# MA 542 REGRESSION ANALYSIS
# SPRING 2018

HW - 6 - Solution Key

1. (Chapter 5 question 25)

1. (1)

$$(X'X)^{-1} = \begin{pmatrix} 0.2 & -0.1 \\ -0.1 & 0.1 \end{pmatrix}$$

(2) b = (10.2,4)

(3) e = (1.8,-1.2,-1.2,1.8,-0.2,-1.2,-2.2,0.8,0.8,0.8)

(4)

$$H = \begin{pmatrix} 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 \\ 0.1 & 0.2 & 0.0 & 0.2 & -0.1 & 0.1 & 0.2 & 0.1 & 0.0 & 0.2 \\ 0.1 & 0.0 & 0.2 & 0.0 & 0.3 & 0.1 & 0.0 & 0.1 & 0.2 & 0.0 \\ 0.1 & 0.2 & 0.0 & 0.2 & -0.1 & 0.1 & 0.2 & 0.1 & 0.0 & 0.2 \\ 0.1 & -0.1 & 0.3 & -0.1 & 0.5 & 0.1 & -0.1 & 0.1 & 0.3 & -0.1 \\ 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 \\ 0.1 & 0.2 & 0.0 & 0.2 & -0.1 & 0.1 & 0.2 & 0.1 & 0.0 & 0.2 \\ 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 \\ 0.1 & 0.0 & 0.2 & 0.0 & 0.3 & 0.1 & 0.0 & 0.1 & 0.2 & 0.0 \\ 0.1 & 0.2 & 0.0 & 0.2 & -0.1 & 0.1 & 0.2 & 0.1 & 0.0 & 0.2 \end{pmatrix}$$

(5) SSE = 17.6

(6)

$$s^2\{b\} = \begin{pmatrix} 0.44 & -0.22 \\ -0.22 & 0.22 \end{pmatrix}$$

(7) $\hat{Y}_h$ when $X_h = 2$ is 18.2

(8) $s^2\{\hat{Y}_h\} = 0.44$

2. (1) $s^2\{b_1\} = 0.22$

(2) $s\{b_0, b_1\} = -0.22$

(3) $s\{b_0\} = \sqrt{0.44} = 0.66$

3. SSR = 160

2. (Chapter 6 question 1)

a) Let $Z_{i1} = X_{i1}$ and $Z_{i2} = X_{i1}X_{i2}$ Then the matrices are:

$$X = \begin{bmatrix} 1 & z_{11} & z_{12} \\ 1 & z_{21} & z_{22} \\ 1 & z_{31} & z_{32} \\ 1 & z_{41} & z_{42} \end{bmatrix} \qquad \text{and} \qquad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}$$

b) Let $Y_i' = log(Y_i)$ Then the matrices are:

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ 1 & x_{31} & x_{32} \\ 1 & x_{41} & x_{42} \end{bmatrix} \qquad \text{and} \qquad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}$$

3. (Chapter 6 question 3)

Yes that is true $R^2$ increases with number of predictors. But practically we can not use all the predictors. We may not know about some of the predictors. Even though all the predictors are known, Some of them may not have a significant increment in $R^2$. Computations for a model with more predictors are difficult and interpretation of the model is getting complex with number of predictors.

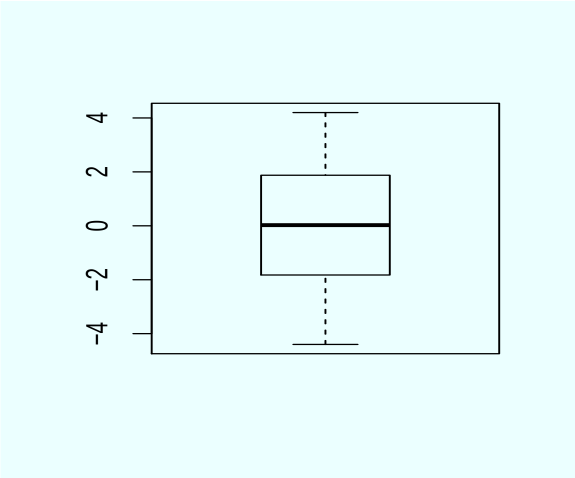
4. (Chapter 6 question 5)

b)

```
> data=read.table("https://netfiles.umn.edu/users/nacht001/www/nachtsheim/Kutner/Chapter%20%206%20Data%
> Y=data[,1]
> X1=data[,2]
> X2=data[,3]
>
> data=data.frame(Y,X1,X2)
> LR=lm(Y~X1+X2)
> LR

Call:
lm(formula = Y ~ X1 + X2)

Coefficients:
(Intercept)           X1           X2
     37.650        4.425        4.375
```

