Lab 2 - Group 16

Eric Börjesson (ericbor@student.chalmers.se)
Ludvig Ekman (eludvig@student.chalmers.se)
Wenli Zhang (wenliz@student.chalmers.se) Tim Grube (gusgruti@student.gu.se)

15.12.2020

A/B/n testing

Question a.

What is the minimum number of customers in each group for the experiment to have the power of 90%? How many monthly customers must the company have to be able to run this experiment in 1 month? What happens with the minimum number of customers needed if the effect size decreases (and power and significance level remain constant)? What happens if the effect size increases?

Based on the background, there are 5 levels (original cover + 4 new ones) of the factor of Cover, so that k is equal to 5. Compared to the control (original cover), the company wants to have at least 8% improvement (f=0.08) with a confidence of 95% (sig.level=0.05) and has the power of 90% (power=0.09) to develop better art covers and run experiments.

```
data <- read.csv("./gotaflix-abn.csv")
pwr::pwr.anova.test(k=5, n=NULL, f = 0.08, sig.level = 0.05, power = 0.9)
```

```
##
##      Balanced one-way analysis of variance power calculation
##
##               k = 5
##               n = 482.3577
##               f = 0.08
##       sig.level = 0.05
##           power = 0.9
##
## NOTE: n is number in each group
```

```
pwr::pwr.anova.test(k=5, n=NULL, f = 0.02, sig.level = 0.05, power = 0.9)
```

```
##
##      Balanced one-way analysis of variance power calculation
##
##               k = 5
##               n = 7703.475
##               f = 0.02
##       sig.level = 0.05
##           power = 0.9
##
## NOTE: n is number in each group
```

```r
pwr::pwr.anova.test(k=5, n=NULL, f = 0.32, sig.level = 0.05, power = 0.9)
```

```
##
##      Balanced one-way analysis of variance power calculation
##
##               k = 5
##               n = 31.05305
##               f = 0.32
##       sig.level = 0.05
##           power = 0.9
##
## NOTE: n is number in each group
```

From the result shown above, we can see that the minimum number of customers in each group is 482 for the experiment to have the power of 90% with a confidence of 95% to have 8% improvement.

As each group has minimum 482 customers, the total number of customers is 2410 having to be able to run this experiment in 1 month.

If we decrease f (effect size) to 0.02, the minimum number of customers significantly increases to 7703.

If we increase f (effect size) to 0.32, the minimum number of customers significantly decreases to 31.

Question b.

In one month, the company collected the data (in long format) in the file gotaflix-abn.csv. Import this data to R and generate descriptive statistics for each group.

First, the file gotaflix-abn.csv is imported and then the function psych::describeBy() is used to generate descriptive statistics for each group (cover). The results are as follows:

```r
data <- read.csv("./gotaflix-abn.csv")
data$Cover <- as.factor(data$Cover)
data$Engagement <- as.numeric(data$Engagement)
psych::describeBy(data, group = data$Cover)
```

```
##
##  Descriptive statistics by group
## group: A
##            vars   n mean  sd median trimmed mad   min  max range  skew kurtosis
## Engagement    1 800 0.16 0.1   0.16    0.16 0.1 -0.14 0.54  0.68 -0.01    -0.06
## Cover*        2 800 1.00 0.0   1.00    1.00 0.0  1.00 1.00  0.00   NaN      NaN
##            se
## Engagement  0
## Cover*      0
## ------------------------------------------------------------
## group: B
##            vars   n mean  sd median trimmed mad   min  max range  skew kurtosis
## Engagement    1 800 0.16 0.1   0.16    0.16 0.1 -0.17 0.52  0.69 -0.01     0.07
## Cover*        2 800 2.00 0.0   2.00    2.00 0.0  2.00 2.00  0.00   NaN      NaN
##            se
## Engagement  0
## Cover*      0
```

```
## ------------------------------------------------------------
## group: C
##            vars   n mean   sd median trimmed mad   min max range skew kurtosis
## Engagement    1 800 0.18 0.11   0.17    0.18 0.1 -0.12 0.5  0.62 0.03    -0.12
## Cover*        2 800 3.00 0.00   3.00    3.00 0.0  3.00 3.0  0.00  NaN      NaN
##            se
## Engagement  0
## Cover*      0
## ------------------------------------------------------------
## group: D
##            vars   n mean  sd median trimmed mad  min  max range  skew kurtosis
## Engagement    1 800 0.16 0.1   0.16    0.16 0.1 -0.2 0.44  0.64 -0.16    -0.11
## Cover*        2 800 4.00 0.0   4.00    4.00 0.0  4.0 4.00  0.00   NaN      NaN
##            se
## Engagement  0
## Cover*      0
## ------------------------------------------------------------
## group: E
##            vars   n mean  sd median trimmed mad   min  max range  skew kurtosis
## Engagement    1 800 0.17 0.1   0.17    0.17 0.1 -0.15 0.48  0.63 -0.04     0.06
## Cover*        2 800 5.00 0.0   5.00    5.00 0.0  5.00 5.00  0.00   NaN      NaN
##            se
## Engagement  0
## Cover*      0
```

Question c.

