Predicting Black Friday Sales via Bayesian Multiple Regression

Applied Bayesian Statistics

RMIT University

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

Table of Contents

[Introduction 3](#_Toc527900760)

[Methodology 4](#_Toc527900761)

[Data Pre-Processing and Descriptive Analysis: 4](#_Toc527900762)

[Model Construction and Implementation: 5](#_Toc527900763)

[Results/Discussion 7](#_Toc527900764)

[Descriptive Analysis: 7](#_Toc527900765)

[MCMC Diagnostics: 11](#_Toc527900766)

[Posterior Estimates: 14](#_Toc527900767)

[Predictions: 16](#_Toc527900768)

[Test data 16](#_Toc527900769)

[MCMC Diagnostics 17](#_Toc527900770)

[Posterior Estimates 19](#_Toc527900771)

[Summary 21](#_Toc527900772)

# 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.

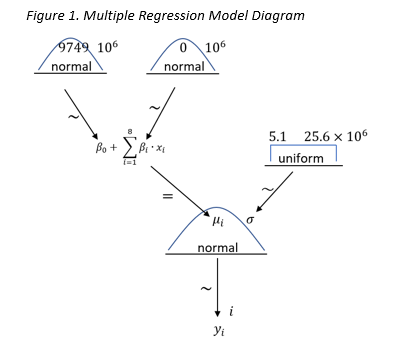
The aggregated dataset contained 5891 observations. Prior to further analysis, the data were summarized, and several descriptive plots were constructed to help gain insight into the relationships between the variables. Following descriptive analysis, all factor variables were converted to binary dummy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

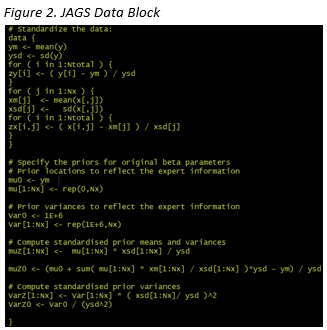
The Bayesian model diagram shown below in Figure 1 was the basis for the analysis conducted.

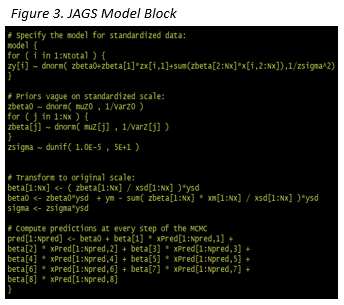


As average purchase price is continuous in nature, each observation was specified to be distributed according to a normal distribution – N(µ, σ). The mean of this distribution, µ, was in turn represented by the multiple linear regression equation, with a β parameter associated with each predictor variable 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 in Figures 2 and 3, 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 and model blocks, all non-binary 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. Descriptive Statistics of Variables.*

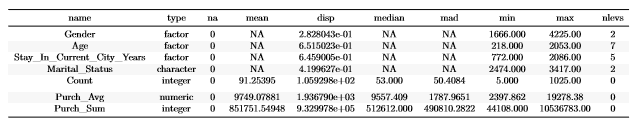


Table 1 above shows the dataset before binary transformations. Gender and age were factors with 2 and 7 different levels, respectively. Count, the number of items bought per person averaged around 91 for each customer, while the average sales per customer (the response variable) had a mean of approximately $9,750.

*Table 2. Head of Processed Dataset.*

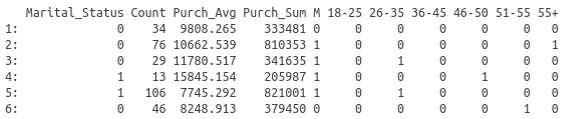


Table 2 shows the head of the dataset and the resulting transformed variables. The ages were broken down into their original binned factors ranging from 18-25 to 55+. Gender became binary under M with 1 being male and 0 being female.

