STAT 5810
Introduction to Data Analysis
Homework #2

Exercise 3.2

$N = 25$, $\bar{Y}_{1\cdot} = 2.16$, $\bar{Y}_{2\cdot} = 2.45$, $\bar{Y}_{3\cdot} = 2.91$, $\bar{Y}_{4\cdot} = 3.00$, $\bar{Y}_{5\cdot} = 2.71$

$$\bar{\bar{Y}}_{\cdot\cdot} = 1/5\,(2.16 + 2.45 + 2.91 + 3 + 2.71)$$

$$= 2.64$$

$$SS_{Trt} = \sum_{i=1}^{5} \left(\bar{Y}_{i\cdot} - \bar{\bar{Y}}_{\cdot\cdot}\right)^2$$

= ~~$25 \times [-0.48^2 + (0.19)^2 + (0.27)^2 + (0.36)^2 + (0.07)^2]$~~

$$= 25\left[(2.16-2.64)^2 + (2.45-2.64)^2 + (2.91-2.64)^2 + (3-2.64)^2 + (2.71-2.71)^2\right]$$

$$= 11.84$$

$SS_E = 153.4$ (Given)

$SS_T = SS_{Trt} + SS_E = 165.24$

$$F = \frac{MS_{Trt}}{MS_E} = \frac{SS_{Trt}/4}{SS_E/120} = \frac{11.84/4}{153.4/120} = 2.31$$

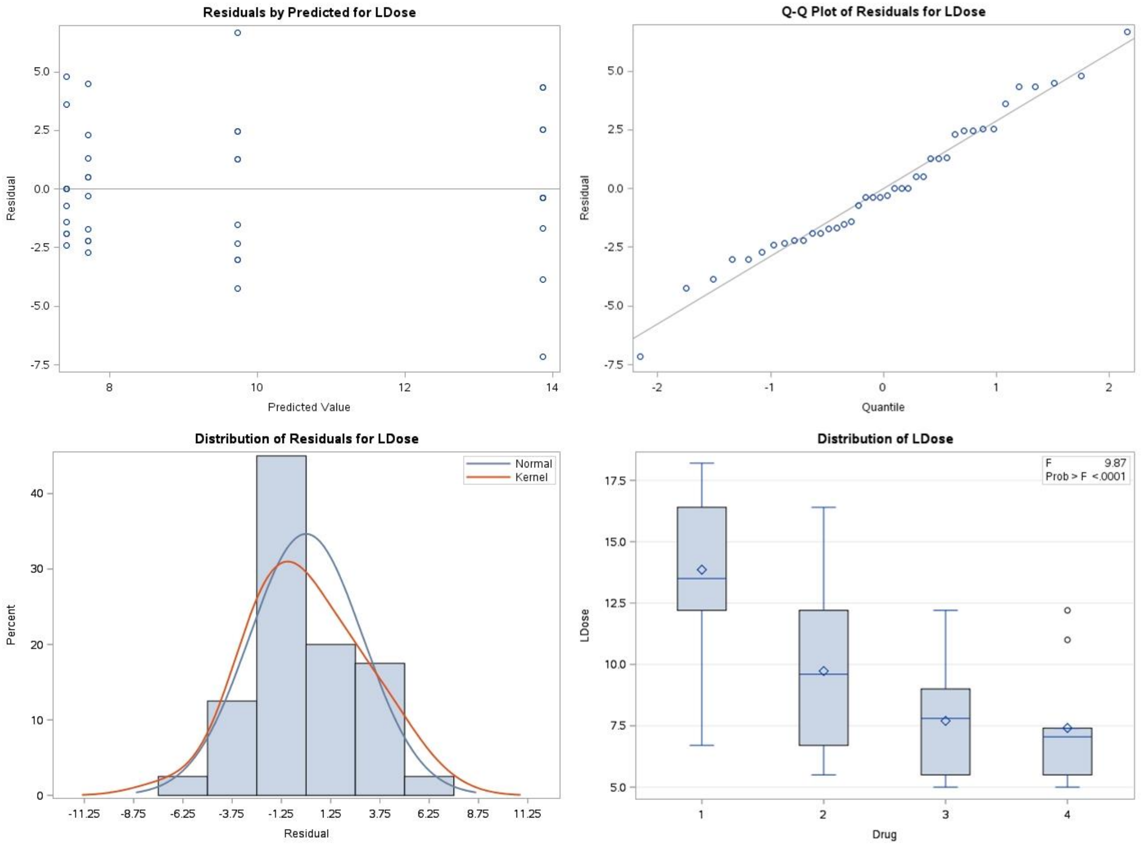
Anova Table

Source	DF	SS	MS	F
Treatment	4	11.843	2.9	2.31
Error	120	153.4	1.27	
Total	124	165.24		

Since $F_{obs} < F_{0.94,4,120}$ or the p-value = 0.061, we cant reject the null hypothesis that the group means are the same

Chapter 6, Exercise 6.3.

