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 using Monte Carlo Markov Chain (MCMC) simulations. 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.

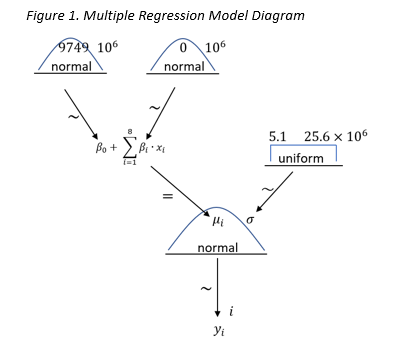
The aggregated dataset contained 5891 observations. Prior to further analysis, the data were summarized and several descriptive plots were constructed to help gain insight into the relationships between the variables. Following descriptive analyses, all factor variables were converted to binary dummy 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.

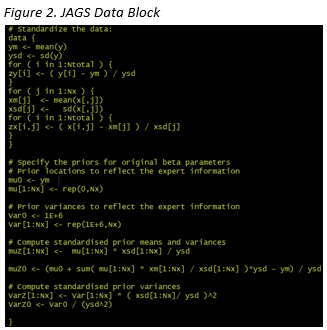
The Bayesian model diagram shown below in Figure 1 was the basis for the analysis conducted.

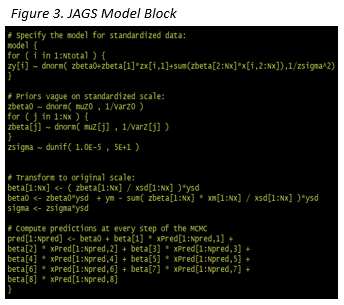


As average purchase price is continuous in nature, each observation was specified to be distributed according to a normal distribution – N(µ, σ). The mean of this distribution, µ, was in turn represented by the multiple linear regression equation, with a β parameter associated with each predictor variable 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.

# Results/Discussion

## Descriptive Analysis:

Table 1 below shows the dataset before binary transformations. Gender and age were factors with 2 and 7 different levels respectively. Count, the number of items bought during black Friday, averaged around 91 for each customer but had a median of 53 showing that a high purchases with lots of items skew the mean.

*Table 1. Descriptive Statistics of Variables.*

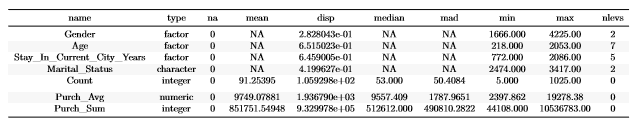
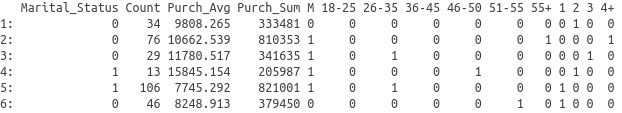


Table 2 below shows the head of the dataset and the resulting transformed variables. The ages were broken down into their original binned factors ranging from 18-25 to 55+. Gender has become binary under M with 1 being male and 0 being female. Purch\_Avg is our response variable representing the average purchase amount by each customer.

*Table 2. Head of Processed Dataset.*



The histogram of average purchase price in figure 4 shows a clear resemblance to the normal distribution. However, it slightly skews right showing that a relatively small amount of people spend very large amounts.

*Figure 4. Histogram of Average Purchase Price per Customer.*



Figure 5 below shows the box plot of average purchase price by gender. A slightly higher median purchase amount for men compared to women can be seen. Also, women appear to have more widely spread outliers that reach higher purchase amounts than any male purchase. This suggests men spend more consistently across higher amounts while the majority of women spend slightly less with a handful of women spend extravagantly.

