# BAYESIAN CUSTOMER LIFETIME VALUE (CLV) MODELING: THE BG-NBD MODEL

VITOR MARQUES RODRIGUES

# BAYESIAN CUSTOMER LIFETIME VALUE (CLV) MODELING: THE BG-NBD MODEL

Monograph presented to the Undergraduate Program in Statistics of the Federal University of Minas Gerais in partial fulfillment of the requirements for the degree of Bachelor in Statistics.

Advisor: Prof. Fábio N. Demarqui

Belo Horizonte

December 2024

# Resumo

Entre os profissionais de marketing, há grande interesse em modelar o comportamento de compra de clientes para fazer previsões sobre compras futuras e monitorar o valor do tempo de vida do cliente (CLV), que é a receita futura que um determinado cliente trará para a empresa. Dado um banco de dados de clientes contendo informações sobre a frequência e o momento das transações, tal feito pode ser alcançado usando o modelo beta-geométrico-NBD, em que a parte beta-geométrica modela o processo de "abandono", enquanto o componente NBD (mistura Poisson-Gamma) captura o comportamento de recompra. Uma vez estimado, o modelo gera uma taxa de compra e uma probabilidade de estar "ativo" para cada cliente, gerando insights interpretáveis e acionáveis. Exploramos a derivação da verossimilhança de tal modelo, discutindo por que uma abordagem Bayesiana produz resultados melhores tanto do ponto de vista prático quanto teórico. Além disso, apresentamos a implementação no Stan e aplicamos o modelo a um banco de dados real.

# Abstract

Among marketing practitioners, there is a great deal of interest in modeling customer purchase behavior in order to make forecasts about future purchases and track the customer lifetime value (CLV), which is the future revenue a given customer will bring to the company. Given a customer database containing information on the frequency and timing of transactions, such a feat can be achieved using the beta-geometric-NBD model where the beta-geometric part models the "dropout" process, whereas the NBD (Poisson-Gamma mixture) component captures the repeat-buying behavior. Once estimated, the model outputs a rate of purchase and a probability of being "alive" for each customer, generating interpretable and actionable insights. We explore the likelihood derivation of such a model, discussing why a Bayesian approach yields better results both from a practical and theoretical standpoint. In addition to that, we showcase the implementation in Stan and apply the model to a real-world database.

# List of Figures

# List of Tables

# Contents

# Chapter 1

# Introduction

A metric of great importance for customer-facing businesses is the Customer Lifetime Value (CLV), the future revenue a customer will bring to the company. In possession of such a value, businesses can extract the most out of their marketing budget by tailoring marketing campaigns according to the CLV of each customer. For instance, a company can identify and thus reward precious customers (i.e. high CLV) resulting in improved retention. Alternatively, the company can spot low and medium CLV customers and profile them to understand what stops them from purchasing more.

Usually, businesses calculate imprecise descriptive statistics to get a sense of the CLV of their customer base and therefore rank their customers based on such subjective measures. Over the past decades, with advancements in machine learning (ML) and increased computational processing power, successful ML applications to predict CLV have been made. Despite these advancements, the high computational burden and the poor interpretability of ML models make them less appealing for practical use. Therefore, many statistical methods and parametric models have been developed to predict the CLV efficiently and make the results interpretable.

Modeling the CLV depends heavily on the business of interest and the assumptions and abstractions made about the buying process. Broadly speaking, the purchase process can be broken into three components, that is:

- **Dropout**: A customer can decide to end their relationship with the business, and therefore no future purchase will be made. The dropout can be observable, for instance, a client can cancel their subscription for a given service, and in this scenario, the business offering the service knows with certainty that the customer left. On the other hand, the dropout can be unobserved, that is, the company can not know with certainty that the customer left, for example, a grocery store

customer who has not made any purchase in the last six months has probably left, but it is impossible to know with certainty if he is going to make future purchases.

Depending on the nature of the dropout, a different statistical approach has to be chosen. In the observable case, the time to "dropout" can be estimated with survival analysis techniques, whereas in the non-observable scenario, the probability of never making a purchase again should be estimated based on past buying behavior.

- **Frequency**: While the customer is active ("dropout" did not happen), they can make recurrent purchases, for instance, a grocery customer can buy approximately every two weeks, whereas another customer can purchase every month instead.

  According to the assumption about this recurrent process, the statistical treatment differs. For example, if it is reasonable to assume that each customer buys at a constant rate, then a homogeneous Poisson process might be a good choice. However, if one desires to account for a varying purchase rate (i.e. maybe the customer buys more frequently now than a year ago, after establishing trust in the business), a more complex process should be chosen.

- **Spend**: For each purchase, a certain monetary value is spent. For instance, a grocery customer can have an average spend of $80 per purchase, whereas a different customer can have a much higher average spend of $150 per purchase.

So, to predict the CLV for the next year for a given customer, one has to combine these components to answer questions such as: "Is the customer active in the following year?", "How many purchases will the customer make in the following year?" and "For each purchase, how much will they spend?". Figure 1.1 presents an example of what such a process looks like.

Many models have been developed over the years, each particularly suited for a given scenario and making different assumptions about these underlying processes. In this work, the focus will be on modeling the "dropout" and frequency process (i.e. ignoring spend) and assuming unobserved "dropout".

One of the first statistical approaches to model customer purchase behavior can be found in [Ehrenberg, 1959] and uses an NBD (Negative Binomial Distribution) counting process to understand the recurrence of purchases. However, this model tends to generate predictions that overestimate the real number of purchases, because it does not account for the "dropout" possibility in its structure. To address this issue