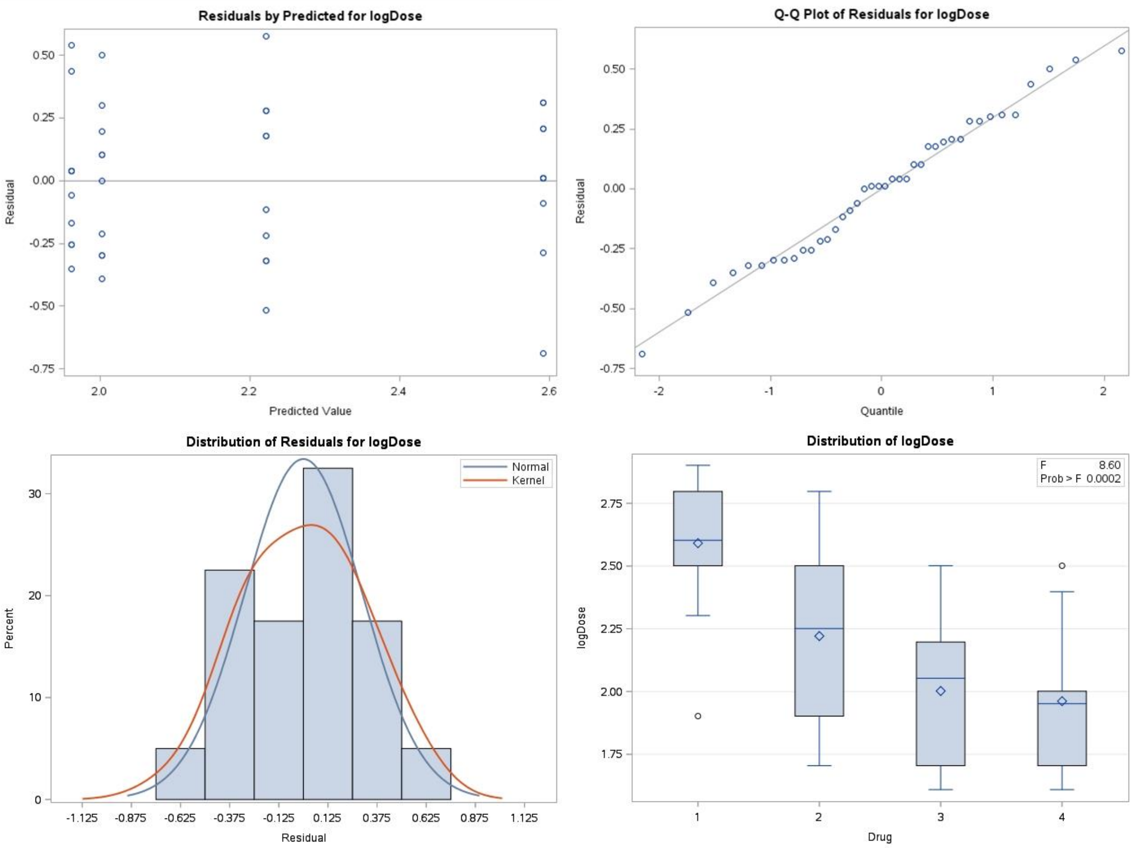
a) Below are the plots for non-transformed data:



- In the residual plot it can be observed that the data points are kind of evenly spaced indicating that the homoscedasticity is less. (Homoscedasticity is proportional to randomness of the data points in the residual plot.
- In the Q-Q plot there is slight deviation of points from the line, indicating good data.
- The histogram is skewed towards right.
- The box plots for 1 and 2 have high ranger and 4 are low range but mean towards the higher side and two outliers.

