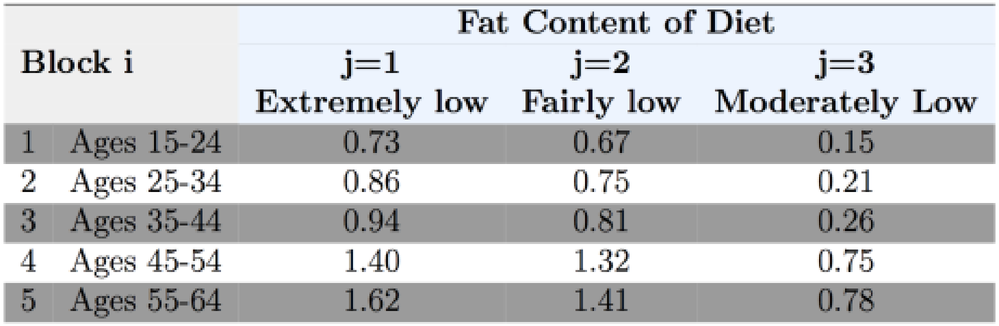
Assignment 10

Jiarong Ye

November 30, 2018

Q1

1. (15 points) Fat in diets. A researcher studied the effects of three experimental diets with varying fat contents on the total lipid (fat) level in plasma. Total lipid level is widely used predictor of coronary heart disease. Fifteen male subjects who were within 20% of their ideal body weight were grouped into five blocks according to age. Within each block, the three experimental diets were randomly assigned to the three subjects. Data on reduction in lipid level (in grams per liter) after the subjects were on the diet for a fixed period of time follow.



Block i		Fat Content of Diet		
		j=1 Extremely low	j=2 Fairly low	j=3 Moderately Low
1	Ages 15-24	0.73	0.67	0.15
2	Ages 25-34	0.86	0.75	0.21
3	Ages 35-44	0.94	0.81	0.26
4	Ages 45-54	1.40	1.32	0.75
5	Ages 55-64	1.62	1.41	0.78

(a)

Why do you think that age of subject was used as a blocking variable?

Because age is a confounding factor and the total fat level might be varied depending on ages, so age should be used as a blocking variable.

(b)

Obtain the residuals for randomized block model $Y_{ij} = \mu + \tau_i + \beta_j + \epsilon_{ij}$ and plot them against the fitted values. Also prepare a normal probability plot of the residuals. What are your findings?

```
In [18]: fat = c(0.73,0.67,0.15,0.86,0.75,0.21,0.94,0.81,0.26,1.40,1.32,0.75,1.62,1.41,0.78)
         diet = rep(c("j1", "j2", "j3"), 5)
         age = c(rep("15-24", 3), rep("25-34", 3), rep("35-44", 3), rep("45-54", 3),
          rep("55-64", 3))
```

```
df = data.frame(fat, diet, age)
df
```