Generate a linear model to fit this experiment data. Write the equation that represents this model. What does the intercept mean? What is the value the model gives when only the coefficient for Cover C equals to 1 and the rest of the coefficients equals to 0?

The linear model below uses the cover as a predictor to estimate the outcome of engagement.

The intercept is the average engagement for the original cover (control version) of 0.16.

When the coefficient for cover C is equal to 1 and the rest of the coefficients is equal to 0, the value of the average engagement is 0.178 (0.16+0.018), which means that using Cover C increases the average engagement.

```
lindata <- lm(Engagement ~ Cover, data)
summary(lindata)
```

```
##
## Call:
## lm(formula = Engagement ~ Cover, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.35922 -0.06652 -0.00211  0.07163  0.38266
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.1603672  0.0036631  43.779  < 2e-16 ***
## CoverB      -0.0006146  0.0051804  -0.119 0.905564
## CoverC       0.0179482  0.0051804   3.465 0.000536 ***
```

```
## CoverD      -0.0021013  0.0051804  -0.406 0.685038
## CoverE       0.0094468  0.0051804   1.824 0.068290 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1036 on 3995 degrees of freedom
## Multiple R-squared:  0.005461,   Adjusted R-squared:  0.004466
## F-statistic: 5.484 on 4 and 3995 DF,  p-value: 0.0002114
```

Question d.

One of the assumptions of a linear model is 'normality'. What needs to be normal? Analyze the normality of the model using a qqplot and then use a Shapiro test. Does the Shapiro-Wilk test agree with the qqplot? Which one do you choose to justify the normality?

The residuals need to be normal.

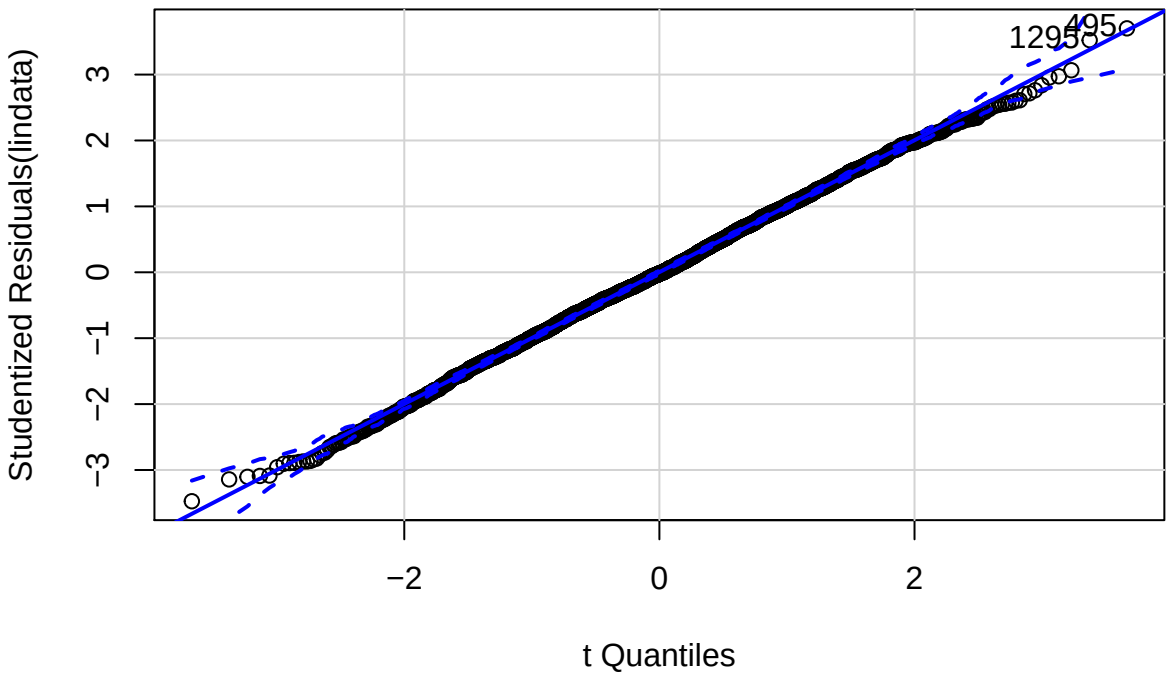
From the following qqplot diagram, we can see that basically most of the points fall close to the straight blue line and inside the dashed interval, so the data are normally distributed. There are some outliers (e.g., the point displayed as 495) and some slight skew on the tails, but the qqplot still shows that the data are normally distributed.