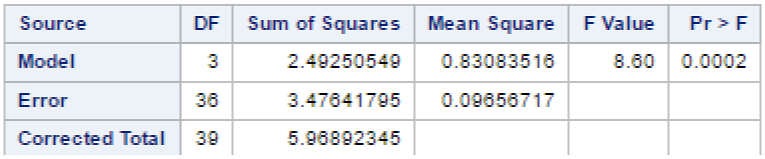
--Next Page--

- Next I tried out different transformations on the data, like square root, cube root, etc. Out of these logarithm yielded the best result. Below are the plots after applying the log transformation:



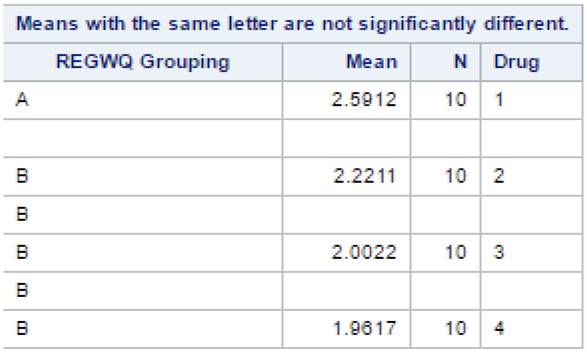
- After the log transformation the data points in the residual plot are more randomly spread out indicating high homoscedasticity.
- The Q-Q plot is very much the same.
- The Histogram is not more balanced with right skewed removed.
- There the less number of outliers in the box plot for 4.

b) On carrying out the F-test on the transformed data, the Pr (or P) value came out be is very small (<0.0002) indicating that the null hypothesis can be rejected. So we cannot explain the differences in means for different delivery services by chance.



Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	2.49250549	0.83083516	8.60	0.0002
Error	36	3.47641795	0.09656717		
Corrected Total	39	5.96892345			

c)



Means with the same letter are not significantly different.			
REGWQ Grouping	Mean	N	Drug
A	2.5912	10	1
B	2.2211	10	2
B			
B	2.0022	10	3
B			
B	1.9617	10	4

- From the above REGWQ table we can conclude that:
  - The mean of Drug 1 is highest, meaning that it is most lethal.
  - Drug 2,3 and 4 lie in the same group B, meaning that they are significantly similar to each other.
  - Drug 1 is alone in group A meaning that it is different from Drug 2,3 and 4.

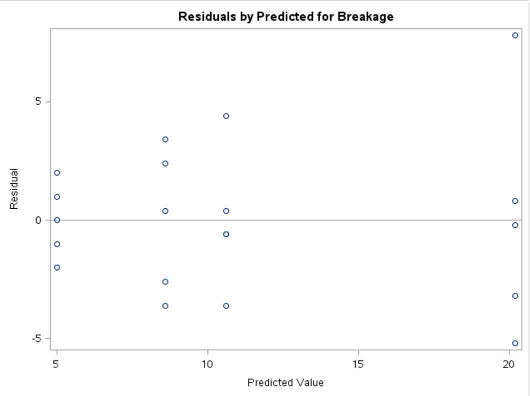
d) Summary:

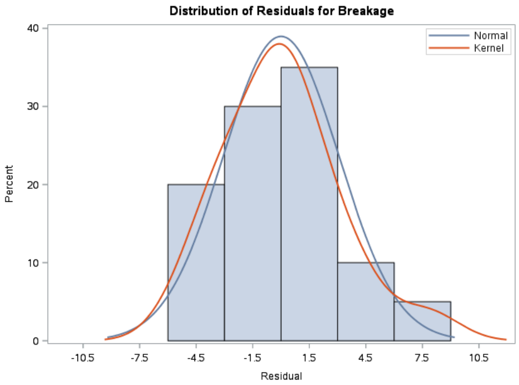
- The log transformation produced the best result of all the transformation that I tried.
- The box plots for Drug 2 and 3 are more spread out in the lower region meaning they have low to medium chance of killing. While box plot for drug 4 in very low level meaning that it is not effective in killing. Drug 1 has most killing effect since it is in the higher region.
- Drug 1 is significantly different from other 3 drugs, because they lie I different groups.
- Drug 1 is the most lethal in killing the genie pigs because of high mean lethality rate.

