Final Exam - Timeslot 2

Xiangyu Ren

This is a take-home exam. You are allowed to use any non-human sources (internet, books, notes, etc), but you are NOT allowed to receive help from or work with any other person. If the instructor feels that cheating may have happened, an oral component will be added to the exam, with students each individually explaining their work. Any cheating will be addressed in accordance with Penn State’s Academic Integrity policies.

For each experiment below, conduct a full analysis of the experimental data, and answer any additional questions as stated in each problem. You must present your answers in a clear manner. That is, only show R code for your final model selection (that is you do not need to show all transformations you try). The answers to the question should be cleanly typed up and easy to read. Your answer should contain all R code used, and you should describe the results of all important hypothesis tests you conduct. You must provide your raw .Rmd code otherwise your exam will not be graded! You must submit the output as a HTML, PDF (preferred) or Word Document file.

## Question 1 (10 Points)

Explain whether or not the following experiment was conducted using completely randomized design.

A researcher wants to determine if greenhouse humidity can affect the height of tomato plants within the greenhouse. With the funding available, she was able to procure 4 different greeenhouses. Of furhter interest, the researcher would like to know if differeneces in soil pH affects the height of a tomato plant. The researcher randomly chooses two of the greenhouses and sets them at a ‘tropical’ humidity level and the remaining two greenhouses are set to a ‘tundra’ humidity level. Each greenhouse houses 20 planters. Within each greenhouse, the scientist selects 10 planters and uses ‘high-acidity’ soil and the remaining 10 planters are given ‘low-acidity soil’. She measures the height of the plants 1 month after planting.

The experiment displayed is not conducted using completely randomized design, because the scientist was picking 4 different greenhouses instead of randomly select from a stack of greenhouses. Also, in the description, it doesn’t say whether the scientist selects 10 planters from 20 planters randomly or not.

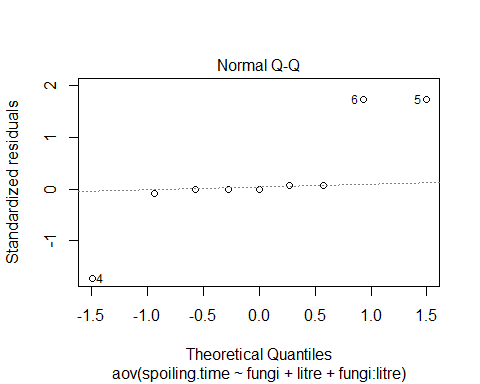
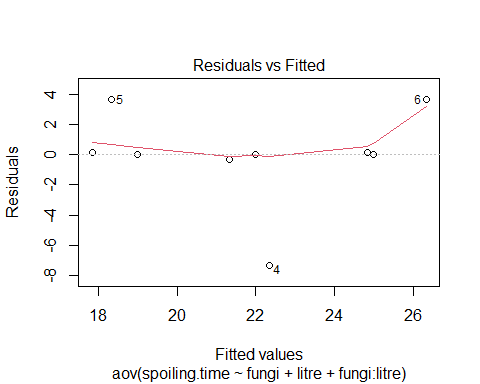
## Question 2 (16 Points)

An experiment is conducted to test whether different fungi will affect the amount of time it takes for milk to spoil. There were 3 litres of milk with fungi A, 3 litres of mlik with fungi B, and 3 litres of mlik with fungi C. 500 mL of milk from each litre was scooped into a bowl and left at room temperature. The time in hours until each bowl spoils is recorded. Conduct a full analysis.