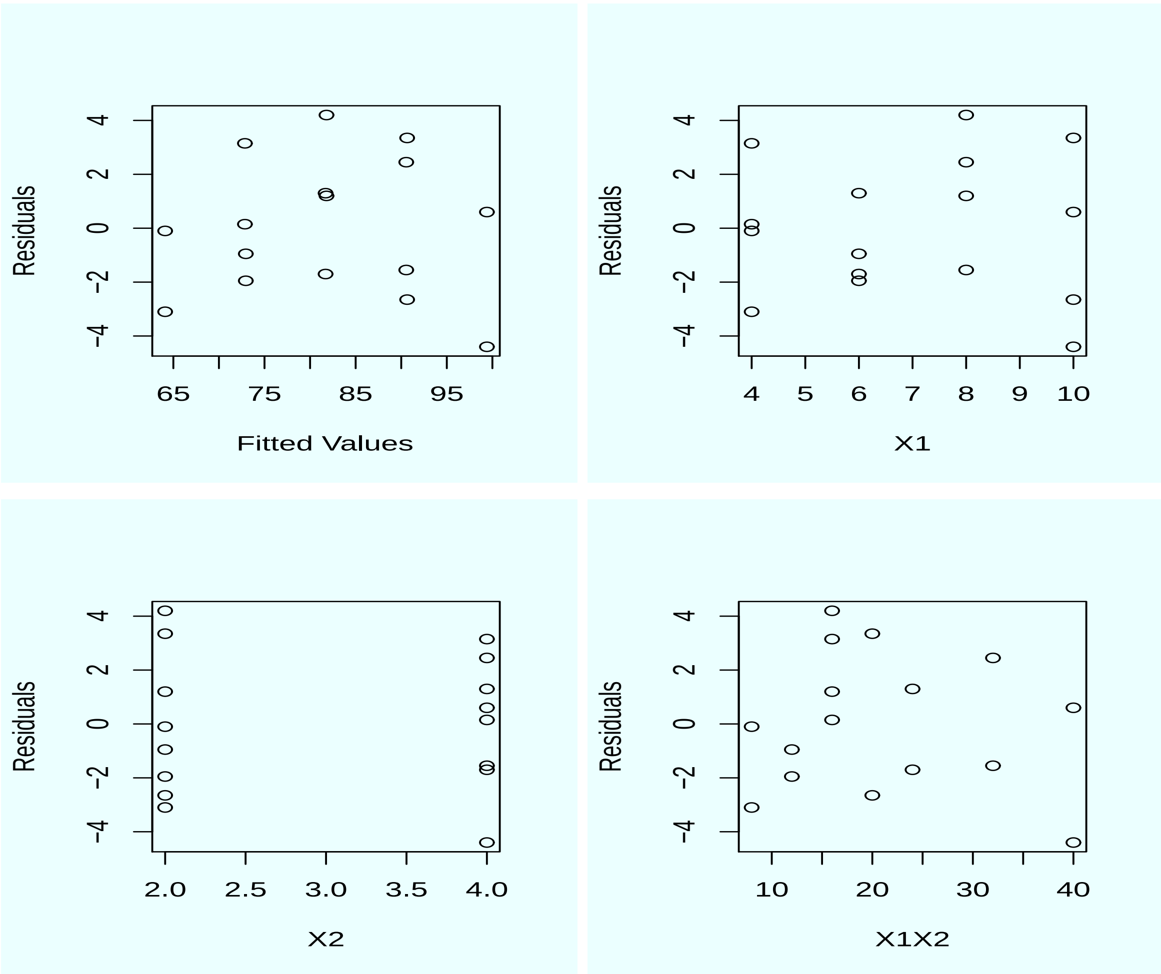
So the estimated Regression function is $\hat{Y} = 37.65 + 4.425X - 1 + 4.375X_2$.

$b_1$ means the change in the mean degree of brand liking per unit increase of moisture content when sweetness is hold constant.