--Next Page--

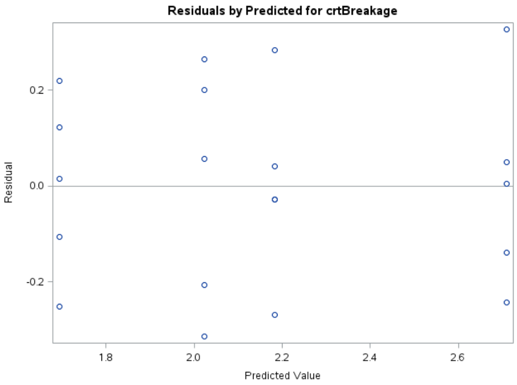
Chapter 6, Exercise 6.4.

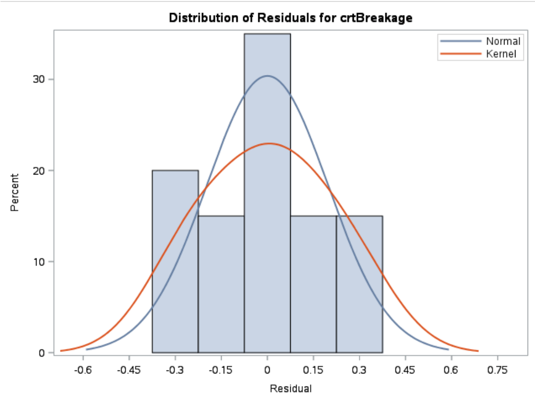
a) Before applying any transformation to the data I got the following residual plot and Q-Q plot. In the left plot we can see that the values are not randomly spread out, meaning that the data has low homoscedasticity and low approximate normality. In the right plot (histogram), we can observe that is skewed towards right. Q-Q plot looked alright, with two outliers.





To fix this I applied **cube root transformation**, I got following plots after the transformation:



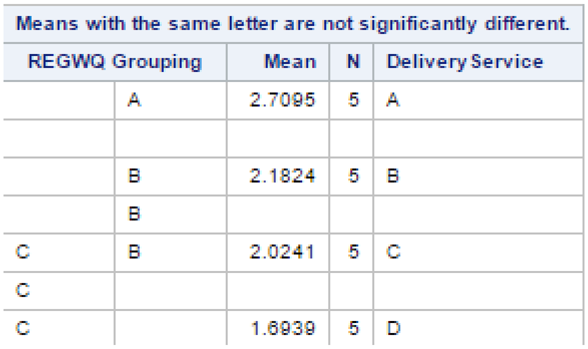


Now in the right residual plot the values are more randomly spread out indicating unbaised homoscedasticity. Even the histogram has become more symmetrical. All in all, we can say that the data has become more homoscedasticity after applying the cube root transformation.

b) On carrying out the F-test on the transformed data, the Pr (or P) value came out be is very small (<0.0001) indicating that the null hypothesis can be rejected. So we cannot explain the differences in means for different delivery services by chance.

c) As the test is significant, I have carried out the REGWQ method, with following conclusions:

- Delivery Service A has highest mean, indicating highest breakage rate.
- Since B and C are under the same grouping B, this indicates they are not significantly different in the breakage rate. Same for C and D.
- A is in a complete different group 'A', meaning that it is very different from B, C and D in breakage rate.



Means with the same letter are not significantly different.

REGWQ Grouping		Mean	N	Delivery Service
	A	2.7095	5	A
	B	2.1824	5	B
	B			
C	B	2.0241	5	C
C				
C		1.6939	5	D

d) Summary: The delivery data is slightly skewed because on residual analysis we found that it has slight homoscedasticity, which we fixed by cube root transformation. Next we did REGWQ analysis, which revealed some interesting facts about pairwise groupings in the data.