## fungi litre spoiling.time  
## 1 A 2 21  
## 2 A 1 25  
## 3 A 3 18  
## 4 B 2 15  
## 5 B 1 22  
## 6 B 3 30  
## 7 C 1 19  
## 8 C 2 22  
## 9 C 3 25

We first give the formula:

model1 = aov(spoiling.time ~ fungi + litre + fungi : litre, data = df)  
plot(model1, which = c(1,2))



anova(model1)

## Analysis of Variance Table  
##   
## Response: spoiling.time  
## Df Sum Sq Mean Sq F value Pr(>F)  
## fungi 2 1.556 0.778 0.0289 0.9718  
## litre 1 8.167 8.167 0.3031 0.6203  
## fungi:litre 2 66.333 33.167 1.2309 0.4071  
## Residuals 3 80.833 26.944

By the two plot, Residual vs. Fitted and QQ-plot, we can see that the residuals seems not to be normal since it has 3 outliers in the QQ-plot. Also, the assumption of constant error variance is been violated, since the residuals are not in squared.

Pick the un-transformed model, since all transformations seems have no improvement on satisfying assumptions of the model.

library(lsmeans)

## Loading required package: emmeans

## The 'lsmeans' package is now basically a front end for 'emmeans'.  
## Users are encouraged to switch the rest of the way.  
## See help('transition') for more information, including how to  
## convert old 'lsmeans' objects and scripts to work with 'emmeans'.

library(multcompView)

## Warning: package 'multcompView' was built under R version 4.0.3

library(multcomp)

## Warning: package 'multcomp' was built under R version 4.0.3

## Loading required package: mvtnorm

## Loading required package: survival

## Loading required package: TH.data

## Warning: package 'TH.data' was built under R version 4.0.3

## Loading required package: MASS

##   
## Attaching package: 'TH.data'

## The following object is masked from 'package:MASS':  
##   
## geyser

library(emmeans)

Give a hypothesis test with

contrast(lsmeans(model1,~fungi), method = "pairwise")

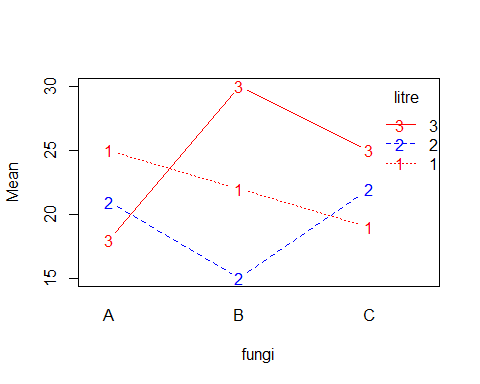
## NOTE: Results may be misleading due to involvement in interactions

## contrast estimate SE df t.ratio p.value  
## A - B -1.000 4.24 3 -0.236 0.9701   
## A - C -0.667 4.24 3 -0.157 0.9865   
## B - C 0.333 4.24 3 0.079 0.9966   
##   
## P value adjustment: tukey method for comparing a family of 3 estimates

cld(lsmeans(model1, ~ fungi : litre), alpha = 0.05)

## fungi litre lsmean SE df lower.CL upper.CL .group  
## A 2 21.3 3 3 11.8 30.9 1   
## C 2 22.0 3 3 12.5 31.5 1   
## B 2 22.3 3 3 12.8 31.9 1   
##   
## Confidence level used: 0.95   
## P value adjustment: tukey method for comparing a family of 3 estimates   
## significance level used: alpha = 0.05

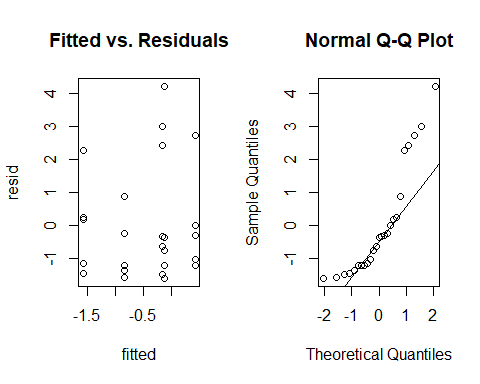
interaction.plot(x.factor = df$fungi, trace.factor = df$litre,response = df$spoiling.time, type = "b",col = c("red","blue"), xlab ="fungi", ylab = "Mean", trace.label = "litre")



We can see that all p-values are larger than 0.05 so we can conclude that no data is significantly different. By the grouping table, we can see that each group is not significantly different from others. Also from the interaction plot, we can see interactions among 3 lines, so there’s a huge chance there’s no significant difference. Therefore, We fail to reject null, and conclude that different fungi will affect the same amount of time it takes for milk to spoil.

