Buy-'Til-You-Die Models for Large Data Sets via Variable Selection

Rafael Dimaano
Advisor: Peter Fader

May 7, 2018

Abstract

In this paper, we analyze a class of probability models known as Buy-'Til-You-Die or BTYD models and lay out a prospectus for using variable selection techniques to adopt these methods to large data sets. The theory of the BTYD models is laid out and frameworks for incorporating regression elements to the model class are developed. We take variable selection procedures as developed for machine learning and extend them to the case of multi-component regression models. We then propose possibilities for formulating a full specification of a BTYD model that can effectively handle several independent variables.

1 Introduction

Although the standard linear model of regression is incredibly effective over a wide variety of applications, interest in the nuances of certain complex problems has led to more sophisticated regression techniques being developed to give a more detailed analysis than otherwise possible. In particular, the problem of analyzing and predicting customer behavior has resulted in the development of the class of Buy-'Til-You-Die models (abbreviated as BTYD) proposed by Fader and Hardie, which can make individual-level claims about customer behavior using only small samples. However, the complex specifications of the models in the class result in estimation procedures that are

statistically and computationally difficult, as well as an inability to incorporate observed variables into the model, Together, these two issues hinder the performance of BTYD models on larger data sets.

One solution to this problem can be found in the statistical learning literature, where a common problem that arises with regards to regression models is selecting which of a thousand or million independent variables should be included or excluded in the model. Since the number of subsets of independent variables is exponential in size to the number of independent variables, it is computationally very inefficient to fit a regression model for every possible subset of variables. Thus, algorithms such as stepwise, stagewise, and streamwise regression have been developed to perform a greedy search over the space of independent variables in polynomial time, yet still obtain near-optimal results in practice.

In this paper, we wish to explore these two topics to develop possible methods for combining the explanatory features of "small-data" models with some simplifying techniques from "big-data" methods. First, we develop the theory of the BTYD model class as a combination of individual components and reason about how such models can be better adopted as multi-part regressions over large data sets. We then turn our attention to variable selection procedures and modify such techniques to account for complex regression models which are built up from multiple regression components.

The benefits of this approach to these problems are twofold. First, this will increase the effectiveness of complex regression models such as the BTYD model class on larger data sets, which will allow their analytical and predictive power to be more effective when applied in practice. Secondly, algorithmic approaches that specifically deal with standard linear regression can now be generalized to apply to complex models, effectively opening up the opportunity to generalize other machine learning meta-algorithms.

Section 2 builds up a class of probability and regression models leading up to the BTYD class. Section 3 discusses possibilities for further introduction of regression coefficients in the BTYD framework and specifies a hypothetical model of interest. Section 4 discusses algorithmic variable selection procedures as currently used in the statistical learning literature. Section 5 then explores how these models can then be adopted to the case of BTYD regressions. Section 6 discusses possible economic and business opportunities for

BTYD modeling. Lastly, section 7 concludes and summarizes all recommendations for future work.

2 Regression Models and the BTYD Framework

A fundamental concept in modern statistics is the standard linear regression, usually known as ordinary least squares or OLS regression. While very commonly used, the standard model assumes an unbounded, real-valued support for the dependent variable. In this section, we present the framework of the generalized linear model as an extension to accommodate varying supports for the dependent variable, and discuss using mixture models to further generalize these models in an empirical Bayesian setting. We then conclude by discussing a class of models known as the Buy-'Til-You-Die models or BTYD, which combine many aspects of the above models and serve as a suitable setting for generalizing variable selection procedures.

2.1 Generalized Linear Models and Poisson Regression

The standard linear model requires several assumptions, an important one being that the response variable y is distributed according to $\mathcal{N}(\mu(\boldsymbol{x}), \sigma^2)$ with $\mu(\boldsymbol{x}) = \boldsymbol{\beta}^\intercal \boldsymbol{x}$. The generalized linear model or GLM, as specified by Nelder and Baker (1972), Wedderburn (1974), relaxes the assumptions placed on the standard model. In particular the GLM allows, with broad limitations, response variables to take on arbitrary distributions provided these are connected to the independent variables by an arbitrary function called a *link function*.

Both ordinary linear regression, as well as logistic regression, fall under the banner of the generalized linear model. In particular, any meta-algorithms that do not rely on the normality and linearity assumptions of the original linear model, such as variable selection algorithms will still work under the GLM framework.

