Project II

Dăscălescu Dana

June 23, 2022

The code is based on the Laboratories work and was supported by the following online resources: [2] [1] [4] [5].

1. Describe at least one experiment where two-way ANOVA (with two between-subjects factors) can be used for data analysis. Data can be generated by you or it can be real data. Specify the factors, the levels and interpret results obtained by applying two-way ANOVA.

   Perform the post-hoc analysis using the appropriate t_tests (this analysis must be performed, even if the result of indicates no differences in your data, just for the sake of practice). For two of the groups in your data, perform in addition a permutation test and comment on the result (compared to the t_test).

   The above experiment is run on the **datarium** [3] package's 'jobsatisfaction' dataset, which contains *58 rows* and *3 columns*: gender, education level and job satisfaction scores.

   The goal of this experiment is to see if there is an interaction between **education level** and **gender** (factors, independent variables) on **job satisfaction score** (dependent variable). In this case, we can see if the introduction term indicates whether the effect of education level (of three possible levels: high school, college, and university) on job satisfaction is influenced by gender (of two possible levels: male, female).

   ### Exploratory Data Analysis

   - **Check how many participants are in each group** We will examine the number of samples for each group and visualise a random sample for each group.

```
> sample_n_by(dataframe.jobsatisfaction, education_level, gender, size=1)
# A tibble: 6    4
  id    gender education_level score
  <fct> <fct>  <fct>           <dbl>
1 9     male   school           4.78
2 32    female school           5.65
3 16    male   college          6.09
4 39    female college          6.52
5 20    male   university       9.28
6 55    female university       8.55
```

```
        school college university
male         9       9         10
female      10      10         10
```

Figure 1: Caption

Each group does not have the same number of people (unbalanced experiment, see Figure 1).

- **Summary statistics**

```
> summary(dataframe.jobsatisfaction)
       id         gender     education_level      score
 1      : 1   male   :28   school     :19     Min.   : 4.780
 2      : 1   female:30   college    :19     1st Qu.: 5.800
 3      : 1               university:20     Median : 6.380
 4      : 1                                 Mean   : 6.963
 5      : 1                                 3rd Qu.: 8.515
 6      : 1                                 Max.   :10.000
 (Other):52
```

Calculate the mean and standard deviation (SD) of the satisfaction score by group:

```
# A tibble: 6 6
  gender education_level variable     n  mean    sd
  <fct>  <fct>           <chr>    <dbl> <dbl> <dbl>
1 male   school          score        9  5.43 0.364
2 male   college         score        9  6.22 0.34
3 male   university      score       10  9.29 0.445
4 female school          score       10  5.74 0.474
5 female college         score       10  6.46 0.475
6 female university      score       10  8.41 0.938
```

- **Visuzalization**

  We will create a box plot of the score grouped by gender and education level (see Figure 2). We can also observe that there are no significant outliers.

Check for assumptions of the ANOVA test

The Analysis of Variance Anova test makes the following data assumptions, which must be met before performing any computation:

- **Outliers** At this step, we are interested in finding any data point that differs significantly from the other observations.

```
> identify_outliers(data=groups, score)
[1] gender          education_level id          score       is.outlier
[6] is.extreme
<0 rows> (or 0-length row.names)
```