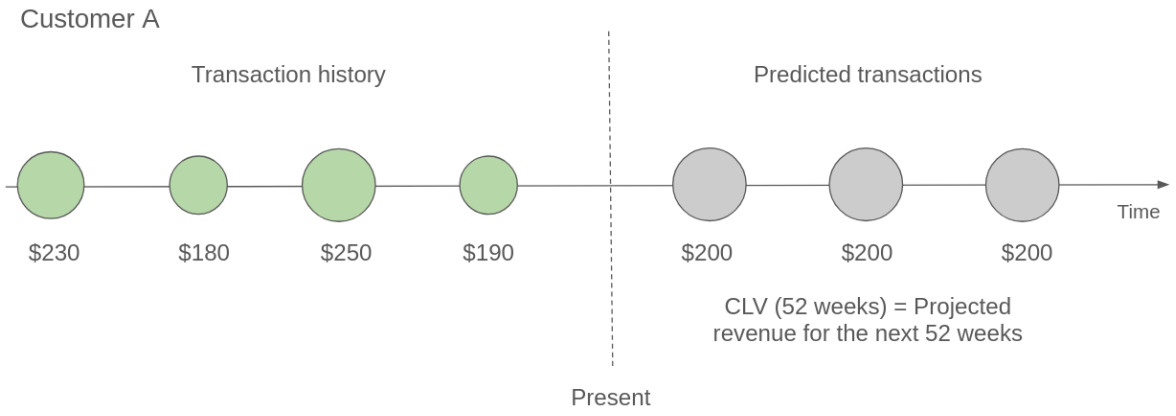


**Figure 1.1.** Example calculation CLV

[Schmittlein et al., 1987] proposed the Pareto/NBD model in which time to "dropout" is modeled using the Pareto (exponential-gamma mixture) timing model, and the recurrence while active modeled with an NBD (Poisson-Gamma mixture) counting model, producing much more accurate results.

Despite being powerful, the parameters of the Pareto/NBD are hard to estimate in practice, so alternative models with better-behaved likelihood functions have been proposed ever since. One such model is the BG/NBD model introduced in [Fader et al., 2005], in which the idea and interpretation are pretty much the same as the Pareto/NBD but much more tractable computationally.

The goal of this work is to discuss the derivation of the BG/NBD model and implement it using a Bayesian approach. In Chapter 2, we derive the likelihood of the model and its variants and discuss the implementation in Stan (see [Stan Development Team, 2024b]). Subsequently, in Chapter 3 we evaluate the model performance on simulated data and compare its variant's performance. In Chapter 4, we apply the model to a real-world database and discuss how to derive quantities of interest. We close this monograph with conclusions and a wrap-up in Chapter 5. Also, we make available the R package "bytd" with functions to fit the models presented in this work (details in Section 2.3).

# Chapter 2

# Methodology

## 2.1 Overview of the BG/NBD model

The BG/NBD model is based on the following assumptions about the buying process outlined in [Fader et al., 2005]

### 2.1.1 Frequency Process

- While active, the time between transactions made by a customer follows an exponential distribution with transaction rate $\lambda$:

$$f(t_j|t_{j-1};\lambda) = \lambda e^{-\lambda(t_j - t_{j-1})},\ t_j > t_{j-1} \geq 0,\ \lambda > 0. \tag{2.1}$$

  That is the same as saying that while active the frequency process follows a Poisson process with rate $\lambda$.

- Heterogeneity in $\lambda$ follows a gamma distribution with pdf (probability density function):

$$f(\lambda|r,\alpha) = \frac{\alpha^r \lambda^{r-1} e^{-\lambda\alpha}}{\Gamma(r)},\ \lambda > 0,\ r > 0 \text{ shape},\ \alpha > 0 \text{ scale}. \tag{2.2}$$

Therefore, the frequency process on an aggregate level can be seen as a Poisson-Gamma mixture, which is mathematically equivalent to a Negative Binomial Counting process, thus the name NBD (Negative Binomial Distribution) for the model.

Another way to interpret the frequency process is through a hierarchical structure. That is to say that each customer has been given a transaction rate $\lambda$ drawn from a gamma-distributed random variable shared across all customers. After that, each cus-

tomer buys according to a Poisson process governed by their assigned transaction rate. This structure highlights that even though each customer has an individual transaction rate, there is some similarity across customers' transaction rates (and thus buying behavior) imposed by this Gamma random process of transaction rate generation.

### 2.1.2 Dropout Process

- After any transaction, the customer becomes inactive with probability $p$, thus the point at which the customer "drops out" is distributed across transactions according to a (shifted) geometric distribution with pmf (probability mass function):

$$P(\text{inactive immediately after } jth \text{ transaction}) = p(1-p)^{j-1}, j = 1, 2, 3, ..., \ 0 \leq p \leq 1. \tag{2.3}$$

- Heterogeneity in dropout:

$$f(p|a,b) = \frac{p^{a-1}(1-p)^{b-1}}{B(a,b)}, 0 \leq p \leq 1, \ a > 0, \ b > 0. \tag{2.4}$$

- The transaction $\lambda$ and $p$ vary independently across customers

So, the dropout process can be seen on an aggregate level as a Beta-Geometric mixture, thus the name BG. Alternatively, one can interpret it hierarchically, as each customer receives an individual $p$ drawn from a Beta distribution shared across all customers. After that, each customer "drops out" according to a Geometric distribution governed by their assigned probability of "dropping out".

### 2.1.3 Observed Data

We can interpret each realization of the BG/NBD process as the transaction data for one particular customer. This observed data is a set of times $t_i \in (0, T]$ indicating when the transactions were made. For a customer who made $x$ repeated transactions, it looks something like:


In practice, the variable $T$ (often called longevity) is the time passed between the first transaction of the customer and the current date. Having calculated that, the time

of every transaction is taken relative to $T$, in which the first transaction takes value 0 and the following ones $t_1, t_2, ...t_x$ in the interval $(0, T]$. Following [Fader et al., 2005], we also assume that the customer is "alive" before $t_1$, that is, at least one repeated transaction takes place.

For example, assume that the present date is 30/01/2024 and Alice made four purchases at 01/01/2024, 07/01/2024, 11/01/2024, 18/01/2024. So, her longevity is $T = 30$ and the $x = 3$ repeated transactions take place at $t_1 = 6, t_2 = 10, t_3 = 17$.

