 0.3967101
## x3 0.8970645 0.1807692 0.5190448 1.0000000 0.7820385
## x4 0.8693865 0.3266632 0.3967101 0.7820385 1.0000000
```

```
#(b)
library(ALSM)
```

```
## Loading required package: leaps

## Loading required package: SuppDists

## Loading required package: car

## Loading required package: carData
```

```
fit1<-lm(y~x1+x2+x3+x4, data=pro)
Anova(fit1)
```

```
## Anova Table (Type II tests)
##
## Response: y
```

```
##             Sum Sq Df F value    Pr(>F)    
## x1         759.83  1 45.2310 1.524e-06 ***
## x2          12.22  1  0.7274   0.40383    
## x3        1064.15  1 63.3465 1.262e-07 ***
## x4         260.74  1 15.5215   0.00081 ***
## Residuals  335.98 20                      
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(fit1)
```

```
## 
## Call:
## lm(formula = y ~ x1 + x2 + x3 + x4, data = pro)
## 
## 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
```

```
#(c)
library(ALSM)
```

```
library("leaps")
BestSub(pro[,2:5], pro$y, num=4)
```

```
##   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
## 1 2 0 0 0 1 2210.6887 0.7558329 0.7452170 110.597414 116.05459 118.49234
## 1 2 1 0 0 0 6658.1453 0.2646184 0.2326452 375.344689 143.61801 146.05576
## 1 2 0 1 0 0 6817.5291 0.2470147 0.2142762 384.832454 144.20941 146.64717
## 2 3 1 0 1 0  606.6574 0.9329956 0.9269043  17.112978  85.72721  89.38384
## 2 3 0 0 1 1 1111.3126 0.8772573 0.8660988  47.153985 100.86053 104.51716
## 2 3 1 0 0 1 1672.5853 0.8152656 0.7984716  80.565307 111.08125 114.73788
## 2 3 0 1 1 0 1755.8127 0.8060733 0.7884436  85.519650 112.29528 115.95191
## 3 4 1 0 1 1  348.1970 0.9615422 0.9560482   3.727399  73.84732  78.72282
## 3 4 1 1 1 0  596.7207 0.9340931 0.9246779  18.521465  87.31433  92.18984
## 3 4 0 1 1 1 1095.8078 0.8789698 0.8616797  48.231020 102.50928 107.38479
## 3 4 1 1 0 1 1400.1275 0.8453581 0.8232664  66.346500 108.63607 113.51157
```

```
## 4 5 1 1 1 1  335.9775 0.9628918 0.9554702   5.000000  74.95421  81.04859
##      PRESSp
## 1 2064.5976
## 1 2548.6349
## 1 7791.5994
## 1 7991.0964
## 2  760.9744
## 2 1449.6001
## 2 2109.8967
## 2 2206.6460
## 3  471.4520
## 3  831.1521
## 3 1570.5610
## 3 1885.8454
## 4  518.9885
```

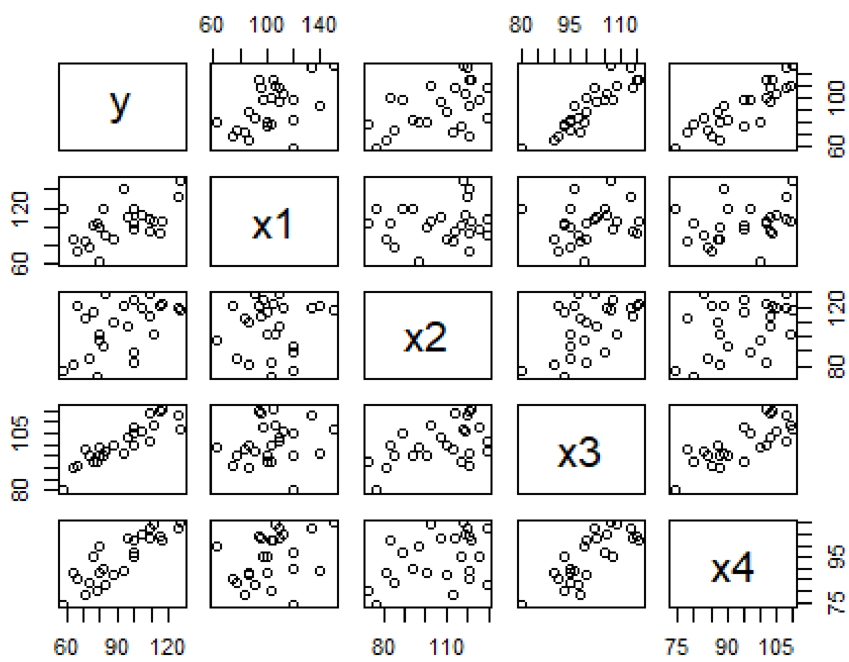
```
##bs<-BestSub(pro[,2:5], pro$y, num=4) #from column 2 to column 5
##bs[which.min(bs[,"Cp"]),"Cp"]  #find the minimun Cp
##bs[which.min(bs[,"AICp"]),"AICp"]
##bs[which.min(bs[,"SBCp"]),"SBCp"]
##bs[which.min(bs[,"PRESSp"]),"PRESSp"]
##colnames(bs)
##library(ALSM)
##plotmodel.s(pro[,2:5], pro$y)
best<-lm(y~x1+x3+x4,data=pro)
summary(best)
```