From the outcome of Shapiro-Wilk test, the value of W is also 0.99943 which is just slightly less than 1 and still indicates that the data is normally distributed. Furthermore, the p-value is 0.27 which is higher than 0.05, so the null hypothesis (data is normally distributed) cannot be rejected.

Both the qqplot and the Shapiro-Wilk test indicate that the residuals are normally distributed.

We can choose both qqplot and Shapiro-Wilk test to justify the normality. However, regarding this experiment the qqplot is more valuable than the Shapiro–Wilk test. A Shapiro–Wilk test is suitable for a small sample size up to 2000 and justifies "normality" based on the p-value. The larger the sample size the more inaccurate it is to determine whether the data is normally distributed according to the p-value. As it is a large sample size of 4000, it is advisable to use a qqplot to justify the normality. The qqplot is able to reveal how the distribution is non-normal according to the characteristics (i.e., skewness, kurtosis, outliers, etc.).

```
qqPlot(lindata)
```




```
## [1]  495 1295
```

```
shapiro.test(data$Engagement)
```

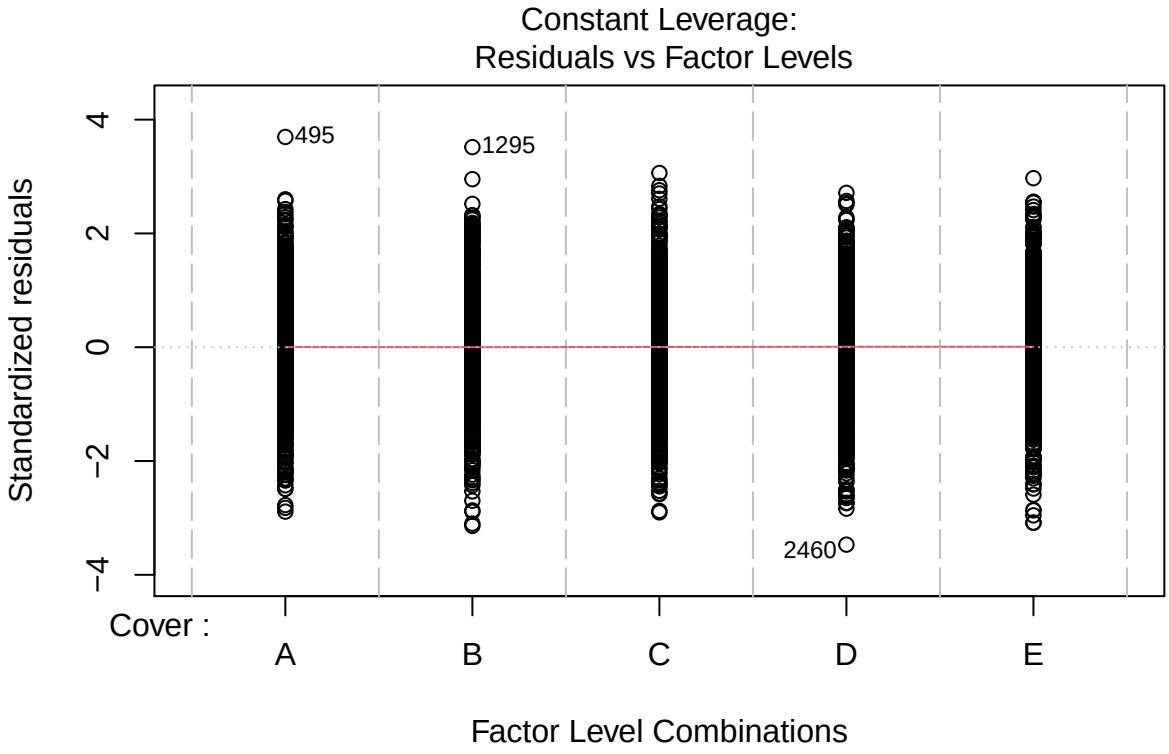
```
##
##  Shapiro-Wilk normality test
##
## data:  data$Engagement
## W = 0.99943, p-value = 0.2713
```

Question e.

