Assignment 4 — Due November 8, 2018

Naiqing Cai

ncai5@wisc.edu

1. Patient satisfaction

(a) Obtain the scatter plot matrix and the correlation matrix. State your key findings.

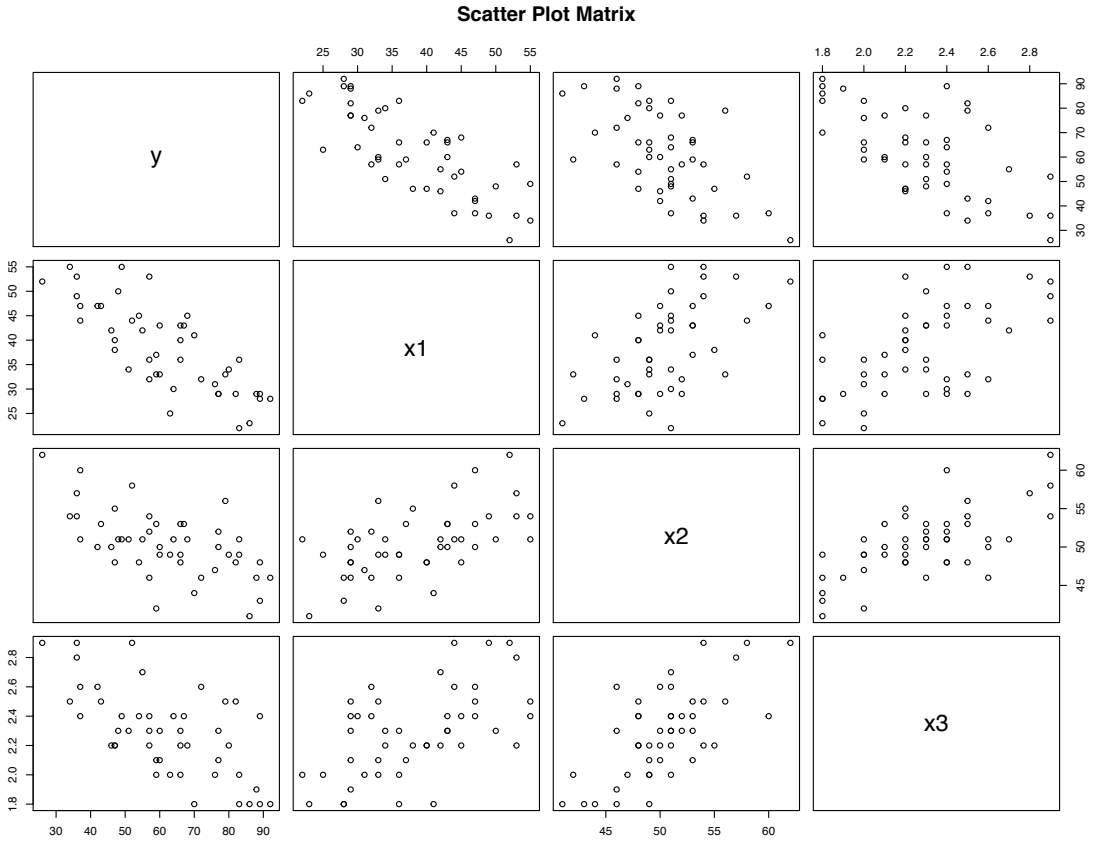

Figure 1.1 Scatter Plot Matrix

	Y	X1	X2	X3
Y	1	-0.7867555	-0.6029417	-0.644591
X1		1	0.5679505	0.5696775
X2			1	0.6705287
X3				1

Table 1.2 Correlation Matrix

Findings：

The correlation coefficients between Y and X1, X2, X3 are all negative.

The correlation coefficients between X1, X2, X3 are all positive.

(b) State the model underlying a multiple linear regression of patient satisfaction on age, severity of illness, and anxiety level. We will consider this model to be the full model. What is the underlying population or the set of underlying populations?

Multiple Linear Regression Model

$$Y = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \beta_3 X_{3,i} + \varepsilon_i,\ \varepsilon_i \overset{iid}{\sim} N(0,\sigma^2),\ i = 1,...,46$$

Y: patient satisfaction

X1: patient's age (in years)

X2: severity of illness

X3: anxiety level

Population

The whole of patient satisfaction, patient's age, severity of illness and anxiety level for each patient.

(c) Obtain the least squares estimates of the regression coefficients.

$$\hat{\beta} = (X'X)^{-1}X'Y,\ i = 0,1,2,3$$
$$\hat{\beta}_i = ((X'X)^{-1}X'Y)_{ii}$$

$$\hat{\beta}_0 = 158.491$$
$$\hat{\beta}_1 = \text{-}1.142$$
$$\hat{\beta}_2 = \text{-}0.442$$
$$\hat{\beta}_3 = \text{-}13.470$$

(d) Test whether severity of illness has any effect on patient satisfaction, while controlling for patient's age and anxiety level.

Hypothesis

$$H_0 : \beta_2 = 0 \ \text{ v.s } \ H_1 : \beta_2 \neq 0$$

Test

$$T^* = \frac{\hat{\beta}_2 - 0}{\sqrt{\hat{\text{var}}(\hat{\beta}_2)}} \sim T_{n-p}$$

$$\Rightarrow T^* = \text{-}0.4420/0.4920 = \text{-}0.898$$

p-value

p=Pr(>|t|)=0.3741>0.05

Conclusion

Thus, we have evidence that we should accept the null hypothesis that severity of illness has no effect on patient satisfaction.

(e) Provide a 95% confidence intervals for the regression coefficient corresponding to severity of illness in the full model.

$$\beta_2 \in [\hat{\beta}_2 - t_{n-p,\alpha/2}\sqrt{\hat{\text{var}}(\hat{\beta}_2)}, \hat{\beta}_2 + t_{n-p,\alpha/2}\sqrt{\hat{\text{var}}(\hat{\beta}_2)}]$$

$$\Rightarrow \beta_2 \in [-1.4348, 0.5508]$$

(f) Provide an unbiased estimate for the error variance and a 95% confidence interval.

