MATH-328 Project

David Oniani

April 10, 2021

## Setup

library(emmeans)  
library(ggplot2)  
library(multcomp)

## Loading required package: mvtnorm

## Loading required package: survival

## Loading required package: TH.data

## Loading required package: MASS

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

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

library(sjmisc)

## Initial Data Exploration and Analysis

# Load the data  
data = read.csv("data.csv")  
  
# Get rid of unwanted columns (will come in handy when designing models)  
data[c("trt\_id", "X\_hue", "X\_bright", "X\_price")] = NULL  
  
# Make hue, bright, and price factors  
data$hue = factor(data$hue)  
data$bright = factor(data$bright)  
data$price = factor(data$price)  
  
# Show first 6 and last 6 rows  
head(data)

## quality hue bright price  
## 1 2.1774 1 1 1  
## 2 4.9506 1 1 1  
## 3 4.8796 1 1 1  
## 4 4.9565 1 1 1  
## 5 6.2517 1 1 1  
## 6 5.0488 1 1 1

tail(data)

## quality hue bright price  
## 223 2.8811 2 2 2  
## 224 2.3638 2 2 2  
## 225 1.6588 2 2 2  
## 226 2.7048 2 2 2  
## 227 3.6134 2 2 2  
## 228 1.1509 2 2 2

# Show the structure of the data  
str(data)

## 'data.frame': 228 obs. of 4 variables:  
## $ quality: num 2.18 4.95 4.88 4.96 6.25 ...  
## $ hue : Factor w/ 2 levels "1","2": 1 1 1 1 1 1 1 1 1 1 ...  
## $ bright : Factor w/ 2 levels "1","2": 1 1 1 1 1 1 1 1 1 1 ...  
## $ price : Factor w/ 2 levels "1","2": 1 1 1 1 1 1 1 1 1 1 ...

# Are factors crossed?  
is\_crossed(data$hue, data$bright)

## [1] TRUE

is\_crossed(data$hue, data$price)

## [1] TRUE

is\_crossed(data$bright, data$price)

## [1] TRUE

# Are factors nested?  
is\_nested(data$hue, data$bright)

## [1] FALSE

is\_nested(data$bright, data$hue)

## [1] FALSE

is\_nested(data$hue, data$price)

## [1] FALSE

is\_nested(data$price, data$hue)

## [1] FALSE

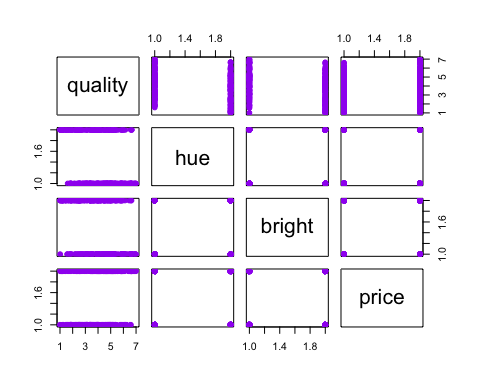
is\_nested(data$bright, data$price)

## [1] FALSE

is\_nested(data$price, data$bright)

## [1] FALSE

# Scatterplot matrix  
plot(data, pch = 20, cex = 1.5, col = "purple")



The data has three categorical variables: hue, bright, and price and one response variable: quality. We will treat hue, bright, and price as factors. The dataset is then suitable for a 3-factor experiment of online shopping. It contains three categorical factors: web page hue (1 = Blue, 2 = Red), Brightness (1 = Dark, 2 = Bright), and Price (1 = Low, 2 = high). Sample sizes are unbalanced. We will analyze when and how the response value - perceived quality - changes with brightness, hue, and price.

Exploratory analysis above prints first and last 6 rows of the data as well as shows the overall structure of the data. We also made hue, bright, and price factor variables. This will help using these variables properly, as factors, in models that we will develop.

Thus, our response variable is quality, which is a perceived quality of the product. Our predictor variables are hue, bright, and price - all of them are categorical variables. Since we do not have any quantitative predictors, we do not have covariates.

Factors hue and bright are crossed. Factors hue and price are also crossed. Factors bright and price are crossed too.

There is no nesting among the factors.

Since the investigators controlled the levels for all factors, all of the factors are fixed. Hence, factors hue, bright, and price are fixed (not random).