The other assumptions of the one-way ANOVA are homoscedasticity of the residuals and independence of the data. Create a scatter plot for to analyze the homogeneity of the variances and them conduct a Levene's test.

If the assumption of homoscedasticity would not be fulfilled, the scatter plot would show that the variance is higher for some x values. For example, higher x values would lead to a higher variance of the residuals. Looking at the following scatter plot, this is not the case. Instead, the variance is very similar and there are just very few outliers.

```
plot(lindata)
```



The Levene test shows, that the F value is a lot lower than the critical F value of 2.37 and the p-value is much higher than 0.05. Therefore, the null hypothesis cannot be rejected, so there is still no evidence that it is not homoscedastic.

```
leveneTest(lindata)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##         Df F value Pr(>F)
## group    4  0.2158 0.9297
##       3995
```

Question f.

Discuss the independence assumption. How can we verify it? Is it part of the design of the experiment or the analysis?

The independence of the residuals is based on the design of the experiment and should be verified there. In this case, the users were randomly allocated into experimental groups and no one is asked twice, so the data should be independent. However, it can be checked additionally with a Durbin-Watson-test for example. The result of the test shows that the p-value is a lot higher than 0.05, so the null hypothesis cannot be rejected and the test indicates as well, that the residuals are independent.

```
durbinWatsonTest(lindata)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1     0.008618299      1.982487   0.566
##  Alternative hypothesis: rho != 0
```

Question g.

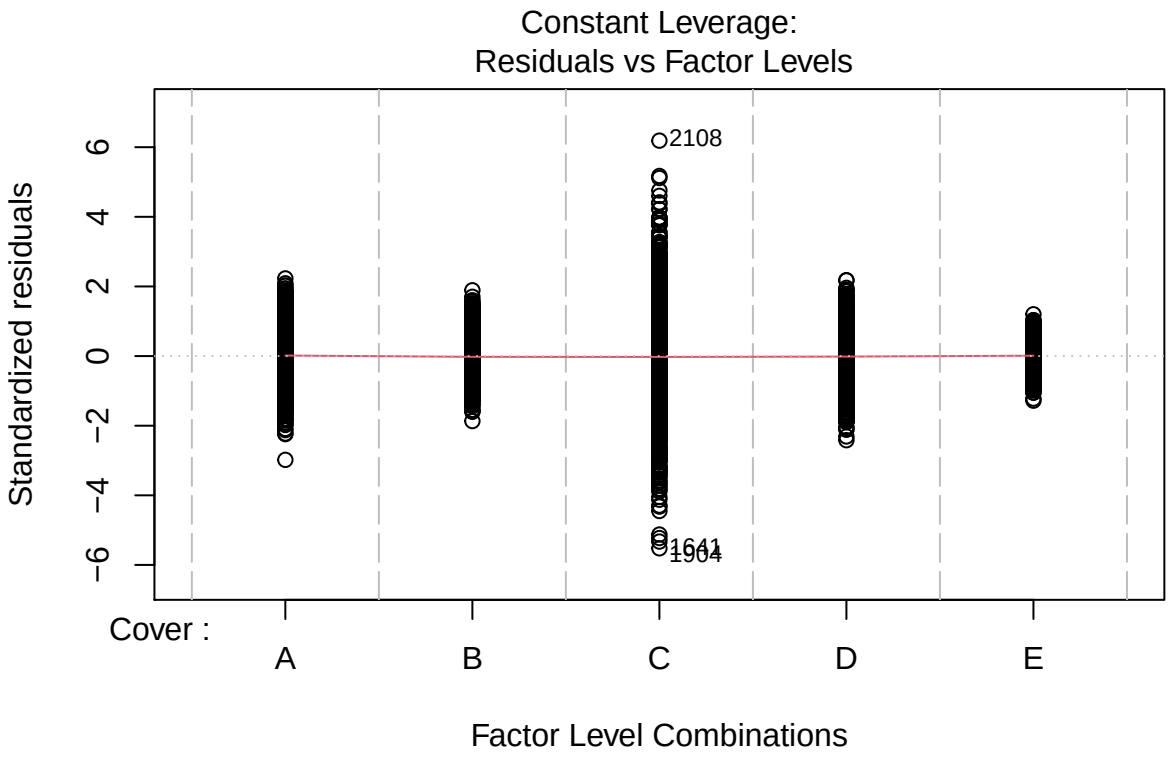
Suppose you have received the data of the gotaflix-abn-modified.csv. Does the data have homoscedasticity?