Unbiased Estimate for the Error Variance

$$\hat{\sigma}^2 = 10.06^2 = 101.2036$$

95% Confidence Interval

$$\frac{(n-4)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-4}$$

$$\chi^2_{n-4,1-\alpha/2} = 61.777$$

$$\chi^2_{n-4,\alpha/2} = 25.999$$

$$\sigma^2 \in [68.805, 163.49]$$

(g) Obtain a point estimate and a 95% confidence interval for the mean satisfaction of patients when Xh1 =35, Xh2 =45 and Xh3 =2.2.

Point Estimate

$$\hat{\mu}_h = X_h^{'}\hat{\beta}, where\ X_h = (1, X_{h1}, X_{h2}, X_{h3})'\ and\ \hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3)$$

$$\Rightarrow \hat{\mu}_h = 69.01$$

95% Confidence Interval

$$\mu_h \in [\hat{\mu}_h - t_{n-p,\alpha/2}\sqrt{\hat{\text{var}}(\hat{\mu}_h)}, \hat{\mu}_h + t_{n-p,\alpha/2}\sqrt{\hat{\text{var}}(\hat{\mu}_h)}]\ where \sqrt{\hat{\text{var}}(\hat{\mu}_h)} = \hat{\sigma}\sqrt{X_h^{'}(X'X)^{-1}X_h}$$

$$\mu_h \in [63.633, 74.388]$$

(h) Obtain a point predictor and a 95% prediction interval for a new patient's satisfaction when Xh1 =35, Xh2 =45 and Xh3 =2.2.

Point Predictor

$$\hat{Y}_h = X_h^{'}\hat{\beta} \Rightarrow \hat{Y}_h = 69.01$$

95% Prediction Interval

$$Y_h \in [\hat{Y}_h - t_{n-p,\alpha/2}\hat{\sigma}_{pred}, \hat{Y}_h + t_{n-p,\alpha/2}\hat{\sigma}_{pred}], where\ \hat{\sigma}_{pred} = \hat{\sigma}\sqrt{1 + X_h^{'}(X'X)^{-1}X_h}$$

$$Y_h \in [48.01, 90.01]$$

(i) Obtain the coefficient of multiple determination R^2.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y}_i)^2}$$

$$R^2 = 0.6822$$

(j) Perform an overall F-test for all predictors in the model. State the assumptions for the F-test and perform model diagnostics.

F-test

```
Analysis of Variance Table

Model 1: y ~ 1
Model 2: y ~ x1 + x2 + x3
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1     45 13369.3
2     42  4248.8  3    9120.5 30.052 1.542e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 1.3 F-test

full model: $Y = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \beta_3 X_{3,i} + \varepsilon_i$

reduced model: $Y = \beta_0 + \varepsilon_i$

$H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ v.s $H_1 : opposite\ of\ H_0$

$$F^* = \frac{(SSR(R) - SSR(F))/(df_R - df_F)}{SSR(F)/df_F} \sim F_{df_R - df_F, df_F}$$

$$\Rightarrow F^* = 30.05$$

p-value=Pr(>|f*|)=1.542e-10<0.05

Thus, we have strong evidence that we should reject null hypothesis, that is there is evidence for non-zero slope.

Assumptions

- A straight line relationship between the response variable Y and the explanatory variable X

$$Y = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \beta_3 X_{3,i} + \varepsilon_i \qquad E(\varepsilon_i)=0$$

- Equal Variance

$$var(\varepsilon_i) = \sigma^2$$

- Independence

$$cov(\varepsilon_i, \varepsilon_{i'})=0, i \neq i'$$

- Normal Distribution

$$\varepsilon_i \sim i.i.d.N(0, \sigma^2)$$

Model Diagnostics

1) Equal Variance Assumption: residual against fitted values

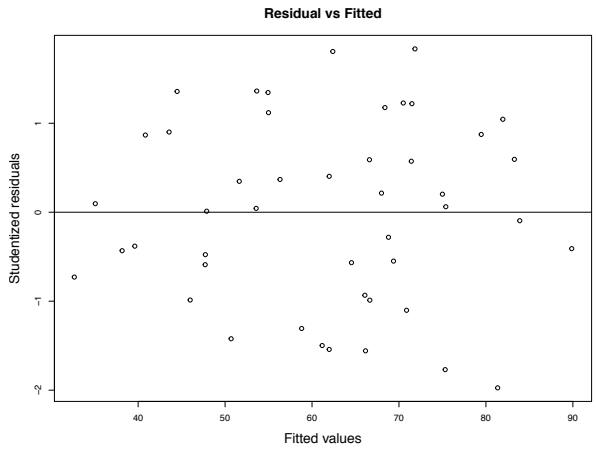

Figure 1.4 Residual against Fitted Values

Based on the plot of residual against fitted values, we can conclude that the residuals do not have equal variance. Thus, the assumption equal variance is not reasonable.

2) Normal Distribution Assumption

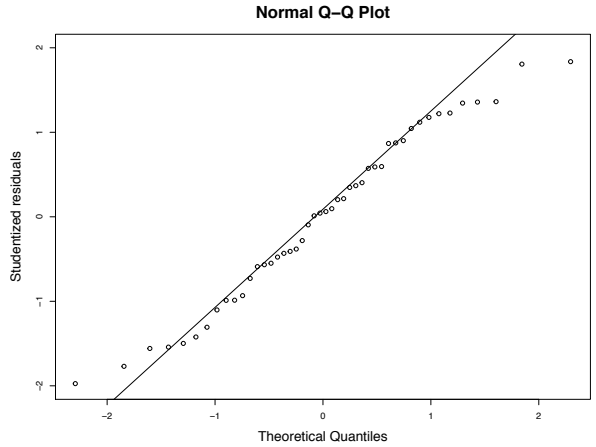

Figure 1.5 QQ Plot of Residuals

Based on the QQ plot of residual above, the residuals are from normal distribution, so the assumption normal distribution is reasonable.

