Final report

Business Analytics
Course: Project Big data (X_400645).

Students:

Yannick Hogebrug – 2626424 - y.r.hogebrug@student.vu.nl
Jesse Schouten – 2621562 - j7.schouten@student.vu.nl

Submission date: 30-06-2018

Introduction

This study was focused on the data analysis of a bedtime procrastination study. Bedtime procrastination is related to sleep problems, to which severe outcomes including memory and health problems have been related by several studies (Kroese et al., 2014). By analyzing personal sleeping data, as well as poststudy questionnaire data filled out by the participant, the following questions were attempted to be answered:

- *Can bedtime procrastination be significantly influenced by experiment?*
- *How well can bedtime procrastination be predicted?*

These questions will be answered during this report. Furthermore, a few visualizations will provide extra insight into the data.

Data description and exploration

A total of 5 participants didn't fill in the questionnaire, so they were excluded from the data in the merged dataframe. The available variables are shown below including a short explanation if necessary:

- *Gender:* male (=1) or female (=2).
- *Age:* years of age of the participant.
- *Chronotype:* 7 point scale if you are more a morning person (1) or an evening person (7).
- *Bp_scale:* bed procrastination scale; the higher, the more procrastination.
- *Motivation:* going to bed on time each night (1 = not motivated, 7 = very motivated).
- *Daytime_sleepiness:* 4-point scale from 0-3; 8 questions, values summed.
- *Self_reported_effectiveness:* do you feel more rested since the intervention (range 0-7).
- *Group:* control- (0) or experimental group (1).
- Delay*_nights:* number of nights a participant delayed their bedtime (range 0-12).
- *Delay_time:* mean time in seconds a participant delayed their bedtime .
- *Sleep_time:* the mean bedtime in seconds.

One of the targets of this research is to predict the variable ´delay_time', as this answers the second research question given in the introduction.

To get an impression of all variables, some descriptive statistics were calculated, which are shown in Table 1.

	count	mean	standard deviation	median	min	max
gender	42	1,5714	0,4949	2	1	2
age	42	31,7381	12,1500	27	18	61
chronotype	42	4,9762	1,8450	5	1	7
bp_scale	42	5,0690	0,9051	5.165	2,67	6,67
motivation	42	4,4524	1,1170	5	1	6
daytime_sleepiness	42	16,0476	3,8480	16	8	26
self_reported_effectiveness	42	2,6190	1,3619	2	1	6
group	42	0,4524	0,4977	0	0	1
delay_nights	42	7,2143	3,2699	8	0	12
delay_time	38	2354,6842	1438,2945	1974	0	5482
sleep_time	38	28822,3158	2848,8786	29190	21929	34644

Table 1: Descriptive statistics of the variables in the dataset

The count stands out in Table 1 as this is not the same for every variable. This is caused by empty cells for the variables 'delay_time' and 'sleep_time'. For the analysis of the variable(s) with an empty cell, it was decided to remove the corresponding row. So, for some tests more data might have been used then others, as it depends on the selected variables.

At least one strange value is spotted: the maximum of the daytime sleepiness, which is equal to 26. This variable is measured on a 4 point scale from 0 to 3, containing 8 questions. This means it could be at most 24. A total of 2 observations turned out to be larger than 24. Despite this, it was decided to include all observations in the analyses. Also, the maximum of age and the maximum of motivation seem to be relatively high compared to the mean and median, indicating there might be some outliers. As these variables didn't turn out to be relevant in this report, this wasn't further investigated.

As part of the data exploration, the delay time between the participant groups was analyzed using Figure 1. The only difference between the groups is the fact that for the experimental group, the lights automatically dim at the intended bedtime. Some difference is spotted in Figure 1, but it is not clear whether this is significant. The lack of observations is, and has to be, kept in mind.

