Unit3 HW

Turner

12/16/2018

## HW Instructions

The weekly HW assignments are designed to accomplish 2 goals for the MSDS student. The first is to provide a series of conceptual and analytical questions so the student can get a feel for their current understanding of the unit. The second goal is to introduce the students to standard functions and routines in R that effectively do the same things that the “Procs” do in SAS.

R and SAS are both wonderful tools and as we go through the assignments, students will begin to recognize very quickly that they both have pros and cons.

The formatting of the HW is as follows:  
1. A series of high level questions will be asked with either short answers or simple multiple choice responses.  
2. Analytical questions will be provided but a short vignette example of how R functions work for a given topic or method will be given. The student will then be asked a follow up question or two based on the output provided.  
3. Thirdly, a new data set will be given to allow the student to gain some experience with a new data set from start to finish.

Solutions to the HW will be provided a day or two after the HW is submitted. It is up to the student to “shore up” any confusion or missunderstanding of a topic. Grading will be based on a combination of correctness, completion, and overall conciseness.

The student may provide there answers in a seperate word document. Just make sure that it is easy to follow and that all questions have been addressed for the grader. You are welcome to use R markdown, but it is not required.

## ANOVA Conceptual questions

1. State the necessary assumptions for Two Way ANOVA analysis to be conducted. Note: That additive versus non additive is not a component of the assumptions

The dependent variable should be approximately normally distributed for each combination of the groups of the two independent variables.

There should be homogeneity of variance for each combination of groups of the 2 independent variables.

The independent variable should be continuous.

The independent variable should contain 2 or categorical groups that are independent on their own.

Each observation should be independent.

No significant outliers.

1. State in words what it means for there to be an “interaction” between two explanatory variables. Note: Do not explain the meaning in terms of a graph with parallel lines.

Interaction between 2 explanatory variables tells us how moderated or modified one of the 2 variables has on the other. In other words, if there is interaction between 2 explanatory variable, then one or each of the variables have affected the effects each or one of them have on the dependent variable.

1. What is the family wise error rate? What is multiple testing and why is it an issue when conducting ANOVA type models such as Two Way ANOVA?

Family-wise error rate is the probability of making one or more type I errors in performing multiple hypothesis test. ANOVA cannot provide detailed information on differences among the various study groups; it can only provide evidence that there are differences or there aren’t any differences.

1. True or False? The overall Type-III sums of squares F-test’s allow the analyst to determine where specific differences lie between levels of the factor.

False. In Type-III, main effects are tested in light of interaction terms as well as in light of other main effects.

## Exercise #1 ACT Scores Revisited

The first step in any analysis is appropriately describing the data both numerically and visually. For a Two Way ANOVA analysis, one of the most helpful visual tools is the mean profile plot (with or without the raw data). The following code reads in the ACT data set from our pre live discussion and provides a handy, modifiable, function that can make a quick summary statistics table really quick.

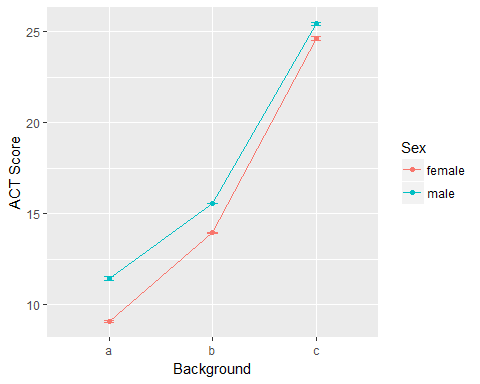
setwd("D:/MSDS6372/HWMark")  
  
ACT<-read.csv("MathACT.csv")  
  
  
#Attaching the data set, creating a function, and creating a summary stats table. Note: In line 44 below, you can add other statistics like median, IQR,etc.  
  
attach(ACT)  
mysummary<-function(x){  
 result<-c(length(x),mean(x),sd(x),sd(x)/length(x))  
 names(result)<-c("N","Mean","SD","SE")  
 return(result)  
}  
sumstats<-aggregate(Score~Background\*Sex,data=ACT,mysummary)  
sumstats<-cbind(sumstats[,1:2],sumstats[,-(1:2)])  
sumstats