```
        Shapiro-Wilk normality test

data:  studres(lm.reg)
W = 0.97109, p-value = 0.304
```

Figure 1.6 Shapiro-Wilk normality test

Based on the Shapiro-Wilk normality test, p-value=0.304>0.05, so the residuals are from normal distribution and the assumption normal distribution is reasonable.

2. Continue to work on the above dataset

(a) Fit a model with severity of illness (X2) only. Test whether the regression coefficient for X2 is zero or not and provide a 95% confidence interval for this regression coefficient. Compare the results with Problem 1(d) and 1(e) in Assignment 4 and explain the results.

Model:

$$Y = \beta_0 + \beta_2 X_{2,i} + \varepsilon_i,\ \varepsilon_i \overset{iid}{\sim} MVN(0,\sigma^2),\ i = 1,...,46$$

Test:

$$H_0 : \beta_2 = 0\ v.s\ H_1 : \beta_2 \neq 0$$

$$\mathrm{T}^* = \frac{\hat{\beta}_2 - 0}{\sqrt{\hat{\mathrm{var}}(\hat{\beta}_2)}} \sim T_{n-2}$$

$$\hat{\beta}_2 = -2.409$$

$$|\mathrm{T}^*| = 5.013 > |T_{n-2,\alpha/2}| = 2.0154$$

$$p = \Pr(>|t|) = 9.23e\text{-}06 < 0.05$$

So, we reject the null hypothesis. In other word, we have evidence that beta2 is not equal to zero.

95% confidence interval:

$$\beta_2 \in [\hat{\beta}_2 - t_{n-p,\alpha/2}\sqrt{\hat{\mathrm{var}}(\hat{\beta}_2)}, \hat{\beta}_2 + t_{n-p,\alpha/2}\sqrt{\hat{\mathrm{var}}(\hat{\beta}_2)}]$$

$$\Rightarrow \beta_2 \in [-3.3778, -1.4407]$$

Comparison:

Compared with the problem 1(d) and 1(e), we find that if we fit a model with severity of illness (X2) only, we should reject the null hypothesis that is beta2 is not equal to zero, which means severity of illness is related with the patient satisfaction. But when we fit the full model with patient's age (X_1) severity of illness (X_2) and anxiety level (X_3), we should accept that beta2 is equal to zero, which means severity of illness is not related with the patient satisfaction. The difference with two results may caused by the correlation between X_1, X_2, X_3.

(b) Compute the VIF values for the explanatory variables in the full model with X1, X2, X3

$$\mathrm{VIF}_k = \frac{1}{1 - R_k^2}, k = 1,2,3$$

$$\mathrm{VIF}_1 = 1.6323$$
$$\mathrm{VIF}_2 = 2.0032$$
$$\mathrm{VIF}_3 = 2.0091$$

(c) Does SSR(X1) equal SSR(X1|X3)? Does SSR(X2) equal SSR(X2|X3)? Explain the results.

```
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)
x1         1 8275.4  8275.4  71.481 9.058e-11 ***
Residuals 44 5093.9   115.8
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 2.1 anova table for SSR(X1)

```
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)
x3         1 5554.9  5554.9  55.158 3.117e-09 ***
x1         1 3483.9  3483.9  34.593 5.434e-07 ***
Residuals 43 4330.5   100.7
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 2.2 anova table for SSR(X1|X3)

Based on the results above, we find that

SSR(X1)=8275.4

SSR(X1|X3)= SSE(X3)-SSE(X1,X3)=7814.4-4330.5=3483.9

So they are not equal.

```
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value   Pr(>F)
x2         1 4860.3  4860.3  25.132 9.23e-06 ***
Residuals 44 8509.0   193.4
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 2.3 anova table for SSR(X2)

```
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)
x3         1 5554.9  5554.9  33.612 7.197e-07 ***
x2         1  708.0   708.0   4.284   0.04451 *
Residuals 43 7106.4   165.3
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 2.4 anova table for SSR(X2|X3)

Based on the results above, we find that

SSR(X2)=4860.3

SSR(X2|X3)= SSE(X3)-SSE(X2,X3)=7814.4-7106.4=708

So they are not equal.

(d) Use the best subsets methods with various model selection criteria (R^2,Ra^2,Cp) and 1 best model for each model size (i.e., set nbest=1 in R function my.regsub).

```
> my.regsub(a[,2:4],y=a[,1],nbest=1,method="exhaustive")
  (Intercept) x1 x2 x3       rsq      rss     adjr2       cp       bic
1           1  1  0  0 0.6189843 5093.915 0.6103248 8.353606 -36.72879
2           1  1  0  1 0.6760864 4330.500 0.6610206 2.807204 -40.36888
3           1  1  1  1 0.6821943 4248.841 0.6594939 4.000000 -37.41593
```

Figure 2.5 Model Selection Criteria with nbest=1

Based on various model selection criteria, we find the best model for each model size.

For size one, the best model is

$$Y = \beta_0 + \beta_1 X_{1,i} + \varepsilon_i,\ \varepsilon_i \overset{iid}{\sim} N(0,\sigma^2),\ i = 1,...,46$$

For size two, the best model is

$$Y = \beta_0 + \beta_1 X_{1,i} + \beta_3 X_{3,i} + \varepsilon_i,\ \varepsilon_i \overset{iid}{\sim} N(0,\sigma^2),\ i = 1,...,46$$

For size three, the best model is

$$Y = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \beta_3 X_{3,i} + \varepsilon_i,\ \varepsilon_i \overset{iid}{\sim} N(0,\sigma^2),\ i = 1,...,46$$

(e) Repeat (d) but with nbest=4. Briefly comment on the choice of nbest. How important is it to change nbest from 1 to 4?