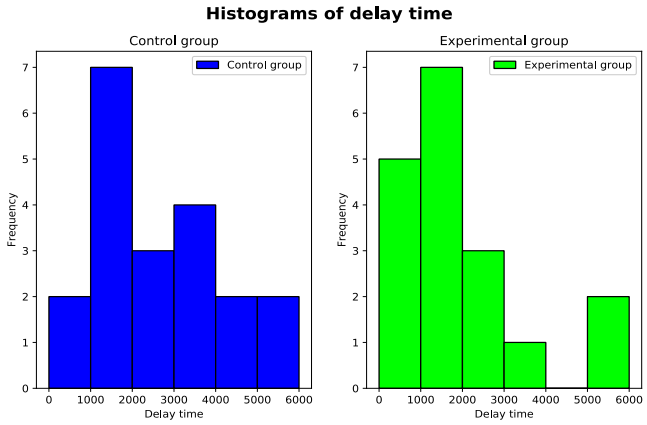

Figure 1: Delay time per participant group

Data analysis

To check whether the experiment had significant effect on sleeping behavior, an analysis of the sleeptime and delay time variables was performed by comparing both the experimental group and the control group.

To determine which test to use, QQ-plots were made, which are shown in Figure 2. These show almost no rough straight lines, except maybe for the sleep time in the control group. However, the lack of observations makes this also doubtful. It indicates normality can't be assumed for the statistical tests: the variables are not from the same location scale family as the normal distribution (University of Iowa, 2016). For this reason, the Wilcoxon rank sum test was used as this is a distribution free test. The Wilcoxon rank sum test has the following hypotheses:

- H0: F=G
- H1: F≠G

The p-values of the performed tests are shown in Table 2. None of the p-values indicate significance, although the delay time is at the verge of rejecting at a 5% level .

	P-value
Delay nights	0.7425
Sleep time	0.8838
Delay time	0.05367

Table 2: Results of Wilcoxon signed rank test for difference in groups

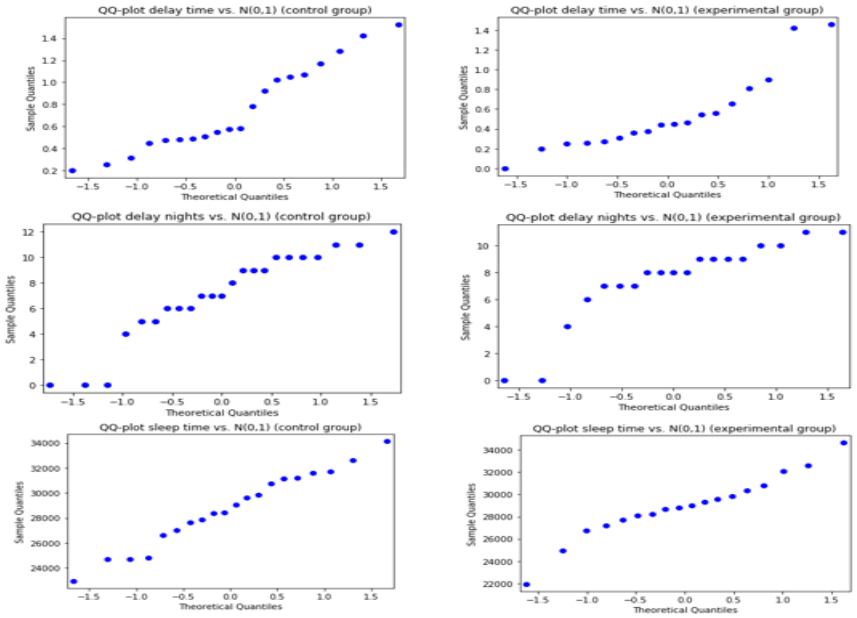

Figure 2: QQ-plots of delay time, delay nights & sleep time per group against N(0,1)

In order to get some insight in correlation between delay time and other variables, we made scatterplots, shown in Figure 3.

