Predicting Black Friday Sales via Bayesian Multiple Regression

Applied Bayesian Statistics

RMIT University

By Arion Evans, Jake Mott, Josh Grosman, and Tim Kirkbride

# Introduction

Black Friday is the name given to the day following the Thanksgiving holiday in the USA on which many major retailors open very early and offer promotional sales on their stock. These promotions typically lead to high sales and even spurts of violence between shoppers and staff on Black Friday – giving the event an infamous reputation. Due to the high amount of sales associated with Black Friday, understanding the kind of shoppers who participate in the event and how much they spend could prove important. This is the main goal of the present report. Specifically, data relating to Black Friday sales will be analyzed in the attempt to predict how much an individual will spend on average, given a selection of predictor variables. This will be accomplished via the construction of a non-informative Bayesian multiple linear regression model. This model will ultimately be utilized to generate predictions over a testing dataset, which will be compared to the true observed values in order to assess the accuracy of the model. Finally, the limitations and overarching implications of the results will be discussed.

# Methodology

This report utilized Black Friday data obtained from Kaggle. A link to the dataset is provided below:

[**https://www.kaggle.com/abhisingh10p14/black-friday**](https://www.kaggle.com/abhisingh10p14/black-friday)

Originally, the dataset contained 537,577 individual observations; each relating to a purchase instance. The features of the raw dataset were as follows:

* User\_ID
* Product\_ID
* Gender
* Age
* Occupation
* City\_Category
* Stay\_In\_Current\_City\_Years
* Marital\_Status
* Product\_Category\_1
* Product\_Category\_2
* Product\_Category\_3
* Purchase

## Data Pre-Processing and Descriptive Analysis:

Initially, features which were deemed ambiguous or unnecessary in the context of this report were dropped. This included the Product\_ID, Occupation, City\_Category and all Product\_Category features. The data were then aggregated by the User\_ID feature (unique for each user) in order to ensure that each individual was represented as one row. In doing so, two new variables were created to represent the average purchase price as well as the count of purchase occasions per shopper. These were calculated by finding the mean of the Purchase variable for each unique User\_ID and summing the number of times an individual appeared in the data, respectively. The new average purchase price feature would act as the response variable.

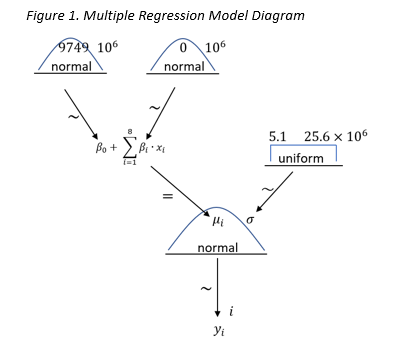
The aggregated dataset contained 5891 observations. Prior to further analysis, the data were summarized and several descriptive plots were constructed to help gain insight into the relationships between the variables. Following descriptive analyses, all factor variables were converted to binary dummy variables to accommodate for the upcoming regression. New columns were made for each level of the Gender, Age, Stay\_In\_Current\_City\_Years and Marital\_Status features. The base level column associated with each factor was then removed.

Finally, 10 observations were selected at random from dataset to be used as the testing data, while the remaining observations were used as the training set.

## Model Construction and Implementation:

R statistical software in unison with JAGS was used to code and execute all analyses in this report.

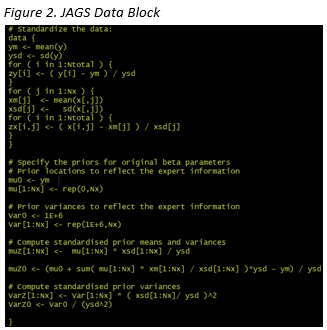
The Bayesian model diagram shown below in Figure 1 was the basis for the analysis conducted.

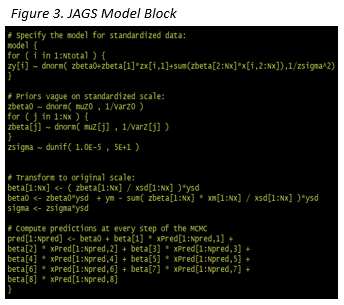


As average purchase price is continuous in nature, each observation was specified to be distributed according to a normal distribution – N(µ, σ). The mean of this distribution, µ, was in turn represented by the multiple linear regression equation, with a β parameter associated with each predictor variable as well as the intercept, β0. A normal distribution was also used to represent the distribution of each β. As this analysis was non-informative, each of these β distributions (with the exception of β0) were centered at 0 with a large variance to reflect the low degree of belief. β0 was centered at the mean of the response variable in accordance with the regression formula. The degree of belief in the average purchase price was represented by σ and was distributed uniformly with a large range to reflect the non-informative nature of the analysis. The apparently arbitrary start and end points of the σ distribution are due to its standardized counterpart being employed in the actual generation of chains.

The above model diagram was converted into JAGS model and data code blocks, which are shown below Figures 1 and 2, respectively.

It should be noted that initially, all available predictor variables were included into the regression analysis. Based on the results of this, non-significant predictors were identified and removed. A new, reduced model was then constructed, which only included the Age, Gender and Count variables as predictors. All code blocks and results presented here relate only to the reduced model.





As shown in the data block, all variables were standardized to reduce potential multicollinearity. Effectively, the standardized model estimates were found using the standardized variables, and then the original coefficient estimates were calculated based on this.

For the final MCMC run, 100,000 iterations were run with 5000 burn-in and 5000 adaption steps and a thinning of 15. Chain diagnostics plots were generated and assessed for each of the original parameter estimates. Furthermore, posterior distributions were generated for the parameter estimates, the model’s R2 value and the predictions.