## Background Sex N Mean SD SE  
## 1 a female 82 9.073171 4.186340 0.05105293  
## 2 b female 387 13.963824 5.000905 0.01292224  
## 3 c female 54 24.629630 4.849806 0.08981122  
## 4 a male 48 11.458333 5.086312 0.10596483  
## 5 b male 223 15.565022 4.888305 0.02192065  
## 6 c male 67 25.432836 5.554752 0.08290675

With the three levels of background and two levels of sex status, the table provides the sample size, mean, standard deviation, and the means standard error for each of the 6 combinations of the two factors combined. This can be used to take a quick look at the data to see if things are making sense. Adding additional summaries like the max, min, and quartiles would be helpful as well.

The above table may not be too aesthetically pleasing. Luckily under the current format of the table, its quite easy to generate a means profile plot to visualize the data. This graphic was most likely a major point of discussion during live session.

library(ggplot2)  
ggplot(sumstats,aes(x=Background,y=Mean,group=Sex,colour=Sex))+  
 ylab("ACT Score")+  
 geom\_line()+  
 geom\_point()+  
 geom\_errorbar(aes(ymin=Mean-SE,ymax=Mean+SE),width=.1)



**HOMEWORK QUESTION**

1. Modify the previous R script so that the summary table also includes the min, the max, and IQR. These functions are all self explanatory…min(x), max(x), IQR(x).

mysummary<-function(x){

result<-c(length(x),mean(x),sd(x),sd(x)/length(x),min(x),max(x),IQR(x))

names(result)<-c("N","Mean","SD","SE",'Min','Max','IQR')

return(result)}

sumstats<-aggregate(Score~Background\*Sex,data=ACT,mysummary)

sumstats<-cbind(sumstats[,1:2],sumstats[,-(1:2)])

sumstats

1. Create another means plot but rather than using the standard errors (SE) to make the error bars. Make it with the raw standard deviations (SD). Which graphic (compared to plot using SE) is more telling about the assumption of equal variances for the ANOVA model? Give a little explanation for your answer.

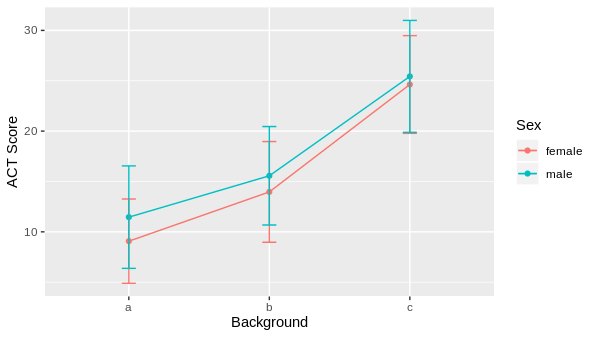
ggplot(sumstats,aes(x=Background,y=Mean,group=Sex,colour=Sex))+

ylab("ACT Score")+

geom\_line()+

geom\_point()+

geom\_errorbar(aes(ymin=Mean-SD,ymax=Mean+SD),width=.1)

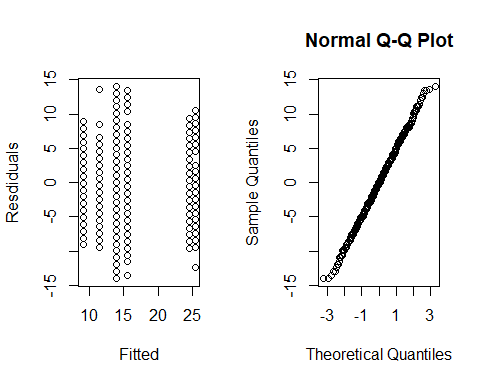
Since the SD is larger than the SE, the range of y-min and y-max of each background is larger than the previous mean plot. Even though that SD gives a wider range, it does not give enough interpretation on error. Using the SD gives a better idea of equal variance because SD is directly related to variance and can access constant variance. It also gives us a sense of the population variability.