A particular case of the GLM that is interesting for our purposes is the *Poisson Regression Model*. The Poisson regression formulation is a specific

case of a GLM, with the distribution being a Poisson distribution over the nonnegative integers and the link function being the natural logarithm, the inverse of which (called the *mean function* in the context of GLMs) is the exponential function.

The model is specified as follows:

- The response variables y follow a Poisson distribution as follows:

$$\mathrm{P}(y \mid \boldsymbol{x}, \boldsymbol{\beta}, \lambda_0) = \frac{\lambda(\boldsymbol{x})^y e^{-\lambda(\boldsymbol{x})}}{y!}$$

  Where $\lambda(\boldsymbol{x})$ is the Poisson rate parameter and:

$$\lambda(\boldsymbol{x}) = \lambda_0 \exp(\boldsymbol{\beta}^\intercal \boldsymbol{x})$$

  Where $\boldsymbol{x}$ is the vector of regressors and $\boldsymbol{\beta}$ is the vector of weights, and λ_0 is a normalizing factor, equivalent to an intercept in standard linear regression.

The Poisson regression and its variations are commonly used in econometrics and the social sciences as a model for count data, where response variables can take on unbounded positive integral values (Cameron and Trivedi, 2013).

2.2 Negative Binomial Regression

Although the Poisson Regression formulation is commonly used for count data, a major restriction inherent in the model is the assumption that the specified mean of the Poisson distribution is equal to its variance. We can relax this assumption by introducing a mixture model formulation of Poisson Regression.

- Starting from the Poisson Regression specification, we generalize this by specifying that λ_0 is now distributed randomly. In particular, we assume that λ_0 follows a Gamma distribution parametrized by shape parameter r and scale parameter α as follows:

$$\mathrm{P}(\lambda_0 \mid r, \alpha) = \frac{\alpha^r \lambda^{r-1} e^{\alpha\lambda}}{\boldsymbol{\Gamma}(r)}$$

Where $\mathbf{\Gamma}(r)$ is the gamma function.

- We then formulate the mixture model as:

$$\mathrm{P}(y \mid \boldsymbol{x}, \boldsymbol{\beta}, r, \alpha) = \int \mathrm{P}(y \mid \boldsymbol{x}, \lambda_0)\, \mathrm{P}(\lambda_0 \mid r, \alpha) d\lambda$$

Thus:

$$\mathrm{P}(y \mid \boldsymbol{x}, \boldsymbol{\beta}, r, \alpha) = \frac{\mathbf{\Gamma}(r+y)}{\mathbf{\Gamma}(r) y!} \left(\frac{\alpha}{\alpha + \boldsymbol{\beta}^\intercal \boldsymbol{x}}\right)^r \left(\frac{\boldsymbol{\beta}^\intercal \boldsymbol{x}}{\alpha + \boldsymbol{\beta}^\intercal \boldsymbol{x}}\right)^y$$

The above regression framework, known as *Negative Binomial Regression* or NBD regression, is particularly interesting for a variety of reasons. The case with no independent variables (simply called the NBD model) has been extensively used for customer behavior modeling since Ehrenberg in 1959. Compared to the standard Poisson regression, rather than having a constant λ for each observation, we can maintain a distribution which is both parametrized by the regressors and refined via Bayesian updating. If each observation $\boldsymbol{x}, y$ represents individuals in a population, then each λ can be interpreted as an unobserved individual-level parameter, distributed according to $r, \alpha, \boldsymbol{x}$.

2.3 Proportional Hazards Regression

Apart from being interested in a count of events happening, we may also be interested in the *time* taken between events. It turns out that a regression framework similar to Poisson regression exists for timing data. Such models are called *Proportional Hazards Models* and they were originally formulated by Cox in 1972:

- We first define the hazard function for a non-negative, real-valued distribution as:

$$\mathrm{h}(t) = \frac{\mathrm{f}(T = t)}{1 - \mathrm{F}(T < t)}$$

Where f is the probability density function and F the cumulative distribution function.

Note that given any hazard function $\mathrm{h}(t)$, we can obtain a corresponding cumulative distribution function through the following formula:

$$\mathrm{F}(t) = 1 - \exp\left(-\int_0^t h(u)du\right)$$

- We then specify the hazard function as:

  $$\mathrm{h}(t \mid \boldsymbol{x}, \boldsymbol{\beta}, \lambda_0) = \lambda_0 \exp(\boldsymbol{\beta}^\intercal \boldsymbol{x}(t))$$

  Where $\boldsymbol{x}$ is the vector of regressors and $\boldsymbol{\beta}$ is the vector of weights, and λ_0 is the exponential baseline hazard parameter.

