Bayesian Binary Regression in Marketing

# A Project to Assess Response Rates in Direct Marketing Campaigns

## Submitted by

## VIJAY ANANTH SUNDARA PANDIAN

## GREGORY ROGNER

## VISHNUPRASAD RAMASUBRAMANI

# Abstract

In this project we attempt to assess the application of Bayesian methods to the marketing field. In the magazine industry making a customer renew his subscription is critical as it has a crucial financial impact. The response variable is binomial as he may either renew the subscription or decline. The predictor variables may be continuous or categorical. In such cases logistic regression provides a way to link the binomial outcome to continuous or categorical predictor variables. Bayesian Binomial regression defers from the normal logistic regression in the sense that it makes use of priors of the coefficients of the predictor variables in arriving at the posterior distribution of probability. In this project we attempt to discover the relationship between age as a predictor variable and successful subscription as the response variable using Bayesian Binary Regression.

Table of Contents

[A Project to Assess Response Rates in Direct Marketing Campaigns 1](#_Toc374881458)

[Submitted by 1](#_Toc374881459)

[Abstract 2](#_Toc374881460)

[Chapter 1. Introduction 4](#_Toc374881461)

[Chapter 2. Project Outline 4](#_Toc374881462)

[Chapter 3. Logistic Regression and BBR 5](#_Toc374881463)

[Chapter 4. Project Scope and Methodology 6](#_Toc374881464)

[Chapter 5. Bayesian Binary Regression in Excel 7](#_Toc374881465)

[Chapter 6. Bayesian Binary Regression with Excel and R 9](#_Toc374881466)

[Chapter 7: Bayesian Binary Regression in Matlab 11](#_Toc374881467)

[Chapter 8.Conclusions 13](#_Toc374881468)

[References: 14](#_Toc374881469)

# Chapter 1. Introduction

1.1 **Context of the Project**: The print magazine industry is very niche in its character. Unlike other ‘products’ the print publications cater to a very narrowly defined segment of customers. Therefore the industry is highly dependent on retaining existing customers.Thus a significant part of the marketing effort in the magazine industy is focused on making subscription renewal very attractive. By bundling long terms subscriptions with discounts and attractive assured gifts, the magazine industry ‘locks-in’ the subscribers. But the schemes come with a price. Every marketing scheme costs significant amount of money and renewal rates tend to be low. And not all schemes come with the guarantee of success. Therefore the marketing manager tries to get an idea of ‘the most effective’ direct marketing scheme. In other words, he tries to understand which one of his schemes generates the maximum response rate.

1.2 **The background**: Readers of a magazine normally pay a subscription fee to read the magazine for a period of twelve months. In order to ensure that the subscriber continues with his subscription after the initial twelve months the marketing efforts usually start after nine months- that is ninety days before the expiry of the existing subscription. Every month the marketing department generates a list of ‘target customers’. There is a well-defined process of direct marketing campaigns. Once a list is ready the marketing department puts together a campaign

The customers are then contacted through a direct marketing campaign consisting of the following:

1. Email
2. Hard copy direct mailers
3. Telecalling agents

The challenge: The schemes have differing response rates defined by the number of customers renewing their subscriptions . The ratio of customers responding to such campaigns tends to be low and varies across diverse marketing campaigns.The challenge is to predict the likely response rate for such direct marketing campaigns and develop a framework for comparing schemes.

# Chapter 2. Project Outline

**2.1Project Objectives:** The project proposes to explore the relationship between the age of the magazine subscriber and the likelihood of renewing his or her subscription. In other words, we seek to answer the question: Does the age of the subscriber influence his behavior towards renewing an existing subscription in any manner? And if it does, how significant is the effect of the subscriber’s age in influencing the renewal behavior?

Using data from the results of a past direct marketing campaign for a magazine, we analyze the behavior of the response variable( subscribed =1 versus not subscribed= 0) as an outcome of the variation in the predictor variable we have chosen which is age and this is a continuous variable

**2.2 Significance of the project:** If we are able to establish a relationship between the predictor variable and the response variable, it may have a significant impact on the campaign management practices in the direct marketing industry. If we are able to identify those subscribers who are more likely to respond, then the campaign investments can be directed towards those customers and thus push up the response rate. This in turn can result in better return on investment for the industry

**2.3 Project Challenge:** It is important to note that we cannot use linear regression to establish the relationship between age and subscription renewal. The reason is that the predictor variable –Age- is continuous while the response variable Subscription renewal is binary. Any attempt to do a direct scatterplot of the variables does not yield a clear picture.

