Rthvik Raviprakash
PUID:0030816481

Final Exam (100)

Note:

- For demonstrating conceptual understanding, you are required to work on the model that is easier to handle or compute, not necessarily the more suitable (or more complicated) model for the dataset. Follow the question description.
- You don't need to check the assumption of a model unless the question asks for it. For example, if the question asks you to make prediction based on a model, you don't need to check the assumption for the model before making prediction.
- For any of the testing (hypothesis test) problem, define Ho/Ha, compute the test statistic, report the exact p value, and state the conclusion. The default alpha value is 5%, unless specify.
- Elaborate your reasoning clearly and show relevant plots, R results, and tables to support your opinion in each step and conclusion.
- Attach the R code along with your answer in a format similar to the homework.
- The data is real, just like the project you are working on. Hence it is possible that even after the remedial method has been done, the model is still not perfect. When this happens, evaluation will be based on the level you execute the methods **covered in Stat512** to improve the model. Don't worry if your model is not perfect, try your best to demonstrate the skill set you learn in this class.

Study the data with a linear analysis and complete the problems. The data set has three continues predictors and two categorical predictors.

Problem 1(10). Summarize data statistics on the continuous variables.
a. (5) What is the mean and standard deviation of Y, X1, X2, X3? What is the sample size?

-

```
> dataJK2 <- read.csv("C:/Users/user/Downloads/dataJK2.csv")
>   View(dataJK2)
> mean(dataJK2$y)
[1] 47.27273
> mean(dataJK2$x1)
[1] 36.18182
> mean(dataJK2$x2)
[1] 39.59091
> mean(dataJK2$x3)
[1] 40.45455
```

The means are as follows:

- Y = 47.27273
- X1 = 36.18182
- X2 = 39.59091
- X3 = 40.45455

```
> sd(dataJK2$y)
[1] 22.70134
> sd(dataJK2$x1)
[1] 17.66254
> sd(dataJK2$x2)
[1] 23.87599
> sd(dataJK2$x3)
[1] 12.42013
```

The standard deviations are as follows:

- Y = 22.70134
- X1 = 17.66254
- X2 = 23.87599
- X3 = 12.42013

- The sample size of the dataset is equal to 22.

b. (5) Complete the following **mean table** by filling in the **? part** with the means of Y for each categories, means for the corresponding rows and columns, and finally the mean for all Y (the grand mean).

| | | X4 | | |
|---|---|---|---|---|
| | | high | low | |
| | loss | 54.8333 | 26.2500 | 43.4 |
| X5 | more | 32.6000 | 63.2857 | 50.5 |
| | | 44.72727 | 49.81818 | 47.27273 |

-

```
> mean.df1 <- aggregate(y ~ x4 + x5, dataJK2, mean)
> mean.df1
    x4   x5        y
1 high less 54.83333
2  low less 26.25000
3 high more 32.60000
4  low more 63.28571
```

```
> mean.df2 <- aggregate(y~x4, dataJK2, mean)
> mean.df2
    x4        y
1 high 44.72727
2  low 49.81818
> mean.df3 <- aggregate(y~x5, dataJK2, mean)
> mean.df3
    x5    y
1 less 43.4
2 more 50.5
```

Problem 2. Consider only the first order model with X1, X2 and X3, perform the following hypothesis.
a. (10) whether X1 can be dropped from the full model.
b. (10) whether X1 can be dropped from the model containing only X1 and X2.

Model with x1, x2 and x3.

```
> mod123 <- lm(y~x1+x2+x3, dataJK2)
> summary(mod123)

Call:
lm(formula = y ~ x1 + x2 + x3, data = dataJK2)

Residuals:
    Min      1Q  Median      3Q     Max 
-21.353 -11.848  -1.457   7.542  42.413 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)   
(Intercept) 55.63321   17.72549   3.139  0.00568 **
x1           0.02878    0.21410   0.134  0.89457   
x2          -0.60816    0.16223  -3.749  0.00147 **
x3           0.36278    0.31182   1.163  0.25985   
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 17.27 on 18 degrees of freedom
Multiple R-squared:  0.5037,    Adjusted R-squared:  0.4209 
F-statistic: 6.089 on 3 and 18 DF,  p-value: 0.004788

> anova(mod123)
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)    
x1         1    9.8     9.8  0.0328 0.8583821    
x2         1 5037.2  5037.2 16.8798 0.0006596 ***
x3         1  403.9   403.9  1.3535 0.2598525    
Residuals 18 5371.5   298.4                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> mod23 <- lm(y~x2+x3, dataJK2)
> summary(mod23)

Call:
lm(formula = y ~ x2 + x3, data = dataJK2)

Residuals:
    Min      1Q  Median      3Q     Max 
-21.539 -11.969  -0.763   7.469  41.870 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)   
(Intercept)  56.7121    15.3898   3.685  0.00157 **
x2           -0.6071     0.1578  -3.848  0.00109 **
x3            0.3608     0.3033   1.189  0.24892   
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 16.82 on 19 degrees of freedom
Multiple R-squared:  0.5032,    Adjusted R-squared:  0.4509 
F-statistic: 9.621 on 2 and 19 DF,  p-value: 0.0013


> anova(mod23)
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)    
x2         1 5045.1  5045.1 17.8278 0.0004613 ***
x3         1  400.4   400.4  1.4147 0.2489221    
Residuals 19 5376.9   283.0                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

a. The model: $Y \sim X1 + X2 + X3$
- By performing a GLT
- Consider a 5% significance level: $\alpha = 0.05$
- $H_0 : \beta_1 = 0 \; and \; H_a : \beta_1 \neq 0$
- Df of the reduced model = 19, df for the full model = 18
- SSR(X1|X2,X3) = SSR(X1,X2,X3) – SSR(X2,X3) = 5450.9 – 5445.5 = 5.4
- The F statistic: $F_s = \frac{\frac{SSE(R)-SSE(F)}{df_R - df_F}}{\frac{SSE(F)}{df_F}} = \frac{\frac{SSR(X1|X2,X3)}{n-p+1-n+p}}{\frac{SSE(X1,X2,X3)}{n-4}} = \frac{\frac{5.4}{1}}{\frac{5371.5}{18}} = 0.01809$
-

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> pf(0.01809, 1, 18, lower.tail = FALSE)
[1] 0.8945005
```