- We will now assume that observations of $\boldsymbol{x}, y$ occur in discrete blocks of time, thus we can write the integral of the hazard rate and simplify it as:

  $$\int_0^t \mathrm{h}(u \mid \boldsymbol{x}(u), \boldsymbol{\beta}, \lambda_0)du = \lambda_0 \sum_{i=0}^{t} \exp(\boldsymbol{\beta}^\intercal \boldsymbol{x}(i)) = \lambda_0 A(t)$$

  Thus we can write the cumulative distribution function of the exponential proportional hazards regression model as:

  $$\mathrm{F}(t \mid \boldsymbol{x}, \boldsymbol{\beta}, \lambda_0) = 1 - e^{-\lambda_0 A(t)}$$

The exponential hazard baseline can further be generalized by using a similar mixture framework as described above:

- Again, we specify that λ_0 follows a Gamma distribution parametrized by shape parameter r and scale parameter α as follows:

  $$\mathrm{P}(\lambda_0 \mid r, \alpha) = \frac{\alpha^r \lambda^{r-1} e^{\alpha\lambda}}{\boldsymbol{\Gamma}(r)}$$

- Using a similar formulation for the mixture model as above. we have:

  $$\mathrm{F}(t \mid \boldsymbol{x}, \boldsymbol{\beta}, r, \alpha) = \int \mathrm{F}(t \mid \boldsymbol{x}, \boldsymbol{\beta}, \lambda_0)\, \mathrm{P}(\lambda_0 \mid r, \alpha)d\lambda_0$$

$$\mathrm{F}(t \mid \boldsymbol{x}, \boldsymbol{\beta}, r, \alpha) = 1 - \left(\frac{\alpha}{\alpha - A(t)} \right)^r$$

This is called the *Pareto II Model with Covariates*. The Pareto II distribution was originally formulated by Lomax (1954).

Note that the behavior induced by the addition of regressors in this model is slightly different than that of the negative binomial regression model above. In particular, the proportional hazards regression model allows for regressors that vary across time as opposed to regressors that just vary across individuals. This means that it will be ideal to use different regressors in statistical models that integrate both regression models as components, as we shall see below.

2.4 The Pareto/NBD BTYD Model

Taken as probability models (without any independent variables or regression coefficients), the above distributions can be combined in a class of models called *Buy-'Til-You-Die Models*, abbreviated as BTYD. These models are commonly used for analyzing and predicting future customer behavior, and involve the integration of various component distributions such as those described above to analyze more complex data and processes.

The Pareto/NBD model as proposed by Schmittlein et al. (1987) is the standard model used for customer behavior analysis. The model applies to continuous time noncontractual purchase settings, where the only observations made are the number of purchases per customer and when these purchases are made. The model specification is as follows:

- Customers exist in one of two states: They are "alive" for a random period of time after which they become permanently "dead".
- While alive, the number of times they purchase is distributed according to a time-varying Poisson distribution with an individual parameter λ:

$$\mathrm{P}(x(t) \mid \lambda) = \frac{(\lambda t)^x e^{-\lambda t}}{x!}$$

- Each customer's unobserved lifetime is exponentially distributed with an individual parameter μ as follows:

$$\mathrm{f}(\tau \mid \mu) = \mu e^{-\mu\tau}$$

- λ itself is distributed according to a gamma distribution with parameters r, α:

$$\mathrm{P}(\lambda \mid r, \alpha) = \frac{\alpha^r \lambda^{r-1} e^{\alpha\lambda}}{\mathbf{\Gamma}(r)}$$

- μ is also distributed according to a gamma distribution with parameters s, β:

$$\mathrm{P}(\mu \mid s, \beta) = \frac{\beta^s \mu^{r-1} e^{\beta\mu}}{\mathbf{\Gamma}(s)}$$

- λ and μ are independent across customers.