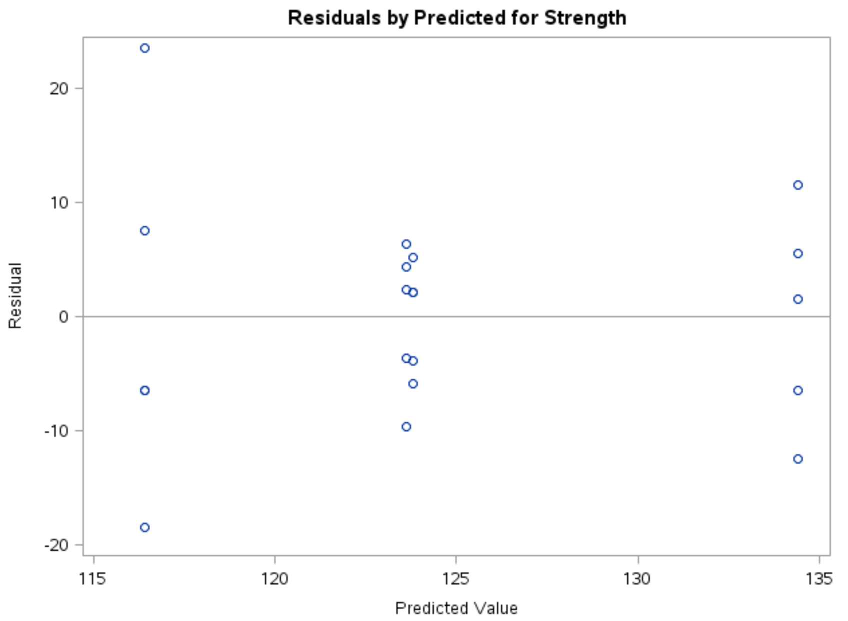
Which delivery service would I recommend? I would recommend D, because it has lesser mean than others, indicating low breakage rate.

--Next Page--

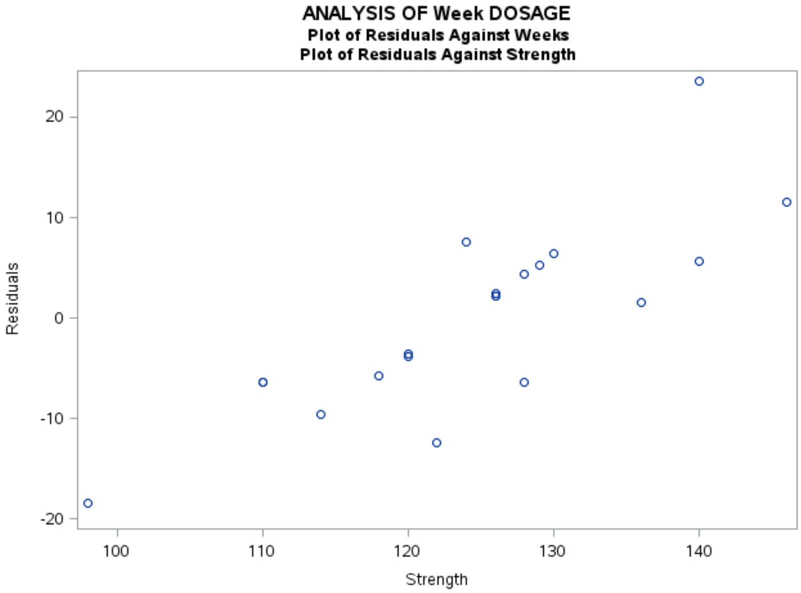
Extra Problem:

a)

Figure below shows the residual plot for the data without any transformation. We can observe that the data points are in a line and are not randomly scattered, indicating low homoscedasticity.

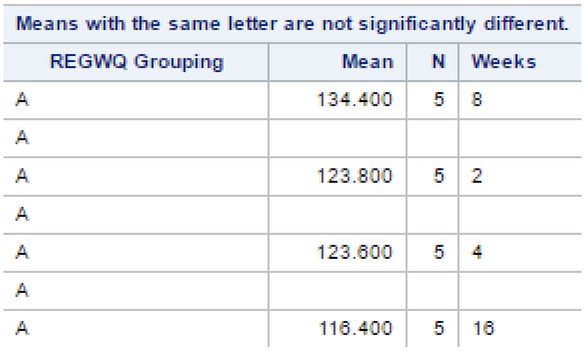


Next I have transformed the data using log transformation. The figure below shows the residual plot for the log transformed data. Here we can see that the data points are now much more randomly scattered indicating high homoscedasticity.



b) On carrying out the F-test on the non-transformed data, the Pr (or P) value came out be 0.0835 indicating that the null hypothesis can be not rejected. While the P value of transformed data came out to be 0.0002, meaning that null hypothesis can be rejected.

c)



Means with the same letter are not significantly different.

REGWQ Grouping	Mean	N	Weeks
A	134.400	5	8
A			
A	123.800	5	2
A			
A	123.600	5	4
A			
A	116.400	5	16

- From the above REGWQ table we can observe that the means of all the polyester is about the same, also all of them lie in the same group A indicating that the polyester do not degrade as the time passes.

e) Summary:

- The log transformation worked best for this dataset.
- From REGWQ we concluded that polyester does not break or loose strength as the time passes.

--End--