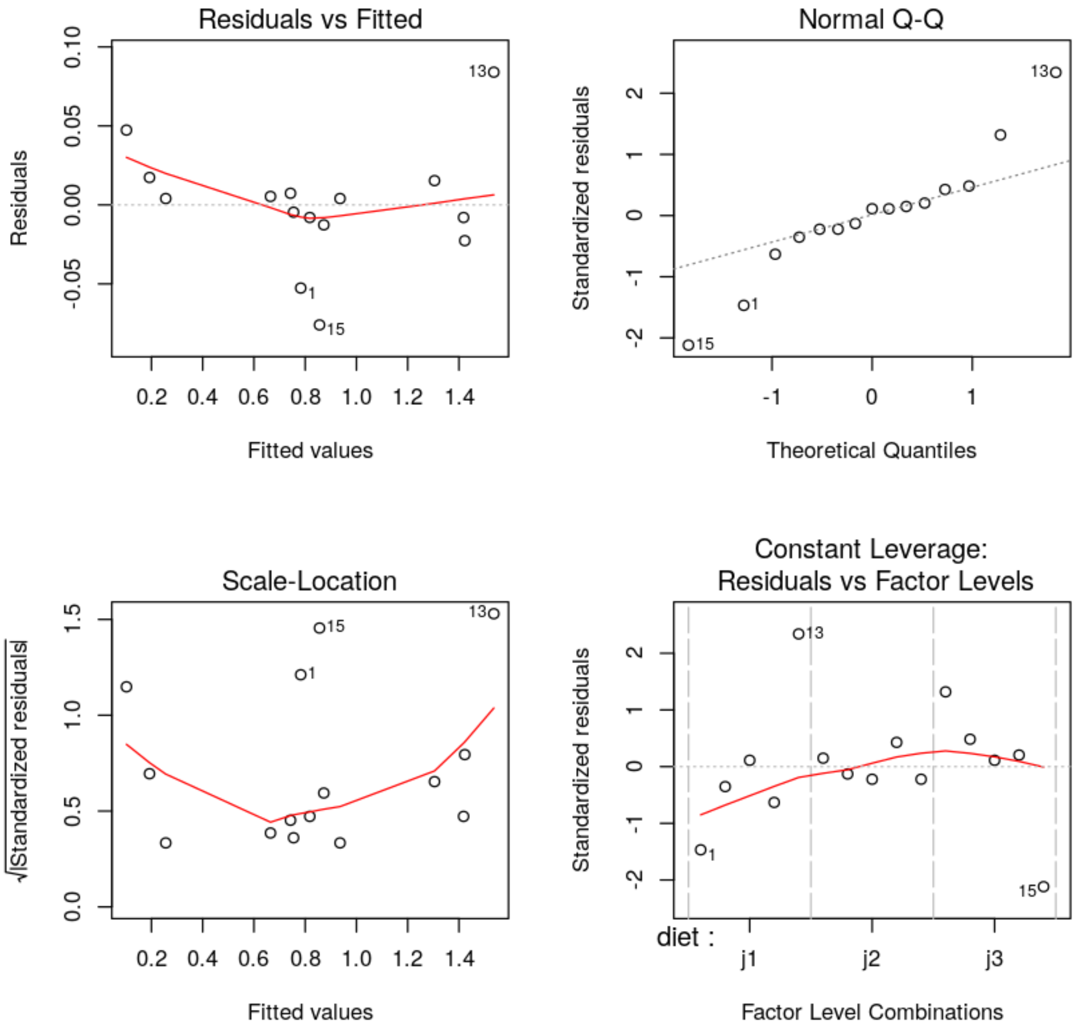
fat	diet	age
0.73	j1	15-24
0.67	j2	15-24
0.15	j3	15-24
0.86	j1	25-34
0.75	j2	25-34
0.21	j3	25-34
0.94	j1	35-44
0.81	j2	35-44
0.26	j3	35-44
1.40	j1	45-54
1.32	j2	45-54
0.75	j3	45-54
1.62	j1	55-64
1.41	j2	55-64
0.78	j3	55-64

```
In [8]: fat.lm = lm(fat ~ diet+age, data=df)
        fat.res = resid(fat.lm)
        diet_1_res = fat.res[c(1, 4, 7, 10, 13)]
        diet_2_res = fat.res[c(2, 5, 8, 11, 14)]
        diet_3_res = fat.res[c(3, 6, 9, 12, 15)]
        diet_1_res
        diet_2_res
        diet_3_res
```

```
 1  -0.0526666666666666

 4  -0.0126666666666666

 7  0.00399999999999987

10  -0.0226666666666668

13  0.0840000000000001



 2  0.00533333333333322

 5  -0.00466666666666661

 8  -0.00799999999999996

11  0.0153333333333334

14  -0.00800000000000009
```

```
 3  0.0473333333333334

 6  0.0173333333333332

 9  0.00400000000000009

12  0.00733333333333337

15  -0.076
```

```
In [9]: par(mfrow=c(2,2))
        plot(aov(fat.lm))
```