```
> my.regsub(a[,2:4],y=a[,1], nbest = 4,method = "exhaustive")
  (Intercept) x1 x2 x3       rsq      rss     adjr2        cp       bic
1           1  1  0  0 0.6189843 5093.915 0.6103248  8.353606 -36.72879
1           1  0  0  1 0.4154975 7814.391 0.4022134 35.245643 -17.04445
1           1  0  1  0 0.3635387 8509.044 0.3490737 42.112324 -13.12698
2           1  1  0  1 0.6760864 4330.500 0.6610206  2.807204 -40.36888
2           1  1  1  0 0.6549559 4613.000 0.6389073  5.599735 -37.46189
2           1  0  1  1 0.4684545 7106.394 0.4437314 30.247056 -17.58453
3           1  1  1  1 0.6821943 4248.841 0.6594939  4.000000 -37.41593
```

Figure 2.6 Model Selection Criteria with nbest=4

When nbest is larger, it will give more different models and there will be larger range of selection of models.

But if n is too larger, all the model selection will appear and no model will be removed by the criteria. So it will be more difficult for ourselves to choose the best model.

Thus, we should choose a suitable nbest which is not too large or too small and then we can have a better choice of model.

(f) Perform a backward elimination with AIC as the model selection criterion in R function step.

```
Start:  AIC=216.18
y ~ x1 + x2 + x3

       Df Sum of Sq    RSS    AIC
- x2    1     81.66 4330.5 215.06
<none>              4248.8 216.19
- x3    1    364.16 4613.0 217.97
- x1    1   2857.55 7106.4 237.84

Step:  AIC=215.06
y ~ x1 + x3

       Df Sum of Sq    RSS    AIC
<none>              4330.5 215.06
- x3    1     763.4 5093.9 220.53
- x1    1    3483.9 7814.4 240.21

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

Coefficients:
(Intercept)           x1           x3
     145.94        -1.20       -16.74
```

Figure 2.7 Backward Elimination with AIC

AIC start with 216.18 and after step one, which getting rid of the x2, AIC changes to 215.06 which is smaller than 216.18. Thus, we should remove x2. And, removing x1 and x3 will not reduce AIC, so we will not remove them.

(g) Repeat (f) but with BIC.

```
Start:  AIC=223.5
y ~ x1 + x2 + x3

       Df Sum of Sq    RSS    AIC
- x2    1     81.66 4330.5 220.55
- x3    1    364.16 4613.0 223.45
<none>              4248.8 223.50
- x1    1   2857.55 7106.4 243.33

Step:  AIC=220.55
y ~ x1 + x3

       Df Sum of Sq    RSS    AIC
<none>              4330.5 220.55
- x3    1     763.4 5093.9 224.19
- x1    1    3483.9 7814.4 243.87

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

Coefficients:
(Intercept)           x1           x3
     145.94        -1.20       -16.74
```

Figure 2.8 Backward Elimination with BIC

BIC start with 223.5 and after step one, which getting rid of the x2, BIC changes to 220.55 which is smaller than 223.5. Thus, we should first remove x2. And when we continue removing x1 and x3, they will not reduce BIC, so we will not remove them.

3. Consider the one-sample problem: Y ~N(μ,σ2),1≤i≤n with the's i.i.d.

(a) Find a design matrix X such that Show that $\hat{\mu} = (X^tX)^{-1}X^tY$.

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} Y_i = (1'1)^{-1}1'Y, where\ 1=(1,...,1)'$$

$$\Rightarrow X=(1,...,1)'$$

(b)

$$\hat{\mu}_\lambda = \arg\min_{\mu} \sum_{i=1}^{n}(Y_i-\mu)^2 + \lambda\mu^2$$

$$\text{Minimizing} \sum_{i=1}^{n}(Y_i-\mu)^2 + \lambda\mu^2$$

$$\text{Differentiating: } -2\sum_{i=1}^{n}(Y_i-\hat{\mu}) + 2\lambda\hat{\mu} = 0$$

$$Finally\ \hat{\mu}_\lambda = \frac{\sum_{i=1}^{n} Y_i}{n+\lambda} = \frac{\hat{\mu}}{1+\lambda/n}$$

(c) Find a design matrix X(λ) and a data vector Y (λ) such that $\hat{\mu}_\lambda = (X(\lambda)^tX(\lambda))^{-1}X(\lambda)^tY(\lambda)$

$$Y(\lambda) = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \\ 0 \end{pmatrix} \text{ and } X(\lambda) = \begin{pmatrix} 1 \\ \vdots \\ 1 \\ \sqrt{\lambda} \end{pmatrix}$$

$$\Rightarrow \left(X'(\lambda)X(\lambda)\right)^{-1}X'(\lambda)Y(\lambda) = \frac{\sum_{i=1}^{n} Y_i}{n+\lambda} = \hat{\mu}_\lambda$$

(d) Generalize this to the constrained regression problem for a vector of non-negative constraints λ = (λ0, ... , λp−1)

$$\hat{\beta}_\lambda = \arg\min\left(\sum_{i=1}^{n}(Y_i - \beta_0 - \sum_{j=1}^{p-1}\beta_j X_{ij})^2 + \sum_{j=0}^{p-1}\lambda_j\beta_j^2\right)$$

$$Minimizing: \sum_{i=1}^{n}(Y_i - \beta_0 - \sum_{j=1}^{p-1}\beta_j X_{ij})^2 + \sum_{j=0}^{p-1}\lambda_j\beta_j^2$$

$$Differentiating: -2(Y_i - \beta_0 - \sum_{j=1}^{p-1}\beta_j X_{ij})X_{ij} + 2\lambda_j\beta_j = 0$$

$$Finally: \hat{\beta}_j = \frac{Y_i - \beta_0 - \sum_{j=1}^{j-1}\beta_j X_{ij} - \sum_{j=j+1}^{p-1}\beta_j X_{ij}}{\lambda_j - X_{ij}}$$

(e) Write a function in R that takes two arguments, one the output of lm, the other a vector λ of length p as above and return βˆλ.