The scatterplot matrix is also shown. It is difficult to assess the relationship between any one of the factors and quality as there are a number of points on both ends of quality vs hue, quality vs bright, and quality vs price plots (three plots on the first row of the scatterplot matrix). This is indicative of some significant interactions between factors/variables (i.e., it is not only hue or only brightness that affects the perceived quality). Relationship between the factors present with hue vs bright, hue vs price, and bright vs price plots (inverse plots bright vs hue, price vs hue, and price vs bright are also present, but they show the same graph, only with flipped axes). Every single one of these plots show groups of points in each of the four corners. This also suggests some interaction.

## Fitting Full Model

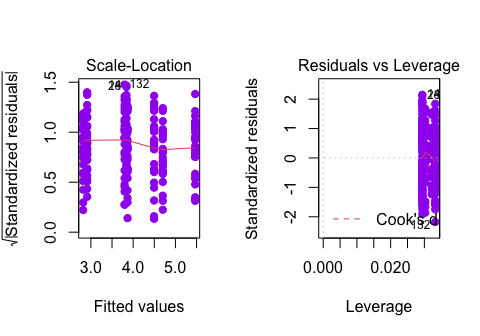
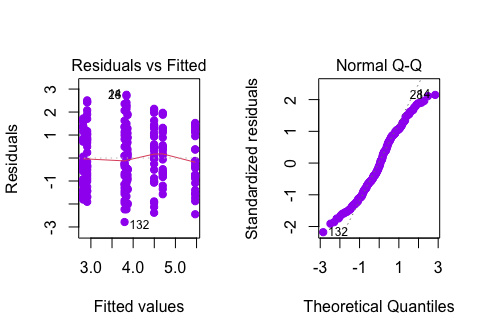
# Fit the initial model with three-way  
fit1 = aov(quality ~ .^2, data)  
  
# Report the summary  
summary(fit1)

## Df Sum Sq Mean Sq F value Pr(>F)   
## hue 1 48.1 48.08 28.448 2.38e-07 \*\*\*  
## bright 1 48.1 48.11 28.465 2.36e-07 \*\*\*  
## price 1 34.1 34.08 20.161 1.15e-05 \*\*\*  
## hue:bright 1 11.7 11.71 6.926 0.00910 \*\*   
## hue:price 1 8.6 8.57 5.068 0.02535 \*   
## bright:price 1 14.1 14.13 8.363 0.00421 \*\*   
## Residuals 221 373.5 1.69   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# Estimates  
fit1$coefficients

## (Intercept) hue2 bright2 price2 hue2:bright2   
## 3.84191422 -0.04488063 0.02511167 1.62618489 -0.91023771   
## hue2:price2 bright2:price2   
## -0.72221749 -0.99778918

# Residual analysis  
par(mfrow = c(1, 2))  
plot(fit1, pch = 20, cex = 1.5, col = "purple")



par(mfrow = c(1, 1))

We cannot do statistical inference for the full model (including the three-way interaction) as we have no replicates. However, we can still fit a model including all two-way interactions.

Residuals vs Fitted plot does not show the constant variance. The red line also has a curvature in two places toward the right tail. The red line forms a hill toward the right tail that is a hint for non-constant variance. Furthermore, ideally, the plot should show a random scattering of the points above and below the reference line at horizontal 0. However, in this case, it does seem like there is a pattern in the data. More specifically, points are aligned across approximately vertical lines. Thus, the constant variance assumption of ANOVA is not met.

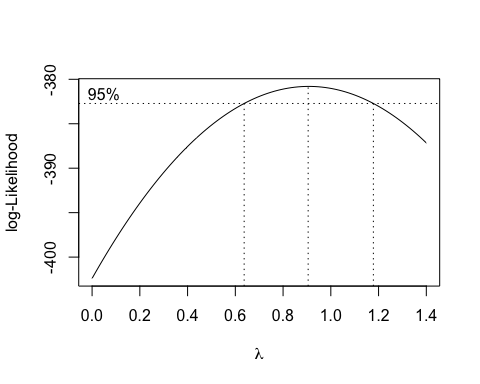
The Normal Q-Q plot shows that most of the points toward both tails of the plot do follow the dotted line. However, at both tails, there is some deviation from the dotted line which suggests some non-normality. This is more of a judgment call, but to me it seems to approximately normal. Hence, the normality assumption for ANOVA is met.

Scale-Location plot does have a decreasing trend in the middle of the plot, but levels off toward the end. The variance does not seem constant. This conclusion matches that of Residuals vs Fitted plot analysis.

Residuals vs Leverage also shows a few outliers, namely points 132, 14, and 28.