- Since the p value = 0.8945005 > 0.05 we fail to reject the null hypothesis at a significance level of 5% we can conclude that there is not enough evidence to support the fact that X1 cannot be dropped. This means X1 can probably be dropped.

```
> mod12 <- lm(y~x1+x2, dataJK2)
> summary(mod12)

Call:
lm(formula = y ~ x1 + x2, data = dataJK2)

Residuals:
    Min      1Q  Median      3Q     Max 
-32.919 -11.582  -1.449   6.234  39.916 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 72.39786   10.41782   6.949 1.27e-06 ***
x1           0.01679    0.21583   0.078 0.938803    
x2          -0.64996    0.15966  -4.071 0.000652 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 17.43 on 19 degrees of freedom
Multiple R-squared:  0.4663,	Adjusted R-squared:  0.4102 
F-statistic: 8.302 on 2 and 19 DF,  p-value: 0.002564
```

```
> anova(mod12)
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)    
x1         1    9.8     9.8  0.0322 0.8595622    
x2         1 5037.2  5037.2 16.5715 0.0006518 ***
Residuals 19 5775.4   304.0                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> mod2 <- lm(y~x2, dataJK2)
> summary(mod2)

Call:
lm(formula = y ~ x2, data = dataJK2)

Residuals:
    Min      1Q  Median      3Q     Max 
-32.991 -11.701  -0.973   6.544  39.606 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  72.9744     7.1380  10.223 2.18e-09 ***
x2           -0.6492     0.1553  -4.179 0.000463 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 17 on 20 degrees of freedom
Multiple R-squared:  0.4662,	Adjusted R-squared:  0.4395 
F-statistic: 17.47 on 1 and 20 DF,  p-value: 0.0004627
```

```
> anova(mod2)
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)    
x2         1 5045.1  5045.1  17.466 0.0004627 ***
Residuals 20 5777.2   288.9                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

b. The model: $Y \sim X1 + X2$

- By performing a GLT
- Consider a 5% significance level: $\alpha = 0.05$
- $H_0: \beta_1 = 0 \text{ and } H_a: \beta_1 \neq 0$
- Df of the reduced model = 20, df for the full model = 19
- SSR(X1|X2) = SSR(X1,X2) – SSR(X2) = 5047 – 5045.1 = 1.9
- The F statistic: $F_s = \frac{\frac{SSE(R)-SSE(F)}{df_R - df_F}}{\frac{SSE(F)}{df_F}} = \frac{\frac{SSR(X1|X2)}{n-p+1-n+p}}{\frac{SSE(X1,X2)}{n-3}} = \frac{1.9}{\frac{5045.1}{19}} = 0.007155$
-

```
> pf(0.007155, 1, 19, lower.tail = FALSE)
[1] 0.9334745
```

- Since the p value = 0.9334745 > 0.05 we fail to reject the null hypothesis at a significance level of 5% we can conclude that there is not enough evidence to support the fact that X1 cannot be dropped. This means X1 can probably be dropped.

Problem 3 (10) Consider the first order model with X1, X2 and X3, simultaneously estimate parameters (beta1, beta2 and beta3) with a confidence level of 75%.

-

```
> mod123

Call:
lm(formula = y ~ x1 + x2 + x3, data = dataJK2)

Coefficients:
(Intercept)           x1           x2           x3
   55.63321      0.02878     -0.60816      0.36278

> summary(mod123)

Call:
lm(formula = y ~ x1 + x2 + x3, data = dataJK2)

Residuals:
    Min      1Q  Median      3Q     Max
-21.353 -11.848  -1.457   7.542  42.413

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 55.63321   17.72549   3.139  0.00568 **
x1           0.02878    0.21410   0.134  0.89457
x2          -0.60816    0.16223  -3.749  0.00147 **
x3           0.36278    0.31182   1.163  0.25985
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 17.27 on 18 degrees of freedom
Multiple R-squared:  0.5037,    Adjusted R-squared:  0.4209
F-statistic: 6.089 on 3 and 18 DF,  p-value: 0.004788
```

- $b_0 = 55.63321$
- $b_1 = 0.02878$
- $b_2 = -0.60816$
- $b_3 = 0.36278$
- B = $t(1 - \frac{\alpha}{2g}, df)$, g = 3
- B = t(0.95833, 18) = 1.833363
-

```
> qt(0.95833,18)
[1] 1.833363
```

The 75% confidence interval for $b_1$ would be $b_1 \pm B * s\{b_1\}$
$s\{b_1\} = 0.21410$, B = 1.833363
$b_1 \pm B * s\{b_1\} = 0.02878 \pm 1.833363 * 0.21410 = 0.02878 \pm 0.392523$
= (-0.363743, 0.421303)

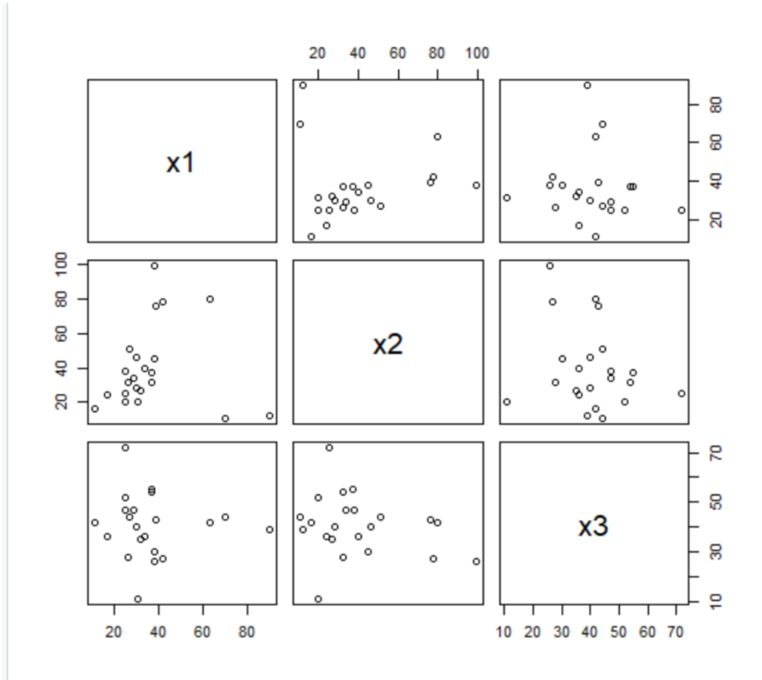
The 75% confidence interval for $b_2$ would be $b_2 \pm B * s\{b_2\}$
$s\{b_2\} = 0.16223$, B = 1.833363
$b_2 \pm B * s\{b_2\} = -0.60816 \pm 1.833363 * 0.16223 = -0.60816 \pm (1.995593)$
= (-2.603753, 1.387433)

The 75% confidence interval for $b_0$ would be $b_3 \pm B * s\{b_3\}$
$s\{b_3\} = 0.31182$, B = 1.833363
$b_3 \pm B * s\{b_3\} = 0.36278 \pm 1.833363 * 0.31182 = 0.36278 \pm 0.571679$
= (-0.208899, 0.934459)

Problem 4 (20) Perform appropriate analysis to diagnose the potential issues with the first order full mode with X1 X2 and X3, improve the model as much as possible with the methods covered in Stat512. You should also consider the assumption checking for your revised model.

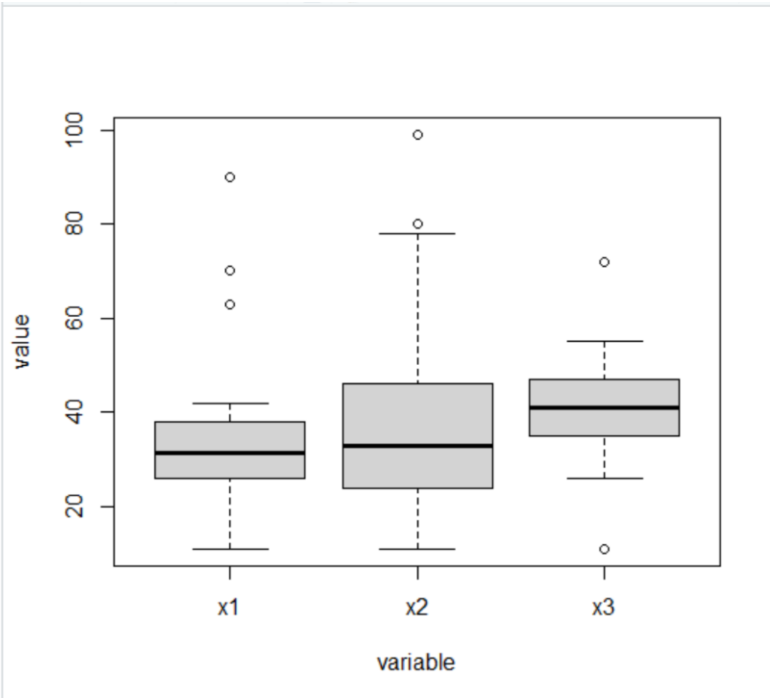
-