*Figure 5. Boxplot of Average Purchase Price per Customer by Gender.*

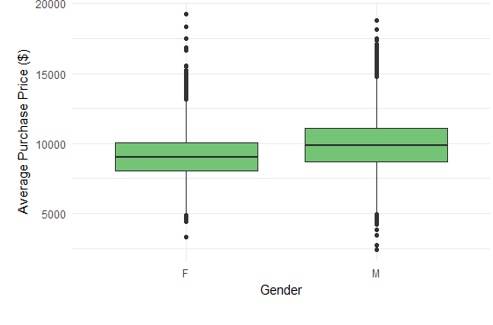
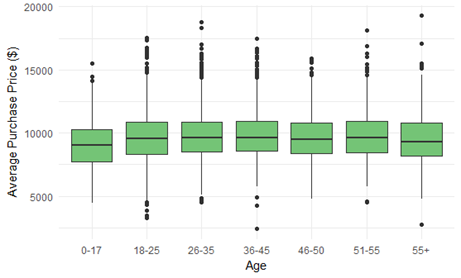


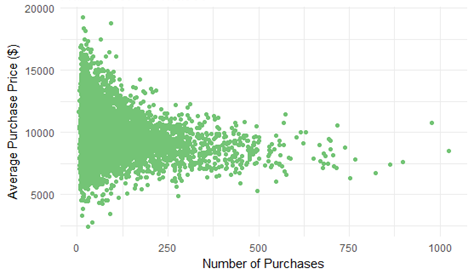
Figure 6 below shows boxplots of average purchase price by age. A slight trend in purchase amounts is evident that peaks around the age group of 34-45. Though there is little variation in the size and location of these boxplots across these bins, age groups 26-35, 51-55 and 55+ show larger reaches of outliers to higher purchase amounts.

*Figure 6. Boxplot of Average Purchase Price per Customer by Age.*



The plot on below in figure 7 shows an unusual relationship. The scatter plot of average purchase price per customer by the number purchases made shows that the more transactions someone does the more average amounts they spend. This suggests that people spend similar amounts in total and that there are two general buying habits; few transactions with very high cost and numerous transactions with low to medium cost.

Figure 7. Scatter Plot of Average Purchase Price per Customer and Number of Purchases.



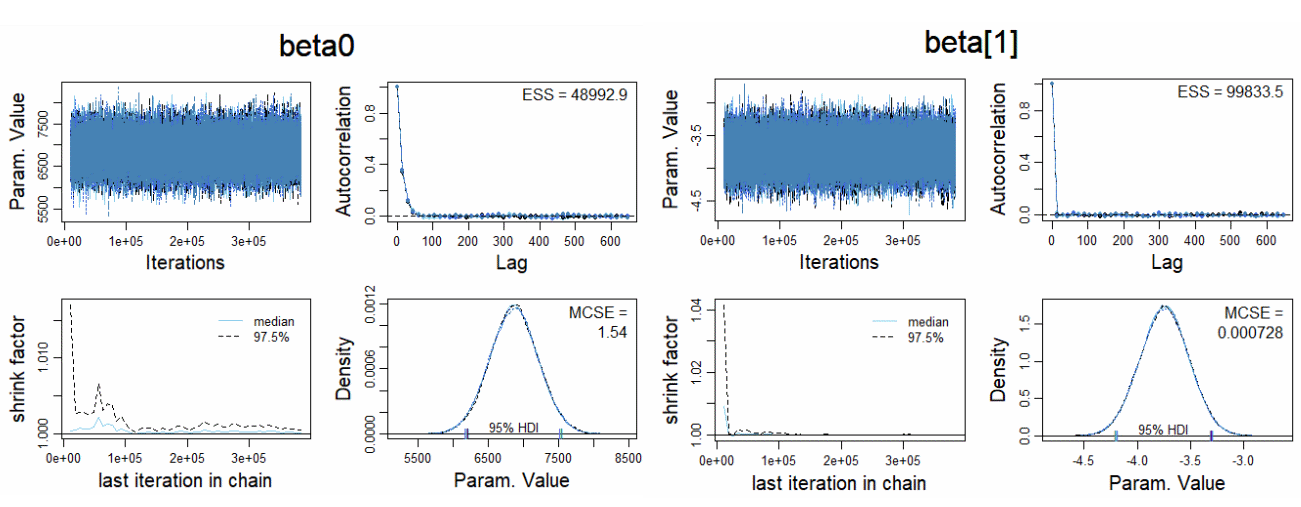
This visualisation below shows average purchase amounts by count faceted by age and coloured by gender. It shows lower densities in the age groups of 0-17 and 55+. Males appear to also spend more overall although some very high transactions done by women show the outliers. As the majority of the dots are blue it shows that black Friday is a spending craze that men participate in more than women. All age ranges show a similar relationship between average purchase amount and the number of transactions.