Import of the data and calculating of a linear model:

```
dataMod <- read.csv("./gotaflix-abn-modified.csv")
dataMod$Cover <- as.factor(dataMod$Cover)
dataMod$Engagement <- as.numeric(dataMod$Engagement)
lindataMod <- lm(Engagement ~ Cover, dataMod)
```

Looking at the following scatter plot, this time there are huge differences between the variances of the different X values, especially for cover C, but also cover E. This is already a strong indication that this data is not homoscedastic.

```
plot(lindataMod)
```



The levene test supports this statement. The F value is a lot higher than the critical F value and the p-value is much smaller than 0.05. Therefore, the null hypothesis has to be rejected and the data is not homoscedastic.

```
leveneTest(lindataMod)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##         Df F value    Pr(>F)
```

```
## group    4  358.91 < 2.2e-16 ***
##       3995
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Question h.

Analyze the data (gotaflix-abn.csv) and discuss which art cover had a better engagement? First, run an Anova test and see if the model is statistically significant. Second, run the Tukey HSD post-hoc test and analyze the results?

The ANOVA table shows that the F value is 5.4844 which is higher than the critical F value of 2.37, so the model is statistically significant. For calculating the F value, the mean squares for the covers (0.0588) and the mean squares for the residuals (0.0107) are determined first and then are divided.

```
Anova(lindata)
```

```
## Anova Table (Type II tests)
##
## Response: Engagement
##           Sum Sq   Df F value    Pr(>F)
## Cover      0.235    4  5.4844 0.0002114 ***
## Residuals 42.884 3995
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The result of the Tukey HSD test (HSD = honestly significant difference) shows that there are three pairs of covers which have a significant difference. Cover C is the only cover which improves the engagement, at least in comparison to the covers A, B and D. There is no significant difference to cover E.

```
TukeyHSD(aov(lindata))
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = lindata)
##
## $Cover
##              diff          lwr          upr     p adj
## B-A -0.0006146171 -0.014751973  0.013522738 0.9999554
## C-A  0.0179482241  0.003810869  0.032085580 0.0048702
## D-A -0.0021013000 -0.016238655  0.012036055 0.9943150
## E-A  0.0094468269 -0.004690529  0.023584182 0.3599648
## C-B  0.0185628413  0.004425486  0.032700197 0.0031621
## D-B -0.0014866829 -0.015624038  0.012650673 0.9985203
## E-B  0.0100614440 -0.004075911  0.024198799 0.2950989
## D-C -0.0200495242 -0.034186880 -0.005912169 0.0010456
## E-C -0.0085013973 -0.022638753  0.005635958 0.4710796
## E-D  0.0115481269 -0.002589229  0.025685482 0.1690835
```

Full factorial experiment

Question a.

How many and what are the experimental groups?

There are 2 different factors and each factor has 2 levels, so there are a total of 4 experimental groups. The groups are:

1. Cover = Character & Summary = Character
2. Cover = Character & Summary = Genre
3. Cover = Genre & Summary = Character
4. Cover = Genre & Summary = Genre

Question b.

The design of this experiment can be seen as a two-way ANOVA. Load the data (gotaflix-2wayANOVA.csv) and fit a linear model. Write the equation of this linear model. What does it mean the intercept value?

The intercept is the model's estimation of the engagement value when the two factors Cover and Summary are set to 0. In this case, that is when *Cover = Character & Summary = Character*.

We have created 3 linear models below: lm1, lm2 and lm3. The model lm1 uses the two independent variables *Cover & Summary* to predict the dependent variable Engagement. The models lm2 and lm3 also include the interaction between the factors and are modeled in two different ways. An inspection of the results below shows that they produce equivalent outcome.

```
twoWayAnova <- read.csv("./gotaflix-2wayANOVA.csv", stringsAsFactors = TRUE)
lm1 <- lm(Engagement ~ Cover + Summary, data = twoWayAnova)
lm2 <- lm(Engagement ~ Cover + Summary + Cover:Summary, data = twoWayAnova)
lm3 <- lm(Engagement ~ Cover*Summary, data = twoWayAnova)
summary(lm1) # alpha = 0.169028
```

```
##
## Call:
## lm(formula = Engagement ~ Cover + Summary, data = twoWayAnova)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.39633 -0.07327 -0.00038  0.07897  0.43081
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.169028   0.003521  48.004   <2e-16 ***
## CoverGenre   -0.008812   0.004066  -2.167   0.0303 *
## SummaryGenre -0.006087   0.004066  -1.497   0.1345
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.115 on 3197 degrees of freedom
## Multiple R-squared:  0.002166,   Adjusted R-squared:  0.001541
## F-statistic: 3.469 on 2 and 3197 DF,  p-value: 0.03126
```

```r
summary(lm2) # alpha = 0.175266
```

```
##
## Call:
## lm(formula = Engagement ~ Cover + Summary + Cover:Summary, data = twoWayAnova)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.39009 -0.07365 -0.00243  0.07861  0.43705
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)               0.175266   0.004061  43.163  < 2e-16 ***
## CoverGenre               -0.021287   0.005742  -3.707 0.000213 ***
## SummaryGenre             -0.018563   0.005742  -3.233 0.001239 **
## CoverGenre:SummaryGenre   0.024951   0.008121   3.072 0.002141 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1148 on 3196 degrees of freedom
## Multiple R-squared:  0.005104,   Adjusted R-squared:  0.00417
## F-statistic: 5.465 on 3 and 3196 DF,  p-value: 0.000958
```

```r
summary(lm3) # alpha = 0.175266
```

```
##
## Call:
## lm(formula = Engagement ~ Cover * Summary, data = twoWayAnova)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.39009 -0.07365 -0.00243  0.07861  0.43705
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)               0.175266   0.004061  43.163  < 2e-16 ***
## CoverGenre               -0.021287   0.005742  -3.707 0.000213 ***
## SummaryGenre             -0.018563   0.005742  -3.233 0.001239 **
## CoverGenre:SummaryGenre   0.024951   0.008121   3.072 0.002141 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1148 on 3196 degrees of freedom
## Multiple R-squared:  0.005104,   Adjusted R-squared:  0.00417
## F-statistic: 5.465 on 3 and 3196 DF,  p-value: 0.000958
```

Question c.

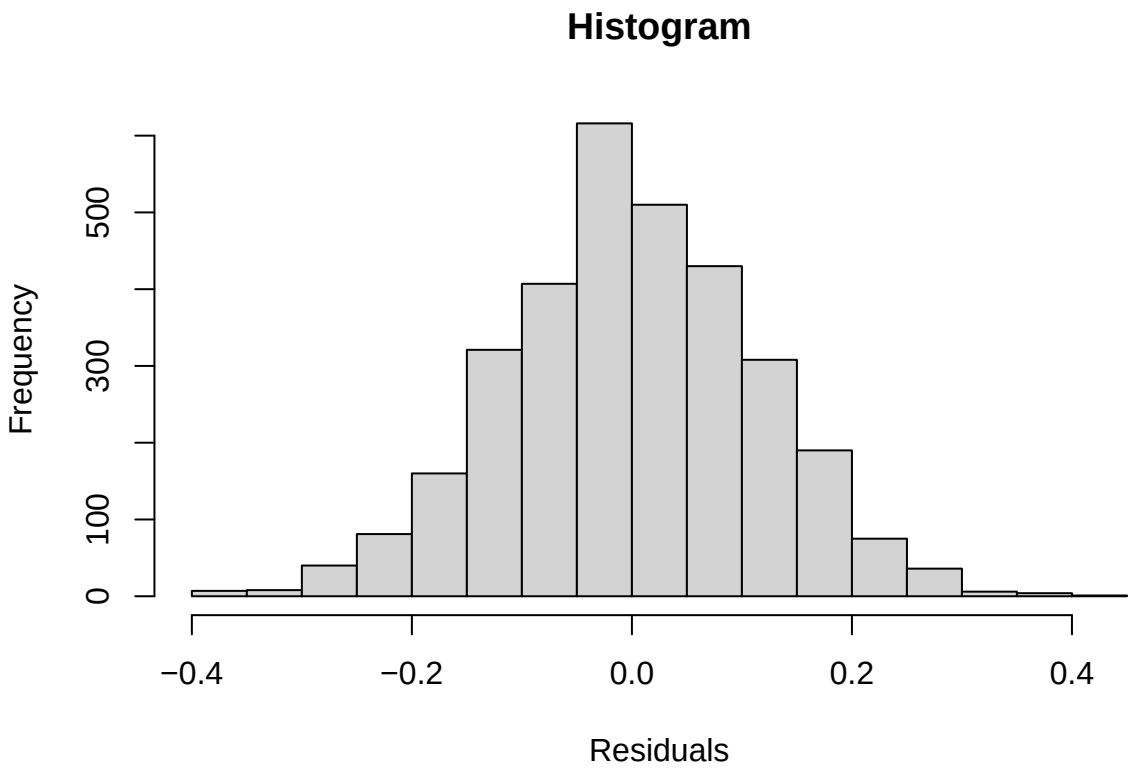
Analyze the assumptions of this two-way ANOVA. Independence, normality and homoscedasticity

Independence - The assumption of independence means that there should not exist any relationship between the observations within and between the groups. This is related to the design of the study and is

not something we can verify from the given data set. However, from the provided description, 50 movies have been randomly selected and users have been randomly assigned to the experimental groups. This would conform with the assumption of independence since both the movies and users are randomly picked. However, the movies have been randomly selected among the top 200 movies and are thereby not a random sample of the population. This could effectively influence the outcome and the result might not be possible to generalize for the entire population, that is all the movies.