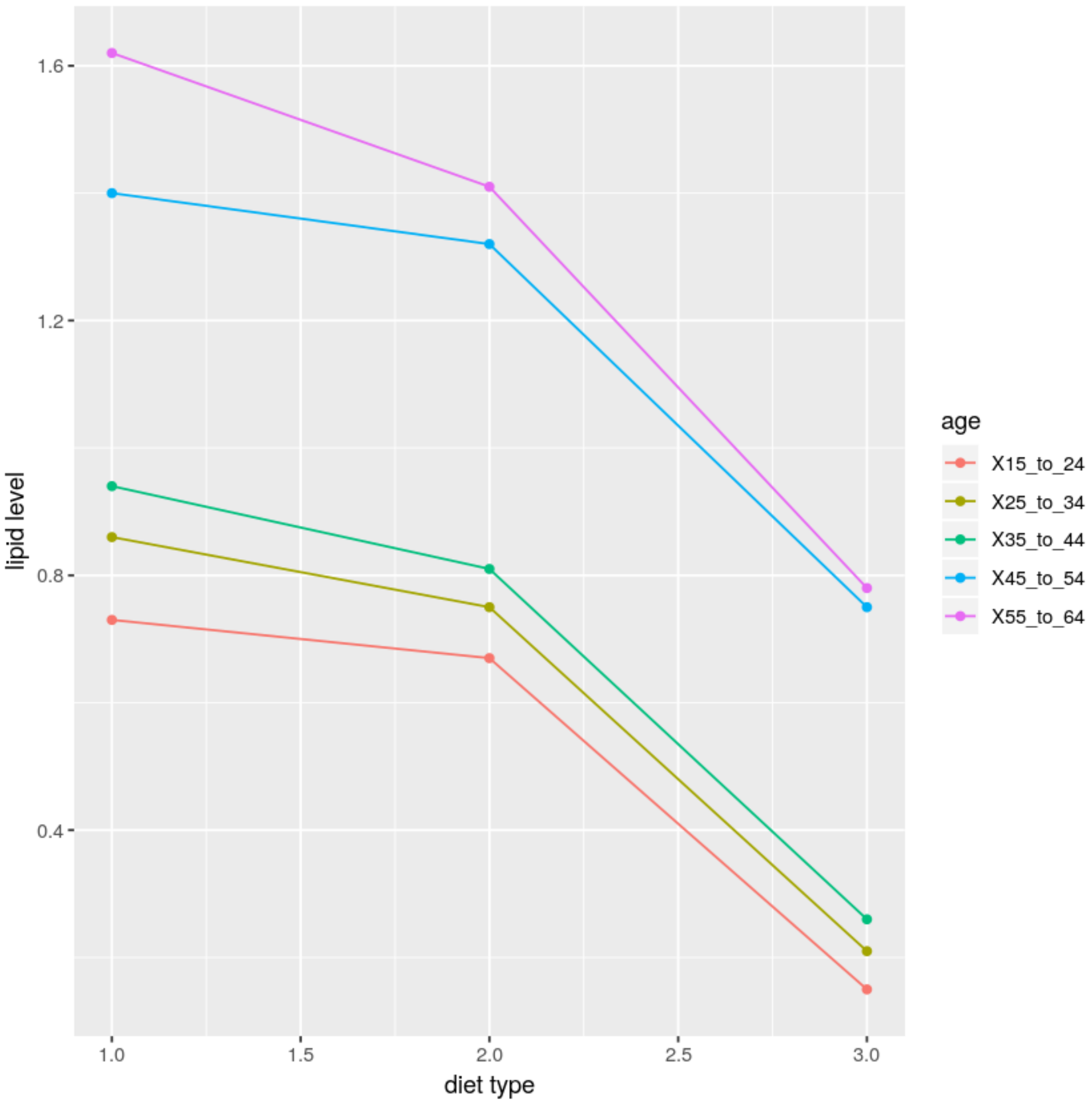
- From the Residual vs. Fitted plot we can see that for each vertical line of points representing a different treatment, the spread on the points appear to be approximately equal, indicating that these 3 treatments have the same variance. So the assumption of constant variance is not violated.

- From the QQ-plot above we could conclude that since not all the points fall on the dotted line, thus the residuals are not normal, it also appears to be heavy tailed.

(c)

(c) Plot the response Y_{ij} by blocks (Present the lipid levels for each kind of diet by block). What does this plot suggest about the appropriateness of the no-interaction assumption here?

```
In [10]: library(ggplot2)
         library(reshape2)

         dataset = "
         diet 15_to_24   25_to_34  35_to_44   45_to_54   55_to_64
         1   0.73     0.86     0.94     1.40     1.62
         2   0.67     0.75     0.81     1.32     1.41
         3   0.15     0.21     0.26     0.75     0.78"

         df = read.table(text=dataset, header = TRUE)

         melted_data = melt(df, id.vars = "diet",
                     variable.name="age", value.name="fat")

         ggplot(melted_data, aes(x = diet, y = fat, color=age)) + geom_point() + geom_line() +
                  xlab('diet type') + ylab('lipid level')
```



The no-interaction assumption is appropriate here since all five lines are approximately parallel.