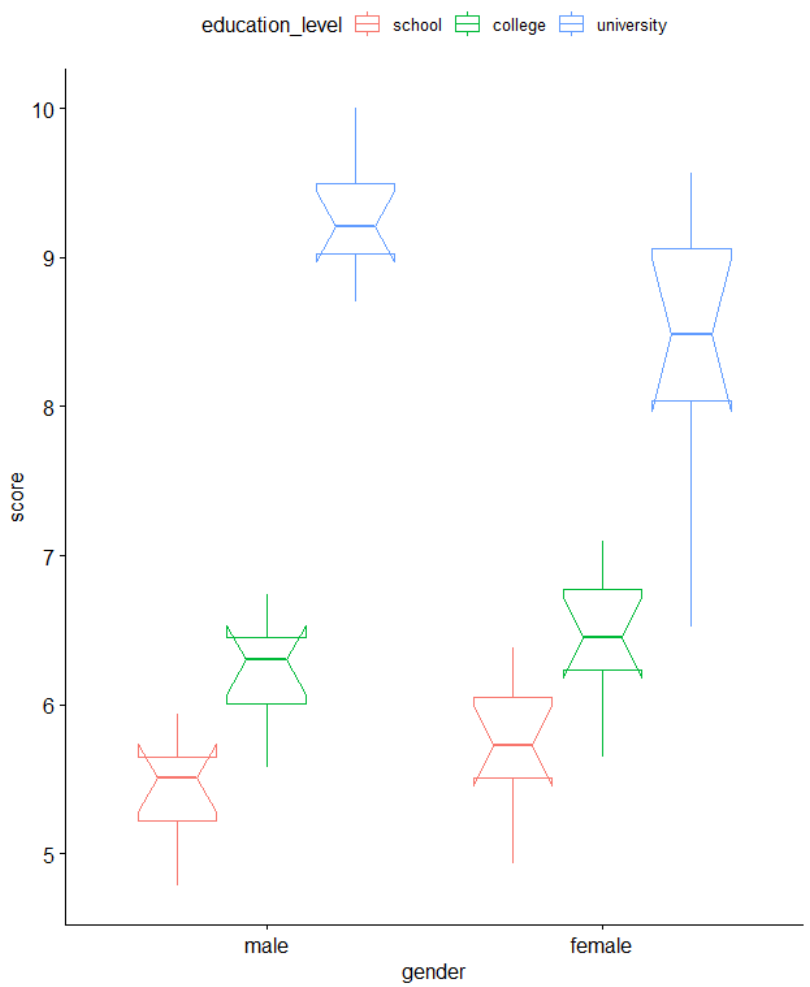

Figure 2: Boxplot of the job satisfaction score grouped by gender and education level

The analysis determined that there are no outliers in our data.

- **Normality**
  Analyze the model residuals to test the normality assumption. The QQ plot and the Shapiro-Wilk test of normality are employed.

  In the first step, we will build a linear model and then create a *QQ* plot of residuals.
  We can assume normality in the QQ plot because all of the points fall roughly along the reference line (see Figure 4). In order to support this conclusion, we will also compute the Shapiro-Wilk test.

```
> linear.model <- lm(score ~ education_level * gender, jobsatisfaction)
> res <- residuals(linear.model)
> ggqqplot(res)
> shapiro.test(res)
```

```
        Shapiro-Wilk normality test

data:  res
W = 0.96786, p-value = 0.1267
```

The obtained $p - value$ is 0.127, which is greater than the choosen $alpha = 0.05$, so we can assume normality since the null hypothesis H_0 (that the data come from a normally distributed population) can not be rejected.

In the following, we will compute the Shapiro-Wilk test for each factor level combination and visualize the QQ plots for each one (See Figure 3). The following code is based on [2]:

```
> shapiro_test(data = groups, score)
# A tibble: 6  5
  gender education_level variable statistic     p
  <fct>  <fct>           <chr>        <dbl> <dbl>
1 male   school          score        0.980 0.966
2 male   college         score        0.958 0.779
3 male   university      score        0.916 0.323
4 female school          score        0.963 0.819
5 female college         score        0.963 0.819
6 female university      score        0.950 0.674
```

According to Shapiro-Wilk's test, the score is normaly distributed for each group.

- **Homogeneity of variances**

  As stated in the lecture (*Lecture 10, Slide 7*), ANOVA requires the assumption of variance homogeneity. We will perform the Barlett test and Levene test for homogeneity.

  The Levene's test is not statistically significant ($p - value > 0.05$). The null hypothesis of Levene's test is that all groups have equal variances. Levene's alternative hypothesis is that at least one pair of groups has unequal variances. As a result, we can assume that the variances in the different groups are homogeneous. Instead the Barlette test is statistical significant (p-value = 3.77e-15). This results are due to the fact that the Levene test is less sensitive to deviations from normality than the Bartlett test.

```
d <- list(as.numeric(jobsatisfaction$gender),
          as.numeric(jobsatisfaction$education_level),
          as.numeric(jobsatisfaction$score))
bartlett.test(d)
```

```
> levene_test(dataframe.jobsatisfaction, score ~ education_level * gender)
# A tibble: 1  4
    df1   df2 statistic      p
  <int> <int>     <dbl>  <dbl>
1     5    52      2.20 0.0686
```