```
> data_sub_x <- dataJK2[c(2,3,4)]
> plot(data_sub_x)
```




- As we can see from the scatterplot above x1 and x2 might be correlated since we see a linear pattern in the scatterplot which we do not see between the other variables. So, there might be a possibility of multicollinearity among the variables.
- If multicollinearity exists it is probably because of the correlation due to the linear pattern observed between x1 and x2.

```
> library(reshape)
> meltData <- melt(data_sub_x)
Using  as id variables
> boxplot(data=meltData, value~variable)
```




- As we can see from the boxplot above there are outliers in the predictors. If the influence of these cases on the linear function is not significant, we can ignore them if not we have to investigate those outliers.

1. There might be a multicollinearity issue.
2. So, as we can see there are outliers present.

```
> mod123 <- lm(y~x1+x2+x3, dataJK2)
> mod123

Call:
lm(formula = y ~ x1 + x2 + x3, data = dataJK2)

Coefficients:
(Intercept)           x1           x2           x3
   55.63321      0.02878     -0.60816      0.36278
```

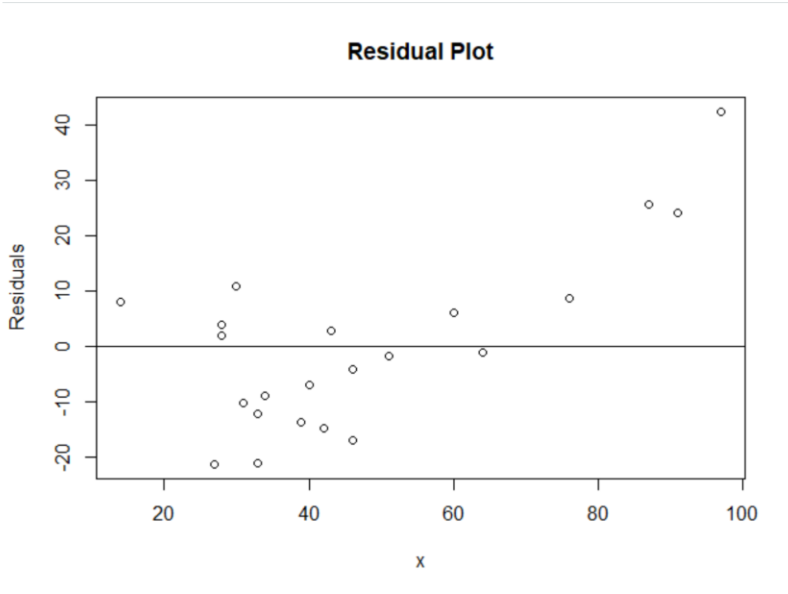
```
> summary(mod123)

Call:
lm(formula = y ~ x1 + x2 + x3, data = dataJK2)

Residuals:
    Min      1Q  Median      3Q     Max 
-21.353 -11.848  -1.457   7.542  42.413 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)   
(Intercept) 55.63321   17.72549   3.139  0.00568 **
x1           0.02878    0.21410   0.134  0.89457   
x2          -0.60816    0.16223  -3.749  0.00147 **
x3           0.36278    0.31182   1.163  0.25985   
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 17.27 on 18 degrees of freedom
Multiple R-squared:  0.5037,	Adjusted R-squared:  0.4209 
F-statistic: 6.089 on 3 and 18 DF,  p-value: 0.004788

> anova(mod123)
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)    
x1         1    9.8     9.8  0.0328 0.8583821    
x2         1 5037.2  5037.2 16.8798 0.0006596 ***
x3         1  403.9   403.9  1.3535 0.2598525    
Residuals 18 5371.5   298.4                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- As we can see from the above anova output only x2 seems to have a significant linear impact on y and not x1 and x3 since the p values for x1 and x3 which are 0.8583821 and 0.2598525 respectively are greater than 0.05.

We assume that the random residuals have constant variance and are normally distributed and are also independent.