Q2

2. (40 points) (By hand) Refer to the Fat in diets problem. Assume that randomized block model is appropriate.

(a)

(a) Obtain the analysis of variance table.

$$\bar{y}_{.j1} = 1.11$$

$$\bar{y}_{.j2} = 0.992$$
$$\bar{y}_{.j3} = 0.43$$

$$\bar{y}_{age1.} = 0.517$$
$$\bar{y}_{age2.} = 0.607$$
$$\bar{y}_{age3.} = 0.67$$
$$\bar{y}_{age4.} = 1.157$$
$$\bar{y}_{age5.} = 1.27$$

$$\bar{y}_{..} = 0.844$$

$$SST_{diet} = 5 \cdot (1.11 - 0.844)^2 + 5 \cdot (0.992 - 0.844)^2 + 5 \cdot (0.43 - 0.844)^2 = 1.32028$$

$$SSBlock = 3 \cdot (0.517 - 0.844)^2 + 3 \cdot (0.607 - 0.844)^2 + 3 \cdot (0.67 - 0.844)^2 + 3 \cdot (1.157 - 0.844)^2 + 3 \cdot (1.27 - 0.844)^2$$
$$= 1.41896$$

$$SSTOTAL = (0.73 - 0.844)^2 + (0.67 - 0.844)^2 + (0.15 - 0.844)^2 + (0.86 - 0.844)^2 + (0.75 - 0.844)^2$$
$$+ (0.21 - 0.844)^2 + (0.94 - 0.844)^2 + (0.81 - 0.844)^2 + (0.26 - 0.844)^2 + (1.4 - 0.844)^2 + (1.32 - 0.844)^2$$
$$+ (0.75 - 0.844)^2 + (1.62 - 0.844)^2 + (1.41 - 0.844)^2 + (0.78 - 0.844)^2 = 2.75856$$

$$SSE = SSTOTAL - SST - SSBlock = 2.75856 - 1.32028 - 1.41896 = 0.01932$$

$$MSB = \frac{SST}{3-1} = 0.660140$$

$$MSBlock = \frac{SSBlock}{5-1} = 0.354740$$

$$MSE = \frac{SST}{(3-1)(5-1)} = 0002415$$

$$F_{diet} = \frac{MSB}{MSE} = \frac{0.660140}{0.002415} = 273.3409$$

$$F_{block} = \frac{MSBlock}{MSE} = \frac{0.354740}{0.002415} = 146.8903$$

```
In [19]: # check my answer
         library(knitr)
         library(lsmeans)
         aov.fat = aov(fat ~ diet+age, data=df)
         kable(anova(aov.fat), format='markdown')
```

```
|          | Df|  Sum Sq|  Mean Sq|  F value| Pr(>F)|
|:---------|--:|-------:|--------:|--------:|------:|
|diet      |  2| 1.32028| 0.660140| 273.3499|  0e+00|
|age       |  4| 1.41896| 0.354740| 146.8903|  2e-07|
|Residuals |  8| 0.01932| 0.002415|       NA|     NA|
```

(b)

Test whether or not the mean reductions in lipid level differ for the three diets; use $\alpha = 0.05$. State the alternatives, decision rule, and conclusion. What is the P-value of the test?

Null Hypothesis:

$$H_0 : \tau_1 = \tau_2 = \tau_3 = 0$$

$$H_a : \text{ there exists at least one } \tau_i \text{ that's not equal to 0}$$

And from the calculation in question part (a), we get the p-value of diet as:

```
In [20]: p_value = 0e+00
         p_value < 0.05
```

TRUE

Since p-value < 0.05, thus we are confident enough to reject the null hypothesis that $H_0 : \tau_1 = \tau_2 = \tau_3 = 0$, indicating that the mean reductions in lipid level differ for the three diets.