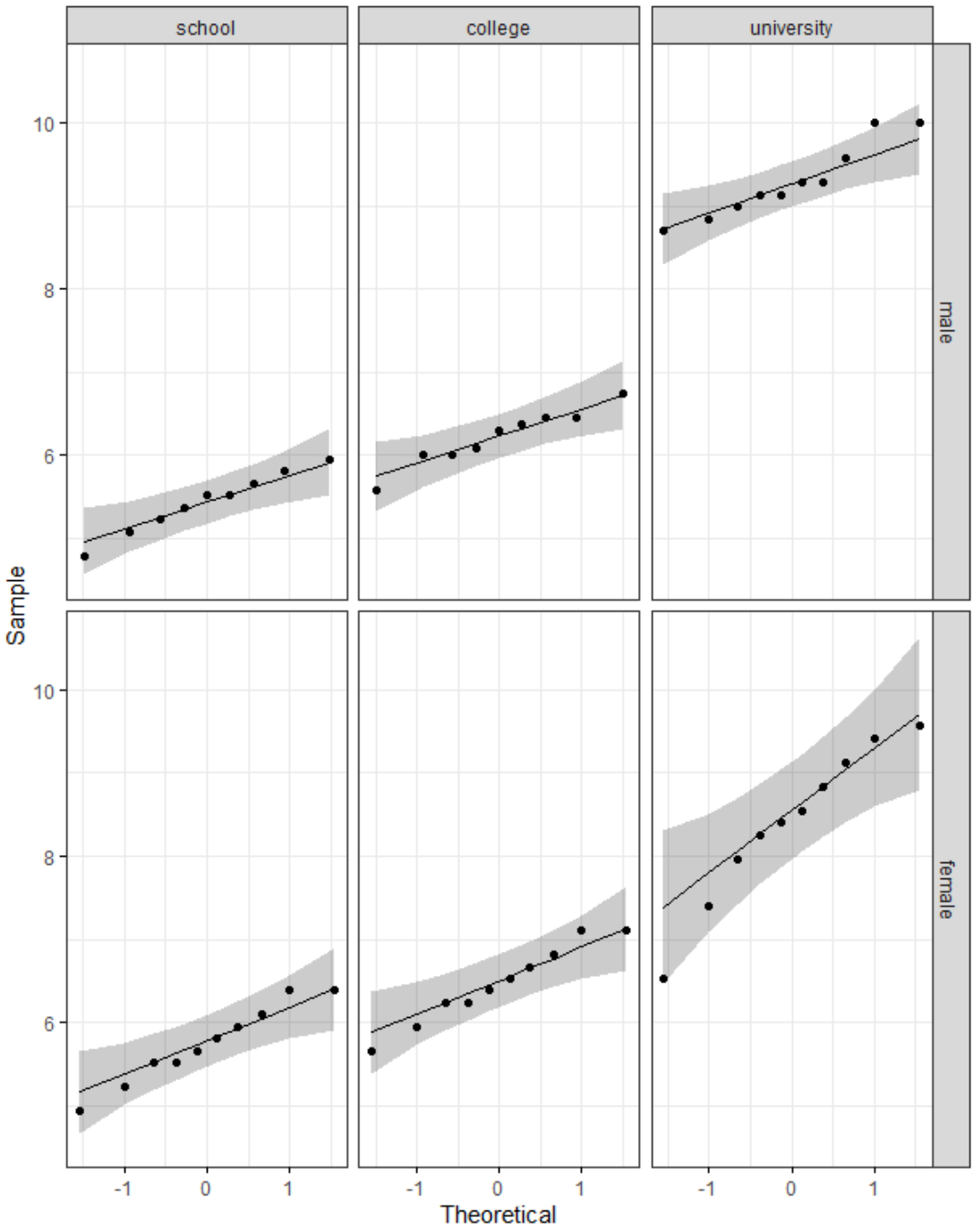

Figure 3: Q-Q plots for comparing the probabilities distributions of each factor level

Two-way (with two between-subjects factors) ANOVA

```
a <- aov(data = jobsatisfaction, formula =  score ~ gender * education_level)
summary(a)
```

```
                        Df Sum Sq Mean Sq F value  Pr(>F)
```

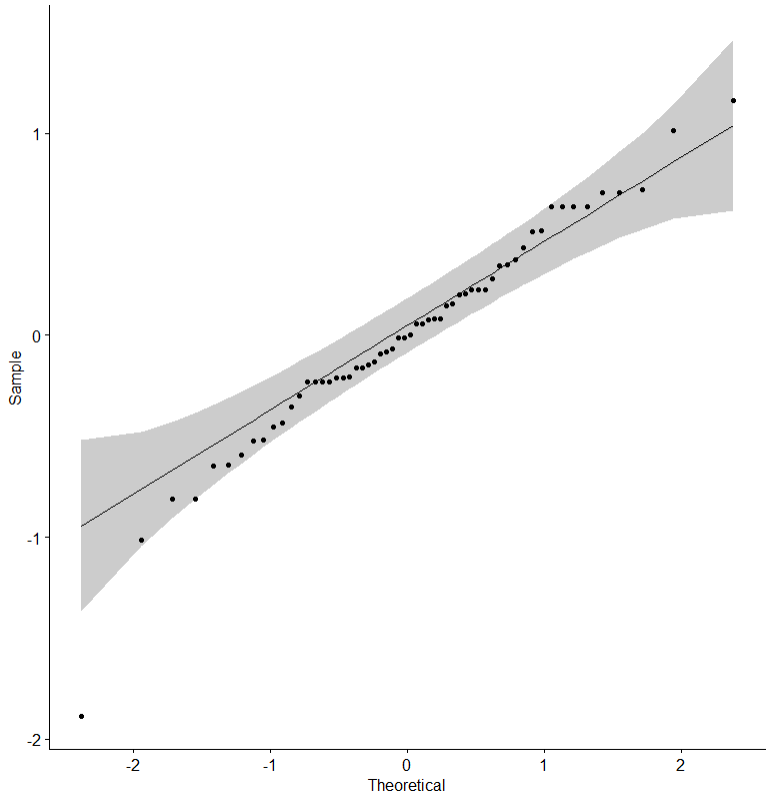

Figure 4: QQplot of the residuals

gender	1	0.54	0.54	1.787	0.18709	
education_level	2	113.68	56.84	187.892	< 2e−16	***
gender:education_level	2	4.44	2.22	7.338	0.00156	**
Residuals	52	15.73	0.30			