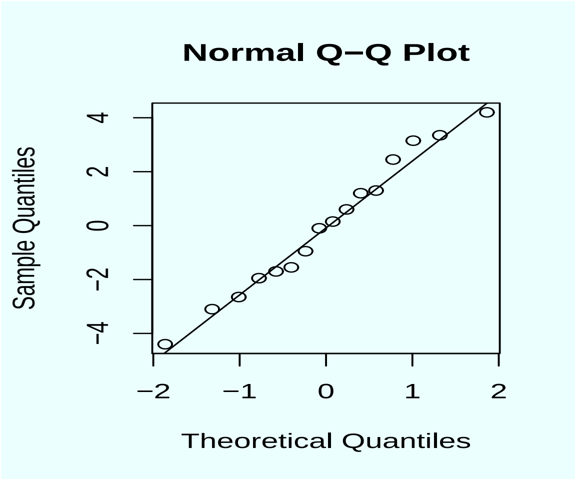
c) 


The box plot is roughly symmetric around 0 and all the residuals are from -4 to 4. So we can assume that the residuals follows a normal distribution. However a normal probability plot is needed to make this sure.

d) Residual plots



Normal probability plot




The plots of residual against fitted values, residuals against $X_1$ and residuals against $X_2$ support the appropriateness of the regression model for the data. Plot of residuals against $X_1X_2$ has no any specific pattern. So the interaction term is not that important in the model. Normal probability plot is approximately linear. So Residuals are proximately normal.

f)

```
> data=read.table("https://netfiles.umn.edu/users/nacht001/www/nachtsheim/Kutner/Chapter%20%206%20Data%
> Y=data[,1]
> X1=data[,2]
> X2=data[,3]
>
> data=data.frame(Y,X1,X2)
> LR=lm(Y~X1+X2)
> LR

Call:
lm(formula = Y ~ X1 + X2)

Coefficients:
(Intercept)           X1           X2
     37.650        4.425        4.375
```

$$H_0 : E[Y] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \quad \text{vs} \quad H_1 : E[Y] \neq \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

$$F^* = \frac{37.3}{5} / \frac{57}{8} = 1.047$$

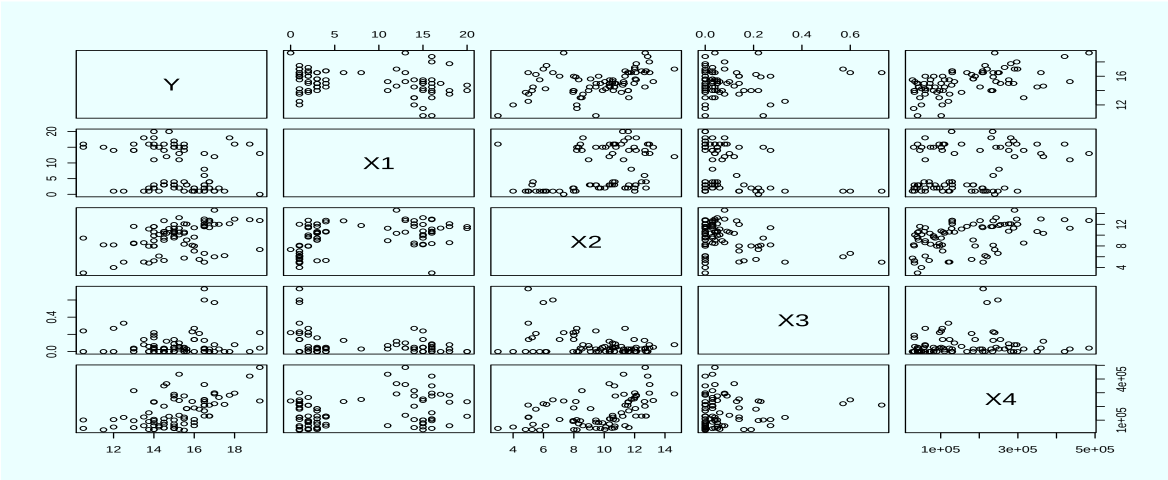
But,

$$F_{0.99,5,8} = 6.63 \quad \text{OR} \quad P - value = 0.453$$

Since $F_{0.99,5,8} = 6.63 > F^* = 1.047$ or $P - value = 0.453 > 0.01$, $H_0$ is not rejected. So the relationship between brand liking and predictors moisture content and sweetness explained well by the regression model.