Given our analysis, let us now perform a Box-Cox analysis in order to find an appropriate power transformation for the response variable.

# Box-Cox analysis for the power transformation  
MASS::boxcox(fit1, lambda = seq(0, 2, 0.7))



From the plot, we can see that the lambda value is approximately 0.9. This means that we have two options:

1. Leave the response as is
2. Raise the response to the 0.9 power

We will try raising the response to 0.9 power and see if it helps. In the end, we will pick the better model.

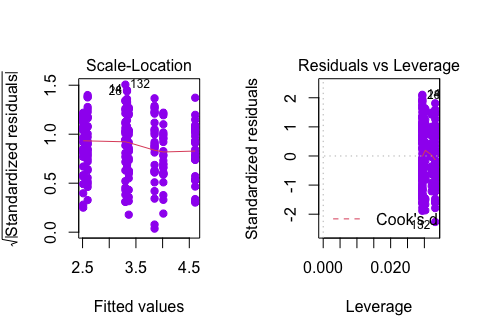
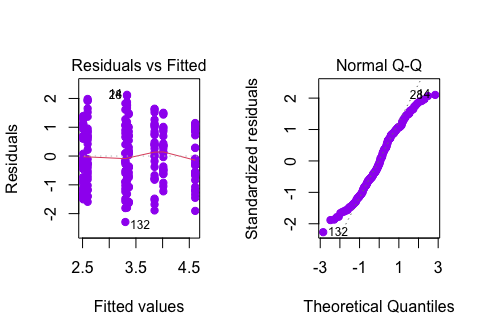
# Refit the model, applying the right power transformation  
fit2 = aov(quality^0.9 ~ .^2, data)  
  
# Summary  
summary(fit2)

## Df Sum Sq Mean Sq F value Pr(>F)   
## hue 1 30.29 30.291 28.711 2.11e-07 \*\*\*  
## bright 1 30.07 30.070 28.502 2.32e-07 \*\*\*  
## price 1 20.74 20.735 19.653 1.46e-05 \*\*\*  
## hue:bright 1 7.74 7.736 7.333 0.00730 \*\*   
## hue:price 1 5.28 5.281 5.005 0.02627 \*   
## bright:price 1 8.96 8.962 8.495 0.00393 \*\*   
## Residuals 221 233.16 1.055   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# Estimates  
fit2$coefficients

## (Intercept) hue2 bright2 price2 hue2:bright2   
## 3.33265938 -0.03049151 0.03297859 1.27795402 -0.74002608   
## hue2:price2 bright2:price2   
## -0.56646754 -0.79454105

# Residual analysis  
par(mfrow = c(1, 2))  
plot(fit2, pch = 20, cex = 1.5, col = "purple")



par(mfrow = c(1, 1))

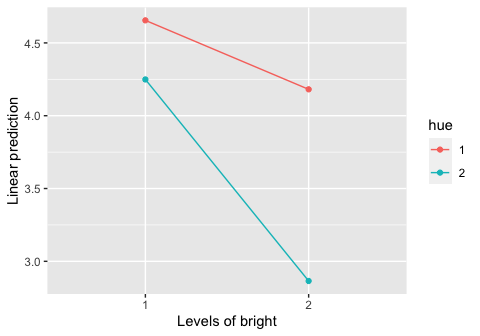
After applying the transformation, all four plots look similar and the summary also shows approximately the same p-values and estimates (coefficients). For this reason, we will stick with our initial model - fit1.

Since all of our interactions are significant, we do not have any insignificant interactions to remove and we do not need to apply a stepwise regression to remove non-significant effects.

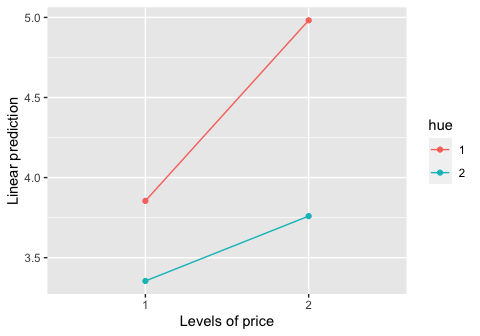
## Results

We start by analyzing the interaction plots of significant effects. Since all of the two-way interactions were significant and we have three factors, we will analyze three plots. This would also be the highest order interactions as due to the nature of the study that collected the data, including three-way interaction would not be appropriate.