Figure 1: scatterplot when the response variable is binary

# Chapter 3. Logistic Regression and BBR

**3.1 Logistic Regression:**Logistic Regression enables us to deal with the problem of finding a relationship between a continuous predictor variable and a dichotomous response variable.It makes use of a link function which takes the form . The equation for the logistic regression is given below:

* ++Where is the intercept and , etc are the predictor variables and, are the respecctive slopes

As a result the probability of the response is derived from the above reaction and is shown below

/1+

In our case, since we are estimating the impact of a single variable, the equations reduce to:

* + where is the intercept and , is the predictor variables and, is the slope
* /(1+)] where P is the probability of an event occurring and 1-p is the probability it not occurring and [p/(1-p )]is the odds ratio

We use the predictor variable to calculate the Likelihood estimate. Then the coefficients of the equation are found by solving the equation by optimizing the Maximum Likelihood Estimate

**3.2 Bayesian Binary Regression:** Bayesian Binary Regression takes the binary regression concept and incorporates a ‘prior’ knowledge of the response variables into the calculation of the posterior probabilities. Bayesian Binary Regression takes the following form:

Here the posterior probability of the response variable is found by multiplying the likelihood by the prior probability and dividing by a normalizing constant. It has been established that the normalizing constant can be dropped from the calculations by Cheng Wei in his doctoral thesis. The equation for posterior probability is then written as follows:

* P(r=1|)= |r=1)

Where P(r=1|) is the posterior probability of the response variable being 1 and |r=1) represents the maximum likelihood function or more appropriately Maximum Posterior Value in Bayesian Statistics.

# Chapter 4. Project Scope and Methodology

**4.1 Project scope:** Our study took data from a magazine subscription campaign. The data contained information on the subscribers identification, his geographic location, tenure , age and other data. A sample is shown below in table1.



Table 1: Sample Data Profile

We decided to measure the impact of age on the response and so we defined the variables as follows:

|  |  |  |
| --- | --- | --- |
| Type of variable | Name | Attribute |
| Predictor | Age | continuous |
| Response | Renewal Status | Binomial Y= 1 and blank = 0 |

**4.2 Methodology:** We approached the project in three distinct stages. Since we had to understand the impact of BBR(Bayesian Binary Regression) we decided to move from the familiar to the unfamiliar. Our reasoning was that it will also give us a comparative platform. Our methodology therefore consisted of the following stages:

* Step1: : Work out a Bayesian binary regression using excel and grouped data
* Step 2: Work out a Bayesian binary regression using R and excel
* Step 3:Work out a Bayesian Binary Regression with Matlab and excel

# Chapter 5. Bayesian Binary Regression in Excel

**5.1Steps in BBR in excel:** Our approach to the project in excel consisted of the following steps:

We took a sample of 30 records for building the logit equation. Using these values and assigning arbitrary values for the values of and in the equation{ +} we constructed the equation. The through the steps outlined below we calculated the Maximum Likelihood Estimate. We then used Excel solver to solve for the values of and .This revised equation was then used to calculate the predicted LR values for all the values in the given data.

Steps in calculating the logistic regression probabilities:

* Construct a logit equation with arbitrary values
* Calculate e^L
* Calculate p = e^L/(1+e^L)
* Calculate Likelihood function p(x)= (p^)\*(1-p)^(1-)
* Calculate MLE
* Use Excel solver to find the ideal intercept and slope for the logit function

**5.2 Data view:** The data grouped on the observed data points for age and the distribution are shown below:



Table 2: Data grouped on observed datapoints on age Figure 2: Chart of distribution of data

The predictor variable ranged from 21 years to 51 years and the number of positive and negative responses at each age point was captured for analysis.

**5.3 Calculating Bayesian Binary regression using Excel:** After doing the logistic regression calculations, we calculated the Bayesian Binary regression through the following method:

* Step1:Group the universe on data points found on the universe
* Step2:Calculate the rate of success for each data point
* Step3:Calculate the mean and SD response rate for the entire group of observations
* Step 4:Form a ‘subjective’ prior with these observations where the prior is a normal distribution with the observed mean and SD of the response rate
* Step 5:Apply the logistic equation to all the values in the universe
* Step 6:Multiply the likelihood by the prior to arrive at the BBR predicted probability

**5.4 Results of Stage 1:**



Table 3:results of BBR and LR with Excel

Figure 3: excel based -BBR and LR compariosn

Our analysis of the results seem to indicate that BBR predicts lower probability values compared to Logistic regression . These conclusions are drawn based on the methodology adopted for both on excel based calculations.

# Chapter 6. Bayesian Binary Regression with Excel and R

**6.1Using R:** In this stage we decided to introduce a newer element into the experiment. Instead of the calculating the Logit function with excel, we fed the grouped data into R. The code snippet used is shown in figure 4.