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



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

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

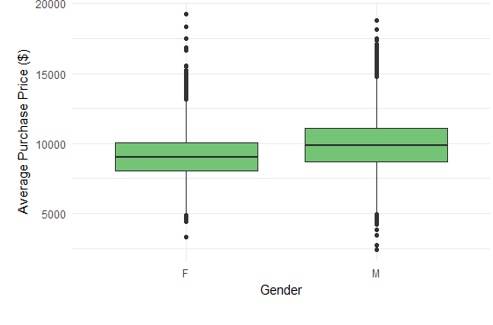


Figure 5 shows the box plot of average purchase price by gender. A slightly higher median purchase amount for men compared to women can be seen. Also, women appear to have more widely spread outliers that reach higher average purchase amounts than any male purchase. This suggests men spend more consistently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

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

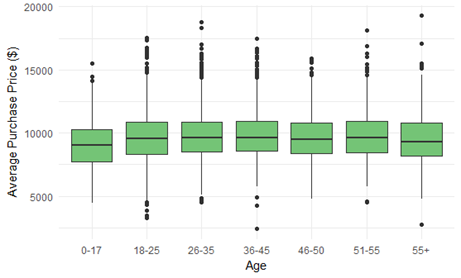
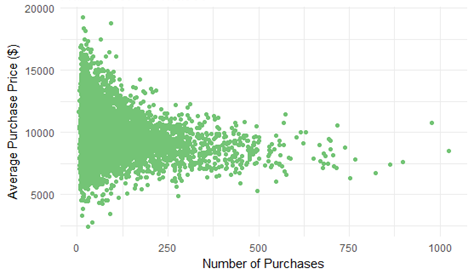


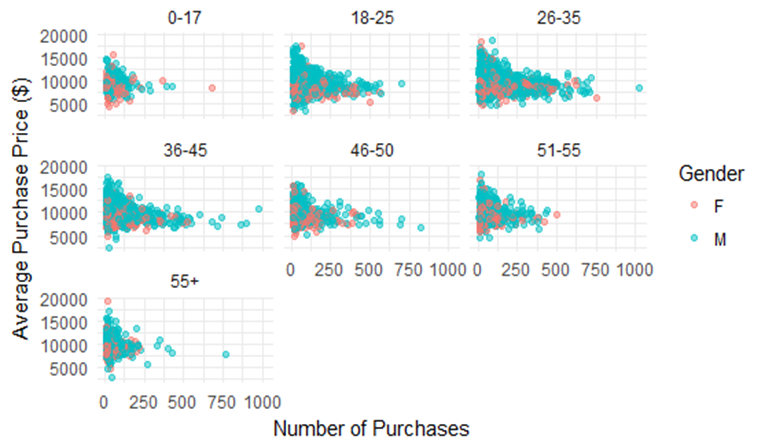
Figure 6 above shows boxplots of average purchase price by the age bins. A slight trend in purchase amounts is evident that peaks around the age group of 36-45. Though there is little variation in the size and location of these boxplots across these bins, age groups 26-35, 51-55 and 55+ show larger reaches of outliers to higher purchase amounts.

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



The plot in Figure 7 shows an unusual relationship. The scatter plot shows that the more transactions someone has, 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.

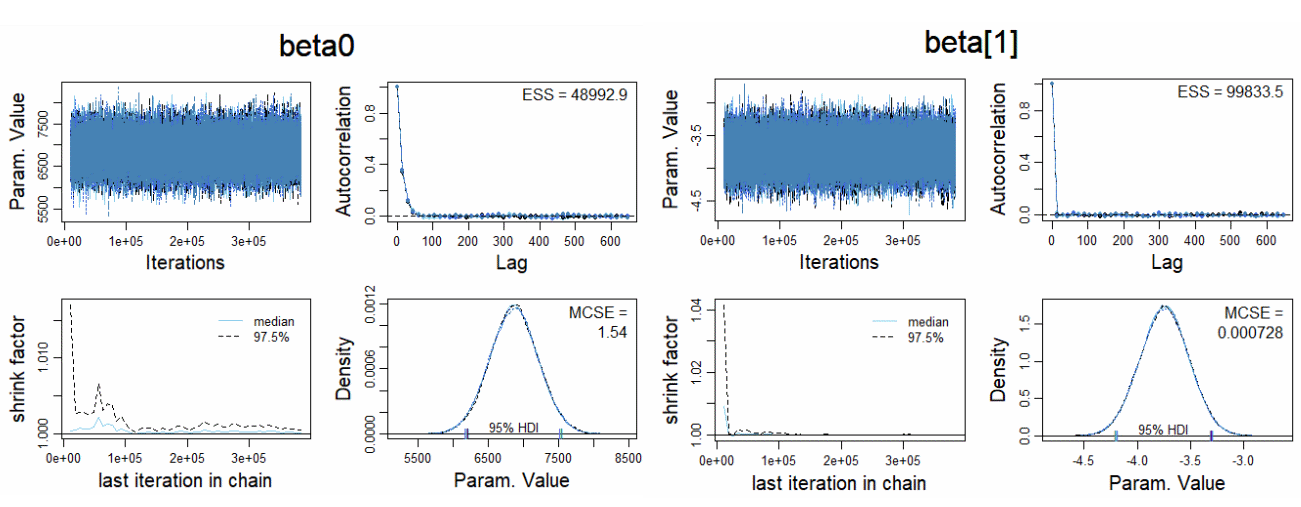
*Figure 8. Average Purchase Price per Customer by Age, Gender and Number of Purchases.*