### 2.1.4 Simulating the Process

In light of the explanations above, given the parameters of the process $r, \alpha, a, b$ and a maximum longevity $T_{max}$, one can simulate $N$ realizations (or rather transaction data for $N$ customers) as follows:

---

BG/NBD data generation algorithm

**Require:** $N, T_{max}, r, \alpha, a, b > 0$
**Ensure:** $W = \{w_1, w_2, ..., w_N\}$

$W \leftarrow \emptyset$
**for** $i \leftarrow 1$ to $N$ **do**
    $\lambda_i \sim Gamma(r, \alpha)$
    $p_i \sim Beta(a, b)$
    $T_i \sim ROUND[Unif(0.8, 1.0)T_{max}]$
    $w_i \leftarrow \{0\}$
    **while** True **do**
        $t_{candidate} \leftarrow ROUND(Exp(\lambda_i)) + MAX(w_i)$
        **if** LENGTH$(t_i) > 1$ and $(Bernouli(p_i)$ or $t_{candidate} \geq T_i)$ **then**
            **break**
        **end if**
        append $t_{candidate}$ to $w_i$
    **end while**
    append $w_i$ to $W$
**end for**
**return** $W$

---

## 2.2 Model Development

In addition to the parameters $r, \alpha, a, b$, the hierarchical structure of the model imposes unobserved $\lambda$ and $p$ for each customer, so-called random effects. One can decide to develop the model at different levels with regard to the treatment of these random effects. The decision as to how to treat them, affects not only the likelihood derivation

and the interpretation, but also the computational approach necessary to estimate the model. In the following subsections, we derive four models with varying treatments for the random effects.

### 2.2.1 Conditioning on both $\lambda$ and $p$

In this approach, we look at the transaction data of a particular customer assuming they have been given an unknown $\lambda$ and $p$. That is, we derive the likelihood at the individual level conditioning on unobserved $\lambda$ and $p$.

Considering a customer who had $x$ repeated transactions in the period $(0, T]$, with transactions occurring at $t_1, t_2, ..., t_x$.


The likelihood, as derived in [Fader et al., 2005], is:

$$L(\lambda, p|X = x, T) = (1-p)^x \lambda^x e^{-\lambda T} + \delta_{x>0} p(1-p)^{x-1} \lambda^x e^{-\lambda t_x}, \tag{2.5}$$

where $\delta_{x>0} = 1$, if $x > 0$, 0 otherwise.

Note that a sufficient statistic for this model is $(X = x, t_x, T)$, that is, the specific timing of each transaction is not necessary. This is a consequence of the strong assumption that the transaction rate of each customer is constant, which drastically simplifies the model. Furthermore, this property is particularly appealing from a computational viewpoint, since one has to carry only three numbers for each customer instead of the whole transaction history, making the model more scalable to large customer bases.

This individual-level model can not be estimated by standard MLE (Maximum Likelihood Estimation) methods. That is because it is conditional on $\lambda$ and $p$, so there is still a layer above in which these latent variables are generated by a $Gamma(r, \alpha)$ and $Beta(a, b)$, respectively. For a customer base of size $N$, to estimate the parameters $r, \alpha, a, b$ and the latent variables $\{\lambda_i\}_{i=1}^N$, $\{p_i\}_{i=1}^N$, one has to rely on EM (Expectation Maximization), or a Bayesian Approach and MCMC (Markov Chain Monte Carlo) methods.

An interesting aspect of this model is that we compute the latent variables $\lambda$ and $p$ for each customer, which can have useful interpretations and applications. For example, one can cluster customers based on their transaction rate, i.e. weekly customers, monthly customers, etc. Furthermore, one can spot loyal recurrent customers ($p \approx 0$), or filter churned customers ($p \approx 1$), tailoring marketing campaigns accordingly.

However, for a base of $N$ customers, it is necessary to estimate $2N+4$ parameters,

which imposes a high computational burden. This might lead to convergence issues and high costs for large databases.

### 2.2.2 Conditioning on $\lambda$

The expression derived in 2.5 is conditional on latent variables $\lambda$ and $p$. We can condition only on the latent variable $\lambda$ by taking the expectation of 2.5 over the distribution of $p$. The likelihood follows:

$$\begin{aligned}
L(a,b,\lambda|X=x,T,t_x) &= \int_{-\infty}^{\infty} L(\lambda,p|X=x,T,t_x)f(p)dp \\
&= \int_{-\infty}^{\infty} L(\lambda,p|X=x,T,t_x)\frac{p^{a-1}(1-p)^{b-1}}{B(a,b)}dp \\
&= \int_{-\infty}^{\infty} (1-p)^x\lambda^x\exp(-\lambda T)\frac{p^{a-1}(1-p)^{b-1}}{B(a,b)}dp+ \\
&\delta_{x>0}\int_{-\infty}^{\infty}(p(1-p)^{x-1}\lambda^x\exp(-\lambda t_x))\frac{p^{a-1}(1-p)^{b-1}}{B(a,b)}dp \\
&= \frac{\lambda^x}{B(a,b)}[\exp(-\lambda T)B(a,b+x)+\delta_{x>0}\exp(-\lambda t_x)B(a+1,b+x-1)]
\end{aligned}$$

Note that there is still the random effect $\lambda$ present, so, similar to the individual-level model case, estimation is only possible via EM or a Bayesian Approach. In this setting, however, for a base of $N$ customers, $N+4$ parameters must be estimated, that is $r, \alpha, a, b$ and latent variables $\{\lambda_i\}_{i=1}^N$. So, one has to estimate $N$ parameters less than the individual-level model at the expense of losing the interpretation and application of the random effect $p$.

### 2.2.3 Conditioning on $p$

Analogous to the previous case, to derive the model conditional only on $p$, we take the expectation of 2.5 over the distribution of $\lambda$. The likelihood follows:

$$
\begin{aligned}
L(r,\alpha,p|X=x,T,t_x) &= \int_{-\infty}^{\infty} L(\lambda,p|X=x,T,t_x)f(\lambda)d\lambda \\
&= \int_{-\infty}^{\infty} L(\lambda,p|X=x,T,t_x)\frac{\alpha^r\lambda^{r-1}\exp(-\lambda\alpha)}{\Gamma(r)}d\lambda \\
&= \frac{\alpha^r(1-p)^x}{\Gamma(r)}\int_{-\infty}^{\infty}\lambda^{x+r-1}\exp(-\lambda(\alpha+T))d\lambda + \\
\delta_{x>0}\frac{\alpha^r p(1-p)^{x-1}}{\Gamma(r)}&\int_{-\infty}^{\infty}\lambda^{x+r-1}\exp(-\lambda(\alpha+t_x))d\lambda \\
&= \frac{\alpha^r(1-p)^x\Gamma(r+x)}{\Gamma(r)}[\frac{1}{(\alpha+T)^{r+x}} + \delta_{x>0}\frac{p}{(1-p)(\alpha+t_x)^{r+x}}]
\end{aligned}
$$

In this setting, for a base of $N$ customers, $N+4$ parameters must be estimated, that is $r,\alpha,a,b$ and latent variables $\{p_i\}_{i=1}^N$. So, one has to estimate $N$ parameters less than the individual-level model at the expense of losing the interpretation and application of the random effect $\lambda$.