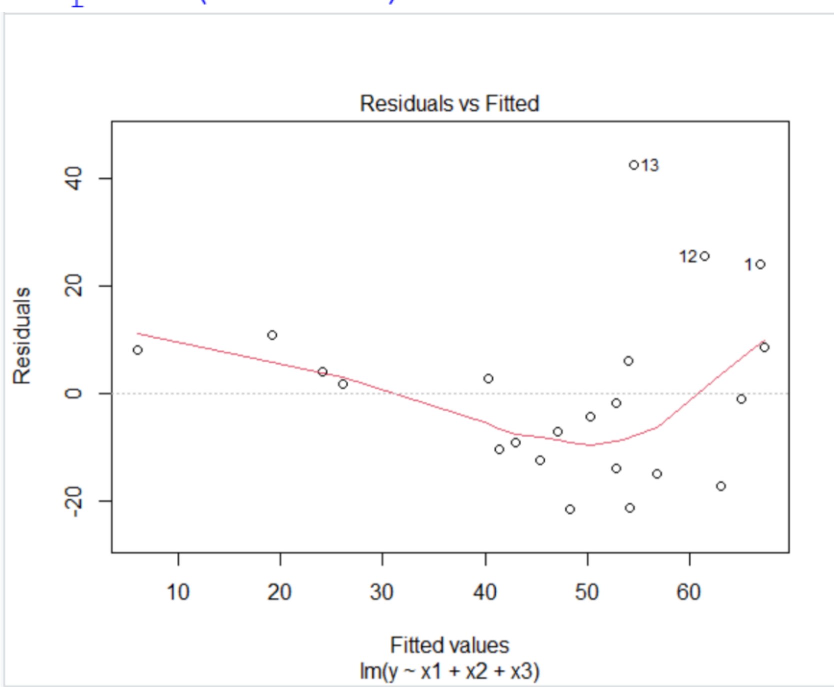
```
> mod123.res <- resid(mod123)
> plot(dataJK2$y, mod123.res, xlab = 'x', ylab = 'Residuals', main = 'Residual Plot' )
> abline(0,0)
```

The residual plot of the model:

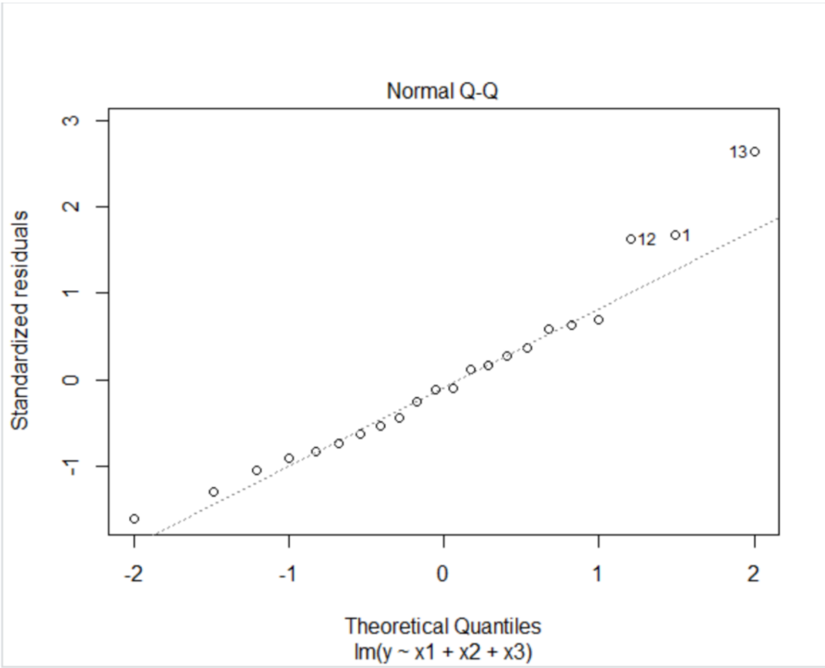



As we can see from the plot there seems to be outliers and the points are also not evenly distributed above and below the line.

```
> plot(mod123)
```




- Based on the plot above there might be some issues with constant variance because of the slightly curved shape of the variables.




- We can see outliers in the Normal qq plot which might be providing us a hint of a violation in the normality of the residuals.

  We need to do further tests in order to find out if these violations are possible.

  Let us perform the Brown Forsythe Test:

$$H_0: residuals\ have\ constant\ variances$$
$$H_a: residuals\ have\ non-constant\ variances$$

```
> library(onewaytests)
> dataJK2$group <- cut(dataJK2$y, 5)
> dataJK2$residual <- mod123$residuals
> bf.test(residual~group, dataJK2)

  Brown-Forsythe Test (alpha = 0.05)
-------------------------------------------------------------
  data : residual and group

  statistic  : 12.81604
  num df     : 4
  denom df   : 9.093385
  p.value    : 0.0008899772

  Result     : Difference is statistically significant.
-------------------------------------------------------------
```

As we can see from the above test the p value is less than 0.05 and we will reject the null hypothesis, so the difference is statistically significant. The constant variance assumption fails in this case.

Now let's perform the Shapiro-Wilk normality test:

$$H_0: The\ Data\ follows\ normal\ distribution$$
$$H_a: The\ Data\ violated\ from\ normal\ distribution$$

-

```
> shapiro.test(dataJK2$residual)

        Shapiro-Wilk normality test

data:  dataJK2$residual
W = 0.93413, p-value = 0.1496
```

The p-value 0.1496 > 0.05 so we do not reject the null hypothesis. So, there is not enough evidence to show that the data violates normal distribution. Hence, the data is mostly normally distributed.

We are not sure if there is a violation on the independence.

Transformation of the model:
Since we see that there is a constant variance violation, let's try and transform the model.

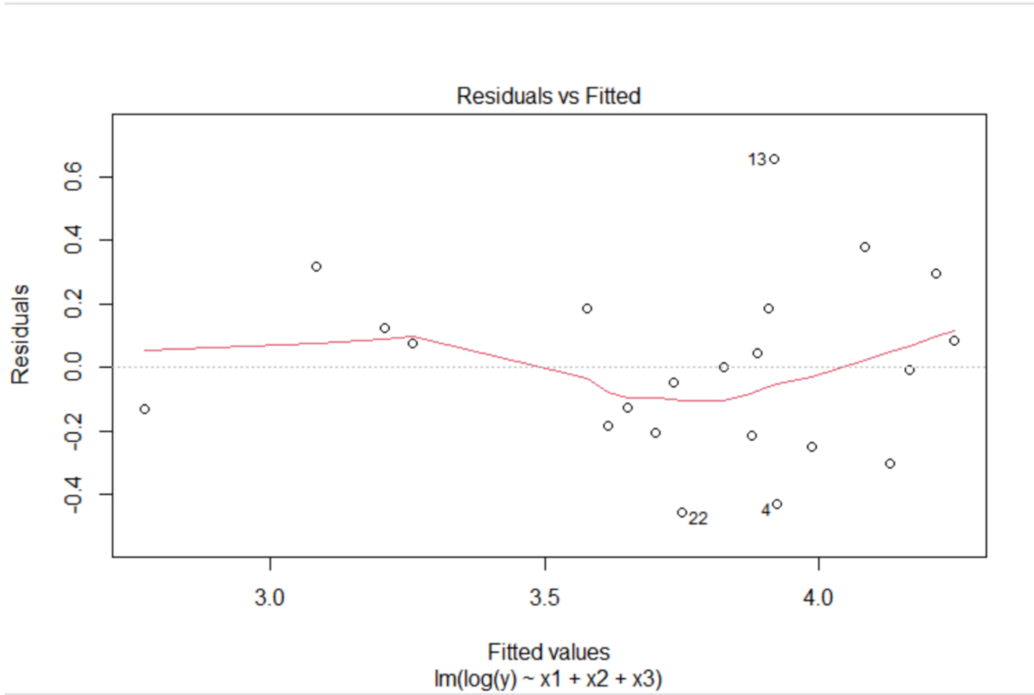
For simplicity let's choose $\lambda = 0$,
The transformation function is: $Y' = Y^{\lambda}\ and\ Y^{\lambda} = \ln(Y)\ when\ \lambda = 0$
The back transformation function would be: $f^{-1}(Y') = e^{Y'}$