Signif. **codes**: 0 *** 0.001 ** 0.01 * 0.05 . 0.1

The row "gender correspond to SS_α and H_α. The row "education_level" corresponding to SS_β and H_β. The row "gender:educatioon_level" correspond to SS_γ and H_γ.

$Pr(> F)$ are the $p - values$ of the $F - test$. According to the values in the above table, we reject H_β and H_γ, but we fail to reject H_α. This means that the factor "education_level" has a significance influence (effect) on the satisfaction level of the employees and that the effect of education_level on job satisfaction is strongly influenced by gender.

Post-hoc analysis

- **Simple main effect** We will investigate the effect of education level factor at each level of gender factor. We will use the overall error term (from the two-way ANOVA) as input in the one-way ANOVA model because we have met the assumptions of the two-way ANOVA.

```
> anova_test(group.gender, score ~ education_level, error=linear.model)
Coefficient covariances computed by hccm()
Coefficient covariances computed by hccm()
# A tibble: 2    8
  gender Effect           DFn   DFd     F         p `p<.05`   ges
* <fct>  <chr>          <dbl> <dbl> <dbl>     <dbl> <chr>   <dbl>
1 male   education_level     2    52 132.  3.92e-21 *       0.836
2 female education_level     2    52  62.8 1.35e-14 *       0.707
```

For both men and women, the simple main effect of "education level" on job satisfaction score was statistically significant.

- **Comparisons using t-test**

  As all $p-value$ are greater than 0.05, we accept (fail to reject) the null hypotheses of normality:

```
> shapiro.test(group1$score)

        Shapiro-Wilk normality test

data:  group1$score
W = 0.98043, p-value = 0.9664

> shapiro.test(group2$score)

        Shapiro-Wilk normality test

data:  group2$score
W = 0.95815, p-value = 0.7788

> shapiro.test(group3$score)

        Shapiro-Wilk normality test

data:  group3$score
W = 0.91572, p-value = 0.3226

> shapiro.test(group4$score)

        Shapiro-Wilk normality test

data:  group4$score
W = 0.96292, p-value = 0.8185

> shapiro.test(group5$score)

        Shapiro-Wilk normality test

data:  group5$score
```

```
W = 0.96295, p-value = 0.8189

> shapiro.test(group6$score)

        Shapiro-Wilk normality test

data:  group6$score
W = 0.95044, p-value = 0.6737
```

Now, we will compare the means of all the combinations of the groups (in total 15). We failed to reject the hypothese that the true difference in means is equal to 0 for the following groups:

- *group*1 and *group*4 (male participants who graduated school and female particpants who graduated school, respectively), as the $p - value = 0.1267$;
- *group*2 and *group*5 (male participants who graduated college and female participants who graduated college, respectively), as the $p - value = 0.2275$

- **Permutation test for two groups from our data**

  We will select *group*1 and *group*6 for this step because they failed the previous test. The average in *group*6 is higher than the average in *group*1, and we want to know if this difference is due to chance or statistically significant.

  We can observe in Figure 5 that the observed difference between the two groups is not due to chance (the red line does not lie well within the permuted values), thus is statistically significant.

2. Linear regression model for at least 2 causes.

In this exercise, we will use the 'marketing' dataset from the **datarium** package [3]. The dataset contains 200 samples, each containing the effect of three advertising media on sales (youtube, facebook, and newspaper). Data include advertising budgets in the thousands of dollars as well as sales data [3].

The goal is to predict the dependent variable,'sales,' using two independent variables, 'youtube' and 'facebook.' We want to look at the relationship between the budgets set aside for social media advertising and sales.