## Exercise #2 Conducting a Two Way ANOVA Analysis in R

Since Two Way ANOVA’s are technically just special cases of multiple linear regression, it’s not too surprising that the same function call is used to build the model. After viewing and exploring the data via Exercise 1. The next step would be to fit a full nonadditive model, check the assumptions of the model, and then examine the type III sums of squares F tables.

The following code fits the nonadditive two way anova model and then produces the first the main residual diagnostics for assumption checking. The syntax for including interaction terms is slightly different so please make note.

model.fit<-aov(Score~Background+Sex+Background:Sex,data=ACT)  
par(mfrow=c(1,2))  
plot(model.fit$fitted.values,model.fit$residuals,ylab="Resdiduals",xlab="Fitted")  
qqnorm(model.fit$residuals)

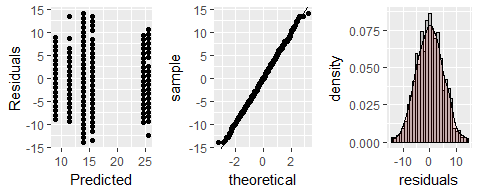


The previous graphics are not very pretty. We can use the ggplot2 package to jazz things up a bit.

library(gridExtra)

## Warning: package 'gridExtra' was built under R version 3.5.1

myfits<-data.frame(fitted.values=model.fit$fitted.values,residuals=model.fit$residuals)  
  
#Residual vs Fitted  
plot1<-ggplot(myfits,aes(x=fitted.values,y=residuals))+ylab("Residuals")+  
 xlab("Predicted")+geom\_point()  
  
#QQ plot of residuals #Note the diagonal abline is only good for qqplots of normal data.  
plot2<-ggplot(myfits,aes(sample=residuals))+  
 stat\_qq()+geom\_abline(intercept=mean(myfits$residuals), slope = sd(myfits$residuals))  
  
#Histogram of residuals  
plot3<-ggplot(myfits, aes(x=residuals)) +   
 geom\_histogram(aes(y=..density..),binwidth=1,color="black", fill="gray")+  
 geom\_density(alpha=.1, fill="red")  
  
grid.arrange(plot1, plot2,plot3, ncol=3)



As discussed in class, the residual diagnostics do not provide any concern about the assumptions of a two way anova analysis. If there were, we would have to address those concerns via a transformation of the response or multiple analysis with and without outliers, etc. Examining the type-III sums of squares F table we have:

library(car)

## Warning: package 'car' was built under R version 3.5.1

## Loading required package: carData

## Warning: package 'carData' was built under R version 3.5.1

Anova(model.fit,type=3)

## Anova Table (Type III tests)  
##   
## Response: Score  
## Sum Sq Df F value Pr(>F)   
## (Intercept) 6750.4 1 276.4610 < 2.2e-16 \*\*\*  
## Background 8045.8 2 164.7564 < 2.2e-16 \*\*\*  
## Sex 172.2 1 7.0542 0.008055 \*\*   
## Background:Sex 37.6 2 0.7709 0.462898   
## Residuals 20876.8 855   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

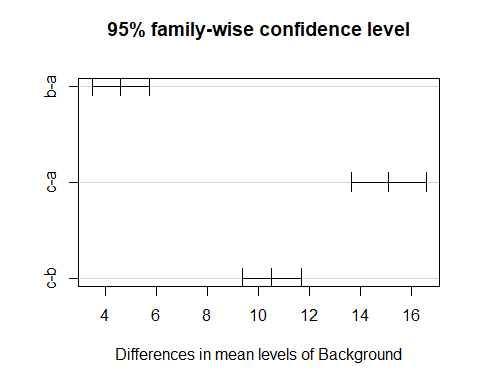
Writing contrasts are a little more cumbersome in R. To help you guys out and alleviate the need to keep track of all of the zero’s and one’s, I’ve wrote a little script that allows you to just specify the contrast that you want in a slightly simpler way. But first lets use some tools that provides a blanket lists of comparisons. Since there is no significant interaction, we just need to examine each factor one at a time. To examine all pairwise comparisons for say “background”, the following script provides the t-test results adjusted for multiple tests using Tukey’s procedure.