These assumptions result in deriving a combination of both an NBD and a Pareto II model, hence the overall model's name. The resulting model is incredibly complex, thus we refer the reader to Fader et al. (2005) for a full treatment of the derivation of the likelihood and other useful quantities. In fact, because of the model's complexity, several simpler models have been developed to serve as alternatives to the Pareto/NBD, such as the BG/NBD model as proposed by Fader et al. (2005).

3 Introducing Independent Variables in the BTYD Framework

Currently, Fader and Hardie (2007) specify the inclusion of independent variables in the Pareto/NBD using a different method than specified above. Rather than introducing a mixture over a constant coefficient in the regression model, they specify that the model parameters themselves are a function of the regression coefficients as follows:

$$\alpha = \alpha_0 \exp(-\boldsymbol{\gamma}_1 \boldsymbol{x}_1)$$

$$\beta = \beta_0 \exp(-\boldsymbol{\gamma}_2 \boldsymbol{x}_2)$$

Although mathematically convenient, such a specification is not ideal for the following reasons:

- The specification only works for time-invariant independent variables. This is a limiting assumption because while the dynamics of the NBD portion of the model are time-invariant, the dynamics of the Pareto portion can be better modeled with the hazard function varying across time.

- The coefficients do not have the same interpretability of the GLM and mixture fomulations specified above. In both Poisson and NBD regression, an additive increase in an independent variable corresponds to a proportional increase of a *specific* individual's latent rate of behavior, whether constant as in the Poisson formulation or governed by a prior as in the NBD formulation. In the Pareto/NBD specification, changes in the variables correspond to changes in the *entire distribution* of possible latent variables, thus each observed individual has a different associated prior.

- The effects of observed heterogeneity among individuals (as borne out by the variables and regression coefficients) cannot be separated from the effects of unobserved heterogeneity (as borne out by the Bayesian prior), which makes the value of observed data difficult to quantify.

Several papers, including Abe (2009), Babkin and Goldberg (2017), and Schweidel and Knox (2013), have attempted to modify the specifications of the Pareto/NBD and other BTYD models further to accommodate time-varying variables separately from the model parameters. As of now there is no proposed model that is able to account for time-varying variables in both the count and timing processes in a BTYD model, independent of the model parameters.

As the derivation of the Pareto-NBD model as-is is already incredibly complex, we will not attempt to derive a new formulation of the Pareto/NBD that allows for independent variables to be incorporated into the individual and time-varying level, although we recommend future work in this direction.

Instead, we will consider the BTYD framework in general and develop a possible methodology to incorporate these variables, independent of any model we can consider. Nevertheless, at the end of this section we will utilize our approach to hone in on a possible model specification similar to one described above that we believe will serve as a good alternative when compared to a full Pareto-NBD regression model.

In the BTYD framework, independent variables can fall into three main categories:

- Variables which differ across individuals but are constant across time.
- Variables which are constant for all individuals but differ across time.
- Variables which differ across both individuals and time.

Note that any variables in the last category can be converted into variables that fall into either of the first two categories by either averaging across individuals or averaging across time.

Since all models in the BTYD framework assume that the process that governs purchase frequency is static over time, it only makes sense to include variables from the first category in the purchase frequency section of a BTYD model. We will also argue that for the process that governs dropout behavior, it makes more sense to use only variables which vary only across time.

- The actual point of a customer's dropout is unobserved, meaning we will not be able to accurately visualize how an individual-level hazard function affects customer dropout directly.
- Estimation itself may be statistically difficult, since the model has to consider every possible point of dropping out and fit an individualized hazard function to accommodate all those possibilities.
- Lastly, it is more computationally efficient to only work with variables with either individual or time dependency. Having variables which depend on both will require computation time $O(nt)$ for calculating the overall likelihood, where n is the number of individuals and t the number of units of time each variable is measured at. By contrast,

an estimation procedure which does not have such variables will require computation time $O(n+t)$, since the data can be pooled and the likelihood for both components of the model computed separately.

With those issues in mind, it makes sense to strongly consider using only individual-varying variables for estimating purchase frequency, and only time-varying variables for estimating dropout. Thus it is recommended to simply convert the variables in the third category above to either individual-varying only or time-varying only.

Nevertheless, if the dynamics of the individual and time-varying variables are significant enough that information is lost when they are flattened into one dimension only, an alternative approach that may work is to segment the customer base into discrete segments. This can easily be done using a hard selection algorithm such as K-Means (Lloyd, 1982). However, an alternative approach that may work better is to estimate a mixture of time series models on the time-varying covariates, then for each individual hard-select the most likely time series model that fits the data as the segment that that individual belongs to.

