STAT523 HW3

Md Muhtasim Billah

2/5/2020

Question 1

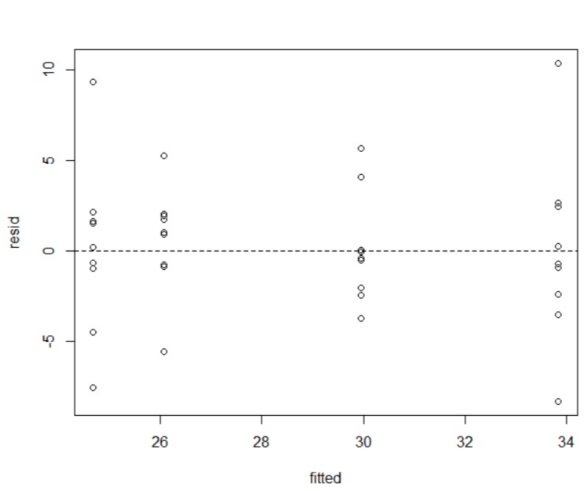
(Chapter 10, Section 10.1, Exercise 6. "Origin of Precambrian Iron Formations") Carry out the ANOVA to investigate if the average Fe percentage for any of the four iron formations is different.

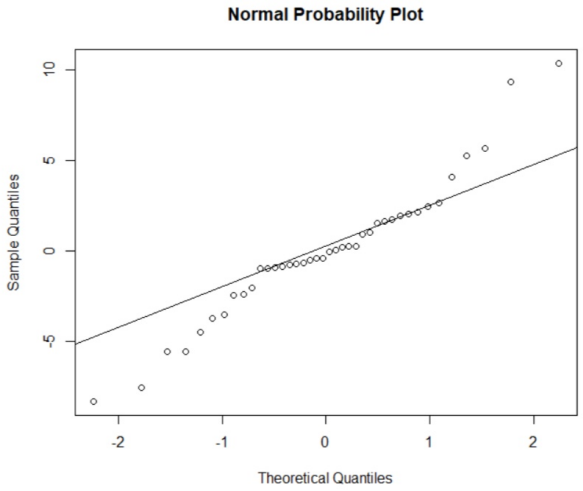
a. How many treatments are there? How many replications for each treatment?
b. State the null and alternative hypothesis for the average Fe percentage for any of the four iron formations.
c. Construct the ANOVA table using the summary statistics provided below. (You can also use R software. The dataset "iron" is on blackboard.)

$$\sum\sum {x_{ij}}^2 = 33882.24$$

Treatment	Mean
Carbonate	$\bar{x}_{1.} = 26.08$
Hematite	$\bar{x}_{2.} = 33.84$
Magnetite	$\bar{x}_{3.} = 29.95$
Silicate	$\bar{x}_{4.} = 24.69$

d. Make a decision about the test and interpret.
e. Do the following diagnostic plots suggest that the model assumptions are inappropriate? Justify each assumption separately.





Answer:

a.

There are four treatments which are the types of iron formation. The treatmets are:

i) Carbonate
ii) Hematite
iii) Magnetite
iv) Silicate

There are 10 replications for each treatment.

b.

We assume that, the population (true) mean for the Fe concentration of the four treatments are μ_1, μ_2, μ_3 and μ_4 respectively. Now, we state the null and alternative (research) hypothesis as below.

Null hypothesis, $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$.
Alternative hypothesis, $H_a : H_0$ is false.

Provided level of significane, $\alpha = 0.01$

c.

Constructing ANOVA table doing step by step calculation.

Given summary statistics,

$$\sum\sum {x_{ij}}^2 = 33882.24$$

Treatment	Mean
Carbonate	$\bar{x}_{1.} = 26.08$
Hematite	$\bar{x}_{2.} = 33.84$
Magnetite	$\bar{x}_{3.} = 29.95$
Silicate	$\bar{x}_{4.} = 24.69$

Here,
Number of treatments, $I = 4$.
Number of replicates in each treatment group, $J = 10$.
Level of significance, $\alpha = 0.01$.