### 2.2.4 No random effects

To derive the model without any latent variable, it is necessary to take the expectation of 2.5 over the distribution of $\lambda$ and $p$. That is:

$$
\begin{aligned}
L(r,\alpha,a,b|X=x,T,t_x) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} L(\lambda,p|X=x,T,t_x)f_\lambda(\lambda)f_p(p)d\lambda dp \\
&= \int_{-\infty}^{\infty} L(r,\alpha,p|X=x,T,t_x)f_p(p)dp \\
&= \frac{B(a,b+x)}{B(a,b)}\frac{\Gamma(r+x)\alpha^r}{\Gamma(r)(\alpha+T)^{r+x}} + \delta_{x>0}\frac{B(a+1,b+x-1)}{B(a,b)}\frac{\Gamma(r+x)\alpha^r}{\Gamma(r)(\alpha+t_x)^{r+x}}
\end{aligned}
$$

In this scenario, there is no latent variable, so MLE is straightforward. Furthermore, for any customer base, we only have to estimate 4 parameters, $r,\alpha,a,b$. These properties make this formulation very simple to work with and this is the reason why it is widely used in practice. [Fader et al., 2005] focus on this model and derive managerial interesting quantities from the estimated parameters, such as the expected number of purchases in the future and so on.

However, this simplification has a price, which is the loss of the random effects $\lambda$ and $p$ and thus of their interpretation.

## 2.3 Implementation

We have implemented the aforementioned models in Stan and interfaced it with R using RStan (see [Stan Development Team, 2024a]). The R package with the complete code can be found at `github.com/vitormarquesr/bytd`. We note below some features and details of the implementation:

- Each model supports a Bayesian estimation procedure in which we impose a Gamma distributed prior for the parameters $r, \alpha, a, b$. By default, these priors are non-informative Gamma(0.01, 0.01), but the shape and scale of each prior can be defined by the user as well. We opted for a Gamma distributed prior because of its richness in shape and the property that its parameters can be easily set in terms of a given mean and variance.

- In addition to the Bayesian estimation approach, the model with no random effects also supports estimation via MLE.

- Users can choose to generate the likelihoods after the model is estimated, which can be used to perform sensitivity analysis and model comparison using the R package loo (see [Vehtari et al., 2024]).

# Chapter 3

# Simulation

After deriving the four different BG/NBD model formulations, the next natural development is to compare them and decide which version is better in practice. To do that, we take a two-step approach.

Firstly, we fit the four models to a synthetic database and compare them according to different aspects, choosing the variation we consider the best. Having done that, we perform a Monte Carlo study with 1000 replicas for the selected model to asses how well it retrieves the real parameters of the process.

All the synthetic databases were generated by Algorithm 2.1.4 with the following parameters:

$$N = 2500;\ T_{max} = 80;\ r = 0.3;\ \alpha = 7;\ a = 0.6;\ b = 3 \tag{3.1}$$

These parameters produce a database with characteristics that mimic a real-world scenario. This data could represent, for example, a cohort of $N = 2500$ customers who started their relationship with the business approximately $T = 80$ weeks ago. Furthermore, these customers have the following characteristics:

- From 2.2, they have on average a transaction rate of $\frac{r}{\alpha} \approx 0.043\ \frac{\text{purchases}}{\text{week}}$ ($\approx 2$ purchases per year) with standard deviation $\frac{\sqrt{r}}{\alpha} \approx 0.078\ \frac{\text{purchases}}{\text{week}}$. This makes sense for a company that sells durable goods, given that customers only have to buy them a few times a year.

- From 2.4, they have on average a "drop out" probability of $\frac{a}{(a+b)} = 0.167$ with standard deviation $\frac{\sqrt{a}}{(a+b)} = 0.215$. This average indicates that many customers have a propensity not to come back after a few purchases. However, there is significant variability in this behavior as pointed out by the standard deviation.

We fit each model using a Bayesian approach with non-informative prior $Gamma(0.01, 0.01)$ for each parameter $r, \alpha, a, b$. We run 4 chains with 2000 iterations each with Stan.

## 3.1 Model Comparison

After fitting the four models to the synthetic database, we compare them according to different aspects.

| Model | Parameter | Real Value | Mean Estimate | $n_{eff}$ | $\hat{R}$ |
|---|---|---|---|---|---|
| No Random Effect ($M_1$) | $r$ | 0.3 | 0.28 | 2229 | 1.00 |
| | $\alpha$ | 7 | 6.55 | 2179 | 1.00 |
| | $a$ | 0.6 | 0.86 | 1759 | 1.00 |
| | $b$ | 3 | 5.50 | 1704 | 1.00 |
| Conditional $\lambda$ ($M_2$) | $r$ | 0.3 | 0.28 | 586 | 1.01 |
| | $\alpha$ | 7 | 6.54 | 997 | 1.00 |
| | $a$ | 0.6 | 0.88 | 2175 | 1.00 |
| | $b$ | 3 | 5.71 | 2002 | 1.00 |
| | $\lambda[1]$ ... $\lambda[2500]$ | | | | |
| Conditional $p$ ($M_3$) | $r$ | 0.3 | 0.28 | 2741 | 1.00 |
| | $\alpha$ | 7 | 6.55 | 2511 | 1.00 |
| | $a$ | 0.85 | 0.88 | 4 | 1.62 |
| | $b$ | 3 | 5.51 | 5 | 1.59 |
| | $p[1]$ ... $p[2500]$ | | | | |
| Conditional $\lambda$ and $p$ ($M_4$) | $r$ | 0.3 | 0.28 | 630 | 1.01 |
| | $\alpha$ | 7 | 6.54 | 917 | 1.00 |
| | $a$ | 0.6 | 0.85 | 26 | 1.10 |
| | $b$ | 3 | 5.51 | 28 | 1.09 |
| | $\lambda[1]$ ... $\lambda[2500]$ | | | | |
| | $p[1]$ ... $p[2500]$ | | | | |

