STAT 614 - HW 6

Due: Thursday, November 5, 2020 in Blackboard by 11:59pm.

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.)

    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?

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 $\leq$ 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.

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.

Note: Because there are missing observations in this data set, the number of residuals is fewer than the number of lead_typ values (which has no missing observations). So, when you try to plot the residuals vs. lead_typ you will receive an error.

I recommend doing one of two things – either (1) plotting the residuals vs. the fitted values (y-hat values) or (2) removing observations that are missing MAXFT values before you fit any models. R ignores missing values in most functions anyway (this is not always a great idea but we are not going to go into missing data methods in this class – look up multiple imputation if you are interested in that topic) so this last choice just makes formal what R defaults to anyways.

For (1), this is essentially the same as plotting the residuals vs. the exposure groups. Because the fitted values correspond to the sample means for each group, all observations in a group have the same "yhat" values. Here is some example R code to accomplish that:

```
lead <- read.csv("lead.csv",header=T)
model.fit <- aov(MAXFT~factor(lead_typ),data=lead)
qplot(x = model.fit$fitted, y = model.fit$residuals, xlab='fitted values(yhat)', ylab='residual',
    main='residuals vs. fitted values (yhat)')
```

Aside: For the multiple regression model (and beyond) we will always look at residuals vs. fitted values AND residuals vs. the predictors in the model, such as lead_typ. For ANOVA and simple linear regression, these give essentially the same plot – just reordered on the x-axis for ANOVA.

The other option in (2) is to create a data set that removes anyone with missing observations. Be careful with this! This is a big data set with 40 variables and many of them might have missing values, so we don't want to remove *every* individual with a missing value on any variable. Right now, we only want to focus on the MAXFT and lead_typ variables. I do this by creating a subset of the lead data set (I call it 'sublead') which extracts MAXFT and lead_typ and *then* using the "complete cases" by taking only those who have both MAXFT and lead_typ values. The complete.cases() function in R was built to do this. Here is my code:

```
lead <- read.csv("lead.csv",header=T)
summary(lead$MAXFT) # note the NA's are missing values!
summary(lead$lead_typ) # no missing here
sublead <- data.frame(MAXFT=lead$MAXFT,lead_typ=lead$lead_typ) # extract only the two vars of interest
summary(sublead)
sublead <- sublead[complete.cases(sublead),]  # keeps on those without missing (NA) values
summary(sublead)
model.fit <- aov(MAXFT~factor(lead_typ),data=sublead)
qplot(x = sublead$lead_typ, y = model.fit$residuals, xlab='exposure group', ylab='residual',
    main='residuals vs. exposure group')
```