Exploratory Data Analysis

- **Descriptive statistics**

  We generate descriptive statistics about our dataset to summarize its central tendency, dispersion, and shape.

```
> summary(marketing)
    youtube          facebook        newspaper          sales
 Min.   :  0.84   Min.   : 0.00   Min.   :  0.36   Min.   : 1.92
 1st Qu.: 89.25   1st Qu.:11.97   1st Qu.: 15.30   1st Qu.:12.45
```

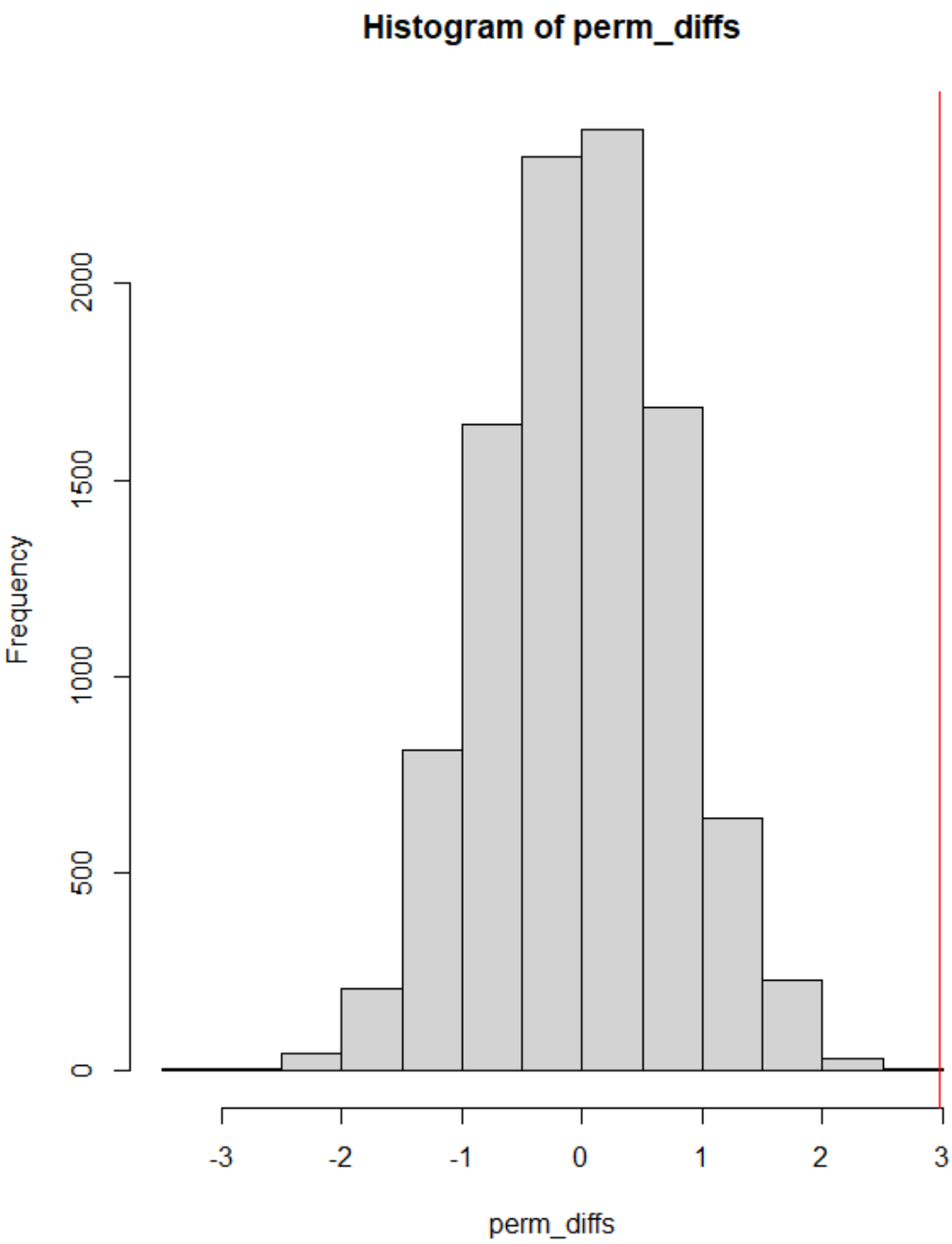

Figure 5: Permutation test

Median :179.70	Median :27.48	Median : 30.90	Median :15.48
Mean :176.45	Mean :27.92	Mean : 36.66	Mean :16.83
3rd Qu.:262.59	3rd Qu.:43.83	3rd Qu.: 54.12	3rd Qu.:20.88
Max. :355.68	Max. :59.52	Max. :136.80	Max. :32.40

- **Exploring the features**

In Figure 6, we can see the relationship between the variables, such as the close linearity relationship between the variables *youtube* and *sales*.