5. (Chapter 6 question 18)

b)




```
> cor(data)
              Y          X1          X2          X3         X4
Y    1.00000000  -0.2502846   0.4137872  0.06652647 0.53526237
X1  -0.25028456   1.0000000   0.3888264 -0.25266347 0.28858350
X2   0.41378716   0.3888264   1.0000000 -0.37976174 0.44069713
X3   0.06652647  -0.2526635  -0.3797617  1.00000000 0.08061073
X4   0.53526237   0.2885835   0.4406971  0.08061073 1.00000000
```

From the scatter plot matrix we can see that taxes ($X_2$) and square footage ($X_4$) have positive linear relationships with rental rates ($Y$) and relation between the age ($X_1$) and Vacancy rate ($X_3$) is week. There is a high correlation between ($X_2$) and ($X_4$). We also can confirm our finding from the correlation matrix.

c)

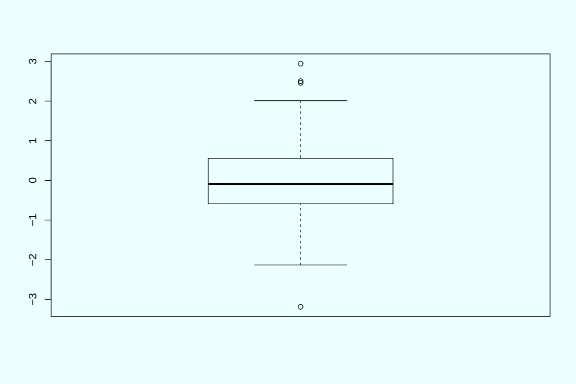
```
> RLR=lm(Y~X1+X2+X3+X4)
> RLR

Call:
lm(formula = Y ~ X1 + X2 + X3 + X4)

Coefficients:
(Intercept)           X1            X2            X3            X4
  1.220e+01   -1.420e-01     2.820e-01     6.193e-01     7.924e-06
```

So the estimated Regression function is

$$\hat{Y} = 12.2 + 0.142X - 1 + 0.282X_2 + 0.6193X_3 + 0.0000008X_4$$

.

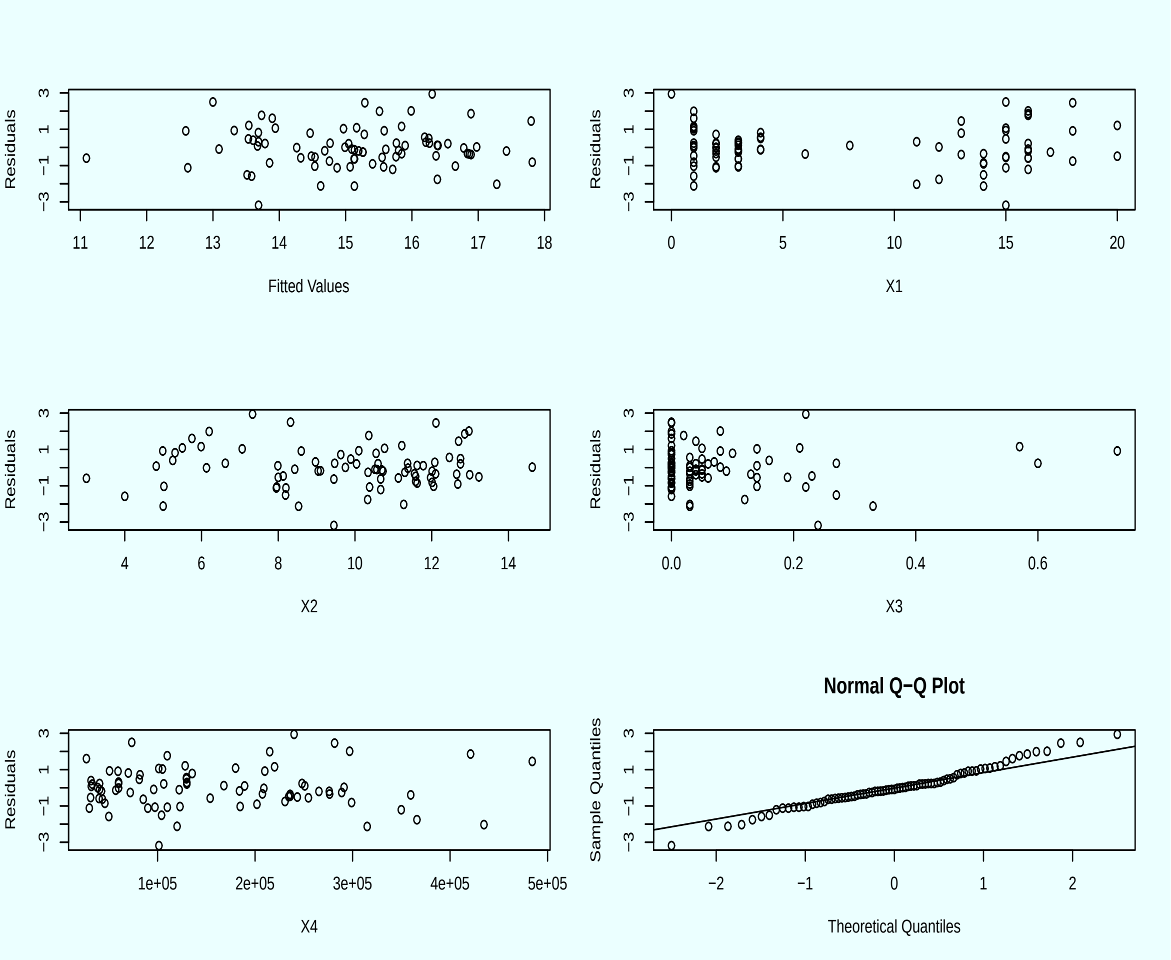
d)