## Question 3 (10 Points)

Below is code from a 1-Way ANOVA data analysis. What do these residuals tell us about the assumptions in the ANOVA model? What, if any, changes would you consider? If there needs to be a transformation, you do not need to state the specific transformation, simply that one is needed or not.

 The residual plot tells that the assumption of constant error variance for the model is been violated, since the residuals plots are not in squared. The QQ-plot tells that it seems not to be normal since the residuals are not in a straight line.

By far, I would recommend to use transformation for the model because with un-transformed model, we can’t get strong enough evidence to show the reliability by these graphs.

## Question 4 (16 Points)

An experiment was conducted to determine whether if front-wheel drive cars or real-wheel drive cars had a better braking time. Braking time will be measured by timing how long it takes each car to break after reaching 100 mph. 10 cars were selected at random from the test cars from the research and development department. Five cars were chosen at random to be modified to be rear-wheel drive, and the remaining 5 cars were modified to be front-wheel drive. It is assumed that a cars weight might be related to how fast they can stop, so each car’s weight in pounds is taken before driving. Conduct a full analysis.

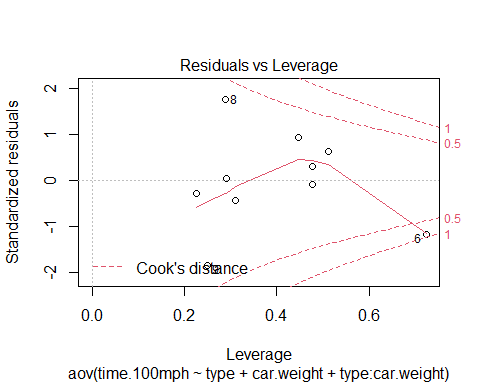
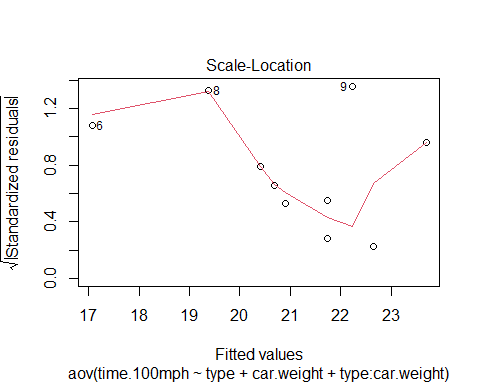
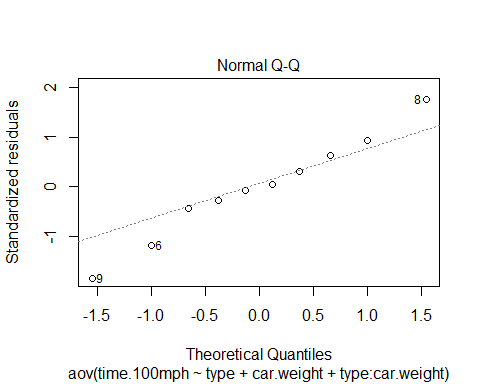
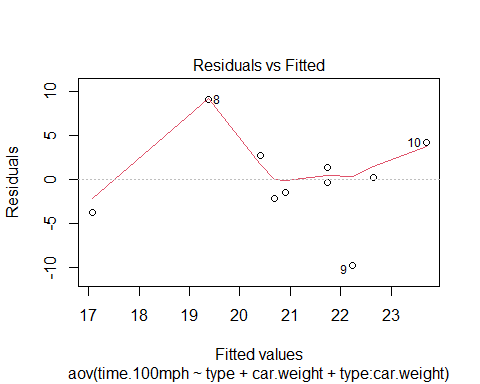
## car car.weight type time.100mph  
## 1 1 2741 front 21.4  
## 2 2 2741 front 23.1  
## 3 3 2835 front 19.4  
## 4 4 2859 front 18.5  
## 5 5 2890 front 23.1  
## 6 6 2986 rear 13.3  
## 7 7 2812 rear 22.9  
## 8 8 2914 rear 28.4  
## 9 9 2825 rear 12.5  
## 10 10 2779 rear 27.9

