ISSN: 2047 - 0398
SPECIAL ISSUE – V|4|N|3|, December | 2014

INCORPORATING HETEROGENEITY INTO COUNT DATA MODELS IN MARKETING

Pedro Jesus Hernandez
Market Analytics.
Av. Brig. Faria Lima, 2954 cj. 102
Itaim- Bibi
04536-020 - São Paulo, SP - Brazil
E-mail: *pedro.fernandez@marketanalytics.com.br*

Delane Botelho
FGV/EAESP
Rua Itapeva, 474, 9o. andar
Bela Vista
01332-000 - São Paulo, SP, Brazil
E-mail: *delane.botelho@fgv.br*

Felipe Zambaldi (*Corresponding Author***)**
FGV/EAESP
Rua Itapeva, 474, 9o. andar
Bela Vista
01332-000 - São Paulo, SP, Brazil
E-mail: *Felipe.zambaldi@fgv.br*

ABSTRACT

This paper deals with two kinds of problems frequently observed in count data collected by scholars in the Consumer Behavior arena: excess of zeros and low homogeneity among sample observations. We estimated seven models using Bayesian methods, but incorporated explanatory variables only in three. The data are from two product categories: cigarette and cocoa. The main conclusions are: the inclusion of unobserved heterogeneity improves the model fit, depending on the data; the inclusion of explanatory variables does not show such improvement, so the consideration of the unobserved heterogeneity may be more important than that inclusion; traditional models applied to count data can be less accurate if they do not take into account the excess of zeros. We discuss the results considering the implications of the models for marketing research.

Keywords: *Heterogeneity, Count Data, Excess of Zeros, Bayesian Methods, Consumer Behavior.*

1. INTRODUCTION

In various empiric studies, in social sciences and elsewhere, the analysis focuses on explaining an essentially limited dependent variable. This limitation can take different forms as a variable with continuous limited distribution to the left or right. Other forms of limitations include variables with a finite number of values, such as the choice of brands, and variables that take non-negative whole values (Maddala, 1983). The latter ones contain data usually referred to as “count data” and occur frequently in marketing contexts, as these describe the number of times one observes a phenomenon or event. Examples include the number of purchases in a specific product category or the number of clients who visit a specific store or a specific internet site during a set period.

In marketing, these types of data represent two basic characteristics, and if treated with traditional models (such as the Poisson Regression model), such characteristics may present real problems for models. These characteristics are: 1) a much higher frequency of zeros than one would expect in any Poisson distribution; 2) low homogeneity among buyers/consumers, a notorious phenomenon encountered by marketing researchers (Allenby, 2012).

The purpose of this article is to discuss these two problems, incorporating, according to Otter et. al. (2004), some less conventional considerations on the count data model in marketing: 1) the recognition that traditional models that don’t account for the presence of zeros adjust more poorly to the data; 2) the fact that, depending on the data, the inclusion of unobserved heterogeneity can significantly improve model fits (which reinforces the idea that, often, explanatory valuables have little real significance); 3) and the fact that often the heterogeneity of the consumers may be enough (or even required) to explain the behavior of the dependent (count) variable.

This introduction presented the justification, the problem and the goals of this study. The next section presents an overview of the literature, followed by the methodology, in which we present the data collection and models under analysis. The fourth section presents and discusses the results. Finally, the article provides the main conclusions of the study, including its implications for marketing research and suggestions for future studies that involve count data.

2. LITERATURE

Count data appear when, during the observation period, a specific event occurs (the data consist of the number of times it occurs) (Morgan, Lenzenweger, Rubin & Levy, 2014). These data include non-negative whole numbers and zero is a possibility. For example, in data collected by optical scanning (Botelho & Urdan, 2003), it is rare to have more than twenty purchases per household in most product categories. The most frequent observation is zero, indicating that no purchase occurred in that category, and the second most frequent observation is one, indicating the purchase of one unit. Franses and Paap (2001) and Maddala (1983).provide general references about count data modeling.