**Table 3.1.** Estimates, convergence and efficiency measures of each model

| Model | Runtime |
|---|---|
| No Random Effect | 4.5 minutes |
| Conditional $\lambda$ | 22 minutes |
| Conditional $p$ | 17 minutes |
| Conditional $\lambda$ and $p$ | 15 minutes |

**Table 3.2.** Runtime of each model

| Model | Elpd | Rank |
|---|---|---|
| No Random Effect | -19273 | 4 |
| Conditional $\lambda$ | -17164 | 2 |
| Conditional $p$ | -19079 | 3 |
| Conditional $\lambda$ and $p$ | -16944 | 1 |

**Table 3.3.** Loo results

From table 3.1, we see that even though the main parameters (i.e. $r, \alpha, a, b$) estimates are relatively the same for each variant, the efficiency measure $n_{eff}$ (effective sample size) and convergence measure $\hat{R}$ indicate problems (see [Stan Development Team, 2024a] for precise definitions of these measures).

In particular, for both the models $M_3$ and $M_4$, the effective sample size of $a$ and $b$ are low, indicating a high autocorrelation in the chains. Furthermore, their $\hat{R} > 1.05$ indicate that chains have not mixed, and the sample generated is unreliable. These issues seem to arise from keeping $p$ as a random effect, which makes it harder to estimate the corresponding parameters $a, b$ from 2.4 a layer above.

Note that for the model $M_2$, the $n_{eff}$ for $r, \alpha$ is much lower than their counterpart in $M_1$. So, leaving $\lambda$ as a random effect also makes it harder to estimate the parameters $r, \alpha$ from 2.2 a layer above. But, in this case, since all $\hat{R} < 1.05$ and their $n_{eff}$ is big enough, the difficulty for the random effect $\lambda$ is manageable. Thus, the only two models that presented no efficiency or convergence issues were the ones with only random effect $\lambda$ or no random effect.

From a different standpoint, if we look at the runtime of the models in 3.2 we see that $M_2$ takes almost 17 minutes more to fit than $M_1$, which is expected given the latter has to estimate 2504 parameters whereas the former only 4. This difference might become even more expressive for large customer bases, however, estimating the transaction rate of each customer is very useful in practice, so this trade-off runtime-interpretability has to be accounted for depending on the application. In our case, since we will be dealing with relatively small customer bases, keeping the random effect $\lambda$ is worth it.

To further guide our quest for the best model, we can look at the Expected Log Predictive Density (ELPD, see [Vehtari et al., 2017] for a precise definition) calculated with R package "loo" (see [Vehtari et al., 2024]). By looking at 3.3 we see that $M_2$ has a higher ELPD than $M_1$, indicating that the latter has a better predictive performance.

After the discussion above, we judge that $M_2$ ("Conditional $\lambda$") offers the best performance and interpretability while keeping estimation tractable and issue-free.

## 3.2 Monte Carlo Study

Having chosen $M_2$ as the best model, it is important to evaluate how well it retrieves the real parameters of the process, therefore, we fit it to 1000 randomly generated databases with parameters 3.1.

| Parameter | Real | Estimate | Std Err | Lower | Upper | Relative Bias | Coverage |
|---|---|---|---|---|---|---|---|
| $r$ | 0.3 | 0.302 | 0.013 | 0.275 | 0.329 | 0.682 | 0.93 |
| $\alpha$ | 7.0 | 7.071 | 0.512 | 6.109 | 8.118 | 1.017 | 0.96 |
| $a$ | 0.6 | 0.639 | 0.151 | 0.415 | 1.001 | 6.627 | 0.95 |
| $b$ | 3.0 | 3.448 | 1.126 | 1.910 | 6.220 | 14.957 | 0.96 |

**Table 3.4.** Monte Carlo results for "Conditional $\lambda$"

From table 3.4, we see that the model retrieves $r$ and $\alpha$ accurately as shown by the small relative bias ($< 2\%$) and standard error of the estimates. However, the parameters $a$ and $b$ present a high relative bias ($> 6\%$) and standard error indicating great uncertainty in these estimates. This is consistent with the previous discussions that the random effect $p$ and corresponding parameters $a$ and $b$ are problematic and difficult to estimate. Therefore, we should be careful using the parameters $a$ and $b$ as their estimates present high uncertainty in customer bases of the size we are working with.

# Chapter 4

# Application

In this analysis, we fit the model "Conditional $\lambda$" to a database of purchases of CDs at the online retailer CDNOW Inc. The dataset consists of weekly data on 2357 customers who made their first purchase in the first quarter of 1997 and spans roughly 78 weeks through June 1998, which is considered the "present". Further information on the dataset can be found in [Fader and Hardie, 2013b].

| Parameter | Mean Estimate | Standard Deviation |
|---|---|---|
| $r$ | 0.28 | 0.01 |
| $\alpha$ | 7.07 | 0.5 |
| $a$ | 0.63 | 0.14 |
| $b$ | 2.91 | 0.84 |

**Table 4.1.** Estimated parameters on the CDNOW database

After fitting the model with the default non-informative Gamma priors, we obtain the estimates in table 4.1. We can use the mean estimates of $r, \alpha, a, b$ to get a rough sense of the heterogeneity of the transaction rate and dropout rate across our customer base.

From 2.2 and 4.1, we can say that our customers' transaction rates are distributed approximately $\lambda \sim Gamma(0.28, 7.07)$. Thus, our average customer makes $\frac{r}{\alpha} = \frac{0.28}{7.07} \approx 0.0396 \ \frac{\text{purchases}}{\text{week}} \approx 2 \ \frac{\text{purchases}}{\text{year}}$. However, the standard deviation of $\frac{\sqrt{r}}{\alpha} = \frac{\sqrt{0.28}}{7.07} \approx 0.0748 \ \frac{\text{purchases}}{\text{week}}$ shows customers are somewhat heterogeneous regarding this behavior. Furthermore, looking at 4.1 we see a high density close to zero, suggesting there are many infrequent buyers in our base. This should be investigated further to understand why some customers buy so infrequently, and hopefully try to engage them more.