So, now we refit the model, and the new model would be:

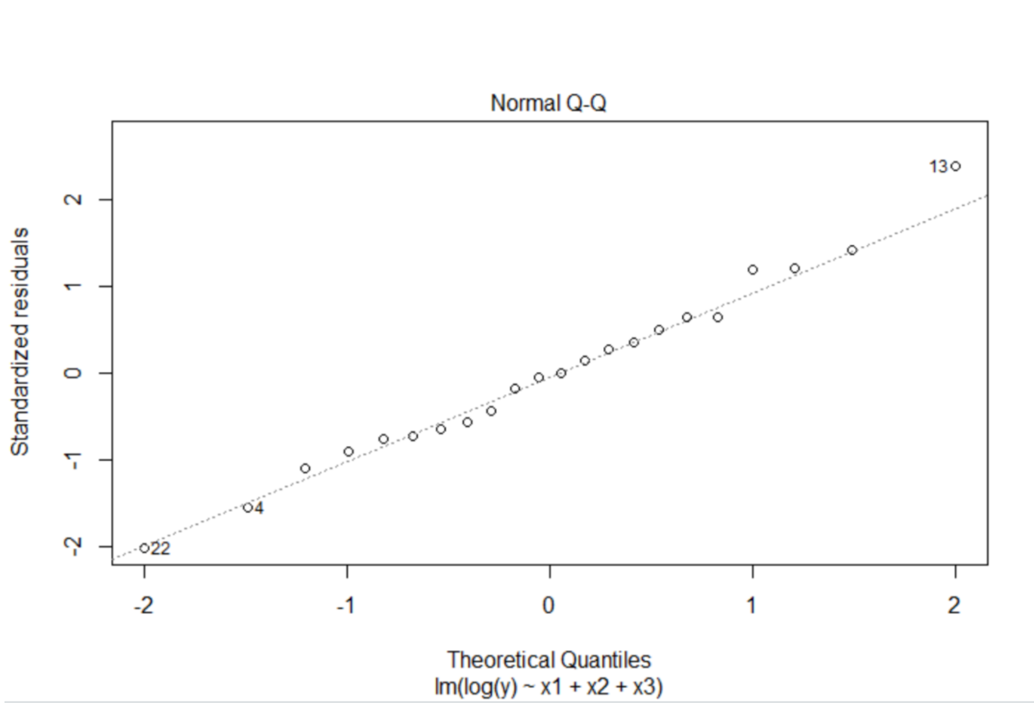
$\text{Ln}(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$

```
> mod123t <- lm(log(y)~x1+x2+x3, dataJK2)
> plot(mod123t)
```

```
> summary(mod123t)

Call:
lm(formula = log(y) ~ x1 + x2 + x3, data = dataJK2)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.45506 -0.19985 -0.00347  0.16888  0.65569 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  3.9121928  0.3024439  12.935 1.49e-10 ***
x1           0.0006625  0.0036531   0.181    0.858    
x2          -0.0142479  0.0027681  -5.147 6.76e-05 ***
x3           0.0093754  0.0053205   1.762    0.095 .  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2948 on 18 degrees of freedom
Multiple R-squared:  0.663,	Adjusted R-squared:  0.6069 
F-statistic: 11.81 on 3 and 18 DF,  p-value: 0.0001647

> anova(mod123t)
Analysis of Variance Table

Response: log(y)
          Df  Sum Sq Mean Sq F value    Pr(>F)    
x1         1 0.00597 0.00597  0.0687   0.79625    
x2         1 2.80153 2.80153 32.2464 2.191e-05 ***
x3         1 0.26977 0.26977  3.1051   0.09502 .  
Residuals 18 1.56382 0.08688                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```




This above shape of the plot seems to show us that the constant variance assumption is satisfied when the model is transformed by the above process, but we still need to check it by the BF test.



The normality assumption seems to be satisfied but we still need to see using the Shapiro-Wilk normality test.

Let us perform the Brown Forsythe Test:

$$H_0: residuals\ have\ constant\ variances$$
$$H_a: residuals\ have\ non-constant\ variances$$

```
> library(onewaytests)
> dataJK2$group <- cut(dataJK2$y, 5)
> dataJK2$residual <- mod123t$residuals
> bf.test(residual~group, dataJK2)

  Brown-Forsythe Test (alpha = 0.05)
-------------------------------------------------------------
  data : residual and group

  statistic  : 6.880715
  num df     : 4
  denom df   : 9.325274
  p.value    : 0.007397547