This type of dependent variable is usually represented by distributions such as Poisson and incorporates explanatory variables that help to understand the behavior of frequency of consumption/purchase, establishing the Poisson Regression models. The Poisson distribution describes rare events, during which a large number of attempts happen and applies to situations in which the event of interest has an homogeneous distribution over the population. If y_i is the occurrence of a random event in a specific interval of time, the probability of the occurrence of y_i is:

$$P\left\langle y_i = y \,\middle|\, x_i \right\rangle = \frac{e^{-\lambda_i}\lambda_i^{y}}{y!} \quad ,$$

y_i = 1, 2, 3, ... , in which λ_i represents the distribution parameter (the average occurrence of y_i) and *e* is the Euler number.

The problem with an excess of zeros occurs naturally in count data: many people do not buy anything from a specific category during the period analyzed by the marketing research or do not visit a specific website during this period [see Mullahy (1997a) for a more in-depth discussion on this phenomenon]. This can be solved, for example, by modifying the Poisson distribution and introducing a Zero Inflated Poisson (ZIP) distribution (for example, to represent the purchases that did not take place during the period under analysis). Explanatory variables can still be included.

Another problem is the low homogeneity among buyers/consumers. Economists try to solve this with econometric models that refer to the "average consumer", which is not very interesting for marketing researchers, as the main interest in marketing is to understand individual behavior of buyers/consumers at a disaggregated level (Rossi & Allenby, 2003). In the econometric literature, the variability among individuals is routinely obtained by including explanatory variables that depend on the buyer/consumer, such as demographic variables. Overall, such heterogeneity may present an unobserved component (Dubé, Hitsch & Rossi, 2010), and the inclusion of explanatory variables becomes insufficient, or even unnecessary, when more solid models can be obtained by considering individual differences (Allenby, 2012; Chandukala, Long-Tolbert & Allenby, 2011). This means that the observed reality is composed of individual components that, when combined, are capable of explaining more complex phenomena. In other words, explanatory variables, considered very valuable in models of aggregate behavior, can lose their importance when heterogeneity among consumers is incorporated (Allenby, 2012).

Incorporating explanatory variables in a model equals accounting for the (observed) heterogeneity of (individual) observations. Intuitively, the presence of heterogeneity means that individuals of the population present differences like, for example, different tastes. Formally, in linear regression model ($y_i = \beta_i' x_i + \varepsilon_i$), heterogeneity can be represented by ξ_i in the equation: $\beta_i = \beta + \xi_i$, by intercept α_i in the equation $y_i = \alpha_i + \beta' x_i + \varepsilon_i$, or absorbed by the error term component ε_i (Wooldridge, 2002).

Unobserved heterogeneity can be accounted for by using *Bayesian* methods. In the 1990s there was a significant increase in the use of these methods in international marketing literature, covering a wide range of problems and data sources (Rossi & Allenby, 2003). This increase assumed that buyers/consumers have different preferences and that companies should consider them when trying to optimize marketing efforts.

All *Bayesian* analyses begin with the specification of the mechanism for data generation or the distribution of the dependent variable, according to unobserved parameters θ, $p\left\langle y \middle| \theta \right\rangle$. Seen as a function of the parameters, this distribution is frequently referred to as the function of verisimilitude $l\theta = p\left\langle y \middle| \theta \right\rangle$. The *Bayesian* principle depends on the fact that the verisimilitude function contains relevant information about the parameters to be estimated. Furthermore, the prior specification of the distribution of the probability of θ, $p(\theta)$ is required. The Bayes theorem (Ross, 2002) provides a mechanism that "translates" the prior assumption of the distribution of probability into a posterior assumption (after obtaining the data).

Bayesian inference, based on the posterior assumption of the distribution of unobserved parameters, is exempt from the use of asymptotic approximations, which is important in models with unrelated data, such as count data (Aregay & Molenberghs, 2015). *Bayesian* methods also consider the estimated parameters as random variables and require less data, because the probabilistic concepts involved reduce the model fit dependency in relation to the amount of data used (Rossi & Allenby, 2003; Lopes & Tsay, 2011). Another advantage is that these allow a relatively more simple estimation than traditional methods, and the comparison between different models is simpler.