From there, there are several ways that an estimation procedure for the BTYD model can proceed. One extreme is to simply estimate a single BTYD model for the whole dataset, with the variables differing by segment. This approach is the simplest but may mask the differences in behaviors between segments. The other extreme is to estimate a BTYD model for each individual segment. This helps separate and delineate differences between segments but this requires many more parameters and is computationally more expensive.

If the segments are such that they explain most of the heterogeneity across the data, an alternative approach will be to instead use a soft segmentation technique such as Gaussian Mixture Models, then replace the use of continuous priors as the mixing distribution in the BTYD framework with finite mixture models. For instance, in the case of the Pareto/NBD model, instead of specifying that λ, μ are distributed according to independent gamma distributions, we instead say that each segment has a specified value of λ, μ and that each individual has a certain probability of belonging in each segment based on their individual variables. Although the estimation procedure for such models may be difficult due to the added parameters and increased

potential for degenerate solutions, such a specification does reduce the complexity of the models themselves and is a nice compromise between utilizing rich probability models versus utilizing rich datasets.

The specification posed above actually poses a strong middle-of-the-road approach to all the ideas we have been discussing so far. First, we relax the assumption that latent parameters are distributed from a continuous parametric distribution and are instead one of a fixed number of values corresponding to segments in the customer base. Second, we find a way to incorporate both individual-level and time-varying variables into a single model specification but restrict them to affecting separate model sections. Lastly, we do not allow for variables to differ in both individuals and time on a per-segment level, but we instead segment the individuals according to differences in these variables. A formal formulation of such a model will be able to introduce the possibility of using large multi-dimensional data sets in a BTYD analysis, and we strongly recommend further work in this direction.

4 Variable Selection Procedures

In this section, we explore algorithmic techniques for selecting which independent variables to include in the model. Clearly, searching all possible subsets for the best combination of variables is infeasible and will result in performing whichever parameter estimation algorithm is used for the model an exponential number of times. Variable Selection Procedures aim to reduce the time needed to find the best set of variables for a model by restricting the subset search to sequentially increasing set sizes through an ordered selection of variables. Variables are included into the model if a certain criterion that scores their fit is above a certain threshold.

We consider three different ways to simplify the search to a smaller set of combinations, each of which presents a different trade-off between computational complexity and goodness-of-fit for the model.

4.1 Stepwise Regression

Stepwise Regression, originally proposed by Efroymson (1960), is the least greedy of the variable selection procedures we will consider. It comes in

two main types, *forward selection*, where the algorithm starts with no independent variables first then fits succeeding ones one-by-one, and *backward elimination*, which starts with a model specified with all independent variables then removes them one-by-one.

The forward selection procedure is as follows:

1. Fit a model with no regressors and evaluate it using a predetermined criterion.

2. For each remaining regressor left out of the model x_i, estimate a new model including x_i. and evaluate it using the same criterion.

3. Of the new models above, select the one which provides the best improvement to the criterion, then repeat the procedure using this model as a baseline, until the model can no longer improve over the threshold with each regressor.

The backward elimination procedure is similar:

1. Fit a model with all regressors and evaluate it using a predetermined criterion.

2. For each remaining regressor in the model x_i, estimate a new model excluding x_i. and evaluate it using the same criterion.

3. Of the new models above, select the one which provides the best improvement to the criterion, then repeat the procedure using this model as a baseline, until the model can no longer improve over the threshold with each regressor.

Both procedures are significantly better than all-subsets regression, as only $O(n^2)$ subsets of the variables are examined instead of $O(2^n)$. However, since estimation complexity usually scales with the number of variables, forward selection should be computationally faster than backward elimination by a significant amount. In particular, for the the complex regression models we have been discussing so far we should see forward selection outperforming backward elimination, as the optimization procedures required by such models are non-convex in nature and scale poorly.

4.2 Stagewise Regression

In order to reduce the time needed to run the estimation procedures in stepwise regression, an alternative will be to fit the model at the start, hold the parameters fixed, and fit each new variable sequentially without changing the parameters estimated beforehand. This forms the bulk of stagewise regression, as follows:

1. Fit a model with no regressors and evaluate it using a predetermined criterion.

2. For each remaining regressor left out of the model x_i, estimate the value of the model coefficient β_i for x_i on the residual $y - \hat{y}$, holding all other parameters fixed, and evaluate it using the same criterion.