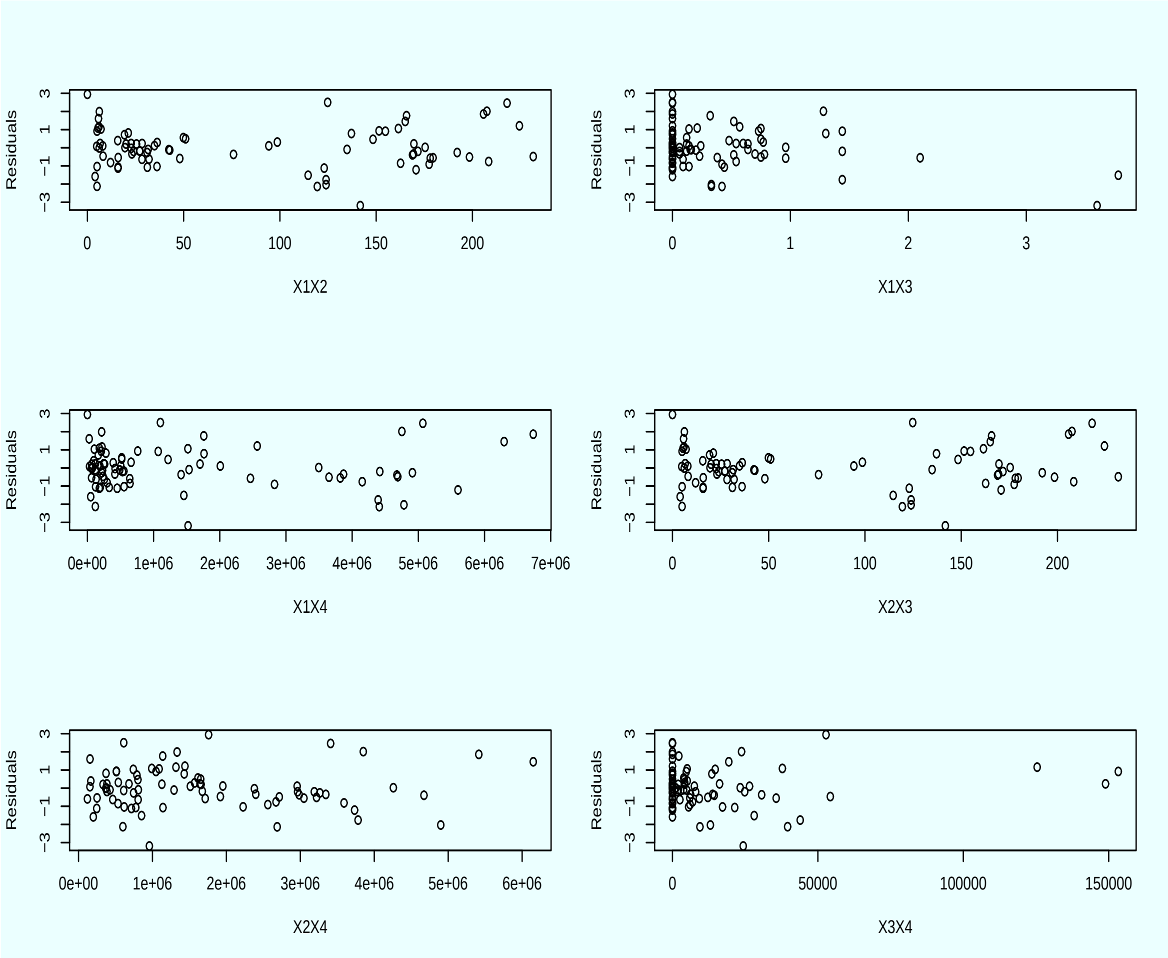
Yes the distribution of the residuals is fairly symmetric.