However, we do pay the price for the advantages of the *Bayesian* approach: the prior need for a distribution of probability has generated criticism for depending on subjective information. However, according to Neelamegham and Chintagunta (1999), the basis for prior information can be objective, based on previous data. Various forms of prior information have been used in Bayesian methods in the recent marketing literature (see Lopes & Tsay, 2011; Allenby 2012), including specialist information (Sandor & Wedel, 2001), theories (Montgomery & Rossi, 1999) and databases comparable to those used in the researches of interest (Ter Hofstede et al., 2002).

3. METHODOS

Every model in this study has been estimated by means of *Bayesian* methods. Their estimation has become possible in recent few years thanks to faster computer processing and the implementation of simulation methods, such as the Monte Carlo methods based on Markov chains (MCMC - *Markov Chain Monte Carlo*) [see Robert and Casella (1999) for a better understanding of these methods]. The computational problem of these methods relates to the calculation of various integrals of the posterior distribution functions. Considering that these integrals can a posterior expectation of a parameter function, simulation methods are the most likely candidate for an approximation.

The models analyzed here have been estimated with the use of the statistical analysis package BUGS (*Bayesian Inference Using Gibbs Sampling*), using MCMC methods. These methods have been applied in social sciences, accounting, economics and marketing (MRC Biostatistics Unit, 2005).

3.1. Data collection

This article analyzed two databases, one that the authors obtained from the website of Wooldridge's book (2000) (on cigarette smoking) and the other was collected directly by the authors in a Brazilian supermarket chain (on purchase of powdered chocolate products). These databases are respectively referred to as 1 and 2.

Database 1 (cigarette consumption) has already been analyzed by many authors, including Wooldridge (2000) and Mullahy (1997), the latter being originally responsible for the data collection. The data are directly available from www.wooldridge.swcollege.com and consist of 807 observations from the *National Health Interview Survey* in the USA. It is a survey with smoking and non-smoking men, in done in 1979 and at the beginning of 1980. The dependent variable is the number of cigarettes smoked in a day. The independent variables are: age (in years); education (numbers of years of schooling); household size (number of members); race (dummy variable: white = 1; 0, not white); annual family income (in thousands of dollars); price of a cigarettes pack (average price during the data collection); and smoking restrictions in restaurants (dummy variable: interviewee lives in a state where smoking in restaurants is restricted = 1; 0, if not). The analysis of these data differs from those previously published and is more comprehensive than what Wooldridge described (2000).

Database 2 (purchase of powdered chocolate products) consists of 3.402 household observations. The data were collected in 2001 in collaboration with the IT department of the largest supermarket chain in the state of Rio de Janeiro, Brazil, which owns 69 stores and in 2001 held 3.6 % of the national market share (Brazilian Association of Supermarkets, 2002). The data refer to purchases made by households where a member possessed a customer rewards card of the analyzed chain. The demographic variables were obtained from the card information over a 35-week period, from July 2000 to March 2001. Purchase variables for the product category and price were obtained from the scanned products and customer rewards cards at the checkout register. Four brands of powdered chocolate products were chosen for analysis; these represent 82% of the retail market sales in that product category. The most demanded package size in the category was chosen, the equivalent of a 500 gram package. The variables are described in the two columns on the right in TABLE 1. Dummy variables were used for the variable "marital status" (SINGLE, MARRIED and DIVORCED), with the category “common-law relationship serving as reference (or is equal to zero). TABLE 1 gives a description of the variables of databases 1 and 2.