Now,
Total value for each treatment can be calculated with the following formula:

$$x_{i.} = \bar{x}_{i.} * J$$

So,

Total Fe concentration for the four treatments will be as below:

$$x_{1.} = \bar{x}_{1.} * J = 26.08 * 10$$

$$x_{2.} = \bar{x}_{2.} * J = 33.84 * 10$$

$$x_{3.} = \bar{x}_{3.} * J = 29.95 * 10$$

$$x_{4.} = \bar{x}_{4.} * J = 24.69 * 10$$

Calculating the values using R, we get the total Fe concentration for the four treatments as:

```
26.08*10
```

```
## [1] 260.8
```

```
33.84*10
```

```
## [1] 338.4
```

```
29.95*10
```

```
## [1] 299.5
```

```
24.69*10
```

```
## [1] 246.9
```

$$x_{1.} = 260.8, x_{2.} = 338.4, x_{3.} = 299.5, x_{4.} = 246.9$$

So, the grand total becomes,

$$x_{..} = x_{1.} + x_{2.} + x_{3.} + x_{4.}$$

$$x_{..} = 260.8 + 338.4 + 299.5 + 246.9$$

```
260.8+338.4+299.5+246.9
```

```
## [1] 1145.6
```

$$x_{..} = 1145.6$$

So,

```
1145.6**2
```

```
## [1] 1312399
```

$$x_{..}^2 = 1312399$$

The correction factor (CF) is given by:

$$CF = \frac{x_{..}^2}{IJ}$$

$$CF = \frac{1312399}{4 * 10}$$

```
1312399/(4*10)
```

```
## [1] 32809.97
```

$$CF = 32809.97$$

Now, using formula 10D from FP, we calculate SST, SSTr and SSE.

i) SST

$$SST = \sum_{i=1}^{I}\sum_{j=1}^{J} x_{ij}^2 - CF$$
$$SST = 33882.24 - 32809.97$$

```
33882.24-32809.97
```

```
## [1] 1072.27
```

$$SST = 1072.27$$

ii) SSTr

$$SSTr = \frac{\sum_{i=1}^{I} x_{i.}^2}{J} - CF$$
$$SSTr = \frac{x_{1.}^2 + x_{2.}^2 + x_{3.}^2 + x_{4.}^2}{J} - CF$$
$$SSTr = \frac{260.8^2 + 338.4^2 + 299.5^2 + 246.9^2}{10} - 32809.97$$

```
(((260.8)^2+(338.4)^2+(299.5)^2+(246.9)^2)/10)-32809.97
```

```
## [1] 509.136
```

$$SSTr = 509.136$$

iii) SSE

$$SSE = SST - SSTr$$
$$SSE = 1072.27 - 509.136$$

```
1072.27-509.136
```

```
## [1] 563.134
```

$$SSE = 563.134$$

Now, Based on the calculated SST, SSTr and SSE values, we can form the ANOVA Table as below.

Source	DF	SS	MS	F
Treamtment	$I-1=3$	$SSTr=509.136$	$MSTr=169.712$	$F=10.85$
Error	$I(J-1)=36$	$SSE=563.134$	$MSE=15.643$	
Total	$IJ-1=39$	$SST=1072.27$		

So, the F-statistic value found from the ANOVA table is,

$$F=10.85$$

Constructing ANOVA table using R

Now, we can get the same F-statistic from a built-in package in R for one-way ANOVA test on the iron dataset posted on the blackboard.

```
mydata=read.csv("/Users/muhtasim/Desktop/iron_data.csv")
out=aov(Fe~Formation,data=mydata)
summary(out)
```

```
##             Df Sum Sq Mean Sq F value  Pr(>F)
## Formation    3  509.1  169.71   10.85 3.2e-05 ***
## Residuals   36  563.1   15.64
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the summary output, we can see that the F-static value from R matches the one from the ANOVA table.

d.

The p-value for our F-statistic is visibale from the summary which can also be found from the built in command in R as below.