We get the formula

model2 = aov(time.100mph ~ type + car.weight + type : car.weight, data = df)  
anova(model2)

## Analysis of Variance Table  
##   
## Response: time.100mph  
## Df Sum Sq Mean Sq F value Pr(>F)  
## type 1 0.025 0.025 0.0007 0.9802  
## car.weight 1 24.976 24.976 0.6684 0.4449  
## type:car.weight 1 6.091 6.091 0.1630 0.7004  
## Residuals 6 224.194 37.366

plot(model2)

 By these graphs, we can see that the assumption of constant error variance is been violated, since not in squared, in QQ-plot residuals not in a straight line show it seems not to be normal distributed. Use un-transformed model since no transformation has improvement on satisfying model assumptions.

Giving hypothesis tests with

## Question 5 (16 Points)

An experiment was conducted to examine differences in migration rates of different bird species in the spring and in the fall. Six hawks and six robins were watched for one day in the spring, and six different hawks and six different robins were watched for one day in the fall. The response variable is the distance traveled in one day of observation. This experiment was conducted over 2 years.

A 2-way crossed ANOVA model was fit to the data, with both season (spring/fall) and species (robin/hawk) as fixed effects. See the R code and output on the next pages.

Using the R output, analyze the results of this experiment. You do NOT need to state any test statistics or p-values, and do NOT need to consider residuals, but you SHOULD clearly describe the results of important hypothesis tests for this analysis, and interpret the results in the context of the experiment. Do NOT interpret any pieces of the output that should not be examined. There may be code run that should not be run.

R code and output for question 5

## season species distance  
## 1 spring Hawk 21.0  
## 2 spring Hawk 19.5  
## 3 spring Hawk 22.5  
## 4 spring Hawk 21.5  
## 5 spring Hawk 20.5  
## 6 spring Hawk 21.0  
## 7 spring Robin 23.7  
## 8 spring Robin 23.8  
## 9 spring Robin 23.8  
## 10 spring Robin 23.7  
## 11 spring Robin 22.8  
## 12 spring Robin 24.4  
## 13 fall Hawk 32.0  
## 14 fall Hawk 30.5  
## 15 fall Hawk 25.0  
## 16 fall Hawk 27.5  
## 17 fall Hawk 28.0  
## 18 fall Hawk 28.6  
## 19 fall Robin 30.1  
## 20 fall Robin 28.9  
## 21 fall Robin 30.9  
## 22 fall Robin 34.4  
## 23 fall Robin 32.7  
## 24 fall Robin 32.7  
## 25 spring Hawk 22.5  
## 26 spring Hawk 26.0  
## 27 spring Hawk 28.0  
## 28 spring Hawk 27.0  
## 29 spring Hawk 26.5  
## 30 spring Hawk 25.2  
## 31 spring Robin 30.6  
## 32 spring Robin 31.1  
## 33 spring Robin 28.1  
## 34 spring Robin 34.9  
## 35 spring Robin 30.1  
## 36 spring Robin 25.5  
## 37 fall Hawk 28.0  
## 38 fall Hawk 27.5  
## 39 fall Hawk 31.0  
## 40 fall Hawk 29.5  
## 41 fall Hawk 30.0  
## 42 fall Hawk 29.2  
## 43 fall Robin 36.1  
## 44 fall Robin 36.6  
## 45 fall Robin 38.7  
## 46 fall Robin 37.1  
## 47 fall Robin 36.8  
## 48 fall Robin 37.1