[Insert TABLE 1 here]

3.2. Count Data Models

The dependent variables have non-negative whole values {0, 1, ...} and there is an excessive number of zeros (beyond what would be considered reasonable for a Poisson distribution). In this case, the Poisson model is modified to fit the situation by incorporating the ZIP model. The model can be conceived in the following manner: a proportion π of the population will be awarded value $y = 0$; the positive numbers of events follows a Poisson distribution with parameter λ:

$$P(Y = y) = I_{(y=0)}(y)\pi + (1-\pi)\left(\frac{\lambda^{y}e^{-\lambda}}{y!}\right),$$

with I being the indicator of the combination $y = 0$ (if the observation for y is 0, $I = 1$, if not, $I = 0$) .

Note that there are no explanatory variables. A way to incorporate these into the model would be to use the Poisson Regression model. In that case Y_i would be the variable indicating the number of events (the number of cigarettes smoked) incurred by individual i during a specific period. At an individual level, the assumption is that Y_i follows a Poisson distribution with an average of λ_i*:*

$$P\left\langle y_i = y \,\middle|\, \lambda_i \right\rangle = \frac{e^{-\lambda_i}\lambda_i^{y}}{y!}$$

The individual average relates to the individual characteristics observed in function $\lambda_i = \lambda_0 e^{x_i'\beta}$. This way the model includes a form of heterogeneity observed in explanatory variables (such as demographic variables, like age, or marketing variables, like price) for the individual i (included in x_i), and an intercept λ_o. An equivalent way of writing this would be $\lambda_i = e^{x_i'\beta}$ and include as first element of x_i the number 1, so the intercept remains $\lambda_0 = e^{\beta_0}$.

If there is an excessive number of zeros, this model can be generalized to take this into account. The result is a Zero Inflated Poisson Regression model. As earlier stated, proportion π corresponds to the absence of events during a specific period and the remaining individuals are characterized by a Poisson Regression model:

$$P(Y_i = y) = I_{(y=0)}(y_i)\pi + (1-\pi)\left(\frac{\lambda^{y_i}e^{-\lambda_i}}{y_i!}\right)$$

With $\lambda_i = \lambda_0 e^{x_i'\beta}$ or $\lambda_i = e^{x_i'\beta}$, in which the first value of x_1' is 1 and $\lambda_0 = e^{\beta_0}$.

So far, heterogeneity has been accounted for by incorporating explanatory variables. However, as we previously described, it is possible that such heterogeneity may not be completely captured by the explanatory variables in this model, so there is a possibility of unobserved heterogeneity. This heterogeneity is represented by a probability of the parameter λ of the Poisson distribution. According to Wooldridge (2002), it is natural to use more flexible distributions in terms of the form, concentrated on real positive numbers, such as Gamma (Γ) (see Aguero-Valverde, 2013).

The resulting model is a Negative Binomial Distribution (NBD), based on the following assumptions:

- the individual level of behavior that is of interest to the study is described as a count variable (Poisson);
- the distribution of parameter λ of the population is characterized by a Gamma distribution, with a $g(r,\alpha)(\lambda)$density function, with parameters r and α.

The aggregated distribution of interest is obtained by analyzing each individual Poisson distribution $P\left\langle X = x \middle| \lambda \right\rangle$ according to density *g(r,α)(λ)*. The final distribution will be a combination of Gamma and Poisson distributions:

$$P\left(X = x\right) = \int_0^{+\infty} \frac{\lambda^x e^{-\lambda}}{x!} \frac{\alpha^r \lambda^{r-1} e^{-\alpha\lambda}}{\Gamma(r)} d\lambda$$, in which Γ represents the Gamma function.

This expression is equal to:

$$P\left(X = x\right) = \frac{\Gamma(r+x)}{\Gamma(r)x!}\left(\frac{\alpha}{\alpha+1}\right)^r \left(\frac{1}{\alpha+1}\right)^x$$

To maintain explanatory variables in the model and, simultaneously, the unobserved heterogeneity, these possibilities can combine in a Negative Binomial Distribution Regression Model.