e)

From the following residual plots, we can see that there is no any model departure. SO the fitted regression model is appropriate for the data.




Since it is hard to see any pattern from the following plots (i.e., plots of residuals vs interaction terms), the interaction terms are not needed in the model.




f) No, The lack of test requires repeat observations for one or more combination of X levels, but this data set has no repeat observations. Further this test is not practical here since there are many combination of X-levels.

g)

[h]

Here since $|T^*| < t_{0.975,79}$, we can conclude that the error variance is constant.

6. (Chapter 6 question 22)
a) If we let $Z_{i1} = X_{i1}$, $\;Z_{i2} = \log X_{i2}$ and $Z_{i3} = X_{i1}^2$, then

$$Y_i = \beta_0 + \beta_1 Z_{i1} + \beta_2 Z_{i2} + \beta_3 Z_{i3} + \epsilon_i$$

b) This can not be transformed to GLR model (because any transformation will trans-

```
> Fitvals=fitted.values(RLR)
> RFV=sort(Fitvals)
> CP=RFV[41]

> #Dividing residuals into two groups.
>
> R1=c(rep(0,40))
> R2=c(rep(0,41))
> j=0
> k=0
>
> for(i in 1:81){
+   if (Fitvals[i]<CP) {
+     j=j+1
+     R1[j]=ResR[i]
+ } else {
+   k=k+1
+   R2[k]=ResR[i]
+ }
+ }
>
> D1=abs(R1-median(R1))
>
> D2=abs(R2-median(R2))
> #Pooled Variance
>  S2=(sum(D1-mean(D1)^2)+sum(D2-mean(D2))^2)/79
> Tstar=(mean(D1)-mean(D2))/(sqrt(S2)*sqrt(1/40+1/41))
> Tstar
[1] 1.694825
> Ttable=qt(0.975,79)
> Ttable
[1] 1.99045
```

form the model parameters).

c) This model can not be transformed to the GLR model.

d) This model can not be transformed to the GLR model.

e) If we let $Y_i' = Y_i$, then

$$Y_i' = \beta_0 + \beta_1 X_{i1} + \epsilon_i$$

7. (Chapter 6 question 24)
a) Least squares criterion: Minimize

$$Q = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_{i1} - \beta_2 X_{i1}^2 - \beta_3 X_{i2})^2$$

By considering all the partial derivatives and setting them to 0, we get

$$\begin{bmatrix} n & \sum_{i=1}^n X_{i1} & \sum_{i=1}^n X_{i1}^2 & \sum_{i=1}^n X_{i2} \\ \sum_{i=1}^n X_{i1} & \sum_{i=1}^n X_{i1}^2 & \sum_{i=1}^n X_{i1}^3 & \sum_{i=1}^n X_{i1}X_{i2} \\ \sum_{i=1}^n X_{i1}^2 & \sum_{i=1}^n X_{i1}^3 & \sum_{i=1}^n X_{i1}^4 & \sum_{i=1}^n X_{i1}^2 X_{i2} \\ \sum_{i=1}^n X_{i2} & \sum_{i=1}^n X_{i1}X_{i2} & \sum_{i=1}^n X_{i1}^2 X_{i2} & \sum_{i=1}^n X_{i2}^2 \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ 3 \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^n Y_i \\ \sum_{i=1}^n X_{i1} Y_i \\ \sum_{i=1}^n X_{i1}^2 Y_i \\ \sum_{i=1}^n X_{i2} Y_i \end{bmatrix}$$