```
##
## Call:
## lm(formula = y ~ x1 + x3 + x4, data = pro)
##
## 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
```

```
#(e)
library(MASS)
library(leaps)
library(lattice)
```

```
##
## Attaching package: 'lattice'

## The following object is masked from 'package:ALSM':
##
##     oneway
```




```
library(ggplot2)
library(caret)

set.seed(123) #set seed for reproducibility

train.control<-trainControl(method="cv", number=5)  #10 fold cross validation

step.model1<-train(y~x1+x3+x4, data=pro, method="leapBackward",
                   tuneGrid=data.frame(nvmax=4),
                   trControl=train.control)
step.model1$results

##   nvmax     RMSE  Rsquared      MAE   RMSESD RsquaredSD    MAESD
## 1     4 4.195349 0.9586264 3.769349 1.587165 0.03628518 1.774073

#(f)
prof<-read.csv("C:/Users/candi/Desktop/STAT 512/proficiencyTest.csv",header=TRUE,se
p = ",")
prof

##      y  x1  x2  x3  x4
## 1   58  65 109  88  84
```

```
## 2   92  85  90 104  98
## 3   71  93  73  91  82
## 4   77  95  57  95  85
## 5   92 102 139 101  92
## 6   66  63 101  93  84
## 7   61  81 129  88  76
## 8   57 111 102  83  72
## 9   66  67  98  98  84
## 10  75  91 111  96  84
## 11  98 128  99  98  89
## 12 100 116 103 103 103
## 13  67 105 102  88  83
## 14 111  99 132 109 105
## 15  97  93  95 106  98
## 16  99  99 113 104  95
## 17  74 110 114  91  78
## 18 117 128 134 108  98
## 19  92  99 110  96  97
## 20  95 111 113 101  91
## 21 104 109 120 104 106
## 22 100  78 125 115 102
## 23  95 115 119 102  94
## 24  81 129  70  94  95
## 25 109 136 104 106 104
```

```
colnames(prof)<-c("y","x1","x2","x3","x4")
library(ALSM)
library("leaps")
BestSub(pro[,2:5], pro$y, 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
```

```
BestSub(prof[,2:5], prof$y, num=1)
```

```
##   p 1 2 3 4      SSEp        r2    r2.adj        Cp      AICp      SBCp
## 1 2 0 0 1 0 1593.9706 0.7886362 0.7794465 74.116237 107.87769 110.31544
## 2 3 1 0 1 0  471.8126 0.9374367 0.9317491  9.154243  79.44265  83.09928
## 3 4 1 0 1 1  385.4536 0.9488880 0.9415863  6.000988  76.38863  81.26413
## 4 5 1 1 1 1  335.1627 0.9555567 0.9466681  5.000000  74.89350  80.98788
##      PRESSp
## 1 2001.7963
## 2  677.2045
```

```
## 3  638.7037
## 4  604.2496
```

```r
fit2<-lm(y~x1+x2+x3+x4,data=prof)

summary(fit2)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3 + x4, data = prof)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.3312 -2.6117 -0.1671  3.0793  6.3339
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -123.69739   11.33348 -10.914 7.12e-10 ***
## x1             0.31204    0.04519   6.905 1.05e-06 ***
## x2             0.07924    0.04574   1.732   0.0986 .
## x3             1.28105    0.23381   5.479 2.31e-05 ***
## x4             0.48220    0.19125   2.521   0.0203 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.094 on 20 degrees of freedom
## Multiple R-squared:  0.9556, Adjusted R-squared:  0.9467
## F-statistic: 107.5 on 4 and 20 DF,  p-value: 3.174e-13
```

```r
Anova(fit2)
```

```
## Anova Table (Type II tests)
##
## Response: y
##           Sum Sq Df F value    Pr(>F)
## x1        799.03  1 47.6802 1.047e-06 ***
## x2         50.29  1  3.0010   0.09861 .
## x3        503.06  1 30.0191 2.309e-05 ***
## x4        106.53  1  6.3571   0.02029 *
## Residuals 335.16 20
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```