The rationale is that the explanatory variables may not completely capture the components of the differences among individuals. To capture these (unobserved) remaining components, λ_0 can represent the combined effect on the average count of unobserved explanatory variables. Therefore, λ_0 will vary among the population according to a Gamma distribution, with parameters *r* and *α*. For Cameron and Trivedi (1986 and 1990), this distribution is known as Negbin II (*Negative Binomial II*) and characterizes as a negative binomial distribution with the presence of explanatory variables. When we consider the specific case of $\beta = 0$, we find a NBD distribution:

$$P\left(Y_i = y\right) = \frac{\Gamma(r+y)}{\Gamma(r)y!}\left(\frac{\alpha}{\alpha + e^{(x_i'\beta)}}\right)^r \left(\frac{e^{(x_i'\beta)}}{\alpha + e^{(x_i'\beta)}}\right)^y$$

Finally, the excessive number of zeros is also taken into account with the Zero-Inflated Negative Binomial Distribution (ZINBD) model. Once again, Π corresponds to the proportion of the population in which the number of events is zero. The remaining observations characterize by a NBD distribution:

$$P\left(X = x\right) = \pi I_{(x=0)}(x) + (1-\pi)\left(\frac{\Gamma(r+x)}{\Gamma(r)x!}\right)\left(\frac{\alpha}{\alpha+1}\right)^r \left(\frac{1}{\alpha+1}\right)^x$$

4. RESULTS

4.1. Cigarette consumption (Database 1)

TABLE 2 presents the values of the parameters estimated by the models for database 1. *β* indicates the parameters of the explanatory variables. The Deviance Information Criterion (DIC) was suggested by Spiegelhalter et al. (1988) and refers to an asymptotic criterion that reflects both the fit of the model as well as the degree of parameterization (complexity). Therefore, it is a statistic to select the best model (the *Bayesian* equivalent of the Akaike Information Criterion –AIC of classic statistics). The model with the lowest DIC value has the highest probability to predict the collection of data and thus is the most suitable model.

[Insert TABLE 2 here]

4.1.1 Analysis of Poisson, ZIP, NBD and ZINBD Distribution Models:

CHART 1 presents the histograms of the observed values of the survey sample and model predictions. The Poisson Distribution model provides the most basic way to describe the data. When analyzing the corresponding histogram, it becomes apparent that it is impossible to adjust the model to the real data, especially when this involves the observed zeros. The graphic analysis corroborates the highest DIC value (17.293), among all models. The Poisson Regression model presents the lowest DIC value (16.923), although still high compared to the others. In Chart 1, there is a better fit in relation to the Poisson Distribution model, but the improvement in the fit of the ZIP model is clearer, as the drop in the DIC value (3.950) reveals. This model indicates that 61.5% of the sample ($\pi = 0.615$) is composed of men that "almost" do not smoke (average daily cigarette consumption of zero) and the remaining 38.5% smokes between 22 and 23 cigarettes per day, an average of ($\lambda_0 = 22.560$).

In principle, the models that recognize the existence of zeros tend to present the best fit (Wooldridge, 2000). However, the NBD distribution model that incorporates unobserved heterogeneity (even without considering the excess of zeros) provides a significantly better fit, as we can tell from the DIC value (2.220). This is because of the flexibility of the model to deal with varying *r* and *α* parameters. If the λ of the Poisson distribution presents a

Gamma distribution, such as in the case of the NBD, the result is r = 0.129 and α= 0.015. Note that the average value of λ becomes 8.600 (r/α), which is equal to the λ value of the Poisson distribution.

[Insert CHART 1 here]

If the NBD distribution model is modified to take into account the excess of zeros, thus resulting in the ZINBD distribution model, it provides the best fit among all models (DIC = 1.052). The observed data were generated by 61.5% of the "almost" non-smoking individuals (π = 0.615) and smokers (38.5%), with an average of 22.56 (r/α) cigarettes per day (a little more than a pack a day). Note that the value of π and of the average (λ) are equal to those of the ZIP distribution model. However, in this last model, all consumers would be considered to have the same average daily consumption, which does not reflect the reality. The ZINBD distribution model is more realistic as it assumes that each consumer is different and has an individual daily consumption average (the distribution of these averages follow a Gamma distribution with parameters r=2.910 and α=0.129). The quality of the fit of the ZINBD distribution model is also available in CHART 1, where it is clear that this is the model best suited to the data.