3. Of the new models above, select the one which provides the best improvement to the criterion, then repeat the procedure using this model as a baseline, until the model can no longer improve over the threshold with each regressor.

Stagewise regression (Efron, Hastie, Johnstone, Tibshirani, et al., 2004) provides the same benefits as stepwise regression, but now since only the new parameters are fit at each step the time taken to estimate each new model is reduced significantly. A major drawback to this formulation however is the fit of the model, as interdependency between parameters can cause parameter estimates to vary significantly with the addition of a new variable in the model. This problem can be a major factor with some of the regression models above, which include an extra parameter in the base model for hyperdispersion which may no longer be necessary with the addition of observed data.

Note that the specification for stagewise regression involves regressing on the residuals, which does not generalize to the regression models above. We can however easily fix this by specifying the estimation procedure to perform maximum likelihood estimation (or any other specified estimation technique) by varying only the new variable coefficient and holding all else fixed, which is fundamentally equivalent.

4.3 Streamwise Regression

Streamwise regression (Zhou, Foster, Stine, and Ungar, 2006) is by far the greediest of the selection techniques we are considering, as it only considers each variable exactly once. The procedure is as follows:

1. Fit a model with no regressors and evaluate it using a predetermined criterion.

2. For each regressor x_i, estimate a new model including x_i, and evaluate it using the same criterion.

3. If the model including x_i is better than the current model over the threshold, set it as the current model. Otherwise, keep the current model and repeat with the next variable.

The asymptotic performance of this procedure is the best out of all the three, with $O(n)$ complexity relative to the number of variables considered. This procedure also re-estimates all the parameters at each step, so it is less likely to commit itself to suboptimal parameter values early on, at the expense of committing to variables quickly. Lastly, the most important feature of this method is that it allows for variables to be streamed in succession, meaning models can be updated continuously as new data arrives in sequential order.

5 Variable Selection in BTYD

In this section we develop possible methods to apply the above selection techniques to the BTYD framework. First we will focus on the possibility of feature generation in the BTYD framework. We then turn our attention to dealing with the issue of having two or more different model components to select variables for.

5.1 Feature Generation

Recall that the BTYD framework has three classes of variables:

- Variables which differ across individuals but are constant across time.

- Variables which are constant for all individuals but differ across time.
- Variables which differ across both individuals and time.

As stated before, variables in the third category pose several problems for BTYD-style models, and thus we can convert them to instances of the first two through averaging and segmentation.

However, an interesting question that arises is which of the first two categories should the variables be converted to. While a simple answer will be to simply convert the variables into both categories, this will effectively increase the number of variables considered in the overall model, which can affect the required estimation time further down the line. To avoid increasing the number of variables being considered, we can simply take the averages across each dimension and assign it to the category where there is a higher variance in the remaining dimension. That way, we maintain the number of variables we are considering for the model, but also make sure the variables are assigned to the class where they will presumably have a more noticeable effect on the data.

5.2 Variable Selection Across Model Components

First we will consider some possibilities for applying variable selection procedures to the standard case where a BTYD model only has a count component and a timing component:

- Since we already categorized our variables as specified above, and the components of the BTYD model each only take one kind of variable, we don't need to specify a way of sorting each variable into their model components. Thus, the sequential selection procedures as described above can work well even in this setting, since we have already made the decision of separating the variables earlier on in the process.
- If we do want to optimize our selection procedures further, one possibility that can work for stepwise or streamwise regression is to focus on adding variables to the model component that will benefit from more observed variables. For instance, in the Pareto/NBD formulation, both the latent count and timing parameters are governed by independent

gamma distributions with shape parameters r, s. Higher values of these parameters represent more concentrated prior distributions. Therefore, it makes sense that adding a variable to a component will most likely increase that component's shape parameter, since we are adding observations that explain the underlying heterogeneity in the data. What we can do in each iteration of the variable selection algorithm is to select which model component has a lower shape parameter then add a variable to that component, effectively prioritizing adding observations to the component which we are less certain about.

We then consider the case where we can now estimate separate regression components across multiple segments:

- In the case of the finite mixture BTYD variation, one simplification that can be done is to restrict the regression coefficients to be the same across segments. This has a nice interpretation in the sense that the mixture of latent variables across segments manifests as another regression term that varies across individuals, and thus the whole framework can be seen as a unified regression.