## Analysis of Variance Table  
##   
## Response: distance  
## Df Sum Sq Mean Sq F value Pr(>F)   
## season 1 501.81 501.81 52.7999 4.661e-09 \*\*\*  
## species 1 236.74 236.74 24.9095 9.935e-06 \*\*\*  
## season:species 1 12.00 12.00 1.2626 0.2672   
## Residuals 44 418.18 9.50   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## NOTE: Results may be misleading due to involvement in interactions

## contrast estimate SE df t.ratio p.value  
## fall - spring 6.47 0.89 44 7.266 <.0001   
##   
## Results are averaged over the levels of: species

## NOTE: Results may be misleading due to involvement in interactions

## contrast estimate SE df t.ratio p.value  
## Hawk - Robin -4.44 0.89 44 -4.991 <.0001   
##   
## Results are averaged over the levels of: season

## contrast estimate SE df t.ratio p.value  
## fall Hawk - spring Hawk 5.47 1.26 44 4.344 0.0005   
## fall Hawk - fall Robin -5.44 1.26 44 -4.324 0.0005   
## fall Hawk - spring Robin 2.02 1.26 44 1.609 0.3842   
## spring Hawk - fall Robin -10.91 1.26 44 -8.667 <.0001   
## spring Hawk - spring Robin -3.44 1.26 44 -2.735 0.0429   
## fall Robin - spring Robin 7.47 1.26 44 5.933 <.0001   
##   
## P value adjustment: tukey method for comparing a family of 4 estimates

## Question 6 (16 Points)

The fishing line experiment was run by C. Reynolds, B. Grunden, and K. Taylor in 1996 in order to compare the strengths (the amount of weight needed for a line to break) of two brands (B) of fishing line exposed to two different levels of stress. Two different reels (R) of fishing line were selected at random from all reels of each brand. From each reel, 4 sections (S) of line were cut from each reel, making 16 sections total. The 4 sections from each reel were randomly assigned to two different stress levels (stressed, “1”; nonstressed, “2”) so that each stress level was assigned to two of the 4 sections from each reel. The strength of each section was measured by hanging weights (W) on the section of line until it broke. Conduct a full analysis.

## R B S W  
## 1 1 1 2 6.70  
## 2 1 1 1 6.40  
## 3 1 1 1 7.20  
## 4 1 1 2 7.00  
## 5 2 2 1 8.10  
## 6 2 2 1 8.90  
## 7 2 2 2 8.00  
## 8 2 2 2 6.10  
## 9 3 2 1 8.00  
## 10 3 2 1 8.00  
## 11 3 2 2 8.75  
## 12 3 2 2 8.50  
## 13 4 1 2 8.50  
## 14 4 1 1 9.50  
## 15 4 1 2 9.70  
## 16 4 1 1 9.40

## Question 7 (16 Points)

Write out an appropriate statistical model for the following experiment. Clearly define the response variable, identify all factors, and state the levels of each factor. Conduct a full analysis.

Consider the cigarette experiment run by J. Edwards, H. Hwang, S. Jamison, J. Kindelberger, and J. Steinbugl in 1996. The goal of this experiment is to determine the factors that affect the length of time that a cigarette will burn. There were three factors of interest:

* “Tar” at two levels, “1=regular” and “2=ultra-light,”
* “Brand” at two levels, “1=name brand” and “2=generic brand”,
* “Age” at three levels, “1=fresh,” “2=24 hour air exposure,” “3=48 hour air exposure.”

12 cigarettes were burned. 6 of the cigarettes were “regular” and 6 were “ultra-light”, and these were selected so that 6 were name brand and 6 were generic brand, and so that 4 of the cigarettes came from each age level. The cigarettes were chosen so that each of the 12 possible combinations of the three factors was found in 1 cigarette. The data are shown below

## tar brand age time  
## 1 1 2 2 301  
## 2 1 1 1 326  
## 3 1 2 3 260  
## 4 1 1 3 290  
## 5 1 1 2 312  
## 6 1 2 1 292  
## 7 2 1 1 329  
## 8 2 1 2 331  
## 9 2 1 3 285  
## 10 2 2 1 306  
## 11 2 2 2 258  
## 12 2 2 3 276