Then the normal equations are:

$$\sum_{i=1}^{n} Y_i = nb_0 + b_1 \sum_{i=1}^{n} X_{i1} + b_2 \sum_{i=1}^{n} X_{i1}^2 + b_3 \sum_{i=1}^{n} X_{i2}$$

$$\sum_{i=1}^{n} X_{i1}Y_i = b_0 X_{i1} + b_1 X_{i1}^2 + b_2 \sum_{i=1}^{n} X_{i1}^3 + b_3 \sum_{i=1}^{n} X_{i1}X_{i2}$$

$$\sum_{i=1}^{n} X_{i1}^2 Y_i = b_0 X_{i1}^2 + b_1 X_{i1}^3 + b_2 \sum_{i=1}^{n} X_{i1}^4 + b_3 \sum_{i=1}^{n} X_{i1}^2 X_{i2}$$

$$\sum_{i=1}^{n} X_{i2}Y_i = b_0 X_{i2}^2 + b_1 X_{i1}X_{i2} + b_2 \sum_{i=1}^{n} X_{i1}^2 X_{i2} + b_3 \sum_{i=1}^{n} X_{i2}^2$$

b) The maximum Likelihood Function is:

$$L(\beta_0, \beta_1, \beta_2, \beta_3, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(\frac{(Y_i - \beta_0 - \beta_1 X_{i1} - \beta_2 X_{i1}^2 - \beta_3 X_{i2})^2}{2\sigma^2}\right)$$

Then the partial derivatives are

$$\frac{dln(L(\beta_0, \beta_1, \beta_2, \beta_3, \sigma^2))}{d\beta_0} = -\frac{1}{2\sigma^2} 2 \sum (Y_i - \beta_0 - b_1 \sum_{i=1}^{n} X_{i1} - \beta_2 \sum_{i=1}^{n} X_{i1}^2 - \beta_3 \sum_{i=1}^{n} X_{i2})(-1) = 0,$$

$$\frac{dln(L(\beta_0, \beta_1, \beta_2, \beta_3, \sigma^2))}{d\beta_1} = -\frac{1}{2\sigma^2} 2 \sum (Y_i - \beta_0 - b_1 \sum_{i=1}^{n} X_{i1} - \beta_2 \sum_{i=1}^{n} X_{i1}^2 - \beta_3 \sum_{i=1}^{n} X_{i2})(-X_{i1}) = 0,$$

$$\frac{dln(L(\beta_0, \beta_1, \beta_2, \beta_3, \sigma^2))}{d\beta_1} = -\frac{1}{2\sigma^2} 2 \sum (Y_i - \beta_0 - b_1 \sum_{i=1}^{n} X_{i1} - \beta_2 \sum_{i=1}^{n} X_{i1}^2 - \beta_3 \sum_{i=1}^{n} X_{i2})(-X_{i1}^2) = 0,$$

$$\frac{dln(L(\beta_0, \beta_1, \beta_2, \beta_3, \sigma^2))}{d\beta_1} = -\frac{1}{2\sigma^2} 2 \sum (Y_i - \beta_0 - b_1 \sum_{i=1}^{n} X_{i1} - \beta_2 \sum_{i=1}^{n} X_{i1}^2 - \beta_3 \sum_{i=1}^{n} X_{i2})(-X_{i2}) = 0,$$

Then the system of equation is

$$\sum_{i=1}^{n} Y_i = nb_0 + b_1 \sum_{i=1}^{n} X_{i1} + b_2 \sum_{i=1}^{n} X_{i1}^2 + b_3 \sum_{i=1}^{n} X_{i2}$$

$$\sum_{i=1}^{n} X_{i1}Y_i = b_0 X_{i1} + b_1 X_{i1}^2 + b_2 \sum_{i=1}^{n} X_{i1}^3 + b_3 \sum_{i=1}^{n} X_{i1}X_{i2}$$

$$\sum_{i=1}^{n} X_{i1}^2 Y_i = b_0 X_{i1}^2 + b_1 X_{i1}^3 + b_2 \sum_{i=1}^{n} X_{i1}^4 + b_3 \sum_{i=1}^{n} X_{i1}^2 X_{i2}$$

$$\sum_{i=1}^{n} X_{i2}Y_i = b_0 X_{i2}^2 + b_1 X_{i1}X_{i2} + b_2 \sum_{i=1}^{n} X_{i1}^2 X_{i2} + b_3 \sum_{i=1}^{n} X_{i2}^2$$

So Least square estimators and the Maximum Likelihood Estimators are same.