4.1.2 Analyses of the Poisson Regression, ZIP and NBD Models:
The parameters of the explanatory variables are in TABLE 1 by β. The values below each parameter displayed in bold represent the confidence intervals of 95%. This means that if zero is within the confidence interval, the parameter is not significantly different from zero. In the Poisson regression model, there are four significant β parameters (for the variables EDUC, RESTAU, LINCOME and AGE). This model only provides a marginally better fit (DIC = 16.923) than the corresponding model without explanatory variables (Poisson distribution).

In the ZIP regression model, the same parameters of the Poisson regression model are significant (for the variables for EDUC, RESTAU, LINCOME and AGE). However, the parameters for AGE and EDUC do not present the expected signal (they are positive, indicating that the higher the age and education level, the greater the cigarette consumption). The improvement in fit is the result of the explicit recognition of excessive zeros (DIC = 3.790).

In the NBD regression model r and α are significant, even with very large confidence intervals. The same four variables of the previous models are significant and the significant parameters present the expected signal. This model does not offer a good fit for the data, because of its high DIC value (16.854).

In every model with explanatory variables, the variables "race" and "(log) price" don't have any significant effect on cigarette consumption. One of the reasons why (log) price isn't significant is because the price only varies among the analyzed American states, and this variation is smaller than the other variables.

4.2. Consumption of powdered chocolate products (Database 2)
Table 3 presents the values of the estimated parameters for each model for database 2. The dependent variable (PURCHASES) is the total number of purchases made in the category of powdered chocolate products by each shopper. As a characteristic of this database, originally all consumers who did not buy a product from the specific category in the analyzed period were excluded, generating data that did not contain any zeros for this variable (all elements of the sample made at least one purchase in the specific category during a 35-week period). Therefore, it would not be appropriate here to use any models that take the excess of zeros (ZIP and ZINBD distribution ad ZIP regression) into account.

[Insert TABLE 3 here]

The average of the variable PURCHASES is 8.24, indicating that during the 35-week period shoppers bought an average of 8.24 packages of powdered chocolate products in the chain of supermarkets. That means that the "average" shopper bought approximately a package every 4.25 weeks. Chart 2 displays the histograms of the variable PURCHASES for the values observed in the sample and those predicted by the Poisson and NBD distribution models (the models with respectively the worst and best fit for the data). The chart shows that the data present a dispersion to the right, indicating that a simple Poisson model would not be the best option.

In Table 3, we note that, just like in database 1, the model that best suits for the data is the NBD model (DIC = 17.617). However, for these data, the DIC differences of each model are much smaller than those for the data on cigarette consumption. Graph 2 also illustrates this minor difference among the models' fits.

[Insert GRAPH 2 here]

For the models that incorporate explanatory variables (Poisson regression and NBD regression), the parameters of the variables are represented by β in TABLE 2. In terms of the Poisson regression model, the only significant parameters are the intercept and the coefficients of the DEPEN variables (the number of dependents of each customer rewards cardholder) and AGE (age of the cardholder). The most important variable for explaining PURCHASES is DEPEN, which has a positive signal (indicating that the higher the number of dependents, the higher the number of purchases of powdered chocolate products).

With regard to the NBD regression model, r and α are significant, even though with very large confidence intervals. The average of the distribution of heterogeneity is 7.14 (r/α). This model has the same significant parameters as the Poisson Regression model (intercept, DEPEN and AGE), with the same signals. Therefore, the two models that incorporate explanatory variables have a similar interpretation.

The comments made in reference to database 1 also apply here. The inclusion of the unobserved heterogeneity improves the fit, although not as much as in the case of cigarette consumption. The NBD distribution model (without explanatory variables) presents the best fit (more than any other model that incorporates explanatory variables). Therefore, there is no justification to include these explanatory variables. This does not mean that there are not other explanatory variables that could provide a relevant explanation.