- A more extreme restriction will be to allow each variable to appear in only one regression each: At each step of the variable selection algorithm we have to choose between which of the segments the variable belongs to. While this approach may result in models that are extremely far from the optimal regression models that can be achieved, it may provide insight into determining which variables are the most significant for which segments, and may be an appropriate choice if the number of variables being considered is incredibly large.

6 Economic and Business Analysis

6.1 Direct Marketing and Customer Centricity

Statistical modeling to analyze and predict customer behavior has been used in some shape or form since Ehrenberg's NBD model in 1959. In particular, such models go hand and hand with direct marketing, a form of advertising

that engages specific, targeted consumers directly through personal media such as phones, emails, and website cookies. Being able to isolate and predict a specific customer's behavior from previous purchases allows companies to act upon their future behavior via advertising.

Over time, the ability to measure customer lifetime value or CLV has improved drastically through having both bigger data sets to analyze and more sophisticated models to forecast with. This had made it possible for businesses to adopt a "customer-centric" approach, as opposed to a product-centric one. In a customer-centric business model, customers serve as the beginning and the end of the business' value, with the business determining which customers will provide them the best value today and in the future and tailoring their entire business model around these customers. A key aspect of any customer-centric strategy is to be able to accurately forecast customer behavior in the future, and thus the use of data collection and statistical modeling is crucial for a customer-centric business to succeed.

Although BTYD-style approaches have been used since Schmittlein et al. in 1987, such approaches have yet to be commonly used in industry. Currently, most techniques to estimate CLV in industry rely on linear models or machine learning approaches, which are suitable for big data but do not provide similarly rich customer insights. These circumstances point to the opportunity for any customer-facing company to adopt BTYD models to gain a quick and immediate analytical advantage over their competitors, as virtually no company has yet to do this, at least publicly.

By extending the BTYD model class to work for larger data sets, businesses can now utilize more information to gain incredibly specific insights about each customer more effectively than ever before. On one hand, these developments will help make the BTYD model more usable in high-tech, big data companies with possibly millions of customers. Instead of seeing customers grouped into segments based on their CLV, each customer gets an exact CLV estimate. This can be used to determine how much to spend on acquisition and retention for each customer and enact specific direct marketing strategies based on their exact value. At the same time, companies which are already using BTYD models in their internal data analyses will now be able to use additional data to get more exact estimates of the value of each of their customers, whether internal or external.

6.2 Use Cases

Companies which can benefit from a big-data BTYD model come in two groups. The first consists of customer-centric companies that have internal data science capabilities, and the second consists of third-party data aggregators or consultancies which provide analytics services for companies without internal data science activities.

6.2.1 Big Data BTYD for Internal Use

In this case, the company is offering a product to customers who they want to analyze and track over time. While they may have limited or even nonexistent data analytics capabilities, what they do have is a steady stream of private data that they can collect about their customers through their interactions with the company whether at point-of-sale or beyond.

Although the company will definitely want to leverage their pool of data using various analytical methods, the BTYD framework best serves as the fundamental building block for these analyses since it directly answers the questions most companies directly care about. It directly answers the questions of what customers will do in the future, and how much they are worth. Thus, it makes a lot of sense for companies to treat such models as the main building block for their analytics capabilities, as it helps center the data collection and analysis into answering the core questions above.

6.2.2 Big Data BTYD as a Service

A company can offer BTYD modeling as a service in two ways. First, the company can center their services exclusively on BTYD models. This was the approach taken by Zodiac, a company founded by the principal researchers of the BTYD models, prior to their acquisition by Nike. Since such models are relatively unknown in industry, being one of only a handful of companies capable of providing analytics and consulting to other companies will provide a great deal of benefit and advantage, especially if a portion of the core team for such a company also do research in new models.

On the other hand, a company can use BTYD models as one offering that is part of a larger customer analytics framework. Here, BTYD mod-

eling specifically makes up a smaller part of a company's business, being one of many services that they can offer. Nevertheless, since many other models usually used in analytics provides only a tenuous connection between data and decision-making, BTYD-style models will help provide a solution that is more straightforward, interpretable, and accurate than other models companies have in their 3rd party analytics toolkit.