*Figure 8. Average Purchase Price per Customer by Age, Gender and Number of Purchases.*

## 

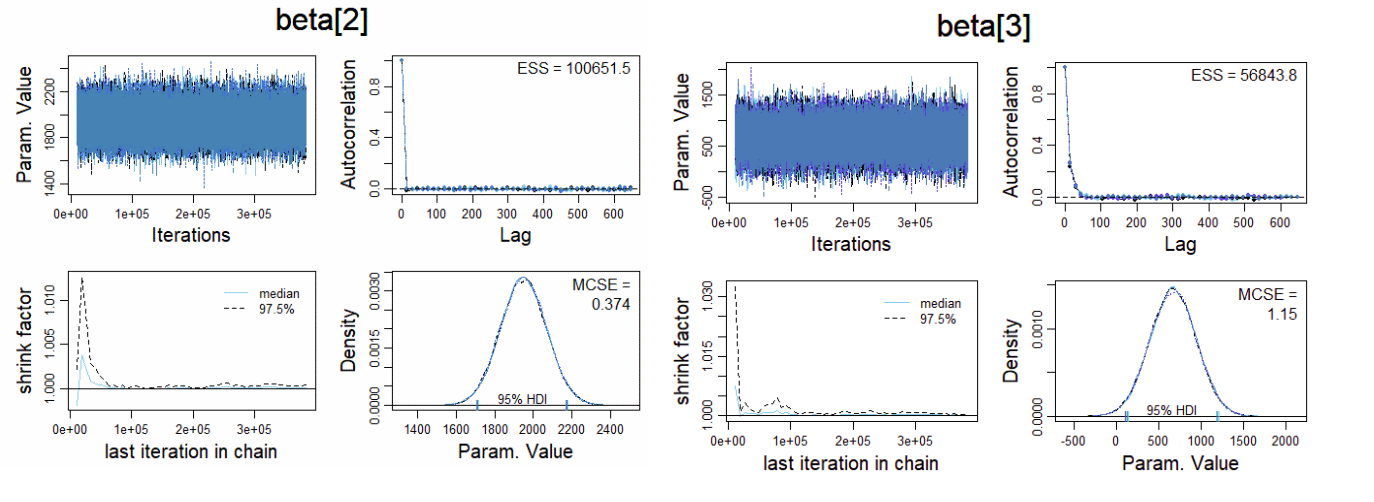
## MCMC Diagnostics:

*Figure 9. MCMC Diagnostic plots for beta0 and beta1*



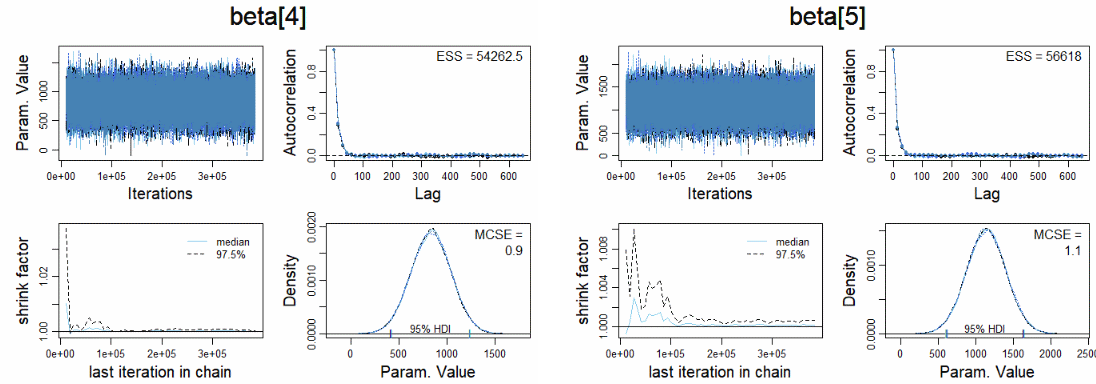
The MCMC diagnostics shown in these graphs are very positive. First off, we can see in the chain plot that the chains have successfully converged and merged as they are overlapping. This indicates that the chains will be representative. The plots also show that the chains will be accurate, as the Monte-Carlo Standard Error (MCSE) value is extremely small, this accuracy is also seen in the almost immediate flattening shrink factor and autocorrelation plots which fall to below 1.1 and 0 extremely quickly.