```
penalize <- function(model,lambda) {
  if (length(model$coefficients) != length(lambda)) {
    return("lambda should have the same parameter length as the model coefficients!")
  }
  x<- rbind(model.matrix(model),diag(lambda))
  y<- c(as.vector(model$residuals+model$fitted.values),rep(0,length(lambda)))
  betalambda<- solve(t(x)%*%x)%*%t(x)%*%y
  return(betalambda)
}
```

4. A soft drink bottler is analyzing the vending machine service routes in his distribution system.

(a) Obtain the scatterplot matrix and the correlation matrix for these three variables. State your key findings.

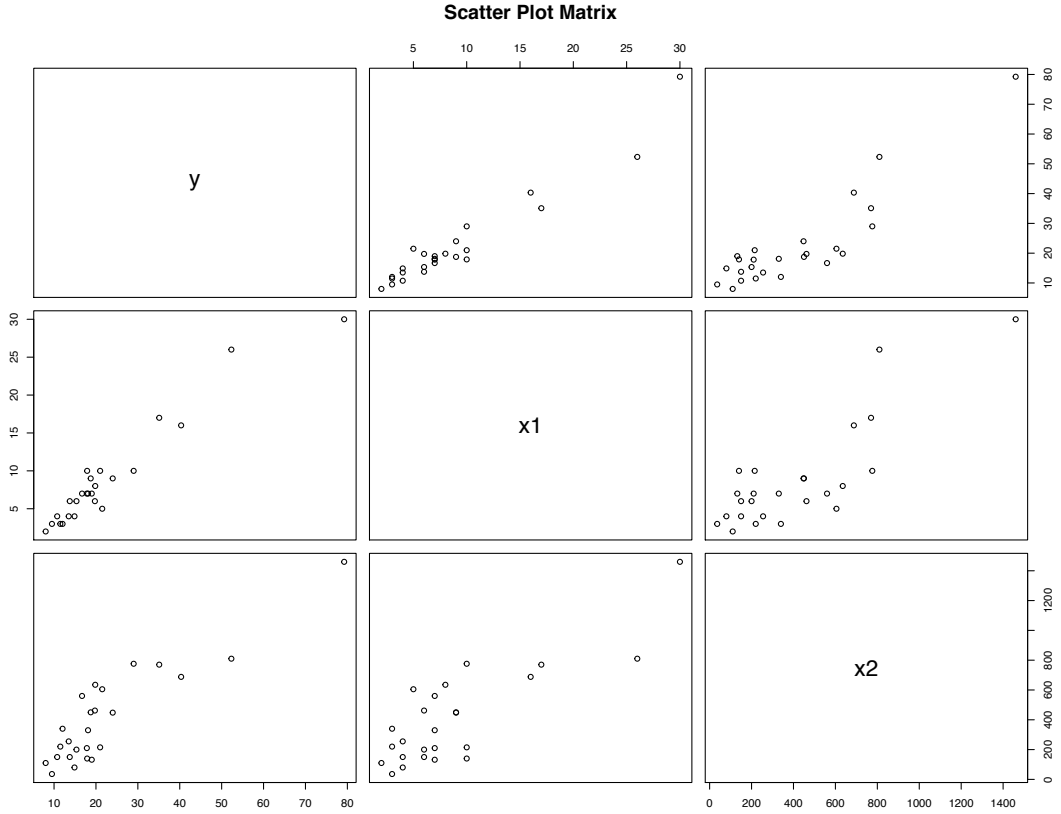

Figure 4.1 Scatter Plot Matrix

	Y	**X1**	**X2**
Y	1	0.9646146	0.8916701
X1		1	0.824215
X2			1

Table 4.2 Correlation Matrix

Findings：

The correlation coefficients between Y and X1, X2 are both positive.

The correlation coefficients between X1, X2 are positive, too.

(b) Fit the linear regression model Y on the two explanatory variables. Report the R summary table from the fitted model.

We fit a model as follows:

$$Y = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \varepsilon_i,\ \varepsilon_i \overset{iid}{\sim} MVN(0,\sigma^2),\ i = 1,...,25$$

And report the R summary table:

```
Call:
lm(formula = b$y ~ b$x1 + b$x2)

Residuals:
    Min      1Q  Median      3Q     Max
-5.7880 -0.6629  0.4364  1.1566  7.4197

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.341231   1.096730   2.135 0.044170 *
b$x1        1.615907   0.170735   9.464 3.25e-09 ***
b$x2        0.014385   0.003613   3.981 0.000631 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.259 on 22 degrees of freedom
Multiple R-squared:  0.9596,    Adjusted R-squared:  0.9559
F-statistic: 261.2 on 2 and 22 DF,  p-value: 4.687e-16
```

Table 4.3 R summary table

(c) Test for an overall regression relation between Y and the two explanatory variables.

```
Analysis of Variance Table

Model 1: y ~ 1
Model 2: y ~ x1 + x2
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     24 5784.5
2     22  233.7  2    5550.8 261.24 4.687e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 4.4 F-test

full model: $Y = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \varepsilon_i$

reduced model: $Y = \beta_0 + \varepsilon_i$

$H_0 : \beta_1 = \beta_2 = 0$ v.s $H_1 : opposite\ of\ H_0$

$$F^* = \frac{(SSR(R) - SSR(F))/(df_R - df_F)}{SSR(F)/df_F} \sim F_{df_R - df_F, df_F}$$

$\Rightarrow F^* = 261.24$

p-value=Pr(>|f*|)=4.687e-16<0.05

Thus, we have strong evidence that we should reject null hypothesis, that is there is evidence for non-zero slope.

(d) Plot the standard (i.e., internally studentized) and the studentized (i.e., externally studentized) residuals versus the fitted values from the model in part (b). State your key findings.