  Result     : Difference is statistically significant.
-------------------------------------------------------------
```

As we can see from the above test the p value is less than 0.05 and we will reject the null hypothesis, so the difference is statistically significant. The constant variance assumption fails in this case.

Now let's perform the Shapiro-Wilk normality test:

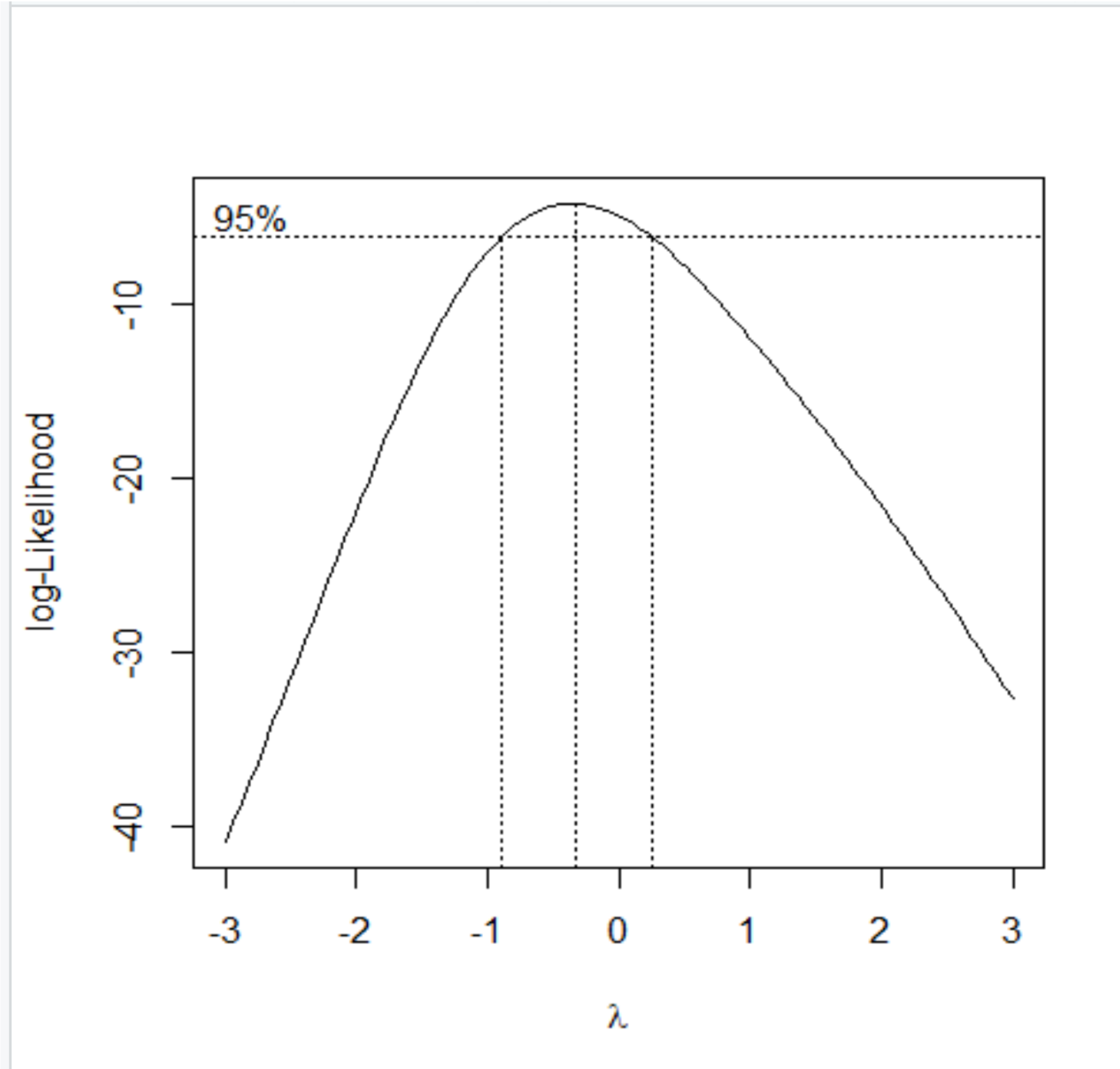
$$H_0: The\ Data\ follows\ normal\ distribution$$
$$H_a: The\ Data\ violated\ from\ normal\ distribution$$

```
> shapiro.test(dataJK2$residualt)

        Shapiro-Wilk normality test

data:  dataJK2$residualt
W = 0.98015, p-value = 0.9183
```

The p-value 0.98015 > 0.05 so we do not reject the null hypothesis. So, there is not enough evidence to show that the data violates normal distribution. Hence, the data is mostly normally distributed.

Now let us try the box-cox transformation:

```
> bcmle <- boxcox(lm(y~x1+x2+x3, dataJK2), lambda = seq(-3,3))
> lambda <- bcmle$x[which.max(bcmle$y)]
> lambda
[1] -0.3333333
```




So, now let us transform $Y' = Y^{-0.333}$

```
> bcmle <- boxcox(lm(y~x1+x2+x3, dataJK2), lambda = seq(-3,3))
> sub_mod <- lm(y^-0.3333~ x1 + x2 + x3, dataJK2)
> dataJK2$residualt <- sub_mod$residuals
> bf.test(residualt~group, dataJK2)

  Brown-Forsythe Test (alpha = 0.05)
-------------------------------------------------------------
  data : residualt and group

  statistic  : 4.377272
  num df     : 4
  denom df   : 8.433382
  p.value    : 0.03366787

  Result     : Difference is statistically significant.
-------------------------------------------------------------
```

Even in this case the difference is not statistically significant, so we need to pick a different value of lambda.

Let's pick lambda = -0.5

Now let us transform and perform the tests again:

```
> sub_mod <- lm(y^-0.5~ x1 + x2 + x3, dataJK2)
> dataJK2$residualt <- sub_mod$residuals
> bf.test(residualt~group, dataJK2)

  Brown-Forsythe Test (alpha = 0.05)
-------------------------------------------------------------
  data : residualt and group

  statistic  : 3.282455
  num df     : 4
  denom df   : 7.868858
  p.value    : 0.07276398

  Result     : Difference is not statistically significant.
-------------------------------------------------------------
```

Let us perform the Brown Forsythe Test:

$$H_0: residuals\ have\ constant\ variances$$
$$H_a: residuals\ have\ non-constant\ variances$$

As we can see from the above test the p value is greater than 0.05 and we will not reject the null hypothesis, so the difference is not statistically significant. This means that there is enough evidence for the constant variance assumption.

Now we see that the difference is not statistically significant.

So, the constant variance assumption is satisfied.

```
> shapiro.test(dataJK2$residualt)

        Shapiro-Wilk normality test

data:  dataJK2$residualt
W = 0.98493, p-value = 0.9746
```

$$H_0: The\ Data\ follows\ normal\ distribution$$
$$H_a: The\ Data\ violated\ from\ normal\ distribution$$

The p-value 0.9746 > 0.05 so we do not reject the null hypothesis. So, there is not enough evidence to show that the data violates normal distribution. Hence, the data is mostly normally distributed.

Let's look at the transformed model:

```
> mod123fit <- lm(y^-0.5~x1+x2+x3, dataJK2)
> summary(mod123fit)