The multivariate visualization depicted in Figure 8 captures the relationship between all predictor variables included in the final model as well as the response variable. Lower point densities are evident in the age groups of 0-17 and 55+, suggesting fewer purchases. Males appear to also spend more overall although some very high transactions done by women show the outliers mentioned previously. As the majority of the dots are blue, this implies that there are more men in the sample than women. All age ranges show a similar relationship between average purchase amount and the number of transactions, however, the 0-17 and 55+ are different once more in that fewer of these individuals have a high purchase count.

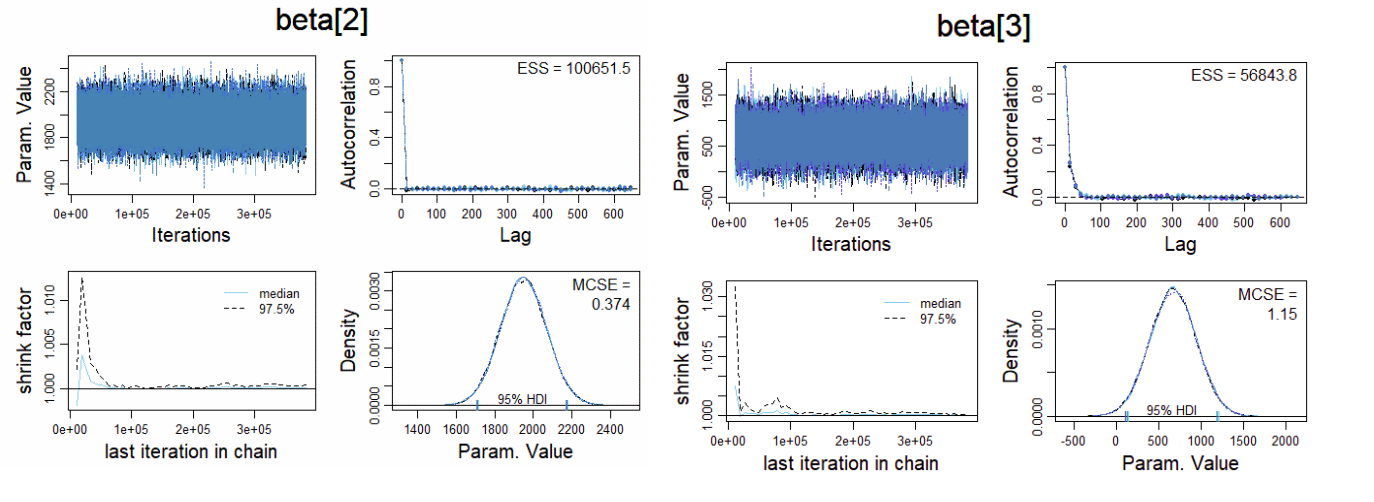
## MCMC Diagnostics:

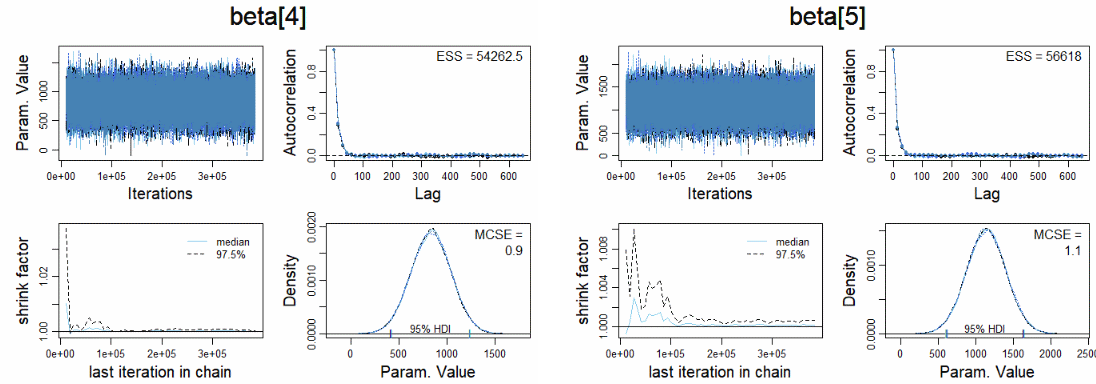
The MCMC diagnostics shown in Figures 9 - 13 are very positive and indicative of a good posterior sample for each beta parameter, as well as sigma and R2. In all cases, chains are shown to have successfully converged and are well mixed. Furthermore, the shrink factor plots can be seen to be far from 1.1 and the distributions are all overlapping, ultimately suggesting representative chains. The Monte-Carlo Standard Error (MCSE) values are very small, and this in unison with the almost immediate flattening of the autocorrelation plots implies good chain accuracy. One potential cause for concern, however, is the relatively low ESS of some beta parameters.

*Figure 9. MCMC Diagnostic plots for Beta0 and Beta1*

In Figure 9, the chains can be mostly seen to be desirable as described above, however the ESS of beta0 is almost half of beta1, and there is some evidence of autocorrelation. However, both are still very large in terms of sample size.

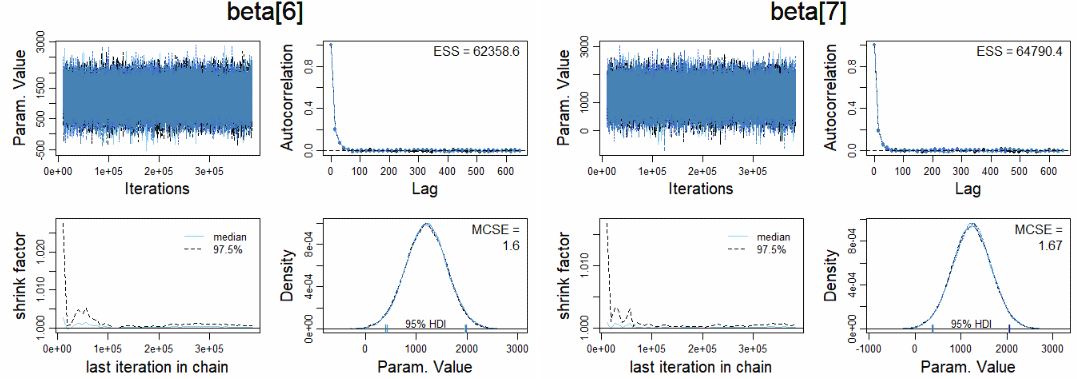
*Figure 10. MCMC Diagnostic plots for Beta2 and Beta3*