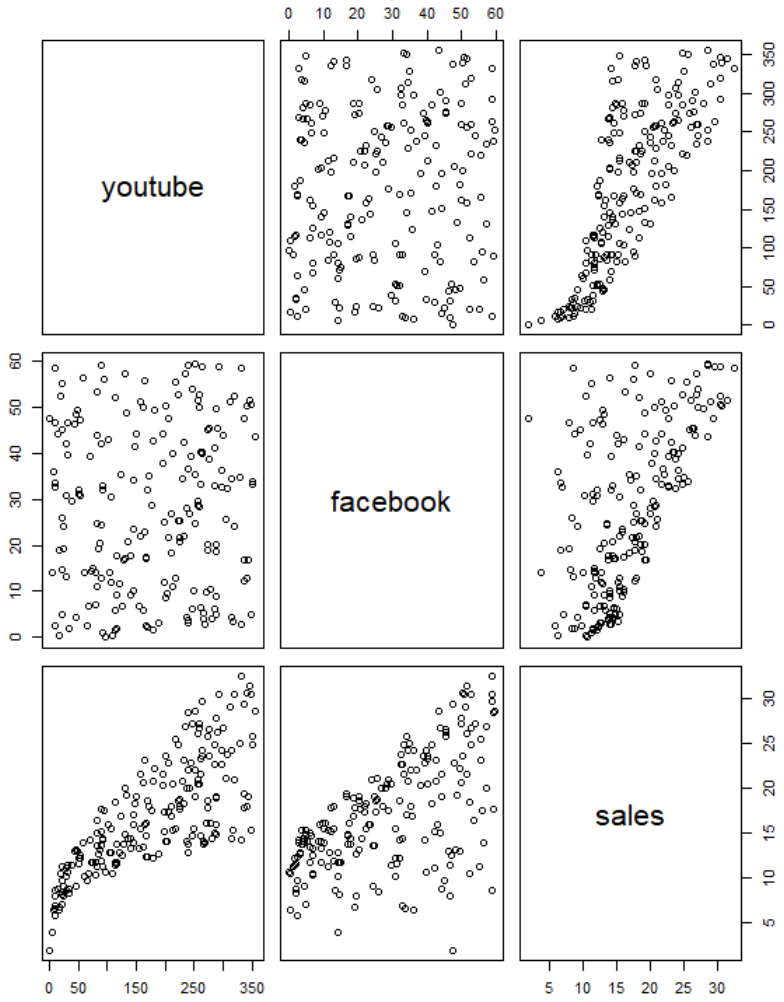

Figure 6: Matrix plot for *Marketing* dataset

Linear Regression Model

We will build a linear regression model (regression plane) for modeling the relation between the *"causes"*, "youtube" and "facebook" independent variables, and "effect", "sales" dependent variable.

```
x <- marketing[,"sales"]
y1 <- marketing[,"youtube"]
y2 <- marketing[,"facebook"]

model = lm(x ~ y1 + y2)

> model
```

```
Call:
lm(formula = x ~ y1 + y2)

Coefficients:
(Intercept)          y1          y2
    3.50532     0.04575     0.18799
```

We obtain that:

- the estimation for β_0 is 3.50532;
- the estimation for β_1 is 0.04575;
- the estimation for β_2 is 0.18799;
- the regression plane is $x = 3.50532 + 0.04575 \cdot y_1 + 0.18799 \cdot y_2$

The summary of the model is shown below:

```
> summary(model)

Call:
lm(formula = x ~ y1 + y2)

Residuals:
     Min       1Q   Median       3Q      Max
-10.5572  -1.0502   0.2906   1.4049   3.3994

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.50532    0.35339   9.919   <2e-16 ***
y1           0.04575    0.00139  32.909   <2e-16 ***
y2           0.18799    0.00804  23.382   <2e-16 ***
---
Signif. codes:  0  ***  0.001  **  0.01  *  0.05  .  0.1

Residual standard error: 2.018 on 197 degrees of freedom
Multiple R-squared:  0.8972,    Adjusted R-squared:  0.8962
F-statistic: 859.6 on 2 and 197 DF,  p-value: < 2.2e-16
```

We fail to reject the overall F-test of the model ($H_0 : \beta_1 = \beta_2 = 0$, since the $p_value = 2.2e-16 < 0.05$.

The coefficients are also tested individually, $H : \beta_1 = 0$, respectively $H : \beta_2 = 0$ with $T - tests$:

- In the row "y_1" above, the $p - value$ is $2e - 16 \Rightarrow H : \beta_1 = 0$ is rejected
- In the row "y_2" above, the $p - value$ is $p - value$ is $2e - 16 \Rightarrow H : \beta_2 = 0$ is rejected

This analysis indicates that the both causes y_1 and y_2 should be retained in the model.