(c)

(c) If there is significant difference in lipid level, how do they differ?

```
In [21]: TukeyHSD(aov(fat.lm),"diet")
```

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = fat.lm)

$diet
        diff        lwr         upr     p adj
j2-j1 -0.118 -0.2068109 -0.02918909 0.0129653
j3-j1 -0.680 -0.7688109 -0.59118909 0.0000000
j3-j2 -0.562 -0.6508109 -0.47318909 0.0000002
```

Q3

3. (45 points) (ANCOVA) A manufacturer of felt-tip markers investigated by an experiment whether a proposed new display, featuring a picture of a physician, is more effective in drugstores than the present counter display, featuring a picture of an athlete and designed to be located in the stationary area. Fifteen drugstores of similar characteristics were chosen for the study. They were assigned at random in equal numbers to one of the following treatments: (1) present counter display in stationary area, (2) new display in stationary area, (3) new display in checkout area. Sales with the present display (X it) were recorded in all 15 stores for a three week period. Then the new display was set up in the 10 stores receiving it, and sales for the next three week period (Y it) were recorded in all 15 stores. The data on sales (in dollars) follow.

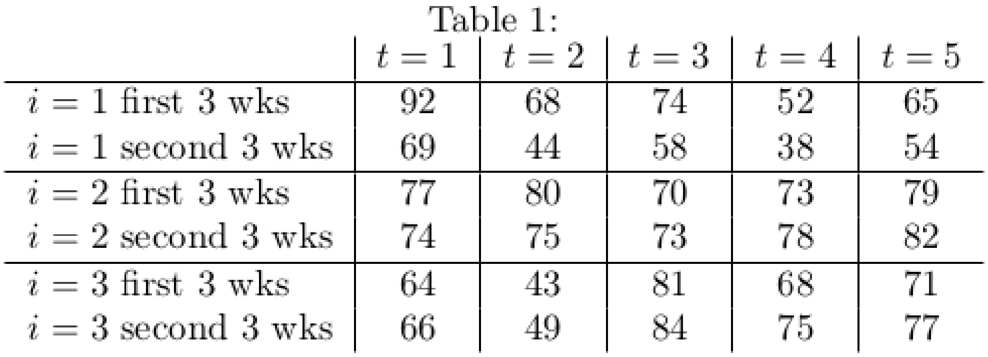


Table 1:

	$t = 1$	$t = 2$	$t = 3$	$t = 4$	$t = 5$
$i = 1$ first 3 wks	92	68	74	52	65
$i = 1$ second 3 wks	69	44	58	38	54
$i = 2$ first 3 wks	77	80	70	73	79
$i = 2$ second 3 wks	74	75	73	78	82
$i = 3$ first 3 wks	64	43	81	68	71
$i = 3$ second 3 wks	66	49	84	75	77

The analyst wishes to analyze the effects of the three different display treatments by means of covariance analysis.

(a)

(a) Obtain the residuals for covariance model $Y_{it} = \mu + \tau_i + \gamma(X_{it} - \bar{X}_{..}) + \epsilon_{it}$.

```
In [22]: x =  c(92,68,74,52,65,
              77,80,70,73,79,
              64,43,81,68,71)
         sales = c(69,44,58,38,54,
                   74,75,73,78,82,
                   66,49,84,75,77)
         time = rep(c("t1", "t2", "t3", "t4", "t5"), 3)
         display = c(rep("i1", 5), rep("i2", 5), rep("i3", 5))
         df = data.frame(x, sales, display, time)
         df
```

x	sales	display	time
92	69	i1	t1
68	44	i1	t2
74	58	i1	t3
52	38	i1	t4
65	54	i1	t5
77	74	i2	t1
80	75	i2	t2
70	73	i2	t3
73	78	i2	t4
79	82	i2	t5
64	66	i3	t1
43	49	i3	t2
81	84	i3	t3
68	75	i3	t4
71	77	i3	t5

```
In [23]: sales.lm = lm(sales ~ I(x-mean(x)) + display, data=df)
         sales.res = resid(sales.lm)
         display_1_res = sales.res[1:5]
         display_2_res = sales.res[6:10]
         display_3_res = sales.res[11:15]
         display_1_res
         display_2_res
         display_3_res
```