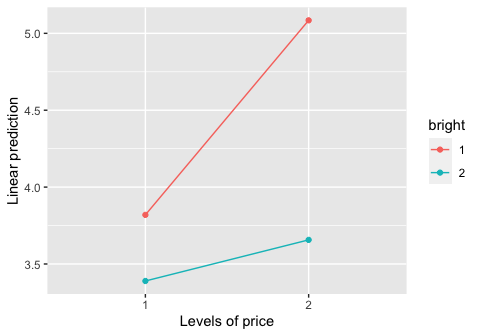
# Interaction plots  
emmip(fit1, hue ~ bright)



emmip(fit1, hue ~ price)



emmip(fit1, bright ~ price)



In terms of interactions plots, all three interaction plots show some degree of interaction between the variables. The lowest degree of interaction seems to be between variables hue and price (lines are relatively parallel with respect to each other). The highest degree of interaction seems to be between variables bright and price (lines are not relatively parallel with respect to each other). That being said, in all three plots, two lines would cross each other if extended and thus, the interactions are significant. Further analysis will show the degree of these interactions in more detail.

# Report the summary  
summary(fit1)

## Df Sum Sq Mean Sq F value Pr(>F)   
## hue 1 48.1 48.08 28.448 2.38e-07 \*\*\*  
## bright 1 48.1 48.11 28.465 2.36e-07 \*\*\*  
## price 1 34.1 34.08 20.161 1.15e-05 \*\*\*  
## hue:bright 1 11.7 11.71 6.926 0.00910 \*\*   
## hue:price 1 8.6 8.57 5.068 0.02535 \*   
## bright:price 1 14.1 14.13 8.363 0.00421 \*\*   
## Residuals 221 373.5 1.69   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# Estimates  
fit1$coefficients

## (Intercept) hue2 bright2 price2 hue2:bright2   
## 3.84191422 -0.04488063 0.02511167 1.62618489 -0.91023771   
## hue2:price2 bright2:price2   
## -0.72221749 -0.99778918

The summary report confirms our conclusions based on interaction plots. Indeed, hue and price interaction seems to be the least significant one (but still significant as the p-value is approximately 0.025 < 0.05). The interaction between bright and price, as observed, is most significant with the p-value of approximately 0.004. The interaction between hue and bright is also significant with the p-value of approximately 0.009.

Variables hue, bright, and price are all significant with the p-values of 2.38e-07, 2.36e-07, and 1.15e-05 respectively.

The estimate of hue is approximately -0.045 which is a negative direction. The estimates for bright and price are 0.025 and 1.626 respective, both having a positive direction.

The intercept is 3.842, which means that the mean quality when all factors are null is 3.842.

cld(emmeans(fit1, ~ hue | bright), letters=LETTERS)

## bright = 1:  
## hue emmean SE df lower.CL upper.CL .group  
## 2 4.25 0.179 221 3.90 4.60 1   
## 1 4.66 0.168 221 4.32 4.99 1   
##   
## bright = 2:  
## hue emmean SE df lower.CL upper.CL .group  
## 2 2.86 0.168 221 2.53 3.20 1   
## 1 4.18 0.175 221 3.84 4.53 2   
##   
## Results are averaged over the levels of: price   
## Confidence level used: 0.95   
## significance level used: alpha = 0.05

cld(emmeans(fit1, ~ hue | price), letters=LETTERS)

## price = 1:  
## hue emmean SE df lower.CL upper.CL .group  
## 2 3.35 0.174 221 3.01 3.70 1   
## 1 3.85 0.174 221 3.51 4.20 2   
##   
## price = 2:  
## hue emmean SE df lower.CL upper.CL .group  
## 2 3.76 0.172 221 3.42 4.10 1   
## 1 4.98 0.169 221 4.65 5.32 2   
##   
## Results are averaged over the levels of: bright   
## Confidence level used: 0.95   
## significance level used: alpha = 0.05

cld(emmeans(fit1, ~ bright | price), letters=LETTERS)

## price = 1:  
## bright emmean SE df lower.CL upper.CL .group  
## 2 3.39 0.174 221 3.05 3.73 1   
## 1 3.82 0.174 221 3.48 4.16 1   
##   
## price = 2:  
## bright emmean SE df lower.CL upper.CL .group  
## 2 3.66 0.169 221 3.32 3.99 1   
## 1 5.08 0.172 221 4.74 5.42 2   
##   
## Results are averaged over the levels of: hue   
## Confidence level used: 0.95   
## significance level used: alpha = 0.05

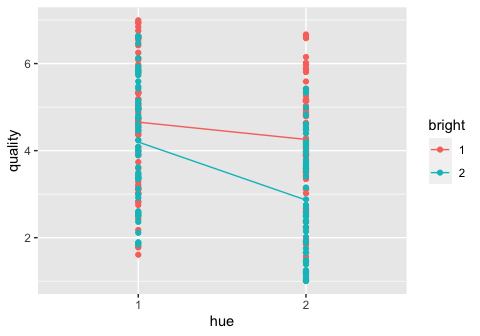
Above find the appropriate pairwise comparisons for the significant effects.