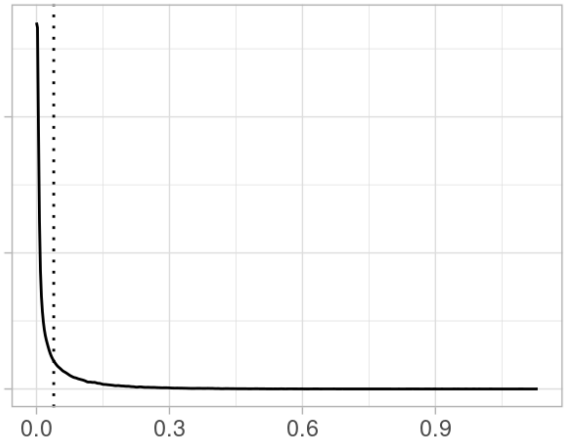


**Figure 4.1.** Transaction rate heterogeneity

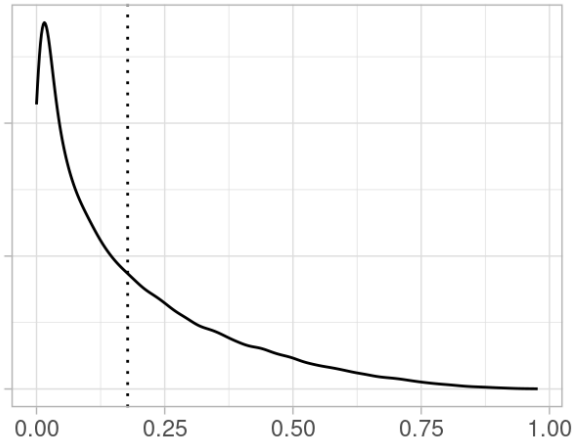


**Figure 4.2.** Dropout rate heterogeneity

From 2.4 and 4.1, we can say our customers' dropout rate are distributed approximately $p \sim Beta(0.63, 2.91)$. Therefore, our average customer has a probability of becoming inactive after any transaction of $\frac{a}{(a+b)} = \frac{0.63}{0.63+2.91} \approx 0.17796$, making on average $\frac{1}{0.17796} = 5.62$ transactions before dropping out. This value could be considered low, indicating problems with customer retention. However, looking at 4.2 we see a high density close to zero, indicating a presence of customers with no propensity to drop out, which is very good. On the other hand, we note a heavy tail going towards the right, meaning a significant part of our customer base has a propensity to drop out after just a few purchases, which points to problems in retention.

It is important to stress that the analysis using the mean estimates of the parameters $r, \alpha, a, b$ serves more as a "back-of-the-envelope" calculation for the macro insights we want. That is because, unlike Maximum Likelihood Estimators, the Equivariance

principle does not apply in a Bayesian setting. Therefore, using point estimates of the parameters to calculate the mean of the heterogeneity of $\lambda$ and $p$ with $\frac{r}{\alpha}$ and $\frac{a}{a+b}$ might lead to imprecise estimates. A better approach is to take the mean of the posterior distribution of these quantities, which is easy to do using posterior samples of $r, \alpha, a, b$. With this approach we get that $\mathbb{E}(\frac{r}{\alpha}) \approx 0.03904 \ \frac{\text{purchases}}{\text{week}}$ and $\mathbb{E}(\frac{a}{a+b}) \approx 0.18068$.

Analyzing the heterogeneity of $\lambda$ and $p$ is great for obtaining macro-level insights into our customer base, however, there are not many concrete actions we can take with this analysis only. To get insights we can act upon, we need individual-level information to identify recurrent customers, customers about to leave our business, and so on.

| Customer Id | $\lambda$ |
|---|---|
| 1981 | 0.470 |
| 1901 | 0.437 |
| 157 | 0.433 |
| 1203 | 0.408 |
| 814 | 0.404 |
| 1955 | 0.374 |
| 813 | 0.374 |
| 558 | 0.365 |
| 1516 | 0.354 |
| 1887 | 0.347 |
| ... | ... |

**Table 4.2.** Top 10 customers with the highest $\lambda$ mean estimates

Remember that the model we fit estimates the random effect $\lambda$, so we can look at this individual-level quantity to rank the customers and filter them according to their purchase rate. From Table 4.2 we see that the customer with highest purchase rate, "Customer 1981", makes on average $0.47 \ \frac{\text{purchases}}{\text{week}} \approx 1.88 \ \frac{\text{purchases}}{\text{month}} \approx 24.44 \ \frac{\text{purchases}}{\text{year}}$, a rate much higher than the average customer of our base, who makes approximately two purchases a year. There are many other customers with purchase rates much higher than the average, and we should reward them for buying with our business so religiously. For example, we can offer the top 10% customers discount coupons, or devise a personalized marketing campaign towards this high-frequency niche. Alternatively, we can try to correlate customers' purchase rates with other variables in our database, hopefully discovering something we can use that drives recurrence. These suggestions highlight the many possibilities we can use $\lambda$ for generating actionable and powerful analysis.

However, $\lambda$ alone does not present the full picture, because a customer might have been a frequent buyer in the past, and thus have a high $\lambda$, but have already ended

their relationship with the business. Therefore, we need a way to guess whether the customer is still active and will make future purchases. The random effect $p$ encodes an individual probability of dropping out after each purchase, so one might think that estimating this random effect might be the way. However, this is not what we want since we are looking for the probability that the customer is "alive" now rather than their intrinsic propensity to drop out. Furthermore, even if computing $p$ was what we wanted, we could not do it since our chosen model does not admit the random effect $p$.

The expression for the probability that the customer is alive given their history and model parameters is derived in [Fader et al., 2008] for the BG/NBD with no random effects. Following an analogous straightforward derivation, the expression 4.1 computes this quantity for the variation with only random effect $\lambda$.

$$P(\text{alive } |x, t_x\ T, \lambda, a, b) = \frac{1}{1 + \delta_{x>0} \frac{a}{b+x-1} e^{-\lambda(T-t_x)}} \tag{4.1}$$

| Customer Id | $P$(alive) |
|---|---|
| 1 | 0.959 |
| 2 | 0.973 |
| 3 | 1.000 |
| 4 | 1.000 |
| 5 | 1.000 |
| 6 | 0.973 |
| 7 | 0.964 |
| 8 | 0.854 |
| 9 | 0.943 |
| 10 | 1.000 |
| ... | ... |

**Table 4.3.** Customers' $P$(alive) mean estimates

From Table 4.3, we see customers with $P$(alive) very close to 1, indicating they are active and likely to purchase again. On the other hand, "Customer 8" has a $P$(alive) of 0.854, which is not that worrying, but might indicate they have not purchased in a while, therefore, it might be worth investigating this client further and possibly reach out to them to prevent them ending their relationship with the business permanently.