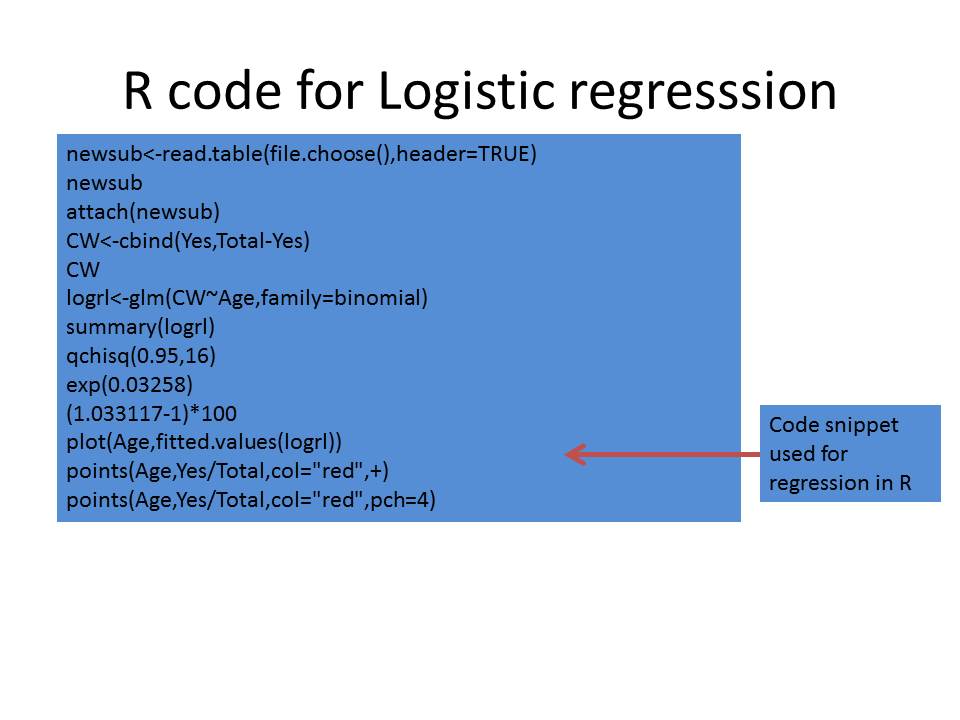


Figure 4:R code snippet for binary regression

The values for the slope and intercept obtained as an output from R where fed into excel. The remaining steps as outlined in the previous stage were repeated for calculating both the logistic regression and the BBR. The results are shown below: 

Table 4: Exceland R based output LR and BBR

In the combined modelling from R and Excel put together, the logistic regression and Bayesian Binary Regression produce a closer value. However, the values from BBR seem to be still on the lower side.

# Chapter 7: Bayesian Binary Regression in Matlab

**7.1 Plotting the Posterior probabilities:** Our approach to this stage consisted of calculating the posterior probabilities using Matlab. We fed the data grouped on data points with the attendant success rate into Matlab. We used the in-built logit function available in Matlab to build the equation and plotted the posterior distribution. The steps we followed are as below:

* Feed in the grouped data
* Attribute a normal prior for the intercept and slope
* Calculate the posterior distribution
* Utilise the slicer sampling for intercept and Slope.

The code used for the Matlab for Bayesian Binary Regression is shown below:

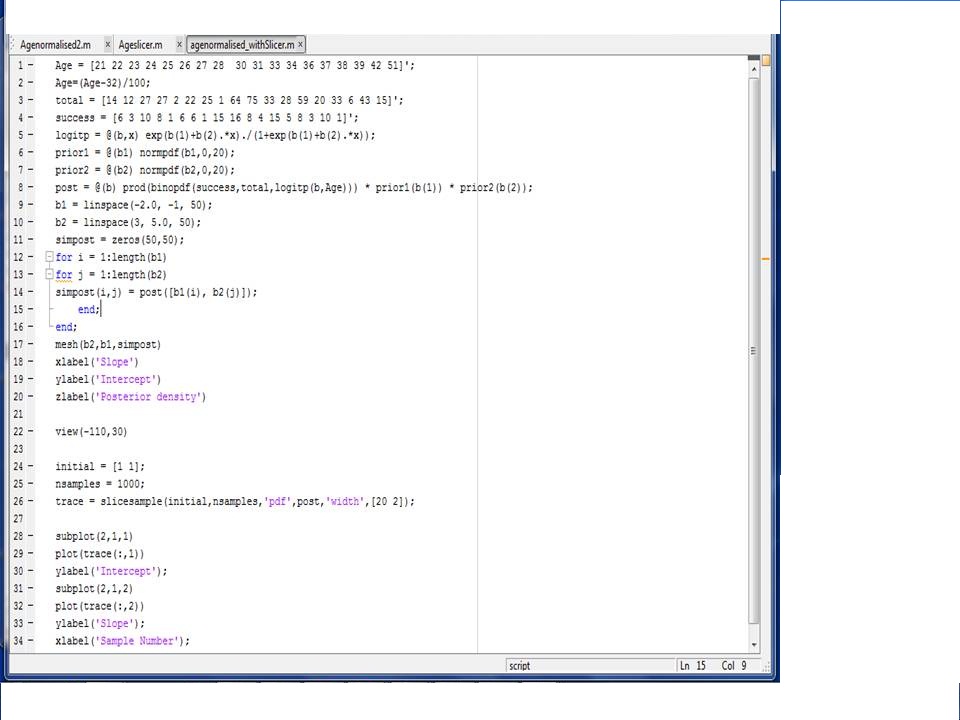


Figure 5 Matlab code for BBR

We were able to generate the posterior probability distribution using Matlab. The output is shown below:

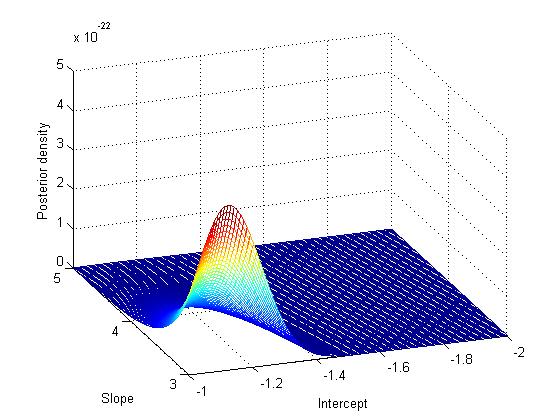


Figure 6:Posterior probability distribution with Matlab

We also did a slice sampling to calculate intercept and slope. The number of samples was set at 1000 and the result is given below:

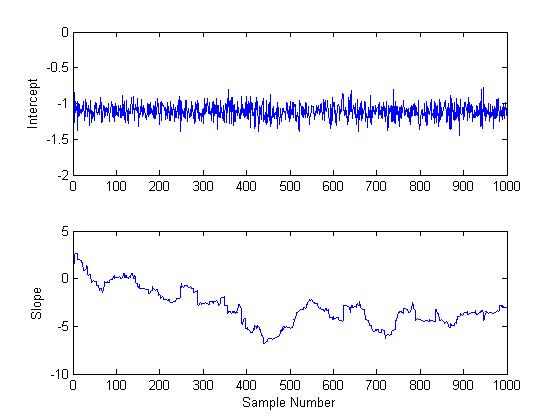


Figure 7:intercept and slope values after 1000 samples

Studying the above diagrams, it appears that the slope is converging towards3 while the intercept is settling between -1 and -1.5. taking the values of the slope at 3.5 and the intercept at -1.25, we attempted to calculate the equation once again.

Figure 8: BBR calculations with intercept and slope from Matlab

# Chapter 8.Conclusions

Before we make an attempt to sum up our conclusions, we believe that a background is in order. The first and foremost point is that we were attempting to find a relationship and did not have any prior idea that there is a relationship between the predictor and the response variables. Secondly we took the sample for a particular period and we do not know what impact the variable of time may have had on the results. Finally we were trying to assess the impact of one variable age on the outcome. In the real life situation, no response variable is a dependent on a single variable. Keeping in mind these factors, we would like to state our conclusions from our observations as follows:

1. Values observed for Bayesian Binary Regression method were lower than the simple logistic regression on all the methods we adopted
2. Values for the intercept and slope were significantly different in each method from excel to R to matlab
3. We did not see a predictability in the values in the sense that the predicted probabilities did not reflect the actual outcome observed in the test samples.
4. All these lead us to believe that there is no relationship between the predictor variable and the outcome variable . Or to put it more precisely, the relationship between the predictor variable (Age) and the outcome variable( RenewalRate) is very weak.

Of course, our discussions with the people in the industry also seem to indicate that there is no direct relationship beween the age of the reader and the tendency to renew the subscription. Unfortunately the data set did not contain other predictor variables. We feel that it would be interesting to repet this study at a later time with more predictor variables such as income and interest/hobbies of the reader as these variables may have a closer relationship to magazine subscription.

# References:

* Bayesian Statistics and Marketing by PE Rossi GM Allenby and R McCulloch, 2005
* Predicting Customer Response to Direct Marketing- A Bayesian Approach by Chen Wei, 2007
* Bayesian Logistic Models for Credit Scoring – A thesis paper by Gregg Weber (Rhodes University)