```
1-pf(10.85,4-1,4*(10-1))
```

```
## [1] 3.196696e-05
```

We see, that the p-value from both method is,

$$p-value=3.2e^{-05}$$

Which is way smaller than our level of significance, $\alpha=0.01$

So, it can be said that we have enough evidence to reject the null and accept the alternative hypothesis. Which means that, among the four treatments, at least two of them have significantly different value of true mean.

e.

For ANOVA, we assume that data are normally distributed with the same variance. These assumptions need to be checked.

Assumption 1: Normality

There are several ways for checking whether the data are normally distributed or not. We can perform a Shapiro-Wilk normality test, Anderson-Darling test or we can draw a normal probability (QQ) plot. We will discuss all three options. The test are done at a level of significance, $\alpha = 0.01$

i) Shapiro-Wilk normality test:

```
Fitted=out$fitted.values
## In ANOVA, the fitted value is the sample average within each group/treatment.
## They are fitting the mu's.
Residuals=out$residuals
shapiro.test(Residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Residuals
## W = 0.95406, p-value = 0.1046
```

As the p-value$>\alpha$, we have enough evidence to retain the null. Which means that the data come from a normal distribution.

ii) Anderson-Darling normality test:

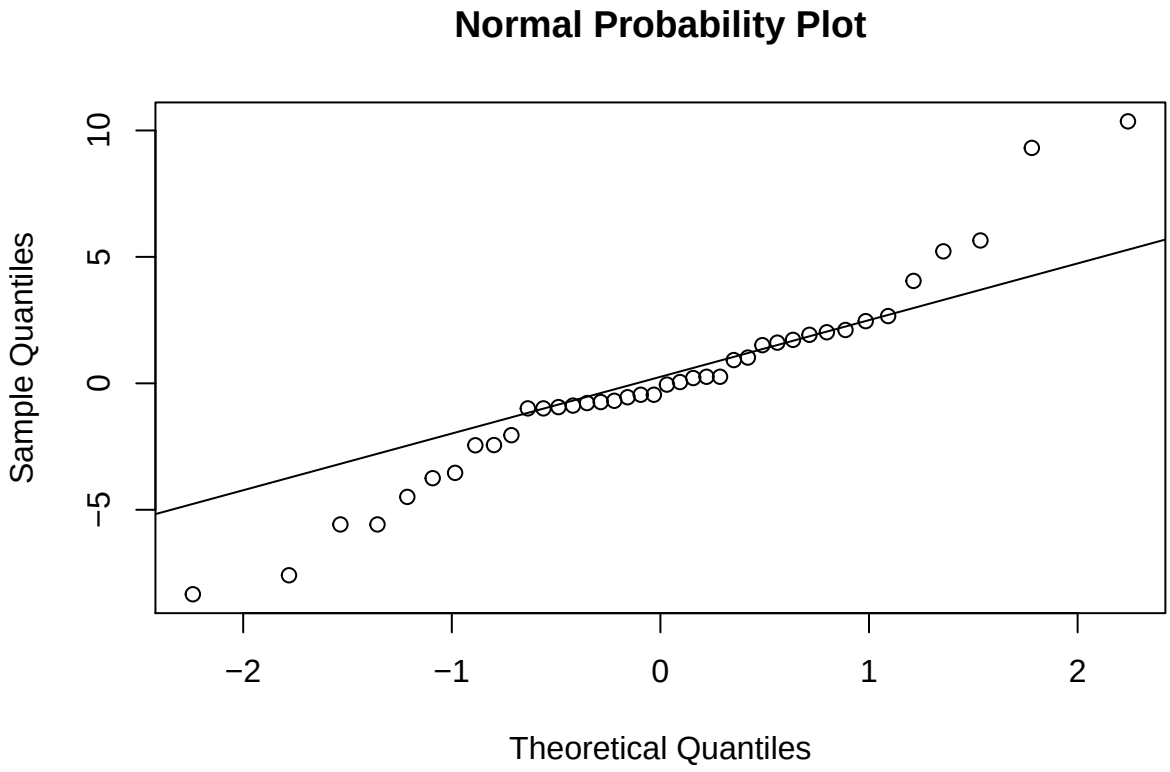
```
library(nortest)
ad.test(Residuals)
```