Looking at the posterior distribution of the individual $P$(alive)s in Figure 4.3, we see there is some uncertainty around these estimates for Customers 1, 2, 7, 8, and 9, whereas we have more certainty for Customers 3, 4, 5, 6 and 10. This ability to quantify uncertainty easily with posteriors could be useful in real-world decision-making and is one of the advantages of taking a Bayesian approach.

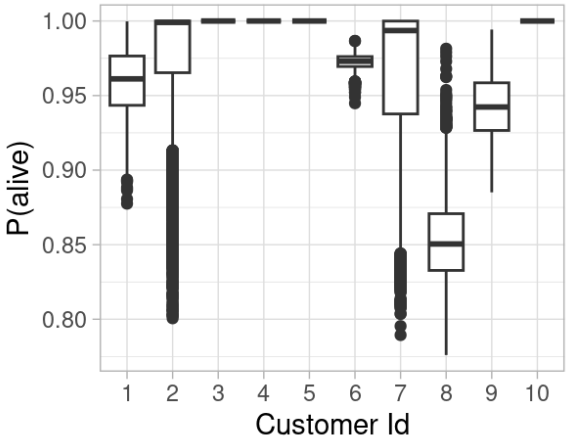


**Figure 4.3.** Customer's P(alive) posterior distribution

In practice, we could stratify our clients according to their $P(\text{alive})$ range and tailor our marketing campaigns to each group. For example, we could use it to filter "Churned" customers (those who have probably left, i.e. low $P(\text{alive})$) away in our marketing campaigns and prevent spending resources on customers who have already left and that would be too expensive to bring back. We highlight that the $P(\text{alive})$ of each customer changes through time and we can always track it to spot beforehand customers about to leave our business and better manage our customer base.

$$\mathbb{E}(Y(t) \mid X = x, t_x, T, \lambda, a, b) = \frac{\frac{a+b+x-1}{a-1}[1 - e^{-\lambda t}\ {}_1F_1(b+x, a+b+x-1, \lambda t)]}{1 + \delta_{x>0}\frac{a}{b+x-1}e^{\lambda(T-t_x)}} \quad (4.2)$$

| Customer Id | $\lambda$ | $P(\text{alive})$ | $\mathbb{E}(Y(52) \mid X = x, t_x, T)$ |
|---|---|---|---|
| 1981 | 0.470 | 0.993 | 19.9 |
| 157 | 0.431 | 0.995 | 18.1 |
| 1203 | 0.408 | 0.988 | 17.6 |
| 813 | 0.375 | 0.991 | 16.0 |
| 1516 | 0.356 | 0.993 | 14.8 |
| 1887 | 0.346 | 0.994 | 14.0 |
| 1017 | 0.290 | 0.982 | 12.5 |
| 2149 | 0.303 | 0.994 | 11.8 |
| 1539 | 0.276 | 0.985 | 11.7 |
| 1242 | 0.269 | 0.984 | 11.5 |
| ... | ... | ... | ... |

**Table 4.4.** Customers' $\mathbb{E}(Y(52) \mid X = x, t_x, T)$ mean estimates in descending order

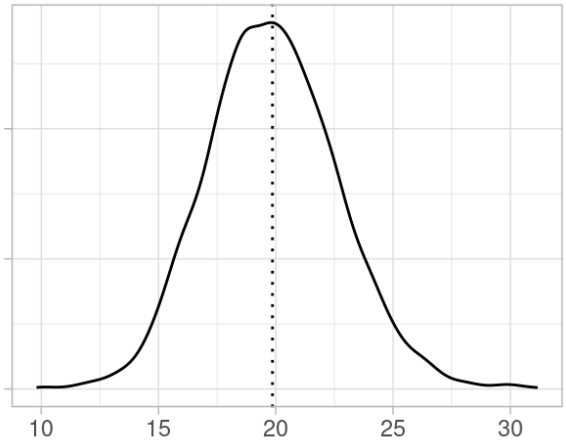


**Figure 4.4.** Posterior distribution of $\mathbb{E}(Y(52) \mid X = x, t_x, T)$ for "Customer 1981"

The quantities $\lambda$ and $P(\text{alive})$ inform us about two different aspects of customer behavior, that is the recurrence of purchases and the dropout. However, we would like to derive a measure that combines these two aspects to predict how many purchases the client will make in the future. Equation 4.2 computes the expected number of purchases in the period $(T, T+t]$ for a client with history $X = x, t_x, T$ and transaction rate $\lambda$ (see Appendix A for a careful derivation).

From Table 4.4 we see the clients with the highest expected number of purchases for the next 52 weeks (roughly a year). This is extremely valuable because we can spot these high-value clients and reward them for their commitment, further cementing their loyalty to the business. On another note, once we model the average spend of each client (see [Fader and Hardie, 2013a] for an easy way to do it), we can estimate the future Customer Lifetime Value (CLV) on an individual level and use it to get a rough sense of how much money the business will make in a given period. More ambitiously, we can also use this prediction to value a company by defining its value as the predicted CLV of its customer base, say for the next three years. These suggestions showcase the many possibilities we can use this "expected number of purchases" to improve decision-making in a business scenario.

Note that since we have used a Bayesian approach, we can take a posterior sample of the quantities of interest. Figure 4.4 shows this sample for the expected number of purchases for "Customer 1981" and thus we can compute a credibility interval with the HDI (*Highest Density Interval*) method.

# Chapter 5

# Conclusion

In this study, we present an overview of the BG/NBD model and apply it to a real-world database, showcasing how its findings can be used to add business value. More specifically, we explore a formulation of the model that computes the random effect $\lambda$ (transaction rate) for each customer and we choose to use a Bayesian approach to deal with these latent variables. We implement the BG/NBD models in Stan and perform a simulation study to assess their performance.