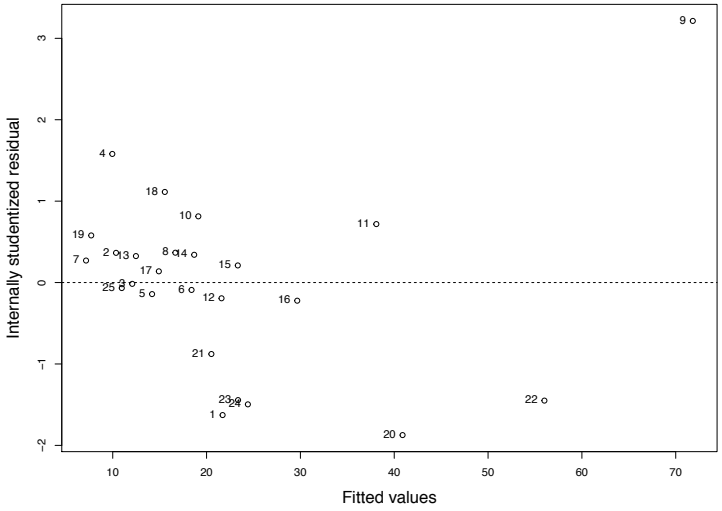

Figure 4.5 Internally Studentized Residuals V.S Fitted Values

key findings:

From the internally studentized residuals, I find that 9th sample may be an outlier in the observations because it's internally studentized residual is a bit large.

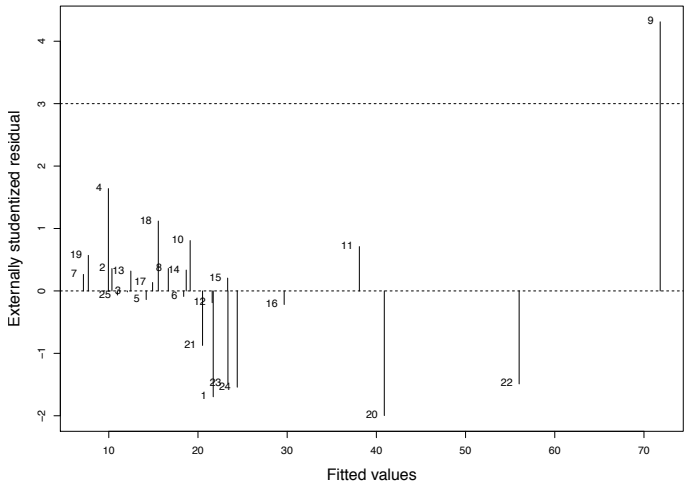

Figure 4.6 Externally Studentized Residuals V.S Fitted Values

key findings:

From the externally studentized residuals, I find that 9th sample may be an outlier in the observations because it's externally studentized residual is a bit large.

(e) For the model in part (b) compute and investigate various regression diagnostics including the leverage values, DFFITS, Cook's distance, and DFBETAS. State your key findings.

1) leverage values:

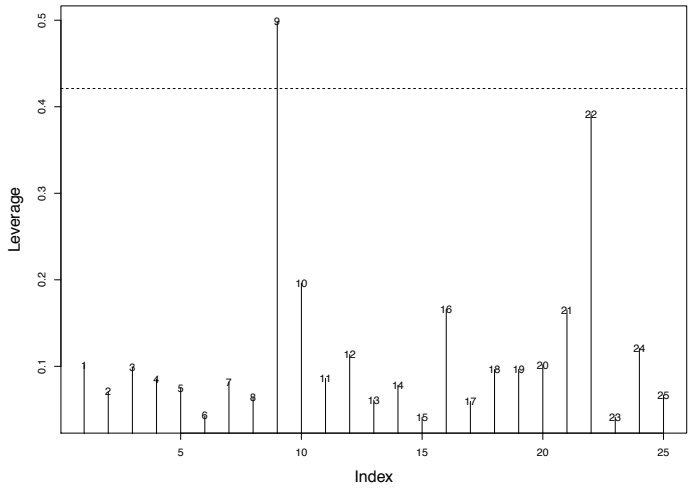

Figure 4.7 Leverage Values

key findings:

From the leverage test above, I find that there is an observation be the outlier in X. It is 9th observation that is an outlier since it has a high leverage.

2) DFFITS:

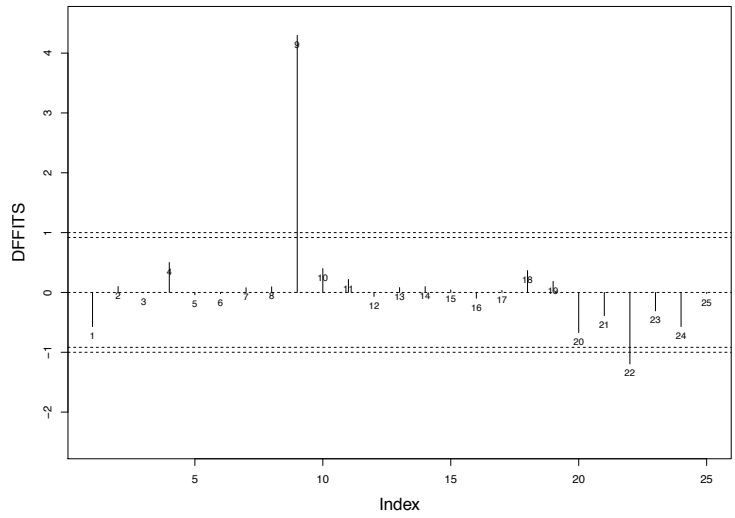

Figure 4.8 DFFITS

key findings:

From the DFFITS test above, I find that there are two outliers. They are 9th observation and 22nd observation that are outliers and they will have great influence on their own fitted value y(hat).