```
##
##  Anderson-Darling normality test
##
## data:  Residuals
## A = 0.75178, p-value = 0.04625
```

As the p-value$>\alpha$, we have enough evidence to retain the null. Which means that the data come from a normal distribution.

iii) Normal probability plot test:

```
###create the normal probability plot
qqnorm(Residuals,main="Normal Probability Plot")
qqline(Residuals)
```



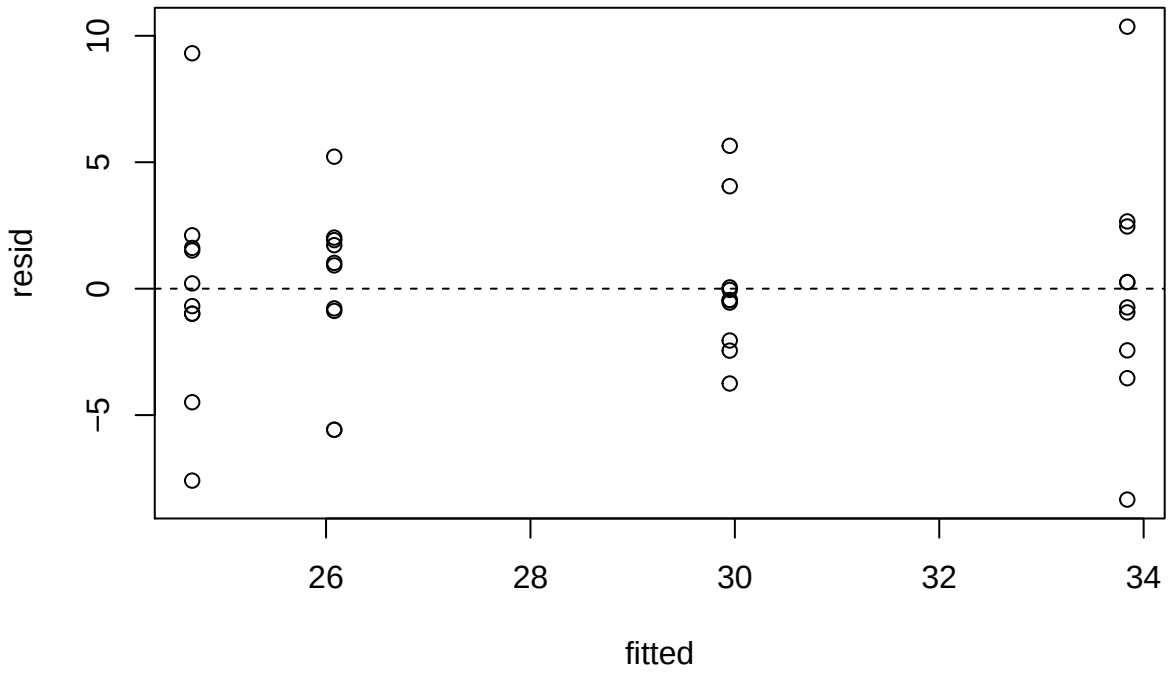
From the normal probabilty plot, we can see a linear trend among the data points which is expected for a normally distributed data. So, it indicates that the data come from a normal distribution.

Since all three methods gives the same outcome, with a level of significance, $\alpha = 0.01$ we can say that the data are normally distributed. Thus, the first assumption required for ANOVA test holds to be true.

Assumption 2: Equal Variance

One of the best ways for determining whether the data has a consistency in their variance is to analyze the residuls vs. fitted plot. The residuals can be calculated by hand for each data point in the way showed in class. But to reduce calculation time, we will use R to do that for us. The residuals vs fitted plot is plotted below.