TukeyHSD(model.fit,"Background",conf.level=.95)

## Tukey multiple comparisons of means  
## 95% family-wise confidence level  
##   
## Fit: aov(formula = Score ~ Background + Sex + Background:Sex, data = ACT)  
##   
## $Background  
## diff lwr upr p adj  
## b-a 4.595334 3.474636 5.716033 0  
## c-a 15.120534 13.655047 16.586021 0  
## c-b 10.525200 9.370655 11.679744 0

The table is helpful for quickly examining the results and getting the p-values and estimates. Its always helpful to visualize.

plot(TukeyHSD(model.fit,"Background",conf.level=.95))

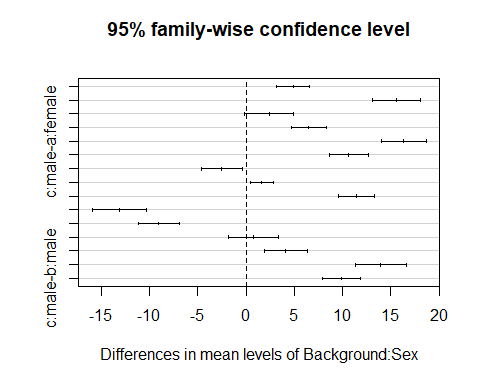


If an interaction is present, you can rinse and repeat the code just using the interaction term instead. This code below is for illustration, it makes no sense to do this on the ACT data set since the interaction F test is not significant.

TukeyHSD(model.fit,"Background:Sex",conf.level=.95)

## Tukey multiple comparisons of means  
## 95% family-wise confidence level  
##   
## Fit: aov(formula = Score ~ Background + Sex + Background:Sex, data = ACT)  
##   
## $`Background:Sex`  
## diff lwr upr p adj  
## b:female-a:female 4.8906536 3.1748785 6.606429 0.0000000  
## c:female-a:female 15.5564589 13.0830154 18.029902 0.0000000  
## a:male-a:female 2.3851626 -0.1797967 4.950122 0.0854058  
## b:male-a:female 6.4918517 4.6691025 8.314601 0.0000000  
## c:male-a:female 16.3596651 14.0354027 18.683927 0.0000000  
## c:female-b:female 10.6658053 8.6155720 12.716039 0.0000000  
## a:male-b:female -2.5054910 -4.6652479 -0.345734 0.0123022  
## b:male-b:female 1.6011981 0.4146282 2.787768 0.0017335  
## c:male-b:female 11.4690115 9.6014652 13.336558 0.0000000  
## a:male-c:female -13.1712963 -15.9710443 -10.371548 0.0000000  
## b:male-c:female -9.0646072 -11.2051648 -6.924050 0.0000000  
## c:male-c:female 0.8032062 -1.7778339 3.384246 0.9492008  
## b:male-a:male 4.1066891 1.8610087 6.352369 0.0000033  
## c:male-a:male 13.9745025 11.3056335 16.643371 0.0000000  
## c:male-b:male 9.8678134 7.9015327 11.834094 0.0000000

plot(TukeyHSD(model.fit,"Background:Sex",conf.level=.95))



As discussed in class, including all possible combinations of comparisons may be too much and of little interest to the actual study at hand. We can manually create the comparisons of interest and manual adjust the p-values through writing contrasts. To help streamline this for you guys, I’ve included a little R script that makes the process a little more automated for you.

The following script allow you to write out your contrasts in a more verbal syntax. I’ll run you through the most tedious scenario. The script can be easily modified to handle simpler situations. First things first, all you need to do is provide some details as to what comparisons you’d like to make. Suppose, that if the interaction was significant, the only meaningful comparisons to make in the analysis comparing males versus females for each level of background.

library(lsmeans) #maybe need eemeans package

## The 'lsmeans' package is being deprecated.  
## Users are encouraged to switch to 'emmeans'.  
## See help('transition') for more information, including how  
## to convert 'lsmeans' objects and scripts to work with 'emmeans'.

contrast.factor<-~Background\*Sex  
mycontrast<-c("amale-afemale","bmale-bfemale","cmale-cfemale")  
dat<-ACT

The above piece of code provides no output, but formats things for the following code to run. The key player here is the “contrast.factor” and the “mycontrast” objects. The contrast.factor piece is just specifying what types of comparisons you would like to make. For example, if we only wanted to compare the background levels we would have just specified “~Background”. The “mycontrast” object is where you get to specify what comparisons you would like to make. For a single factor, you just simply write out the factor levels you want to compare with a subtraction between them. For an interaction type comparison the syntax depends on what was used in the contrast.factor object. In our example, background is listed first, so when making comparisons the levels of background are concatenated to the levels of Sex before subtracting which combinations you want to compare.