3) Cook's distance:

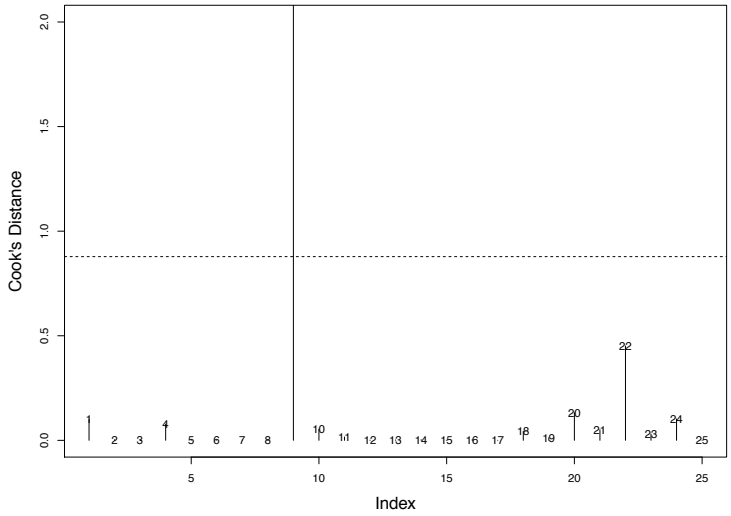

Figure 4.9 Cook's distance

key findings:

From the Cook's distance test above, I find that there is an outlier. It is 9th observation and it will have a great influence on all n fitted values.

4) DFBETAS:

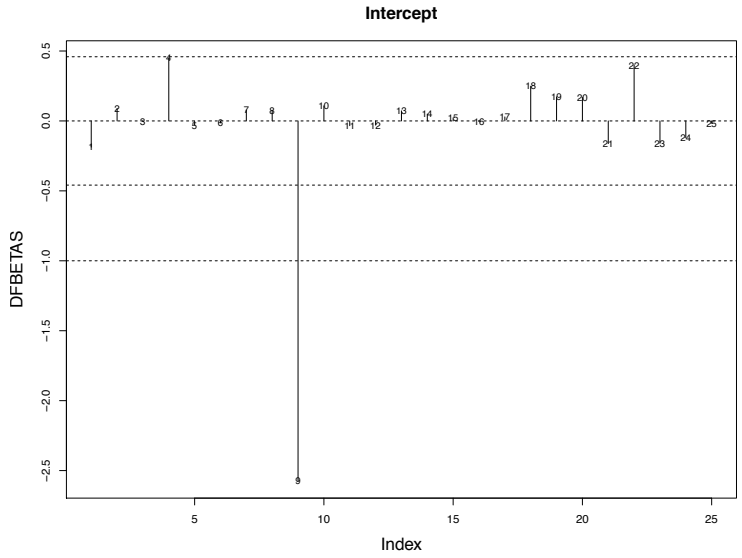

Figure 4.10 DFBETAS

key findings:

From the DFBETAS test above, I find that there is an outlier. It is 9th observation and it will have a great influence on its own regression coefficient.

(f) Based on your results in part (e), select a final model. Provide the R summary table and discuss the differences between the results here and the results in part (b).

Final Model:

Based on the results in part (e), I select a final model which remove the 9th sample because it is an outlier and it will have a great influence on the model.

R summary table:

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

Residuals:
    Min      1Q  Median      3Q     Max
-4.0325 -1.2331  0.0199  1.4730  4.8167

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.447238   0.952469   4.669 0.000131 ***
x1          1.497691   0.130207  11.502 1.58e-10 ***
x2          0.010324   0.002854   3.618 0.001614 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.43 on 21 degrees of freedom
Multiple R-squared:  0.9487,    Adjusted R-squared:  0.9438
F-statistic: 194.2 on 2 and 21 DF,  p-value: 2.859e-14
```

Table 4.11 R summary table

Comparison:

Compared with the r summary table in part b, I find that the range of residuals is smaller, so the standard error in part f is smaller than that in part b. Also, the estimates of coefficients change as well. This is caused by the removal of the outlier.

Appendix:

```
#1(a)
a=read.table("patient.txt",header = TRUE)
pairs(a,main="Scatter Plot Matrix")
cor(a$y,a$x1)
cor(a$y,a$x2)
cor(a$y,a$x3)
cor(a$x1,a$x2)
cor(a$x1,a$x3)
cor(a$x2,a$x3)

#1(c)
lm.reg=lm(data=a, y~x1+x2+x3)
summary(lm.reg)
confint(lm.reg)

#1(d)
a=read.table("patient.txt",header = TRUE)
lm.reg=lm(data=a, y~x1+x2+x3)
summary(lm.reg)

#1(f)
qchisq(0.025,42)
qchisq(0.975,42)
42*(10.06^2)/qchisq(0.025,42)
42*(10.06^2)/qchisq(0.975,42)

#1(h)(g)
lm.reg=lm(data=a, y~x1+x2+x3)
predict.lm(lm.reg, newdata=data.frame(x1=35, x2=45, x3=2.2),
                interval = "confidence")
predict.lm(lm.reg, newdata=data.frame(x1=35, x2=45, x3=2.2),
                interval = "prediction")
```

```
#1(j)
## for testing H0: beta1=beta2=beta3=0
ls.fit0=lm(data=a,y~1)
ls.fit=lm(data=a,y~x1+x2+x3)
anova(ls.fit0,ls.fit)

# cheking correlation betw vars
pairs(data=a,~y+x1+x2+x3)
lm(data=a,x1~x2+x3)

## check assumption
# install.packages("MASS");library(MASS)
# install.packages("lmtest");library(lmtest)
install.packages("MASS")
library(MASS)
install.packages("lmtest")
install.packages("nortest")
library(nortest)
library(lmtest)
library(base)
## fit vs res ##
plot(lm.reg$fitted.values ,studres(lm.reg),xlab="Fitted values",
     ylab="Studentized residuals",
     main="Residual vs Fitted",cex.lab=1.5,cex.main=1.5)
abline(h=0);abline(h=3,lty=2);abline(h=-3,lty=2)

