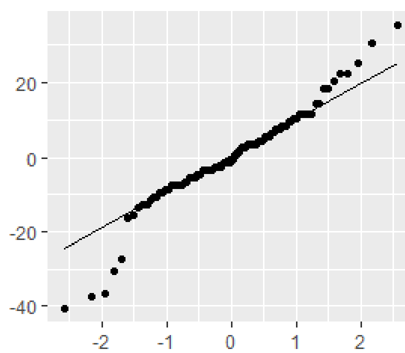
STAT 614 - HW 6
By Sihyuan Han

Instructions: Please type your solutions in a separate document and upload the document in Blackboard as a pdf. Include supporting work (plots, etc.) when appropriate, but do not copy all computer output. Select only relevant output. I will not be collecting syntax for this assignment.

1. Revisit the analysis of the data from the study of effects of exposure to lead on the psychological and neurological well-being of children from the previous HW. (Recall that the data are given in the lead.csv data set from HWs 1 and 5.)

   a. Use a **residual analysis** to assess whether the **assumptions** of the ANOVA model (on the untransformed, full data set) are met. What remedies do you *recommend and why* (Note: you will use a nonparametric method in the next part of this problem so there is no need to complete the analysis using your recommended remedy). (See my notes on the next page for addressing the missing observations on the MAXFT variable.)

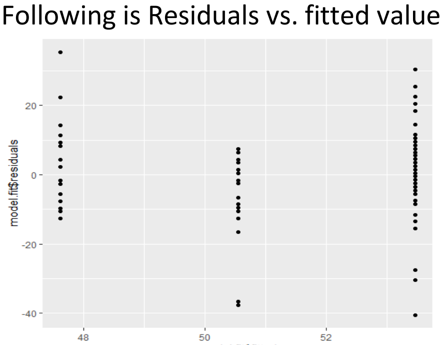
   Ans:
   Assumptions:
   1. Normal distributed (residuals)
   2. Equal variance (residuals)
   3. Influential outliers assessed (residuals)
   4. Independent observations with groups (study design)
   5. Independent samples (between groups) (study design)




```
        Shapiro-Wilk normality test

data:  model.fit$residuals
W = 0.94812, p-value = 0.0006708
```

From the QQ plot of residuals, we can see it is skewness from normality. Shapiro test also indicates that p-value is 0.0006708 which is < 0.05.
Following is Residuals vs. fitted value




```
Analysis of Variance Table

Response: MAXFT
          Df  Sum Sq Mean Sq F value   Pr(>F)
lead_typ   2  1600.1  800.04  5.2773 0.006692 **
Residuals 96 14553.8  151.60
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Based on the plot that there are some potential outliers, also the variances are different.
So we may conclude that we should use transformation methods (log, square) or nonparametric method to do this test.

b. Use **nonparametric methods** to address the research questions of interest. Restate how the hypotheses of interest from the last HW will be addressed using the nonparametric methods. Carry out the analysis (on untransformed data) and clearly state the conclusions. **Conduct two-sided pairwise comparisons using the Bonferroni method**. How would you **conduct a one-sided test**?

Ans:

Overall test

```
        Kruskal-Wallis rank sum test

data:  MAXFT by lead_typ
Kruskal-Wallis chi-squared = 10.587, df = 2, p-value = 0.005024
```

H0: mean MAXFT score is the same for all groups, group1 = group2 = group3

Ha: The distribution of at least one group is shifted away from the other groups, not all means are equal (at least one pair of means is not equal)

We can see from the Kruskal-Wallis test that the p-value is 0.005024 <0.05, so we can conclude that there is enough evidence to reject the null hypothesis, that at least one pair of means isn't equal.

```
        Pairwise comparisons using Wilcoxon rank sum test with continuity correction

data:  lead$MAXFT and lead$lead_typ

  1     2
2 0.004 -
3 0.562 0.718

P value adjustment method: bonferroni
```

RQ1- Normal blood-lead levels will have higher average MAXFT scores

H0: mean MAXFT group1 = group2

Ha: mean MAXFT group1 > group2

We can see from the result that the p-value is 0.004, one sided p-value will be 0.004/2 = 0.002 which is < 0.05, since it's Bonferroni so use 0.05 to test. There is enough evidence to reject H0, so we have 95% confident that there is evidence to conclude the mean of MAXFT score group1 is higher than group2.

RQ2- Control group (group1) is higher than previously exposed group (group3), on average

H0: mean MAXFT group1 = group3

Ha: mean MAXFT group1 > group3

The p-value is 0.562, so one sided p = 0.562/2 = 0.281 > 0.05, so we have not enough evidence to reject H0, we can say that there is no enough evidence to indicate that the control group (group1) is higher than previously exposed group (group3), on average with 95% confidence interval.

RQ3- Previously exposed populations will have "recovered" compared to a currently exposed population

H0: mean MAXFT group2 = group3

Ha: mean MAXFT group2 ≠ group3

From the test, it shows that the p-value is 0.718 which is > 0.05, so we have not enough evidence to reject null hypothesis, which we can conclude that there is no enough evidence to say previously exposed populations will have recovered compared to currently exposed population with 95% confidence interval.

2. From The Statistical Sleuth, Third Edition, Chapter 5, problem 17. Note that to get the p-value you will need to find the probability from an F-distribution. In R the function `pf(x,numdf, denomdf)` gives the probability P(X ≤ x) from an F(numdf, denomdf) distribution, where `numdf` denotes the numerator degrees of freedom (between groups df) and `denomdf` the denominator degrees of freedom (within groups df) of the F-statistic.

Ans:

```
> pf(3.5,7,24,lower.tail = FALSE)
[1] 0.009941808
```

17. Display 5.20 shows the start of an analysis of variance table. Fill in the whole table from what is given here. How many groups were there? Is there evidence that the group means are different?

$7+1=8$

Because of p-value < 0.05, so we have enough evidence to say at least one pair of groups mean is different!

DISPLAY 5.20 Incomplete ANOVA table for Exercise 17

Source	d.f.	Sum of squares	Mean square	*F*-statistic	*p*-value
Between groups	? 7	35819	? 5117	? 3.5	? 0.009941808
Within groups	24	35,088	? 1462		
Total	31	70,907			

$31-24 = 7$

$70907-35088 = 35819$

$F(7, 24)$

numerator df

denom df

$$F = \frac{\frac{70907-35088}{31-24}}{\frac{35088}{24}} = \frac{5117}{1462} = 3.5$$

Mean square

3. In comparing 6 groups a researcher notices that the sample mean for the 6th group, $\bar{y}_6$, is the largest and that the sample mean for the 3rd group, $\bar{y}_3$, is the smallest. The researcher then decides to test that $\mu_6 = \mu_3$. Is it appropriate to conduct this test? Or, can any of the multiple comparison methods be used to test this hypothesis? If so, which method? If it is not appropriate, explain why not.

Ans:

Before we start the test we should check the assumptions as following,

1. Normal distributed (residuals)
2. Equal variance (residuals)
3. Influential outliers assessed (residuals)
4. Independent observations with groups (study design)
5. Independent samples (between groups) (study design)

then we can further decide what kind of test we should use. We can use PostHocTest if assumptions are all meet, or else we can consider to use transformation method, nonparametric method or further check if they have influential outliers.