The following code is something I wrote that takes the information you specified above and creates a clean table of results with bonferroni adjusted p-values. This script can be reused over and over, just changing the initial starting script is all that is required.

#Running a loop that determines the appropriate 0's and 1's for each   
#contrast specified above.  
library(limma)  
final.result<-c()  
for( j in 1:length(mycontrast)){  
contrast.factor.names<-gsub(" ", "", unlist(strsplit(as.character(contrast.factor),split = "\*", fixed = T))[-1])  
contrast.factor.2 <- vector("list", length(contrast.factor.names))  
for (i in 1:length(contrast.factor.names)) {  
 contrast.factor.2[[i]] <- levels(dat[, contrast.factor.names[i]])  
}  
new.factor.levels <- do.call(paste, c(do.call(expand.grid,   
 contrast.factor.2), sep = ""))  
temp.cont<-mycontrast[j]  
contrast2 <- list(comparison = as.vector(do.call(makeContrasts,   
 list(contrasts = temp.cont, levels = new.factor.levels))))  
  
contrast.result <- summary(contrast(lsmeans(model.fit,   
 contrast.factor), contrast2, by = NULL))  
  
final.result<-rbind(final.result,contrast.result)  
}  
#Cleaning up and applying bonferroni correction to the number  
#of total comparisons investigated.  
final.result$contrast<-mycontrast  
final.result$bonf<-length(mycontrast)\*final.result$p.value  
final.result$bonf[final.result$bonf>1]<-1  
  
final.result

## contrast estimate SE df t.ratio p.value bonf  
## amale-afemale 2.3851626 0.8980349 855 2.656 0.0081 0.0241653060  
## bmale-bfemale 1.6011981 0.4154379 855 3.854 0.0001 0.0003743458  
## cmale-cfemale 0.8032062 0.9036651 855 0.889 0.3743 1.0000000000

**HOMEWORK QUESTION**

1. Consider comparing the mean ACT scores of males versus females specifically for background A. Compare the outputs from the Tukey comparison result table to that of the output generated from my manual contrast maker. Is the estimated differences the same? Can you explain why are the adjusted p-values different for the two result tables? One would suggest that we reject the null while the other would have us to fail to reject. (This is just a conceptual thinking question. The interaction term is not significant for this data analysis.)

When comparing the estimated differences to the previous difference, the three contrast were exactly equal. For the Tuckey results, Tuckey was used and for the manual contrast maker Bonferroni was used. Bonferroni is more conservative than Tuckey so you are more likely fail to reject than reject for Bonferroni.

## Exercise #3

Lets examine the dta Exercise 13.17 from the statistical sleuth book. The data set is easily accessible in R via the following package.

library(Sleuth3)

## Warning: package 'Sleuth3' was built under R version 3.5.1

head(ex1317)

## Iridium Strata DepthCat  
## 1 75 Limestone 1  
## 2 200 Limestone 1  
## 3 120 Limestone 2  
## 4 310 Limestone 2  
## 5 290 Limestone 3  
## 6 450 Limestone 3

1. Provide a means plot of the data. Use this along with any additional information to comment on whether an addative or nonadditive model is probably the most appropriated. If it is not obvious that is okay just do your best.

library(Sleuth3)

head(ex1317)

mysummary<-function(x){

result<-c(length(x),mean(x),sd(x),sd(x)/length(x),min(x),max(x),IQR(x))

names(result)<-c("N","Mean","SD","SE",'Min','Max','IQR')

return(result)