**Again, there is a potential problem with beta3 in relation to the smaller ESS and autocorrelation in the chain.

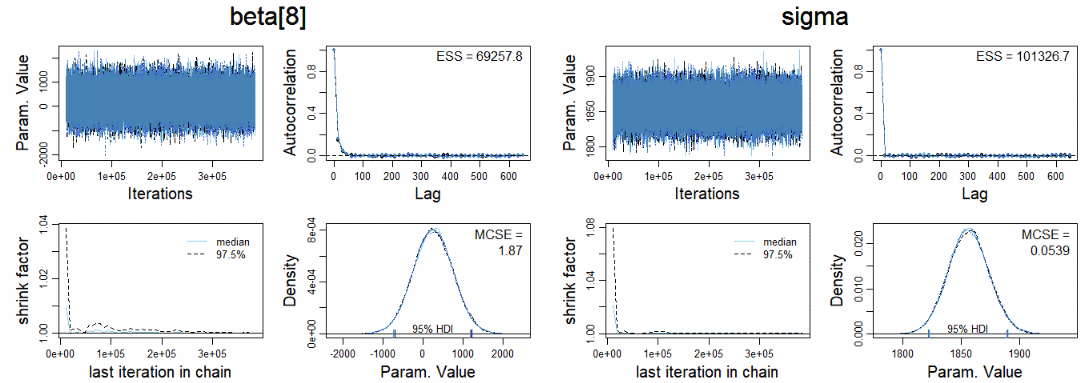


*Figure 11. MCMC Diagnostic plots for Beta4 and Beta5*

The diagnostic plots in Figure 11 are once more mostly very positive for beta4 and beta5, 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.

*Figure 12. MCMC Diagnostic plots for Beta6 and Beta7*

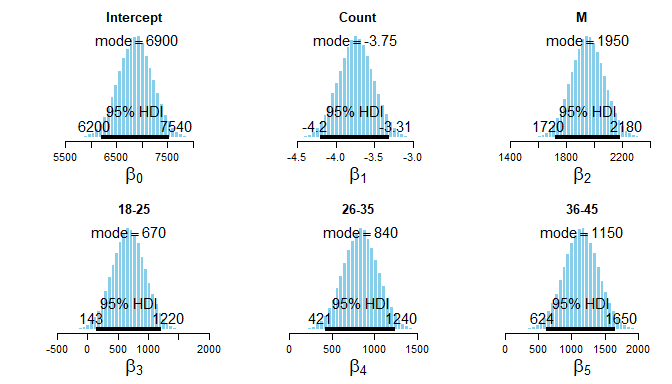
The ESS values for beta6 and beta7 shown above are somewhat higher than those seen previously and might also be resolved with a higher degree of thinning.

*Figure 13. MCMC Diagnostic plots for Beta8 and Sigma*

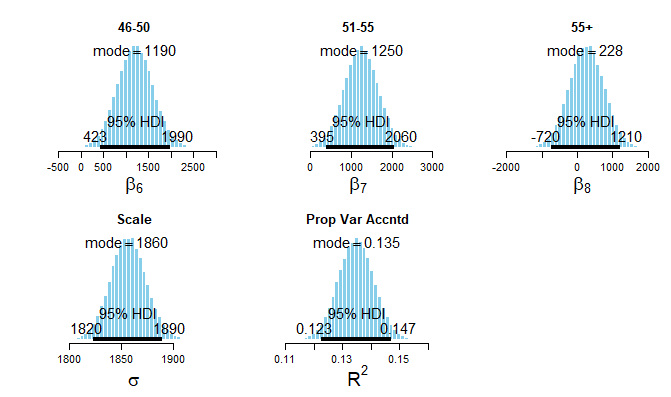
Finally, the diagnostic plots for beta8 are similar to those observed previously for the rest of the coefficients while the plots for sigma show extremely well mixing chains, low autocorrelation and MCSE, and high ESS.

## Posterior Estimates:

The posterior estimates for the beta values here show that the Intercept, Count, M and age brackets for 18-25, 26-35 and 36-45 are all significant (as the HDI does not contain 0). The most unexpected mode value that is shown in the plot below is the count variable, with a mode of -3.75, this value indicates that a marginal increase in the amount of products purchased of one (i.e. buying one more product) will decrease the value of goods purchased by $3.75. This would not be expected, however this could indicate that purchases that only include one item will be bigger ‘one off’ type purchases of expensive goods, while purchases that include multiple products will be a larger collection of cheaper goods. Another significant pattern that is consistent with the descriptive analysis is the gradual increase in the mode as the age brackets increase. In the plot below an increase in the mode value is seen as the age of the purchaser increases, this is also seen in the mode values for the age brackets 46-51 and 51-55 in Figure 15.