Let us first analyze emmeans with hue and bright interaction. We see that the mean response (quality value) is the lowest with high brightness and red hue (2.86). The mean response is the highest when brightness is low and hue is blue (4.66). When bright = 1 (low brightness) and hue = 2 (red), the emmean is 4.25 and when bright = 2 (high brightness) and hue = 1 (blue), the emmean is 4.18.

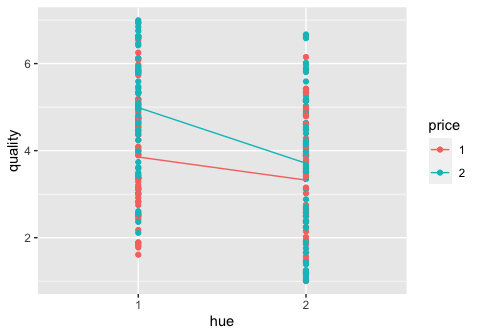
In hue - price interaction, the highest emmean value is achieved when price is high and hue is blue, the value is 4.98. The lowest emmean value is when price = 1 (low price) and hue = 2 (red), the value is 3.76. When price is low and hue is blue, the emmean value is 3.85. When price is high and hue blue, the emmean is 3.76.

In price - bright interaction, the biggest emmean of 5.08 is when bright = 1 and price = 2 (high brightness and low price). The lowest emmean value is when price is low and brightness is high (price = 1, bright = 2). When both brightness and price are low, the emmean is 3.82 and when both price and brightness are high, the emmean is 3.66.

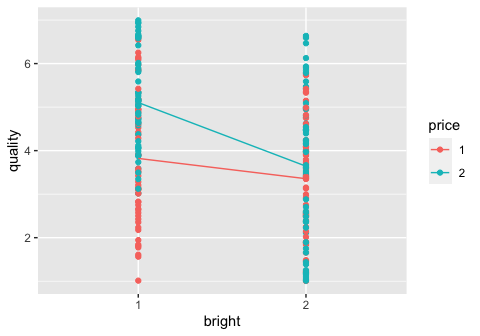
# Hue - Bright  
qplot (hue, quality, data = data, color = bright) +   
 stat\_summary(fun = mean, geom = "line", aes(group = bright))



# Hue - Price  
qplot (hue, quality, data = data, color = price) +   
 stat\_summary(fun = mean, geom = "line", aes(group = price))



# Bright - Price  
qplot (bright, quality, data = data, color = price) +   
 stat\_summary(fun = mean, geom = "line", aes(group = price))



Finally, we state several conclusions. In order for me to explain the conclusions, I will use analogies - high brightness (think blinding website/packaging/etc), high hue (think same as high brightness, with blue being less bright/darker and red being brighter/lighter/blinding in extreme cases), high price (think something that might have high quality, but that would depend on interaction with hue and brightness).

* Products with high brightness and high hue have the lowest perceived quality - makes sense, blinding brightness and high hue does not leave good impressions
* Products with low brightness and low hue have the highest perceived quality - makes sense low brightness and low hue probably means more elegant website and is therefore, more attractive in general)
* Products with high price and low hue have the highest perceived quality - makes sense, low hue would probably mean more elegant website and high price would mean that the product is probably high quality (hence, higher perceived quality)
* Products with low price and high hue have the lowest perceived quality: makes sense, low price on already not very elegant looking website might be very attractive.
* Products with high price and low brightness have the highest perceived quality - makes sense, products presented with low brightness and high price will probably look more elegant and attract more users
* Products with high low and high brightness have the highest perceived quality - makes sense, products presented with blinding colors and low price are not very attractive

Furthermore, we conclude that it is enough to just know one of the factors (e.g., know brightness or price) in order to predict the perceived quality. Every factor - hue, bright, and price - had significant interactions as well.