```
 1  -1.79728464419479

 2  -6.7635767790262

 3  2.22799625468167

 4  0.592228464419472

 5  5.74063670411986
```

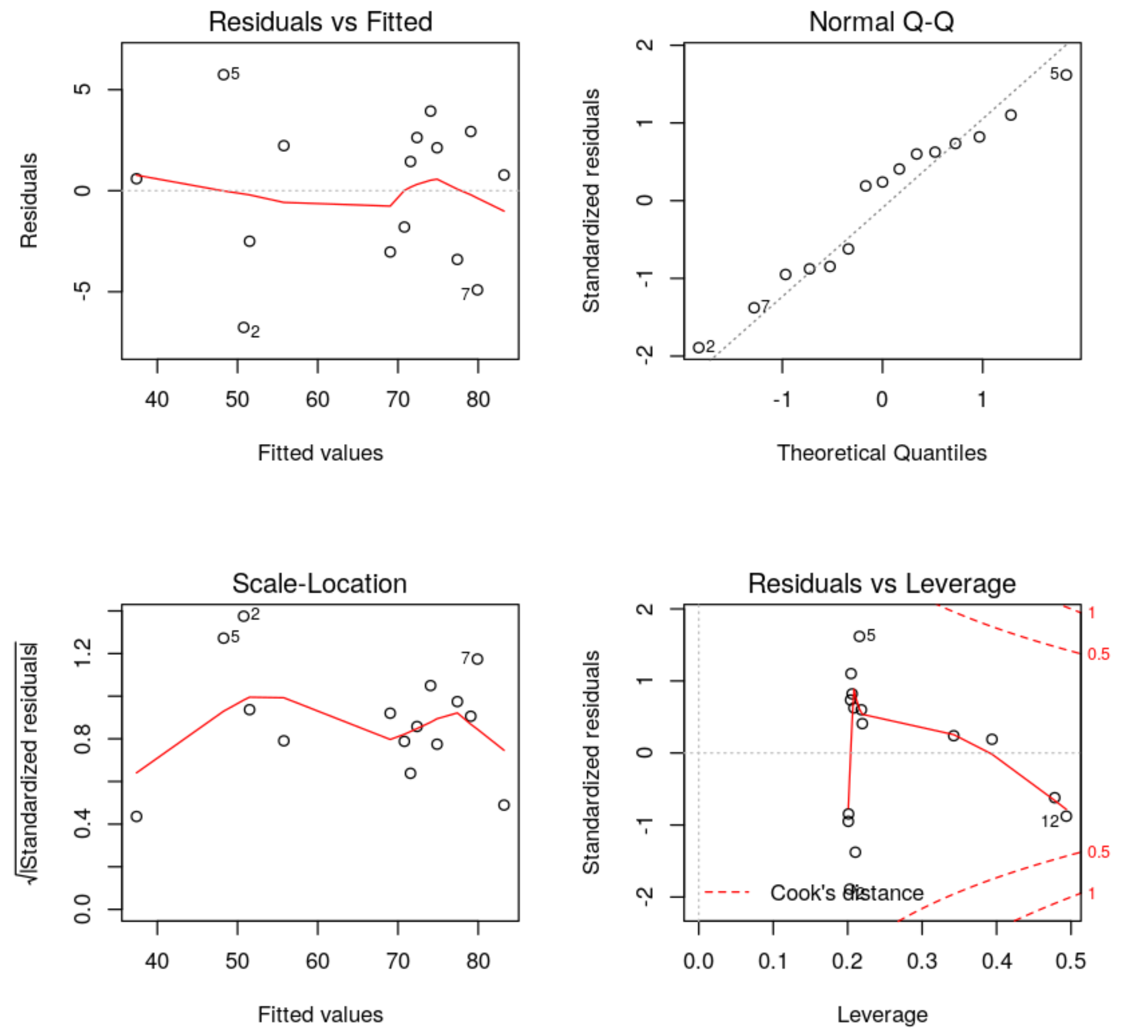
```
 6  -3.40168539325843

 7  -4.90589887640449

 8  1.44147940074906

 9  3.93726591760299

10  2.92883895131086
```

```
11  -3.0313670411985

12  -2.50187265917605

13  0.778089887640463

14  2.62968164794008

15  2.12546816479401
```

(b)

(b) For each treatment, plot the residuals against the fitted values. Also prepare a normal probability plot of the residuals.

```
In [24]: par(mfrow=c(2,2))
         plot(aov(sales.lm))
```



10

- From the Residual vs. Fitted plot we can see that for each vertical line of points representing a different treatment, the spread on the points appear to be approximately equal, indicating that these 3 treatments have the same variance. So the assumption of constant variance is not violated.
- From the QQ-plot above we could conclude that since almost all the points fall on the dotted line, thus the residuals are normal.

(c)

Assume ANOVA with equal slope models, i.e., $Y_{it} = \mu + \tau_i + \gamma(X_{it} - \bar{X}_{..}) + \epsilon_{it}$. Test for whether the slope is significant. Conduct the test using α = 0.05. State the alternatives, decision rule, and conclusion. What is the P-value of the test?

Null Hypothesis:

$$H_0 : \gamma = 0$$

Alternative Hypothesis:

$$H_a : \gamma \neq 0$$

```
In [25]: aov.sales = aov(sales ~ I(x-mean(x))+display)
         kable(anova(aov.sales), format='markdown')
```

```
|               | Df|   Sum Sq|    Mean Sq|  F value|  Pr(>F)|
|:--------------|--:|--------:|----------:|--------:|-------:|
|I(x - mean(x)) |  1| 1317.789| 1317.78912| 82.11456| 2.0e-06|
|display        |  2| 1397.281|  698.64046| 43.53394| 5.9e-06|
|Residuals      | 11|  176.530|   16.04818|       NA|      NA|
```

Since p value is smaller than $\alpha = 0.05$, thus we are confident enough to reject the null hypothesis and reach the conclusion that the slope is significant

(d)

(d) Fit the full and reduced regression models and test for treatment effects: use $\alpha = 0.05$. State the alternatives, decision rule, and conclusion. What is the P-value of the test?

Full Model:

$$Y_{it} = \mu + \tau_i I_{it} + \gamma x_{it} + \beta_i I_{it} x_{it} + \epsilon_{it}$$

Null hypothesis:

$$H_0 : \beta_1 = \beta_2 = \beta_3 = 0$$

Alternative Hypothesis:

$$Ha : \text{ not all } \beta_1, \beta_2 \text{ and } \beta_3 \text{ equal to zero}$$

```
In [26]: aov.full = aov(sales ~ I(x-mean(x))+display+I(x-mean(x))*display)
         kable(anova(aov.full), format='markdown')
```

```
|                        | Df|     Sum Sq|    Mean Sq|    F value|    Pr(>F)|
|:-----------------------|--:|----------:|----------:|----------:|---------:|
|I(x - mean(x))          |  1| 1317.78912| 1317.78912| 81.6807634| 0.0000083|
|display                 |  2| 1397.28092|  698.64046| 43.3039592| 0.0000241|
|I(x - mean(x)):display  |  2|   31.32929|   15.66464|  0.9709444| 0.4151222|
|Residuals               |  9|  145.20068|   16.13341|         NA|        NA|
```

Since p value of I(x - mean(x)):display is larger than $\alpha = 0.05$, thus we are not confident enough to reject the null hypothesis, indicating the fact that $\beta_1 = \beta_2 = \beta_3 = 0$.

Reduced Model:

$$Y_{it} = \mu + \gamma x_{it} + \epsilon_{it}$$

Null hypothesis:

$$H_0 : \gamma = 0$$

Alternative Hypothesis:

$$Ha : \gamma \neq 0$$

```
In [27]: aov.reduced = aov(sales ~ I(x-mean(x)))
         kable(anova(aov.reduced), format='markdown')
```

```
|               | Df|   Sum Sq|   Mean Sq|  F value|    Pr(>F)|
|:--------------|--:|--------:|---------:|--------:|---------:|
|I(x - mean(x)) |  1| 1317.789| 1317.7891| 10.88521| 0.0057558|
|Residuals      | 13| 1573.811|  121.0624|       NA|        NA|
```

Since p value of I(x - mean(x)) is smaller than $\alpha = 0.05$, thus we are confident enough to reject the null hypothesis, indicating the fact that $\gamma \neq 0$.