*Figure 14. Posterior Estimates for beta values 0-5*

*Figure 15. Posterior Estimates for beta values 6-8, sigma and R2*



This increasing mode value trend ends with the 55+ age bracket, this end of the trend is not too inconsistent with our descriptive analysis however the beta 8 posterior estimate is the only posterior element which should not be considered significant (this estimate should not be considered significant as the 95% HDI ranges from -720 to 1210). While the sigma posterior estimate is satisfactory, the R2 value of 0.135 (with an HDI of 0.123-0.147) is relatively low. This indicates that the model only captures approximately 13.5% of the variation in the dependent variable. This could be attributed to many factors. A possible explanation is that other variables beyond the dataset exist that better explain the variability in the average purchase amount or perhaps a different non-linear relationship exists between the present variables and the dependent variable which would be better suited.

## Predictions:

The model to be used to obtain predictions is shown below:

### Test data

The observations in Table 3 were randomly selected from the dataset to serve as a test dataset for the model. The average purchase amount ranges from $6,560 to $11,541 and the number of items from 13 to 298. Most observations are from male customers which is to be expected as the majority of the observations in the original dataset are from male customers.

*Table 3. Test dataset*

A screenshot of a cell phone

Description generated with high confidence

### MCMC Diagnostics

The chains produced by the simulation once again appear to be in good health (Figures 16-17). The chains for all predictions are mixing well with the chains being almost indistinguishable from each other in both the time series and density plots. Additionally, the shrink factors are very close to 1 for all predictions and the autocorrelation drops to almost 0 pretty fast for most predictions. The ESS ranges from 49,654 to 99,875 while all MCSE’s apart from those for predictions 5 and 7 are below 1. Overall, apart from some issues with low ESS values (also observed in the chains for the model coefficients), the chains are representative and accurate.

*Figure 16.* A picture containing map, screenshot

Description generated with high confidenceA picture containing screenshot

Description generated with very high confidence*MCMC diagnostic plots for predictions 1-4*

A picture containing screenshot

Description generated with very high confidenceA picture containing screenshot

Description generated with very high confidence

*Figure 17.* A picture containing screenshot

Description generated with high confidenceA picture containing screenshot

Description generated with very high confidenceA picture containing screenshot

Description generated with very high confidence*MCMC diagnostic plots for predictions 5-10*

A picture containing screenshot

Description generated with very high confidenceA picture containing screenshot, map

Description generated with high confidenceA picture containing screenshot

Description generated with very high confidence

### Posterior Estimates

The posterior distributions and the point and interval estimates for the predictions are seen in Figure 18 and Table 4. As expected from the very low R2 value, the model is performing very poorly in terms of its predictive capability. Only 3 out of 10 HDI intervals capture the true value of the average purchase amount and large error margins can be seen in most predictions. Furthermore, it can be observed that the model gives relatively consistent predictions (around $8,500 to $9,500 for males and $7,000 to $8,000 for females) for all test observations. This would explain why the model is unable to capture the large variation seen in the dependent variable.

*Table 4. Model predictions against actual values*

A screenshot of text

Description generated with very high confidence

*Figure 18. Posterior Estimates for the predictions*

A close up of a map

Description generated with high confidence

# Summary

Bayesian methods were used to construct a simple linear regression model in order to predict the average purchase amount of customers during Black Friday. The independent variables used in the model were the number of items purchased, the gender, and the age group of the customer. The dependent variable was set to be normally distributed with a mean value equal to its estimate from the model. Non-informative priors were given to all the parameters of the model. The resulting model is shown below:



According to this, only the number of items has a negative relationship to the average purchase amount. Furthermore, it is suggested by the model that males spend $1,950 more than females (keeping all other variables constant) while there is also an increasing trend as age increases until 55. Parameter estimates were all found to be significant apart from the coefficient for while the chains produced by the Monte Carlo simulation method were all found to be representative and accurate. Finally, with an R2 value of 0.135, the model was found to be quite inaccurate in predicting the average customer purchase amount. This was attributed to the potential existence of other more correlated explanatory variables not present in the dataset or the existence of a different non-linear relationship between the dependent and independent variables in the model.