Normality - Below is a histogram of the residuals. A visual inspection of the histogram indicates that the residuals are normally distributed and thus the normality assumption is met.

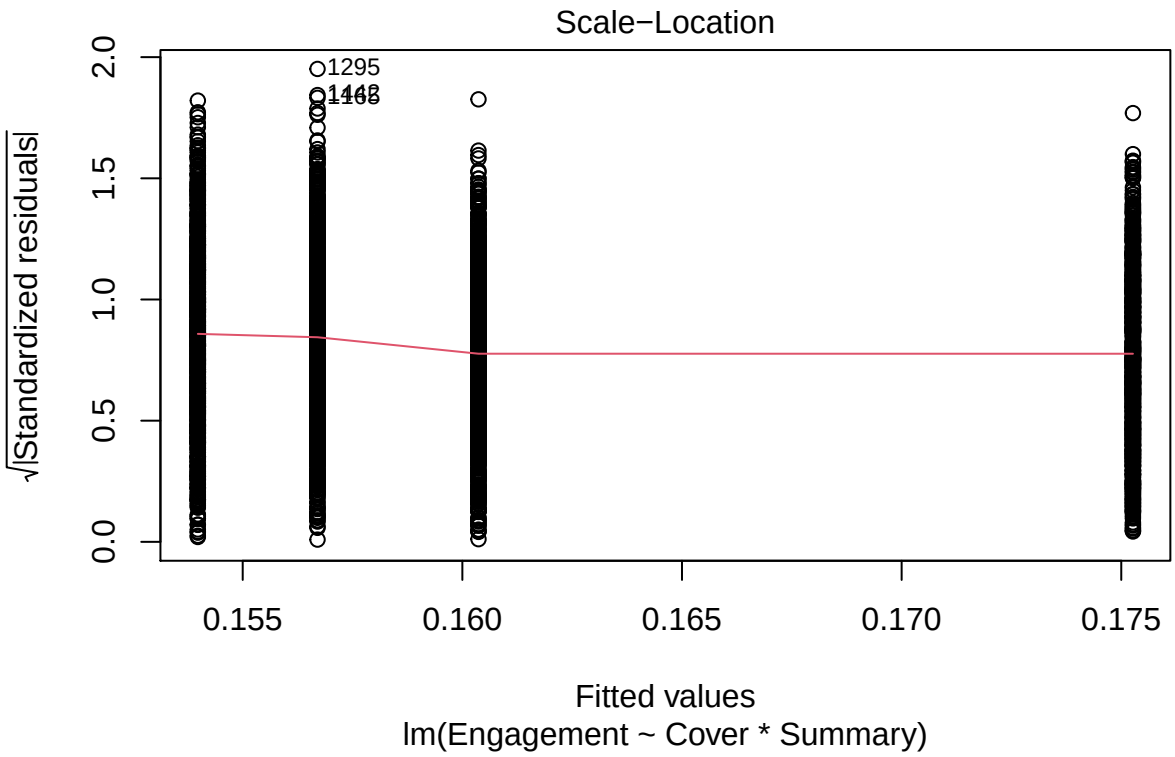
```
# Normality of residuals - looks good
hist(lm3$residuals, main = "Histogram", xlab = "Residuals")
```



Homoscedasticity - We have used a scatter plot to check for homoscedasticity and also compared against the results of a Levene's test.

A visual inspection of the below scatter plot indicates that there is a clear difference between the variances for the different values on the x-scale. The lower the fitted values for the engagement level, the higher the standardized residuals become. This indicates that the data is not homoscedastic but instead heteroscedastic.

```
plot(lm3,3)
```



The Levene's test supports the statement of the data being heteroscedastic. The F value is higher than the the critical value and the p-value is lower than a significance value of 0.05, so the null hypothesis of equal variances can be rejected. However, because of its robustness to the violation of assumptions especially with large sample sizes like in this case, we should not rely too much on it.

```
leveneTest(lm3)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##         Df F value   Pr(>F)
## group    3  15.873 3.04e-10 ***
##       3196
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Question d.

Present the ANOVA table. What is statistically significant? What are the effect sizes? What conclusion can you make about this experiment?