5. FINAL REMARKS

This article has presented a discussion on the relationship among variables that contain count data, frequently collected by marketing researchers, many times containing excesses of zeros. Based on the results, the combination of unobserved heterogeneity and the explicit recognition of the excessive zeros make the ZINBD distribution model the best model (the one that best fits or adapts to the data) for database 1.

Depending on the data, the inclusion of unobserved heterogeneity significantly improves the quality of the model fits. The recognition of heterogeneity or the recognition of the intrinsic differences among buyers/consumers in the two databases may be more important than the inclusion of explanatory variables. In fact, for both databases, the NBD model without the inclusion of such variables (the NBD distribution model) provides a better fit than the model that does include them (the NBD regression model). This shows marketing researchers that the inclusion of explanatory variables is not always a proper solution for explaining the behavior of a dependent variable. However, it is very important to find explanatory variables that really matter.

Explanatory variables are important because generally they involve marketing mix variables and therefore serve as a tool for marketing professionals in influencing a dependent variable that often relates to demand (such as in the case of the two databases analyzed in this article). However, researchers should question the inclusion of any variables that cannot improve the fit of a purely random model, such as the NBD.

The results indicate that the NBD or ZINBD distribution models (in case of excessive zeros) might be used as a benchmark for competing models. Suggestions for future research include their use on other marketing data, such as the number of visits to a website, for example. In the case of purchasing powdered chocolate products, other explanatory variables for purchases in that category should be analyzed to understand how much of the behavior of the variable is explained by other more significant variables (such as the presence of advertising, for instance) or by the buyer's heterogeneity.

It may also be useful to analyze the existence of a number of segments of different buyers/consumers by using other heterogeneity distributions, such as the mix of normal distributions. Also, along the lines of a finite (although unknown) number of segments, non-parametric methods, such as the *Dirichlet Process Prior* (de Valpine & Harmon-Threatt, 2013; Bunn, 1979) should be tried in marketing research.

REFERENCES

Aguero-Valverde, J. (2013). Full Bayes Poisson gamma, Poisson lognormal, and zero inflated random effects models: Comparing the precision of crash frequency estimates. *Accident Analysis & Prevention*, *50*, 289-297.

Allenby, G. M. (2012). Modeling marketplace behavior. *Journal of the Academy of Marketing Science*, *40*(1), 155-166.

Aregay, M., Shkedy, Z., & Molenberghs, G. (2015). Comparison of Additive and Multiplicative Bayesian Models for Longitudinal Count Data With Overdispersion Parameters: A Simulation Study. *Communications in Statistics-Simulation and Computation*, *16*(2), in press.

Botelho, D. (2005). Decomposição da elasticidade-preço no varejo com uso de dados escaneados. *Pesquisa Operacional*, *25*, 201-217.

Bunn, D. W. (1979). The synthesis of predictive models in marketing research. *Journal of Marketing Research, 16*, 280-295.

Cameron, A. C., Trivedi, P. K. (1986). Econometric models based on count data: comparisons and applications of some estimators. *Journal of Applied Econometrics, 1*, 29-53.

Cameron, A. C., Trivedi, P. K. (1990). Regression-based tests for over-dispersion in the Poisson model, *Journal of Econometrics, 46*, 347-364.

Chandukala, S., Long-Tolbert, S., Allenby, G. (2011) A threshold model for respondent heterogeneity. *Marketing Letters, 22(2)*, 133-146.

Congdon, P. (2001). *Bayesian statistical modeling*. New York: John Wiley.

de Valpine, P., & Harmon-Threatt, A. N. (2013). General models for resource use or other compositional count data using the Dirichlet-multinomial distribution. *Ecology, 94*(12), 2678-2687

Dubé, J., Hitsch, G. J., Rossi, P. E. (2010) State dependence and alternative explanations for consumer inertia. *The RAND Journal of Economics, 41(3)*, 417-445.

Franses, P. H., Paap, R. (2001). *Quantitative models in marketing research.* Cambridge: University Press.