}

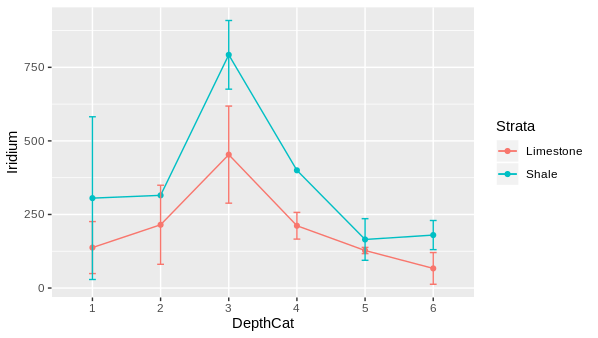
sumstats<aggregate(Iridium~DepthCat\*Strata,data=ex1317,mysummary)

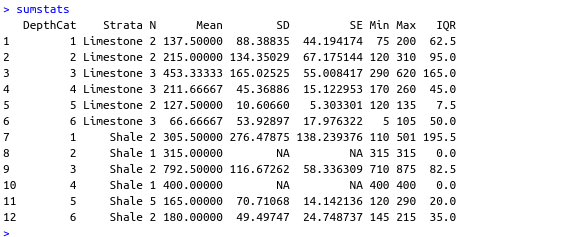
sumstats<-cbind(sumstats[,1:2],sumstats[,-(1:2)])

sumstats

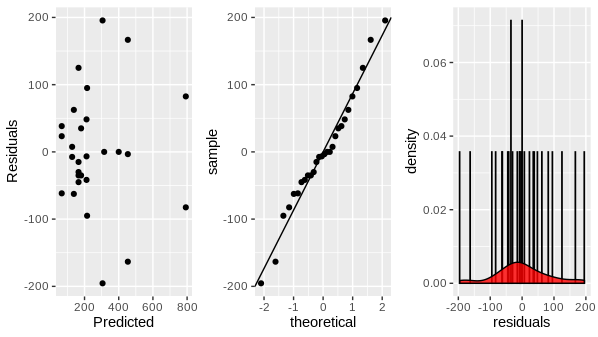
ggplot(sumstats,aes(x=DepthCat,y=Mean,group=Strata,colour=Strata))+ ylab("Iridium")+ geom\_line()+ geom\_point()+

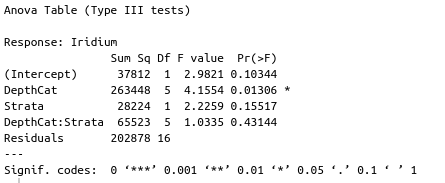
geom\_errorbar(aes(ymin=Mean-SD,ymax=Mean+SD),width=.1)



As you can see above, there are uneven of variance from the plot. additive model is more appropriate for this problem. SD variability is large.

1. Fit a nonadditive 2 way anova model to the data set and provide the residual diagnostics. Comment on the appropriateness of the current anova fit.





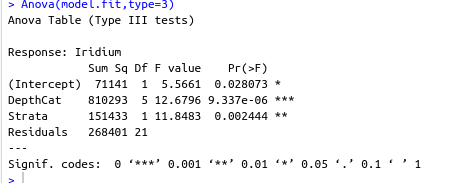
The plots and anova model above shows us that the model isn’t good. A transformation might be needed. The intercept also has a bad p-value.

1. Provide the type 3 ANOVA F-tests. Answer the following question using the table. Do the potential changes in mean Iridium by strata depend on the depth?

No it doesn’t because the F-value for the interaction is very low. The P-Value is more than 0.05 for the interaction.

1. Using multiple testing techniques, determine what factors (or combinations) contribute to changes in mean iridium.

Using additive model indicates that the DepthCat and the Strata both contribute independently in change in mean iridium. Additive model is a better model for this problem without transformation as we can see below.



Using nonadditive model, you can see the combinations of factors that contributes to the changes in Iridium is significantly low.

Checking each combination effects using nonadditive model might not be the best way to check factors that affects changes in Iridium. Below is the print out of the contrast between levels using nonadditive. The combinations with p-value less than 0.05 are combinations that contribute to change in iridium.