Goodness-of-fit

Now we'll look at how well our model fits our data. Our analysis will be based on the residuals, which are as follows:

- First measure explored is that of "*variance explained*": R^2.

```
> model <- lm(x ~ y1 + y2)
> summary(model)

Call:
lm(formula = x ~ y1 + y2)

Residuals:
     Min       1Q   Median       3Q      Max
-10.5572  -1.0502   0.2906   1.4049   3.3994

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.50532    0.35339   9.919   <2e-16 ***
y1           0.04575    0.00139  32.909   <2e-16 ***
y2           0.18799    0.00804  23.382   <2e-16 ***
---
Signif. codes:  0  ***  0.001  **  0.01  *  0.05  .  0.1

Residual standard error: 2.018 on 197 degrees of freedom
Multiple R-squared:  0.8972,    Adjusted R-squared:  0.8962
F-statistic: 859.6 on 2 and 197 DF,  p-value: < 2.2e-16
```

The value of *Multiple* $R - squared$ is 0.8972, indicating that 89.72% of variability can be explained through the linear model. Finally, by this metric, the model fits the data well.

- Normality of residuals:

```
> shapiro.test(residuals(model))

        Shapiro-Wilk normality test

data:  residuals(model)
W = 0.91804, p-value = 4.19e-09
```

We obtained a $p - value$ of $4.19e - 09$ after analyzing the model residuals to test the normality assumption, so we reject the null hypothesis that the residuals are normally distributed.

Only with small sample size does the violation of the normality assumption become an issue. Because of the central limit theorem and the fact that the F- and t-tests used for hypothesis testing and forming confidence intervals are quite resistant to modest deviations from normality, the assumption is less significant for large sample sizes.

- Homoscedasticity - constant residual variance throughout the range of the predicted values:

In the Figure 7, there is no greater variance in residuals despite the presence of certain outliers, implying that the homoscedasticity criteria is met.

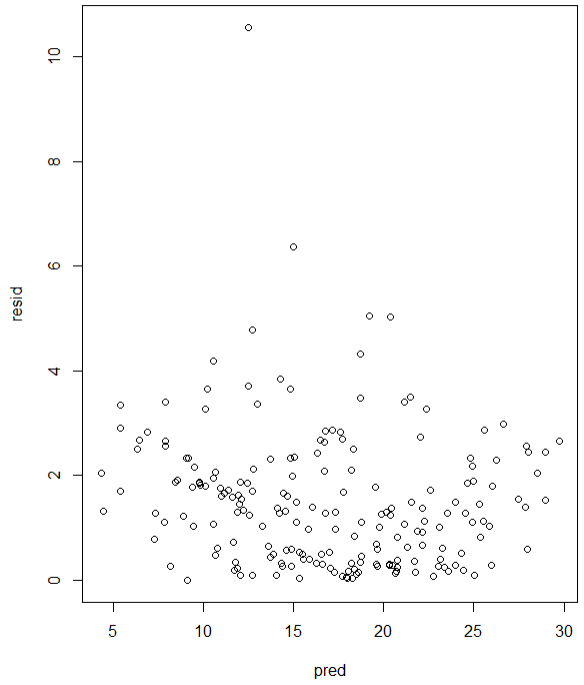

Figure 7: Homoscedasticity

Our analysis shows that the model does not fit the data well, which may indicate that there is a "cause" missing from the linear equation ($cause -> effect$), implying that other models are worth considering.

We also examined the variable 'newspaper' as a cause (see Appendix), however the second model performed just slightly better, indicating that a pre-processing stage before to building the model may be relevant.

References

[1] *A guide to using post hoc tests with ANOVA Retrieved June 23, 2022, from https://www.statology.org/anova-post-hoc-tests/".*

[2] *Comparing Multiple Means In R. Retrieved June 21, 2022, from https://www.datanovia.com/en/lessons/anova-in-r/.*

[3] Alboukadel Kassambara. *Data Bank for statistical analysis and visualization [R package datarium version 0.1.0]. The Comprehensive R Archive Network. Retrieved June 21, 2022, from https://CRAN.R-project.org/package=datarium.* URL: `https://CRAN.R-project.org/package=datarium`.

[4] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2013. URL: `http://www.R-project.org/`.

[5] Antoine Soetewey. *ANOVA in R. Retrieved June 23, 2022, from https://statsandr.com/blog/anova-in-r/.*