```
###create the residual plot
plot(Fitted, Residuals,xlab="fitted",ylab="resid")
abline(h=0,lty=2)
```



From the plot, we can see the residual values for all 10 replicates for each treatment plotted along the vertical direction. The horizontal axis denotes the mean values for different treatments. From looking at the vertical spread of the variance of the replicates, it can be said that the four treatments have same variance. So, the second assumption required for ANOVA test also holds to be true.

Question 2.

An experiment was run to compare $I = 3$ types of boxes (treatment) with respect to compression strength. Each treatment is replicated 7 times.

a. Complete the ANOVA table by filling in the blanks with the correct values.

Source	Df	SS	MS	F	P-Value
Treatment			0.0668		
Error			0.0151		
Total					

b. Let μ_i be the true average compression strength for box type $i = 1, 2, 3$. Construct the Tukey's interval for each pair of the $\mu_i - \mu_j$. Use the sample mean from the table.

Treatment	Mean
1	$\bar{x}_{1.} = 21.714$
2	$\bar{x}_{2.} = 21.525$
3	$\bar{x}_{3.} = 21.750$
Total	$\bar{x}_{..} = 21.682$

c. Apply the T Method and draw the underscore plot for comparing the treatments. Briefly explain your results.

Answer:

a.

The complete ANOVA table is given below where the number of treatments, $I = 3$ and number replicates, $J = 7$.

Source	DF	SS	MS	F	p-value
Treamtment	$I - 1 = 2$	$SSTr = 0.1336$	$MSTr = 0.0668$	$F = 4.4238$	0.0273718
Error	$I(J - 1) = 18$	$SSE = 0.2718$	$MSE = 0.0151$		
Total	$IJ - 1 = 20$	$SST = 0.4054$			

```
1-pf(4.4238,3-1,3*(7-1))
```

```
## [1] 0.0273718
```

b.

For multiple comparisons in ANOVA (equal reps for each treatment), Tukey's formula for silmulataneous calculations of CI for the treaments is given by the following formula.

$$(\bar{x}_i - \bar{x}_j) \pm Q_{\alpha,I,I(J-1)}\sqrt{\frac{MSE}{J}}$$

Where, the margin of error, w is denoted as,

$$w = Q_{\alpha,I,I(J-1)}\sqrt{\frac{MSE}{J}}$$

Here,

Level of significance, $\alpha = 0.05$.
Number of treatments, $I = 3$.
Number of replicates, $J = 7$.
From ANOVA table, $MSE = 0.0151$.

Q is called the Tukey's adjustment and can be calculated in R as below.

```
qtukey(1-0.05,3,18)
```

```
## [1] 3.609304
```

Thus,

$$Q = 3.61$$

So, the margin of error becomes,

$$w = 3.61 \times \sqrt{\frac{0.0151}{7}}$$
$$w = 0.1677$$

Now, from the provided values of the sample means, we can calculate the CIs for the each pair of $\mu_i - \mu_j$.

$\mu_i - \mu_j$	$\bar{x}_{i.} - \bar{x}_{j.}$	$(\bar{x}_{i.} - \bar{x}_{j.}) \pm w$
$\mu_1 - \mu_2$	$21.714 - 21.525 = 0.189$	$(0.3567, 0.0213)$
$\mu_1 - \mu_3$	$21.714 - 21.750 = -0.036$	$(0.1317, -0.2037)$
$\mu_2 - \mu_3$	$21.525 - 21.750 = -0.225$	$(-0.0573, -0.3927)$

From the above table, we can say with 95% confidence that the treatment pair (1-3) doesn't have any significant difference in their true mean (since 0 is within their CI) but for the other two treatment pairs (1-2) and (2-3), their is significane difference (since 0 is not within their CI).

c.

For applying T-method, we first write the samples means in an increasing order as below.

Treatment	2	1	3
Mean	21.525	21.714	21.750

Now, we add w to $\bar{x}_{2.}$ and get,

$$w + \bar{x}_{2.} = 0.1677 + 21.525 = 21.6927$$

Since, 21.6927 doesn't surpass either $\bar{x}_{1.}$ or $\bar{x}_{3.}$, there is significant difference between the treatment pairs (2-1) and (2-3) which is in line with our conclusion from part b of this problem.

Again, we add w to $\bar{x}_{1.}$ and get,

$$w + \bar{x}_{1.} = 0.1677 + 21.714 = 21.8817$$

Since, 21.8817 surpasses either $\bar{x}_{3.}$, it means that there is no significant difference between the treatment pair (1-3) which is also in line with our conclusion from part b of this problem.

The final underscore plot is given below.