The work we have developed demonstrates how Statistical modeling can be used to model the Customer Lifetime Value (CLV) in a way that produces results that can be applied to improve CRM (Customer Relationship Management) strategies.

While the BG/NBD model provides a simple and powerful way to model the CLV, it has some limitations. For example, it has difficulty modeling heavy buyers properly because, according to the model assumptions, they have more opportunities to drop out (i.e., dropout can occur after each purchase). Another shortcoming is that the input data must contain customers of approximately equal $T$ (longevity), otherwise we face estimation issues. Future work may propose more complex and flexible models to address these issues.

# Appendix A

# Derivation of $E(Y(t) \mid X = x, t_x, T)$

Let $Y(t)$ be the number of purchases made in the period $(T, T+t]$. We are interested in computing the conditional expectation $\mathbb{E}(Y(t)|X = x, t_x, T, \lambda, a, b)$.

From [Fader et al., 2005] we have that:

$$\mathbb{E}(Y(t) \mid X = x, t_x, T, \lambda, p) = \frac{p^{-1}(1-p)^x \lambda^x e^{-\lambda T} - p^{-1}(1-p)^x \lambda^x e^{-\lambda(T+pt)}}{L(\lambda, p \mid X = x, t_x, T)} \tag{A.1}$$

Since we do not work with the random effect $p$ in this particular formulation, we take the expectation of A.1 over the distribution of $p$, updated to take account for $X = x, t_x, T$:

$$\mathbb{E}(Y(t) \mid X = x, t_x, T, \lambda, a, b) = \int_0^1 \mathbb{E}(Y(t) \mid X = x, t_x, T, \lambda, p) f(p \mid X = x, t_x, T, \lambda, a, b) dp \tag{A.2}$$

By Bayes theorem, it follows that:

$$f(p \mid X = x, t_x, T, \lambda, a, b) = \frac{L(\lambda, p \mid X = x, t_x, T) f(p \mid a, b)}{L(\lambda, p \mid X = x, t_x, T)} \tag{A.3}$$

From A.1 and A.3 in A.2, we have:

$$\mathbb{E}(Y(t) \mid X = x, t_x, T, \lambda, a, b) = \frac{W - Z}{L(\lambda, a, b \mid X = x, t_x, T)} \tag{A.4}$$

where

$$W = \int_0^1 p^{-1}(1-p)^x \lambda^x e^{-\lambda T} f(p \mid a, b) dp = \frac{B(a-1, b+x)\lambda^x e^{-\lambda T}}{B(a, b)}$$

and

$$
\begin{aligned}
Z &= \int_0^1 p^{-1}(1-p)^x \lambda^x e^{-\lambda(T+pt)} f(p \mid a,b) dp \\
&= \frac{\lambda^x e^{-\lambda T}}{B(a,b)} \int_0^1 p^{a-2}(1-p)^{b+x-1} e^{-\lambda p t} dp
\end{aligned}
$$

letting $q = 1 - p$ (and thus $dp = -dq$), and recalling that the *Moment Generating Function* of $R \sim Beta(\alpha, \beta)$ is $M_R(t) = {}_1F_1(\alpha, \alpha+\beta, t)$, where ${}_1F_1$ is the *confluent hypergeometric function of the first kind*, we have:

$$
\begin{aligned}
Z &= \frac{B(a-1,b+x)\lambda^x e^{-\lambda(T+t)}}{B(a,b)} \int_0^1 \frac{(1-q)^{a-2} q^{b+x-1}}{B(b+x,a-1)} e^{\lambda t q} dq \\
&= \frac{B(a-1,b+x)\lambda^x e^{-\lambda(T+t)}}{B(a,b)} M_{Beta(b+x,a-1)}(\lambda t) \\
&= \frac{B(a-1,b+x)\lambda^x e^{-\lambda(T+t)}}{B(a,b)} \; {}_1F_1(b+x, b+x+a-1, \lambda t)
\end{aligned}
$$

Substituting $W$ and $Z$ in A.4, we get:

$$
\mathbb{E}(Y(t) \mid X = x, t_x, T, \lambda, a, b) = \frac{\frac{a+b+x-1}{a-1}[1 - e^{-\lambda t} \; {}_1F_1(b+x, a+b+x-1, \lambda t)]}{1 + \delta_{x>0} \frac{a}{b+x-1} e^{\lambda(T-t_x)}} \quad \text{(A.5)}
$$

# Bibliography

[Ehrenberg, 1959] Ehrenberg, A. S. (1959). The pattern of consumer purchases. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 8(1):26--41.

[Fader and Hardie, 2013a] Fader, P. S. and Hardie, B. G. (2013a). The gamma-gamma model of monetary value. *February*, 2(2013):1--9.

[Fader and Hardie, 2013b] Fader, P. S. and Hardie, B. G. (2013b). Notes on the cdnow master data set.

[Fader et al., 2005] Fader, P. S., Hardie, B. G., and Lee, K. L. (2005). Counting your customers the easy way: An alternative to the pareto/nbd model. *Marketing Science*, 24(2):275--284.

[Fader et al., 2008] Fader, P. S., Hardie, B. G., and Lee, K. L. (2008). Computing p(alive) using the bg/nbd model. Research Note available via http://www.brucehardie.com/notes/021.

[Schmittlein et al., 1987] Schmittlein, D. C., Morrison, D. G., and Colombo, R. (1987). Counting your customers: Who-are they and what will they do next? *Management Science*, 33(1):1--24.

[Stan Development Team, 2024a] Stan Development Team (2024a). RStan: the R interface to Stan. R package version 2.32.6.

[Stan Development Team, 2024b] Stan Development Team (2024b). Stan modeling language users guide and reference manual. version 2.35.

[Vehtari et al., 2024] Vehtari, A., Gabry, J., Magnusson, M., Yao, Y., Bürkner, P.-C., Paananen, T., and Gelman, A. (2024). loo: Efficient leave-one-out cross-validation and waic for bayesian models. R package version 2.8.0.

[Vehtari et al., 2017] Vehtari, A., Gelman, A., and Gabry, J. (2017). Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and computing*, 27:1413--1432.