```
lm3.anova <- anova(lm3)
lm3.anova
```

```
## Analysis of Variance Table
##
## Response: Engagement
##                 Df Sum Sq  Mean Sq F value   Pr(>F)
## Cover            1  0.062 0.062116  4.7092 0.030075 *
## Summary          1  0.030 0.029642  2.2472 0.133952
## Cover:Summary    1  0.125 0.124515  9.4399 0.002141 **
## Residuals     3196 42.156 0.013190
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#stargazer::stargazer(lm3.anova, header = F)
```

By analyzing the above results using the p-value, we can conclude which independent variables are statistically significant. Below is a summary of each independent variable including the interaction effect, using a significance level of 0.05.

- Cover: p-value = 0.030075 $<$ 0.05. It is significant.
- Summary: p-value = 0.133952 $>$ 0.05. It is not significant.
- Interaction between Cover and Summary: p-value = 0.002141 $<$ 0.05. It is significant.

Effect Size - The effect size, r-squared, can be calculated using:

- SS(A) / SS(T)
- SS(A) = factor A sum of squares
- SS(T) = total sum of squares

We have only calculated the effect size of the factors that were statistically significant, which was Cover and the interaction between Cover and Summary. If we have huge data sets, it is easy to find statistical significance (that is how the model works) but looking at the size of the effect tells us if this factor is relevant at all.

```
TSS <- sum(lm3.anova$`Sum Sq`)
SScover <- 0.062
SScoverSummary <- 0.125
r2cover <- SScover / TSS
r2coverSummary <- SScoverSummary / TSS
c(r2cover, r2coverSummary)
```

```
## [1] 0.001463213 0.002950027
```

Cover and the interaction between Cover and Summary are statistically significant which means that they impact the output value for Engagement. However, the effect size is low for both Cover (0.15%) and the interaction (0.30%). To further understand the effects, we perform a Tukey HSD test which measures the difference in means for each pair.

```
TukeyHSD(aov(lm3))
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
```

```
## Fit: aov(formula = lm3)
##
## $Cover
##                          diff         lwr          upr     p adj
## Genre-Character -0.008811645 -0.01677315 -0.000850142 0.0300747
##
## $Summary
##                          diff         lwr         upr     p adj
## Genre-Character -0.006087055 -0.01404856 0.001874448 0.1339517
##
## $`Cover:Summary`
##                                           diff          lwr           upr
## Genre:Character-Character:Character -0.02128739 -0.036047699 -0.0065270815
## Character:Genre-Character:Character -0.01856280 -0.033323108 -0.0038024911
## Genre:Genre-Character:Character     -0.01489870 -0.029659009 -0.0001383913
## Character:Genre-Genre:Character      0.00272459 -0.012035718  0.0174848992
## Genre:Genre-Genre:Character          0.00638869 -0.008371618  0.0211489989
## Genre:Genre-Character:Genre          0.00366410 -0.011096209  0.0184244085
##                                         p adj
## Genre:Character-Character:Character 0.0012203
## Character:Genre-Character:Character 0.0067932
## Genre:Genre-Character:Character     0.0468911
## Character:Genre-Genre:Character     0.9647097
## Genre:Genre-Genre:Character         0.6818224
## Genre:Genre-Character:Genre         0.9196953
```

The Tukey test confirms that the cover is statistically significant, as a cover of Character is predicted to decrease the mean for the engagement level approximately by 0.0088. The test further confirms that the Summary is not a significant effect.

An analysis of the interaction effects by looking at the comparison of all the different pairs indicates that there exist three statistically significant cases (p-value is less than 0.05). The three significant cases all include a comparison with one of them having the Character as both Cover and Summary, and the expected engagement level is lower for all cases. However, since the other cases are not significant, we can not determine which of the combinations is predicted to yield the highest engagement level. We can therefore only conclude that having the Character in both the Cover and Summary, will likely reduce the engagement level.

Conclusion - Both the Cover and the interaction effect between the Cover and Summary are statistically significant. The pairwise analysis from the Tukey test suggested that any combination other than having the character in both the Cover and Summary will give a higher engagement level.

Contributions

All four of us attended the lab. During the lab, we were able to go through the questions of the first task, discussed them and wrote down our first findings. We could not solve all aspects and only had a short look at task 2, so we split up the group: Wenli and Tim concentrated on the first task and Ludvig and Eric on the second task. After working on these parts separately, all of us checked the whole document again, so that everybody knows about all final results.