Appendix

```
library(tidyverse)
library(ggplot2)
library(rstatix)
library(ggpubr)

set.seed(0)

# 1 b)
data("jobsatisfaction", package = "datarium")
dataframe.jobsatisfaction <- data.frame(jobsatisfaction)
head(dataframe.jobsatisfaction)
summary(dataframe.jobsatisfaction)
sample_n_by(dataframe.jobsatisfaction, education_level, gender, size = 1)

groups <- group_by(dataframe.jobsatisfaction, gender, education_level)
get_summary_stats(groups, score, type = "mean_sd")

ggboxplot(dataframe.jobsatisfaction, x = "gender", y = "score", notch = TRUE,
          color="education_level")

identify_outliers(data=groups, score)

linear.model <- lm(score ~ education_level * gender, data = dataframe.jobsatisfaction
res <- residuals(linear.model)

ggqqplot(res)
shapiro.test(res)
shapiro_test(data=groups, score)

ggqqplot(data = dataframe.jobsatisfaction, x = "score", ggtheme = theme_bw()) + facet

d <- list(as.numeric(dataframe.jobsatisfaction$gender),
          as.numeric(dataframe.jobsatisfaction$education_level),
          as.numeric(dataframe.jobsatisfaction$score))
bartlett.test(d)

levene_test(dataframe.jobsatisfaction, score ~ education_level * gender)

a <- aov(data = jobsatisfaction, formula =  score ~ gender * education_level)
summary(a)

group.gender <- group_by(jobsatisfaction, gender)
group.education_level <- group_by(jobsatisfaction, education_level)
groups['gender'] == 'male', groups['education_level'] == 'school', 'score']
anova_test(group.gender, score ~ education_level, error=linear.model)
```

```
group1 <- groups[groups['gender'] == 'male' &
    groups['education_level'] == 'school', 'score']
group2 <- groups[groups['gender'] == 'male' &
    groups['education_level'] == 'college', 'score']
group3 <- groups[groups['gender'] == 'male' &
    groups['education_level'] == 'university', 'score']
group4 <- groups[groups['gender'] == 'female' &
    groups['education_level'] == 'school', 'score']
group5 <- groups[groups['gender'] == 'female' &
    groups['education_level'] == 'college', 'score']
group6 <- groups[groups['gender'] == 'female' &
    groups['education_level'] == 'university', 'score']

shapiro.test(group1$score)
shapiro.test(group2$score)
shapiro.test(group3$score)
shapiro.test(group4$score)
shapiro.test(group5$score)
shapiro.test(group6$score)

t.test(group1, group2, alternative="two.sided", mu=0, paired=FALSE, var.equal=TRUE)
t.test(group1, group3, alternative="two.sided", mu=0, paired=FALSE, var.equal=TRUE)
t.test(group1, group4, alternative="two.sided", mu=0, paired=FALSE, var.equal=TRUE)
t.test(group1, group5, alternative="two.sided", mu=0, paired=FALSE, var.equal=TRUE)
t.test(group1, group6, alternative="two.sided", mu=0, paired=FALSE, var.equal=TRUE)
t.test(group2, group3, alternative="two.sided", mu=0, paired=FALSE, var.equal=TRUE)
t.test(group2, group4, alternative="two.sided", mu=0, paired=FALSE, var.equal=TRUE)
t.test(group2, group5, alternative="two.sided", mu=0, paired=FALSE, var.equal=TRUE)
t.test(group2, group6, alternative="two.sided", mu=0, paired=FALSE, var.equal=TRUE)
t.test(group3, group4, alternative="two.sided", mu=0, paired=FALSE, var.equal=TRUE)
t.test(group3, group5, alternative="two.sided", mu=0, paired=FALSE, var.equal=TRUE)
t.test(group3, group6, alternative="two.sided", mu=0, paired=FALSE, var.equal=TRUE)
t.test(group4, group5, alternative="two.sided", mu=0, paired=FALSE, var.equal=TRUE)
t.test(group4, group6, alternative="two.sided", mu=0, paired=FALSE, var.equal=TRUE)
t.test(group5, group6, alternative="two.sided", mu=0, paired=FALSE, var.equal=TRUE)

mean(group1$score)
mean(group6$score)
initial_difference <- mean(group6$score) - mean(group1$score)

permutation_test_func <- function(x, sizeA, sizeB)
{
  n <- sizeA +sizeB
  idx_b <- sample(1:n, sizeB)
  idx_a <- setdiff(1:n, idx_b)
  return(mean(x[idx_b]) - mean(x[idx_a]))
}
```

```r
perm_diffs <- rep(0, 10000)
x <- c(group1$score, group6$score)
for(i in 1:10000)
  perm_diffs[i] = permutation_test_func(x, length(group1$score), length(group6$score))

hist(perm_diffs)
abline(v=initial_difference, col = 'red')


# 2)
data(marketing, package="datarium")
summary(marketing)

df <- marketing[, !(colnames(marketing) %in% c("newspaper"))]
plot(df)

x <- marketing[,"sales"]
y1 <- marketing[,"youtube"]
y2 <- marketing[,"facebook"]
y3 <- marketing[, "newspaper"]

model = lm(x ~ y1 + y2)
model
summary(model)

model <- lm(x ~ y1 + y2)
summary(model)

shapiro.test(residuals(model))

residuals1 <- abs(residuals(model))
predictions1 <- predict(model)
plot(predictions1, residuals1)

model2 = lm(x ~ y1 + y2 + y3)
model2

summary(model2)

model2 <- lm(x ~ y1 + y2 + y3)
summary(model2)

shapiro.test(residuals(model2))

residuals2 <- abs(residuals(model2))
predictions2 <- predict(model2)
plot(predictions2, residuals2)
```