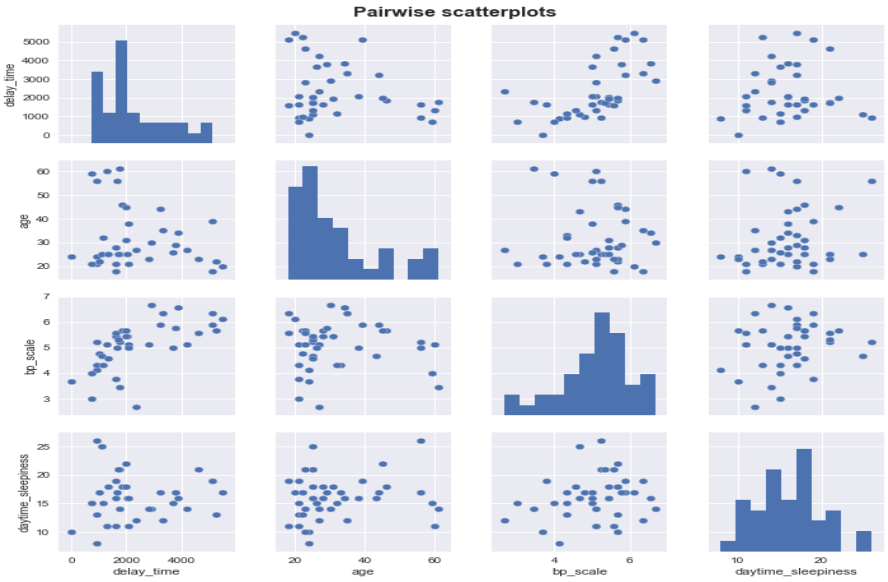

Figure 3: Pairwise plots of delay_time vs. three other variables

The variables 'Delay time' and 'bp_scale' seems to be correlated the most, as there could be drawn a straight line relatively easy through the cloud of points. Kendall's correlation coefficient and Pearson's correlation coefficient were used to test for correlation. The values of these coefficients are at least -1 and at most 1, indicating a perfect negative or positive correlation respectively. A value close to 0 indicates no correlation. It is noted Pearson just checks linear correlation, and Kendall's checks more kinds of relationships (Statistics solutions, 2018).

	Correlation coefficient (r)
Delay_time vs bp_scale (Pearson)	0.6118
Delay_time vs age (Kendall)	-0.02746
Delay_time vs daytime_ sleepiness (Pearson)	0.08328

Table 3: Correlation tests

Table 3 shows that the correlation coefficient between delay time and bp scale is the highest. For the other two variables, the values are pretty close to 0.

For the multiple regression model, information of prior analysis was used as much as possible to predict delay time. This resulted in a model with the following explanatory variables:

- ***bp_scale:*** a relative high correlation with delay time was found earlier.
- ***group:*** a possible statistical significance was found as a p-value just slightly higher than 0.05 can be seen in Table 2.
- ***chronotype:*** by context it would make sense evening people are more likely to delay bedtime.

Figure 4 provides some more insight in the relationship between delay time, chronotype and group. The fact that the blue points seem to lay relatively higher than the orange points in the plot stands out. Also, the delay time seems to be higher on average as the chronotype rises.

The final linear model including the three mentioned variables resulted in a R^2-value of 0.475.

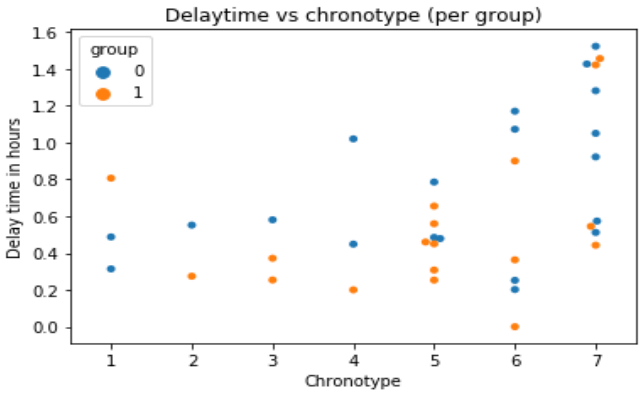

Figure 4: Plot of remaining variables in multiple linear regression model

Discussion

The focus was on two research questions, as mentioned in the introduction section. To answer these questions, data was analyzed using some statistical tests and coefficients.

The first question was about the differences between the two investigated groups to measure the influence of the experimental setup. This was tested with the help of the Wilcoxon rank-sum test. For all the tested variables, the null hypothesis was rejected. However, the p-value of the variable 'delay-time' was on the verge of rejecting by a significance level of 5%. Because of the small number of observations, it is difficult to draw strong conclusions.

Besides this, correlation between delay time and three other variables was analyzed. The one that striked out was the variable 'bp_scale' with an correlation value of 0.61. Using the guide that Evans (1996) suggested for the absolute value of the correlation coefficient r, this correlation was considered to be strong. As shown in the data analysis part, the other variables were not highly correlated.

In the continuation of the study, a linear regression model was constructed to predict the delay time. Here, the obtained knowledge from the prior analysis was used. The variables bp_scale, group and chronotype were included in the model. This resulted in a R^2-value of 0.475, which means that 47,5% of the variance in the dependent variable (delay time) can be explained by the variance of the independent variables (bp_scale, group and chronotype). In contrast to the value of r, there is no specific 'good' or 'bad' value for R^2, this really depends on the context (Nau, n.d.).

Although the R^2-value did increase after adding the variables 'group' and 'chronotype', these variables didn't seem to have significant effect on the model (p-values of respectively 0.108 and 0.064). For a linear model it is always preferred to include as less variables as possible, which will

prevent problems like collinearity (Enders, n.d.). It seemed that only the variable bp_scale would be sufficient in the model.

Altogether, it is difficult to draw strong conclusions, based on this analysis. This is especially associated with the lack of observations. In further research, it is advised to collect data from more participants or combine the outcome of that research with the one discussed in this report. Furthermore, variables in the model were partially based on intuition. In future research one could for example use the step-up strategy for determining the 'best' linear model.
Finally, more data of the participants could be collected. The sleep behavior of someone could also be influenced by the extent of effort that day or the smartphone use before sleeping (Scutti, 2017).

Conclusion

The following conclusions can be drawn from this report:

- There are no clear reasons to assume that bedtime procrastination could be significantly influenced by experiment. Conducted tests did not show any significance between the groups, although one of them was on the verge of rejecting (p-value of 0.053).
- With the help of a linear regression analysis, it was concluded that bp_scale seemed to be sufficient for predicting the delay time. However, these model doesn't seem really strong, regarding the small R^2-value. It is difficult to draw a strong conclusion due to the lack of observations and because there is no real 'good' or 'bad' value for R^2, though.

References

Enders, F.B. (n.d.). *Collinearity.* Retrieved 27-06-2018, from https://www.britannica.com/topic/collinearity-statistics

Evans, J. D. (1996). *Straightforward statistics for the behavioral sciences.* Pacific Grove, CA: Brooks/Cole Publishing

Kroese, F. M., De Ridder, D. T., Evers, C., & Adriaanse, M. A. (2014). *Bedtime procrastination: introducing a new area of procrastination.* Frontiers in psychology, 5, 611.

Nau, R. (n.d.). *What's a good value for R-squared?* Retrieved 26-06-2018, from https://people.duke.edu/~rnau/rsquared.htm

Scutti, S. (2017). *Your smartphone may be hurting your sleep.* Retrieved 27-06-2017, from https://edition.cnn.com/2016/11/09/health/smartphones-harm-sleep/index.html

Statistics solutions. (2018). *Correlation (Pearson, Kendall and Spearman),* from http://www.statisticssolutions.com/correlation-pearson-kendall-spearman/

University of Iowa. (2016). *QQ-plots and PP-plots, from https://homepage.divms.uiowa.edu/~luke/classes/STAT4580/qqpp.html*