## res QQ ##
qqnorm(studres(lm.reg),ylab="Studentized residuals",
       ylim=c(-2,2),cex.lab=1.5,cex.main=1.8)
qqline(studres(lm.reg))

## Shapiro -Wilk test
shapiro.test(studres(lm.reg))
```

```
#2(a)
a=read.table("patient.txt",header = TRUE)
m=lm(a$y~a$x2)
summary(m)
confint(m)
qt(0.025,44)
qt(0.975,44)
#2(b)
install.packages("car");library(car)
a=read.table("patient.txt",header = TRUE)
lm.reg=lm(data=a, y~x1+x2+x3)
vif(lm.reg)
#2(c)
# SSR(X1) & SSR(X1|X3)
anova(lm(data=a,y~x1))
anova(lm(data=a,y~x3+x1))
# SSR(X2) & SSR(X2|X3)
anova(lm(data=a,y~x2))
anova(lm(data=a,y~x3+x2))
#2(d)
install.packages("leaps")
library(leaps)
my.regsub <- function(matrix,y,nbest,method,nvmax=8){
   temp <- regsubsets(matrix,y,nbest=nbest,method=method,nvmax=nvmax)
   temp.mat <- cbind(summary(temp)$which,
                     summary(temp)$rsq,summary(temp)$rss,
                     summary(temp)$adjr2,summary(temp)$cp,
                     summary(temp)$bic)
   dimnames(temp.mat)[[2]] <- c(dimnames(summary(temp)$which)[[2]],
                                "rsq", "rss", "adjr2", "cp", "bic")
   return(temp.mat)
}
my.regsub(a[,2:4],y=a[,1],nbest=1,method="exhaustive")
my.regsub(a[,2:4],y=a[,1], nbest = 4,method = "exhaustive")
```

```
#2(f)
fit=lm(data=a,y~x1+x2+x3)
step(fit,direction = "backward",trace=1)
fitm=lm(data=a,y~x1+x3)
summary(fitm)

#2(g)
n=46
fit=lm(data=a,y~x1+x2+x3)
step(fit,direction = "backward",trace=1,k=log(n))
fitm=lm(data=a,y~x1+x3)
summary(fitm)
```

```
#3
penalize <- function(model,lambda) {
   if (length(model$coefficients) != length(lambda)) {
      return("lambda should have the same parameter length as the model coefficients!")
   }
   x<- rbind(model.matrix(model),diag(lambda))
   y<- c(as.vector(model$residuals+model$fitted.values),rep(0,length(lambda)))
   betalambda<- solve(t(x)%*%x)%*%t(x)%*%y
   return(betalambda)
}
```

```
#4(a)
b=read.table("softdrink.txt",header = TRUE)
pairs(b,main="Scatter Plot Matrix")
cor(b$y,b$x1)
cor(b$y,b$x2)
cor(b$x1,b$x2)


#4(b)
lm.reg=lm(b$y~b$x1+b$x2)
summary(lm.reg)
```

```
#4(c)
## for testing H0: beta1=beta2=0
b=read.table("softdrink.txt",header = TRUE)
ls.fit0=lm(data=b,y~1)
ls.fit=lm(data=b,y~x1+x2)
anova(ls.fit0,ls.fit)
# cheking correlation betw vars
pairs(data=b,~y+x1+x2)
lm(data=b,x1~x2)

#4(d)
index=1:length(b$x1)
par(mfrow=c(1,1))
plot(data=b,lm.reg$fitted, stdres(lm.reg), xlab="Fitted values",
     ylab="Internally studentized residual",cex.lab=1.5)
abline(h=0, lty=2)
text(data=b,lm.reg$fitted, stdres(lm.reg), labels=index, cex=1, pos=2)
plot(data=b,lm.reg$fitted, studres(lm.reg), type = "h", xlab="Fitted values",
     ylab="Externally studentized residual",cex.lab=1.5)
abline(h=c(0,-3,3), lty=2)
text(data=b,lm.reg$fitted, studres(lm.reg), labels=index, cex=1, pos=2)

#4(e)
## leverage
lm.reg.hats = hatvalues(data=b,lm.reg)
plot(lm.reg.hats, type = "h", ylab = "Leverage",cex.lab=1.5)
text(lm.reg.hats, labels = index, cex = 1)
abline(h=2*4/19, lty = 2) # h=2 times p / n =2 times 4 / 19

## DFFITS
lm.reg.dffits = dffits(lm.reg)
plot(lm.reg.dffits, type = "h", ylab = "DFFITS", ylim = c(-2.5,4.5),cex.lab=1.5)
text(lm.reg.dffits, labels = index, cex = 0.8, pos = 1)
abline(h = c(-1,-2*sqrt(4/19), 0, 2*sqrt(4/19), 1), lty = 2) # specify your own h
```

```
## cook's distance
lm.reg.cooksD = cooks.distance(lm.reg)
plot(lm.reg.cooksD, type = "h", ylab="Cook's Distance",ylim=c(0,2),cex.lab=1.5)
text(lm.reg.cooksD, labels = index, cex = 1)
abline(h=qf(0.50,4,15), lty=2) #check whether D_i > f_0.5,p,n-p
## dfbetas
lm.reg.dfbetas = dfbetas(lm.reg)
plot(lm.reg.dfbetas[,1], type = "h", ylab = "DFBETAS", xlab = "Index", main =
"Intercept",cex.lab=1.5,cex.main=1.5)
text(lm.reg.dfbetas[,1], labels = index, cex = 0.8)
abline(h=c(-1, -2/sqrt(19), 0, 2/sqrt(19), 1), lty = 2) # specify your own threshold

#4(f)
b_new=read.table("softdrink_new.txt",header = TRUE)
lm.reg=lm(data=b_new,y~x1+x2)
summary(lm.reg)
```