*Figure 10. MCMC Diagnostic plots for beta2 and beta3*



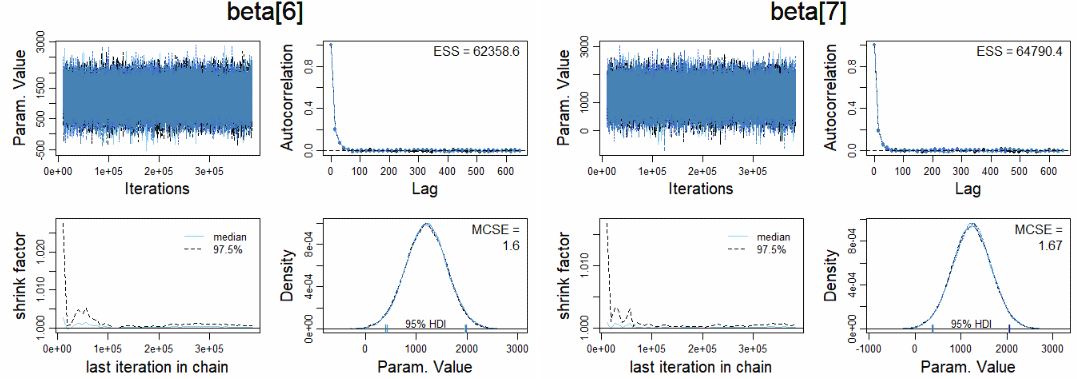
The representation and accuracy is also seen in the MCMC diagnostic plots for beta2 and beta3. One potential problem in beta3 is the smaller Effective Sample Size (ESS) value. Although it is quite high at around 50,000, beta2’s ESS value of around 100,000 (which is approximately the sample size) is better, this ESS value shows that there is potential for some autocorrelation.

*Figure 11. MCMC Diagnostic plots for beta4 and beta5*

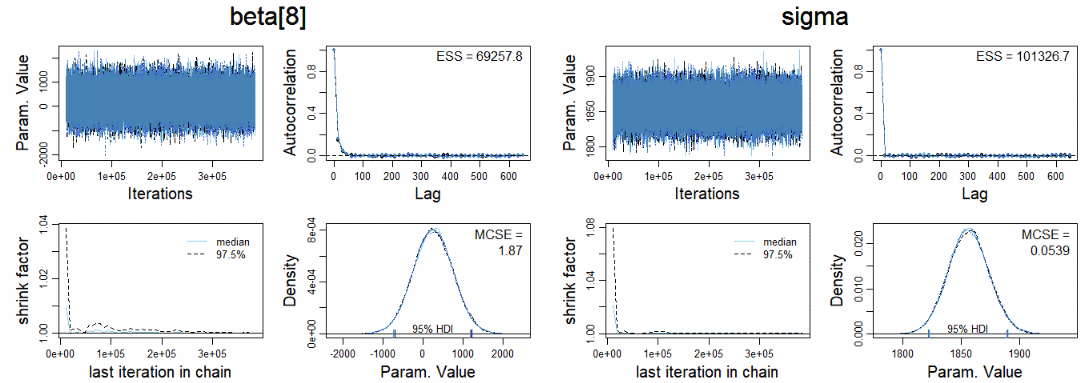


Again the diagnostic plots are mostly very positive for beta 4 and 5, with the main problematic feature being the relatively low ESS value. To increase this value and remove potential autocorrelation, perhaps a larger thinning value could be used. This change in the thinning value would not be necessary however as the autocorrelation plot suggests no significant autocorrelation to remove (This is also seen in the beta6 and 7 plots).

*Figure 12. MCMC Diagnostic plots for beta6 and beta7*



*Figure 13. MCMC Diagnostic plots for beta8 and Sigma*



## Posterior Estimates:

## Prediction Estimates: