# Bayesian Data Analysis and Cognitive Modelling Final Project

Regina Gerber
rfunk@uni-osnabrueck.de

April 2019

## 1 Introduction

The concepts of Bayesian probabilistic modelling gain increasing popularity, in both, the scientific research community, and its commercial application. Although the Bayesian framework already exists for a while, only the rise of computational power allowed the practical application of it. The Bayesian framework depicts a powerful and robust concept for statistical analysis and simulation, compared to the widespread classical (frequentist) approach, i.e. it is not dependent on prerequisite assumptions about the nature of the parameters. The power of the Bayesian framework is settled within its general principles: first, the degree of uncertainty can be quantified and, second, observed data is used to update prior information. Both principles are explicitly included within the modelling process (Lee  Wagenmakers, 2005). In other words, while in the classical framework the parameters are assumed to be fixed, in the Bayesian view all unknown probabilities are treated as uncertain, and, therefore, are described by a probability distribution (Smeets  Schoot, 2018).

As a practice to my personally learned Bayesian analyzing techniques, I will explore, and analyze a commercial data set drawn from the website kaggle (Dagdoug, 2018) as a final project. Specifically, here the problem is a regression problem where I will try to predict the dependent variable (the amount of purchase) with the help of the information contained in the other variables. The aim is to deepen the understanding of the methods and ideas that were applied during the course of the seminar "Bayesian Data Analysis and Cognitive Modelling" (taught by Prof. Dr. Michael Franke in Wintersemester 2018/19 at the University Osnabrück) by creating a summary or practical manual to the course. I will conduct the analysis using the open-source statistical software R (version 3.4.3), in combination with the R Studio environment.

The course of the analysis will be following: First, I will explore the data. I will begin with some visualization. Subsequently, I will transform the data prepare it for further analysis. Afterwards, I will analyse the data in from both points of view, the classical frequentist approach, and the Bayesian ap-

proach. There will be two main research questions which will provide the common thread: firstly, is there a difference in means of the predicted variable between the two gender groups (female vs male), and secondly, which predictor variables explain the predicted variable best. The first question I will investigate in terms of both approaches, the classical statistical t-test, and by means of a Bayesian model. The second question aims to construct and compare different types of models under the Bayesian approach. For this purpose I will concentrate on exploring the functionalities of the brms package in R developed by Paul Bürkner (Bürkner, 2017).

## 2 The Data Set

The data set under investigation is a sample of the transactions made in a retail store on Black Friday sale. The store wants to know better the customer purchase behaviour against different products. Specifically, here the problem is a regression problem where we are trying to predict the dependent variable (the amount of purchase) with the help of the information contained in the other variables. The data set contains 12 different variables and 537.577 rows or data points. The predicted or dependent variable is "Purchase" and of the type metric. Following predictor or independent variables are contained in the data set: "User_ID" (nominal), "Product_ID" (nominal), "Gender" (binomial), "Age" (7-level ordinal), "Occupations" (nominal/ordinal?), "City_Category" (nominal), "Stay_in_Current_City" (count), "Marital_Status" (binomial), "Product_Category_1" (nominal), "Product_Category_2" (nominal), "Product_Category_3" (nominal). As we can see, a large amount of the data set variables is masked, that is, those variables cannot be interpreted directly, i.e. "Product_ID", "Occupation", "City_Category", and "Marital_Status". The aim of our analysis will be to predict the dependent variables "Purchase" by the combination of the other (independent) variables. However, for my purposes I will drop a part of the variables, since my focus will lie within the methods and not the data per se [1].

[1] Instead of an arbitrary or uninformed dropping it is more sensible to discard single factors that yield less to the explanation of the data. One commonly used methods to find those factors is to conduct a Principle Component Analysis, which is a mathematical procedure of dimension reduction that transforms a number of possible correlated variables into a smaller number of uncorrelated variables (called principle components). Yet, since our data is strongly heterogeneous, that is, it contains continuous and categorical variables, we are not allowed to use Principle Component Analysis (PCA)(used for continuous variables only). Instead, we can use Factor Analysis for Mixed Data, which is a principle component analysis method to explore data with both, continuous and categorical variables (for further information see FAMD FactorMineR package in R) (Kassambara, 2019). However, this procedure is theoretically quite complex and, additionally, quite costly in terms of processing power. Therefore, I skipped it out and arbitrarily dropped some factors that may be less informative.

# 3 Visualizing the Data

As a first step towards data analysis it is important to explore the data in order to make yourself familiar with it. Therefore, I will first have a quick look on the structure and the range of values, and afterwards inspect the data set via plotting it in different ways. One way to do so is to use the R package collection tidyverse, which provides the functionality to model, transform and visualize the data in an integrated way (Wickham  Grolemund, 2017). I will check for missing values, check the different variables (and their possible levels), and plot some histogram, density, and box plots. Additionally, I will include a quantile-quantile plot, to check if the data behaves normally distributed, as well as a visual check for collinearity between independent variables. The implementation and the results of this part can be observed in following included R file: BDACM_Final_Project-rgerber_part1.rmd.

# 4 Analyzing the Data

The visualization of the data shows, that the data is highly heterogeneous in multiple ways. First, it does not seem to follow a normal distribution, which is not surprising, since the dependent (predicted) variable "Purchase" is strictly positive. Additionally, it has become apparent, that the errors of the variables are not normally distributed, but heavily tailed. Therefore, a more flexible model has to be constructed, which does not presume normality and homoscedasticity of residuals.

The brms package of R guarantees this flexibility though the possibility to assume various different types of distribution functions and let you specify a wide range of priors. The brm() function estimates the actual posterior samples. It makes use of the Hamiltonian Monte Carlo samples algorithm (MCMC) (via STAN) to approximate the posterior distribution. Therefore, we need to specify some more parameters, i.e. how many iterations we want the MCMC to run, how many chains we want to run, how many iterations we want to discard per chain (called the warm-up or burnin phase), and the initial values for the different chains for the parameters of interest (or just random)(Smeets  Schoot, 2018).

As we have seen the predictor variable does not seem to be drawn from a normal distribution for two reasons: 1st it is heavily tailed (as demonstrated in the QQ-plot in the visualization part), 2nd it is truncated by zero, because it contains only strictly positive values. As a consequence I have to specify the model distribution (by default it will assume a Gaussian Distribution with the identity link function). Therefore, I will explore different models with different family distributions, which could adapt to the given data set. For our case one could use family distribtutions, which fit "Survival Models" (Bürkner, 2019), which are: lognormal, Gamma, weibull, and inverse Gaussian. For simplicity, I will merely try out the most common one: lognormal, Gamma, and inverse Gaussian.

I will try to construct different models with increasing complexity. For a first exploration of the impact of each independent variable on the dependent one, I will first construct some fixed (main) effect models, also called population-effect models. Afterwards, I will continue to more complex models including interaction, random slopes, and random intercepts (mixed effect structures). As a final step I will compare the models against each other by means of different methods, i.e. the Bayes factor, and Leave-one-out cross-validation. The implementation and the results of this part can be observed in following included R file: BDACM_Final_Project-rgerber_part2.rmd.

## 4.1 Encountered Problems

Throughout the analysis it has come to several complications. First of all, there was a huge problem regarding the infeasible running time of the sampling process. As a first step I changed the default settings (2000 iterations, 4 chains, warm up of 1000, improper flat prior)(Bürkner, 2017) towards a shortening the MCMC process in terms of number of iterations, and amount of chains. These changes seemed to affect the running time in a positive direction, however, now another problem occurred: the chains seemed not to converge (according to the Rhat statistics), and an error warning for divergent transitions after warm up occured. Following the recommendation of the Stan Development Team to prevent divergend transitions the *adapt_delta* parameter was adapted to 0.99, that is the step size which controls the resolution of the sampler (Stan Development Team, 2018). However, this change had the effect that the running time increases again.

The final solution to decrease the running time considerably was to reduce the data set under investigation to 2000 data point. The chosen fraction was randomly sampled without replacement to avoid any biases. This shortening lead to a feasible amount of running time.

Another problem I encountered was that the MCMC chains did not converge. Assuming this failure as an indicator for an improper model, I have tried out to construct models from different family distribution, resulting in choosing the lognormal family distribution as my best option. Another possibility to modify the models could be to actively chose different priors and check for fitting the data. However, I did not carry out the investigation because of constraints in resources.

During the course of analysis, especially while constructing the different fixed effect models, it became apparent, that most factors do not contribute, or contribute very little to the predicted variable. Thus, the aim to increase model complexity pointed out to be redundant on this specific data set. Concurrently, there was not no possibility to explore methods for model comparison, since only one model appeared roughly adequate to model the chosen data.

# 5 Results and Conclusion

Despite some troubles, including, (1) the running time, and, (2) the data set itself, I have reached the personal aim to familiarize myself with the visualization and analysis from the Bayesian point of view at least partly. Although, I could not perform all planned steps of the analysis, in particular in constructing models of different complexity, and the comparison of different models, I have deepen my understanding of the (multiple) regression analysis within the Bayesian framework.

Returning to the research questions we can draw following conclusions: Regarding the first hypothesis "Is there a difference in means of the predicted variable between the two gender groups?" the results show, that, the frequentist t-test indicate a significant difference between means, whereas, the Bayesian analysis indicate a rather small effect. It should be considered, that the t-test assumes a normal distribution of the errors as a prerequisite, which is not the case. Therefore, the results are rather untrustworthy. On the other hand the Bayesian analysis could have lead to more convincing result by setting better fitting priors. Regarding the second hypothesis "Which predictor variables explain the predicted variable best?" the only possible conclusion I can draw from the analysis is that the predictor variable "Gender" could have a small effect on the predicted variable "Purchase". For all other variables, i.e "Age", and "Marital_Status", there seem to be no evidence that they contribute at all. As an additional possibility one could further take into account the impact of the independent variable "Product_ID" on the dependent variable, which I left out in this analysis.

# 6 References


Bürkner, P. (2017). Brms: An R Package for Bayesian Multilevel Models Using Stan. Journal of Statistical Software, 80(1). doi:10.18637/jss.v080.i01

Bürkner, P. (2019, February 14). Estimating Multivariate Models with brms. Retrieved March 05, 2019, from https://cran.r-project.org/web/packages/brms/vignettes/brms _families.html

Dagdoug, M. (2018, July 25). Black Friday. Retrieved February 04, 2019, from https://www.kaggle.com/mehdidag/black-friday

Kassambara, A. (2019). FAMD - Factor Analysis of Mixed Data in R: Essentials - Articles - STHDA. Retrieved from http://www.sthda.com/english/articles/31-principal-component-methods-in-r-practical-guide/115-famd-factor-analysis-of-mixed-data-in-r-essentials/

Lee, M. D., Wagenmakers, E.-J. ( 2005). Bayesian statistical inference in psychology: Comment on Trafimow (2003). Psychological Review, 112, 662–668.

Stan Development Team. (2018, April 13). Brief Guide to Stan's Warnings. Retrieved from https://mc-stan.org/misc/warnings.html#divergent-transitions-after-warmup

Smeets, L., Schoot, R., Van de (Eds.). (2018, July 11). BRMS-started. Retrieved March 05, 2019, from https://www.rensvandeschoot.com /tutorials/brms-started/

Wickham, H., Grolemund, G. (2017). R for data science: Import, tidy, transform, visualize and model data. Beijing: OReilly.