Regardless if a company chooses to offer general or specific analytics services, such companies will find that BTYD modeling can be easily packaged into standard business models. One approach which may be effective for companies with high technical capabilities will be to package the models in a Software-as-a-Service business model. In this model, customers pay a licensing fee to be able to utilize special analytics software that either replaces or complements their existing customer relationship management software. This business model has been successfully used for similar business-to-business analytics services before, and can serve as a suitable vehicle for adopting BTYD models. For companies which rely more on soft capabilities, a standard consulting model can be used in tandem with other services. This business model is best suited if BTYD modeling is only one out of several capabilities a company can offer, since said company can leverage their human expertise in determining the best course of action when consulting for a given client.

7 Conclusions and Recommendations

In this paper, we explored the development of the BTYD class of models and developed a general framework to include regression coefficients into such models. We also looked at a set of variable selection techniques and explored possible ways to adopt them into the BTYD framework. Together, both topics serve as a prospectus for ways to adopt the BTYD framework, traditionally estimated on small data sets, to better handle cases where there are many possible variables that may be included in the regression model. Other learning techniques that can be generalized to this regression framework, such as pre-model feature engineering and the use of alternative specifications for latent variables, should definitely be looked at in future work.

While there have already been models that deal with incorporating inde-

pendent variables in BTYD models one way or another, the framework presented in Section 3 should be suitable enough to eventually derive a model specification that can serve as a standard for the BTYD framework with variables. Of note is the consideration that the continuous mixture assumption can be relaxed in order to simplify the model to an extent where additional variables are feasible. While this is so far an approach that isn't taken by other BTYD models, we feel that such a relaxation can allow for observed data to play a larger role in the modeling process.

After a proper model specification, we think that an important direction for future work is to be able to empirically test these hypotheses on large data sets. Although large data sets that describe customer behavior on an incredibly granular level are undoubtedly used for machine learning algorithms at large companies, such data sets are generally unavailable for academic use. Furthermore, most of the canonical data sets being analyzed in the BTYD community do not have enough size and dimensionality to warrant a more "big-data" approach to customer behavior analysis. All-in-all, empirical performance of BTYD models (statistical as well as computational) should be studied under a wide range of possible inputs, varying in both population and variable size.

References

Abe, M. (2009). "counting your customers" one by one: A hierarchical bayes extension to the pareto/nbd model. *Marketing Science 28*(3), 541–553.

Babkin, A. and I. Goldberg (2017). Incorporating time-dependent covariates into bg-nbd model for churn prediction in non-contractual settings.

Cameron, A. C. and P. K. Trivedi (2013). *Regression analysis of count data*, Volume 53. Cambridge university press.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological) 34*(2), 187–220.

Efron, B., T. Hastie, I. Johnstone, R. Tibshirani, et al. (2004). Least angle regression. *The Annals of statistics 32*(2), 407–499.

Efroymson, M. (1960). Multiple regression analysis. *Mathematical methods for digital computers*, 191–203.

Ehrenberg, A. S. (1959). The pattern of consumer purchases. *Applied Statistics*, 26–41.

Fader, P. S. and B. G. Hardie (2007). Incorporating time-invariant covariates into the pareto/nbd and bg/nbd models.

Fader, P. S., B. G. Hardie, and C.-Y. Huang (2005). A note on deriving the pareto/nbd model and related expressions. *accessed July 31*, 2005.

Fader, P. S., B. G. Hardie, and K. L. Lee (2005). "counting your customers" the easy way: An alternative to the pareto/nbd model. *Marketing science 24*(2), 275–284.

Lloyd, S. (1982, March). Least squares quantization in pcm. *IEEE Transactions on Information Theory 28*(2), 129–137.

Lomax, K. S. (1954). Business failures: Another example of the analysis of failure data. *Journal of the American Statistical Association 49*(268), 847–852.

Nelder, J. A. and R. J. Baker (1972). *Generalized linear models*. Wiley Online Library.

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

Schweidel, D. A. and G. Knox (2013). Incorporating direct marketing activity into latent attrition models. *Marketing Science 32*(3), 471–487.

Wedderburn, R. W. (1974). Quasi-likelihood functions, generalized linear models, and the gauss—newton method. *Biometrika 61*(3), 439–447.

Zhou, J., D. P. Foster, R. A. Stine, and L. H. Ungar (2006). Streamwise feature selection. *Journal of Machine Learning Research 7*(Sep), 1861–1885.