Call:
lm(formula = y^-0.5 ~ x1 + x2 + x3, data = dataJK2)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.041856 -0.014172 -0.000811  0.012991  0.034666 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.1443537  0.0223605   6.456 4.49e-06 ***
x1          -0.0000475  0.0002701  -0.176   0.8624    
x2           0.0011803  0.0002046   5.767 1.82e-05 ***
x3          -0.0007912  0.0003934  -2.011   0.0595 .  
---
Signif. codes:  
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.02179 on 18 degrees of freedom
Multiple R-squared:  0.7131,    Adjusted R-squared:  0.6653 
F-statistic: 14.91 on 3 and 18 DF,  p-value: 3.997e-05
```

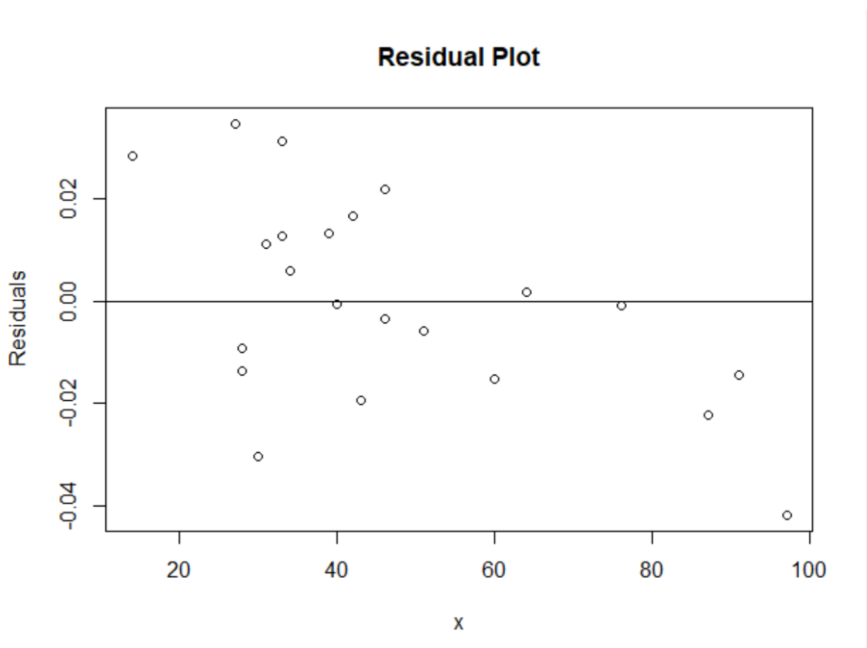
```
> anova(mod123fit)
Analysis of Variance Table

Response: y^-0.5
          Df    Sum Sq   Mean Sq F value    Pr(>F)    
x1         1 0.0000496 0.0000496  0.1045   0.75016    
x2         1 0.0192747 0.0192747 40.5885 5.32e-06 ***
x3         1 0.0019210 0.0019210  4.0453   0.05951 .  
Residuals 18 0.0085479 0.0004749                      
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The refitted model is better with a higher multiple R-squared.
The multiple R-squared for the original model = 0.5037 and for the refitted model = 0.7131

Now there is no violation on the variance or the normality.

The residual plot of the transformed model:

```
> mod123fit.res <- resid(mod123fit)
> plot(dataJK2$y, mod123fit.res, xlab = 'x', ylab = 'Residuals', main = 'Residual Plot' )
> abline(0,0)
```




Here we don't see that many outliers and the points seem to be evenly distributed above and below the line.

Problem 5 . a. (10) Compute AIC, BIC, and PRESSP to compare the following two models.

- The model on the first order terms for X1 and X2 and the interaction term X1X2.
- The model on the first order terms for X1, X2 and X3

Do they all yield the same better model? If not, explain.

- The model on the first order terms for X1 and X2 and the interaction term X1X2. Let's call it model 1 in this problem.
-
- $AIC_P = n * \ln(SSE_P) - n * \ln(n) + 2p$

$$SBC_P = n * \ln(SSE_P) - n * \ln(n) + \ln(n) * p$$

```
> anova(dataJK2.mod12)
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value   Pr(>F)
x1         1    9.8     9.8  0.0309 0.862497
x2         1 5037.2  5037.2 15.9025 0.000863 ***
x1:x2      1   73.8    73.8  0.2330 0.635116
Residuals 18 5701.6   316.8
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> dataJK2.mod123 <- lm(y~x1+x2+x3, dataJK2)
> anova(dataJK2.mod123)
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)
x1         1    9.8     9.8  0.0328 0.8583821
x2         1 5037.2  5037.2 16.8798 0.0006596 ***
x3         1  403.9   403.9  1.3535 0.2598525
Residuals 18 5371.5   298.4
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> dataJK2['X1X2'] = dataJK2['x1']*dataJK2['x2']
> bs <- BestSub(dataJK2[c(2,3,7)], dataJK2$y, num = 1)
> bs
  p 1 2 3     SSEp        r2    r2.adj        Cp     AICp     SBCp   PRESSp
1 2 0 1 0 5777.228 0.4661768 0.4394857 0.2388219 126.5540 128.7361 6805.750
2 3 0 1 1 5735.379 0.4700438 0.4142589 2.1067030 128.3941 131.6672 7087.019
3 4 1 1 1 5701.581 0.4731668 0.3853613 4.0000000 130.2640 134.6282 7951.174
```

AIC = 130.2640, BIC or SBCp = 134.6282, PRESSp = 7951.174

- The model on the first order terms for X1, X2 and X3. Let's call it model 2 in this problem.

```
> bs3 <- BestSub(dataJK2[c(2,3,4)], dataJK2$y, num = 1)
> bs3
  p 1 2 3     SSEp        r2     r2.adj       Cp     AICp     SBCp   PRESSp
1 2 0 1 0 5777.228 0.4661768 0.4394857 1.359698 126.5540 128.7361 6805.750
2 3 0 1 1 5376.865 0.5031709 0.4508731 2.018066 126.9740 130.2471 7453.768
3 4 1 1 1 5371.474 0.5036691 0.4209473 4.000000 128.9519 133.3161 8422.497
```

AIC = 128.9519, BIC or SBCp = 133.3161, PRESSp = 8422.497

As we can see we have the following:
- The AIC is lower for model2 compared to model1.
- The BIC is lower for model2 compared to model1.
- The PRESSp is lower for model1 compared to model2.

So, based on this we might say that since both AIC and BIC are lower for model2 compared to model1, model2 would be the better model.
If all AIC, BIC and PRESSp were lower for one model we could have concluded that model is better.
Here, since the PRESSp is lower for the model other than the model which has its AIC and BIC lower I would say that we cannot just conclude that one model is better than the other.

Here it depends on what we are trying to do when we want to pick the model:

So, if we are trying to predict we need to pick the model with the lower PRESSp which would be model 1 in this case.
But if we are trying to look at the impact of the predictors on the model we pick the one with the less AICp and BICp.

b. (10) Select the model that you think is better to predict the mean response value, then predict the mean response for the following case, at a confident level of 99%.

| x1 | x2 | x3 |
|---|---|---|
| 45 | 36 | 45 |

- Here since we are trying to predict the mean response, we pick the model with the lower PRESSp which would be model1. So, I think that the better model in this case would be model1 which is Y~X1+ X2+ X1*X2
- $\alpha = 0.01$
- X1*x2 = 1620 when x1 = 45 and x2 = 36

```
> dataJK2.mod12 <- lm(y~x1+x2+x1*x2, dataJK2)
> new <- data.frame(1,45,36,1620)
> ci.reg(dataJK2.mod12, new, type = 'm', alpha = 0.01)
  X.Intercept. x1 x2 x1.x2      Fit Lower.Band
1            1 45 36  1620 50.50609   37.28991
  Upper.Band
1   63.72228
```

The confidence interval for the mean response would be (37.28991, 63.72228)

Problem 6. Consider the impact of X4 and X5 on Y.

a. (10) Perform a test for the significant interaction effect between X4 and X5 on Y. If the interaction effect is significant, use your own words to describe the how X4 and X5 interactively affecting Y. (Hint, use the mean table in problem 1).

-

```
> dataJK2.mod45 <- lm(y~x4*x5, dataJK2)
> summary(dataJK2.mod45)

Call:
lm(formula = y ~ x4 * x5, data = dataJK2)

Residuals:
    Min      1Q  Median      3Q     Max 
-23.286 -12.696  -0.925   8.562  36.167 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)    54.833      7.354   7.457 6.57e-07 ***
x4low         -28.583     11.627  -2.458  0.02432 *  
x5more        -22.233     10.907  -2.038  0.05648 .  
x4low:x5more   59.269     15.698   3.776  0.00138 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 18.01 on 18 degrees of freedom
Multiple R-squared:  0.4604,    Adjusted R-squared:  0.3704 
F-statistic: 5.118 on 3 and 18 DF,  p-value: 0.009798
```

```
> anova(dataJK2.mod45)
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value   Pr(>F)
x4         1  142.5   142.5  0.4393 0.515842
x5         1  214.6   214.6  0.6613 0.426713
x4:x5      1 4625.0  4625.0 14.2547 0.001385 **
Residuals 18 5840.2   324.5
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The proposed model would be: $Y_{i,j,k} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{i,j} + \epsilon_{i,j,k}$

Where $\mu = the\ grand\ mean, estimated\ by\ Ybar$

- $\alpha_i\ is\ the\ main\ effect\ of\ belonging\ to\ level\ i\ of\ factor\ A, estimated\ by\ Y_i bar - Y\ bar..$
- $\beta_i\ is\ the\ main\ effect\ of\ belonging\ to\ level\ j\ of\ factor\ B, estimated\ by\ Y_j bar - Y\ bar..$
- $(\alpha\beta)\_i, j\ is\ the\ interaction\ effect\ of\ belonging\ to\ both\ i\ an\ dj\ estimated\ by\ Y_{i,j} bar - Y\ bar..$

The hypothesis test:

$$H_0: all(\alpha\beta_{i,j}) = 0, H_a: not\ all(\alpha\beta_{i,j}) = 0$$

We reject the null hypothesis since the p value is less than 0.05

The p -value = 0.001385 < 0.05 and hence the interaction effect is significant.

If there is no interaction effect the values of the estimates would be equal to 0 but since there is an interaction effect the values of the estimates are not equal to 0 as we can see from the table above.

b. (5) With the ANOVA method, compute the 95% confidence interval for the following difference, respectively:

D1= The difference in the mean of Y when (X4=high, X5=less) and (X4=high, X5=more)

D2= The difference in the mean of Y when (X4=low, X5=less) and (X4=low, X5=more)

-

```
> library(gmodels)
> modelq6 <- lm(y~x4:x5+0, dataJK2)
> summary(modelq6)

Call:
lm(formula = y ~ x4:x5 + 0, data = dataJK2)

Residuals:
    Min      1Q  Median      3Q     Max
-23.286 -12.696  -0.925   8.563  36.167

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
x4high:x5less   54.833      7.354   7.457 6.57e-07 ***
x4low:x5less    26.250      9.006   2.915 0.009249 **
x4high:x5more   32.600      8.056   4.047 0.000757 ***
x4low:x5more    63.286      6.808   9.296 2.72e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 18.01 on 18 degrees of freedom
Multiple R-squared:  0.9026,    Adjusted R-squared:  0.881
F-statistic: 41.72 on 4 and 18 DF,  p-value: 7.171e-09


> anova(modelq6)
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)
x4:x5      4  54146 13536.4   41.72 7.171e-09 ***
Residuals 18   5840   324.5
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> val_d1 <- c(1,0,-1,0)
> estimable(modelq6, val_d1)
           Estimate Std. Error  t value DF   Pr(>|t|)
(1 0 -1 0) 22.23333   10.90721 2.038407 18 0.05647629
> qt(1-(0.05/2), 18)
[1] 2.100922
```

- The 95% confidence interval for D1 would be:
- $estimate \pm t * std\_error$
- $22.23333 \pm 2.100922 * 10.90721$
- $22.23333 \pm 22.9151974$
- The interval: (-0.681844, 45.1485274)

```
> val_d2 <- c(0,1,0,-1)
> estimable(modelq6, val_d2)
            Estimate Std. Error   t value DF   Pr(>|t|)
(0 1 0 -1) -37.03571   11.29004 -3.280389 18 0.00415737
>
```

- The 95% confidence interval for D2 would be:
- $estimate \pm t * std\_error$
- $-37.03571 \pm (2.100922) * 11.29004$
- -37.03571 $\pm$ (23.719493)
- The interval: (-60.7551934, -13.316217)

c. (5) With the ANOVA method, compute the 95% confidence interval for D1-D2

Where D1 and D2 are described in b.

How is your result related to a?

-

```
> val_6c <- c(1,-1,-1,1)
> estimable(modelq6, val_6c)
              Estimate Std. Error  t value DF     Pr(>|t|)
(1 -1 -1 1) 59.26905   15.69816 3.775541 18 0.001384848
```

```
> qt(1-(0.05/2), 18)
[1] 2.100922
```

- The 95% confidence interval for D1 – D2 would be:
- $estimate \pm t * std\_error$
- $59.26905 \pm (2.100922) * 15.69816$
- 59.26905 $\pm$ (32.9806097)
- The interval: (26.2884403, 92.2496597)

In ‘a’ we concluded that there is a significant interaction, and here as we can see in the interval there is no 0, so we can see that none of the differences are equal to each other. In this way this result is related to ‘a’.