Koop, G. (2004). *Bayesian econometrics.* New York: John Wiley.

Lopes, H., Tsay, R.S. (2011) Particle filters and Bayesian inference in financial econometrics. *Journal of Forecasting, 30(1)*, 168-209.

Maddala, G. S. (1983). *Limited dependent and qualitative variables in econometrics.* Cambridge: University Press.

Montgomery, A. L.; Rossi, P. E. (1999). Estimating price elasticities with theory-based priors. *Journal of Marketing Research, 36*, 413-423.

Morgan, C. J., Lenzenweger, M. F., Rubin, D. B., & Levy, D. L. (2014). A hierarchical finite mixture model that accommodates zero-inflated counts, non-independence, and heterogeneity. *Statistics in medicine*, 33(13), 2238-2250. DOI: 10.1002/sim.6091

Mullahy, J. (1997a) Heterogeneity, excess zeros and the structure of count data models. *Journal of Applied Econometrics, 12*, 337-350.

Mullahy, J. (1997b) Instrumental-variable estimation of count data models: applications to models of cigarette smoking behavior. *The Review of Economics and Statistics, 79*, 586-593.

Neelamegham, R., Chintagunta, P. A. (1999). Bayesian model to forecast new product performance in domestic and international markets. *Marketing Science, 18*, 115-136.

Robert, C. P., Casella, G. (1999). *Monte Carlo statistical methods.* New York: Springer.

Ross, S. (2002). *A first course in probability.* 6th Ed. Upper Saddle River: Prentice Hall.

Rossi, P. E., Allenby, G. M. (2003). Bayesian statistics and marketing. *Marketing Science, 22*, 304-328.

Sandor, Z., Wedel, M. (2001). Designing conjoint choice experiments using managers' prior beliefs. *Journal of Marketing Research, 28*, 430-444.

Ter Hofstede, F., Kim, Y., Wedel, M. (2002) Bayesian prediction in hybrid conjoint analysis. *Journal of Marketing Research, 34*, 253-261.

Wooldridge, J. M. (2000). *Introductory econometrics: a modern approach.* South-Western College Publishing.

Wooldridge, J. M. (2002). *Econometric analysis of cross section and panel data.* Cambridge: MIT Press.

Zellner, A. (1971). *An introduction to bayesian inference in econometrics.* New York: John Wiley.

TABLE 1 - Description of the variables of databases 1 and 2

Database 1		Database 2	
Variable	**Description**	**Variable**	**Description**
CIGS	Number of cigarettes smoked per day	PURCHASES	Number of purchases in the category
PRICE	Price	DEPEN	Number of dependents
WHITE	race	AGE	Age
AGE	age	SEX	Sex
INCOME	income	SINGLE	Dummy variable (1= single)
EDUC	education	MARRIED	dummy variable (1=married)
RESTAU	Smoking restriction in restaurants	DIVORC	dummy variable (1=divorced)
LINCOME	Income logarithm	INCOME	Family income in three levels
LPRICE	Price logarithm		

ISSN: 2047 - 0398
SPECIAL ISSUE – V|4|N|3|, December | 2014

TABLE 2 – Estimated parameters for each model (database 1)

Parameters	Poisson Distrib.	Poisson Regression	ZIP Distrib.	ZIP Regression	NBD Distrib.	NBD Regression	ZINBD Distrib.
λ_0	**8.600** 8.487:8.686	**0.420** -0.465:1.924	**22.560** 22.09:23.15	**0.752** -0.429:2.595			
r					**0.129** 0.111:0.146	**8.349** 0.002:66.53	**2.910** 2.413:3.492
α					**0.015** 0.011:0.018	**3.431** 0.00:27.84	**0.129** 0.106:0.156
π			**0.615** 0.582:0.649	**0.615** 0.582:0.648			**0.615** 0.580:0.648
β_EDUC		**-0.045** -0.054:-0.036		**0.034** 0.025:0.043		**-0.045** -0.0534:-0.037	
β_WHITE		**-0.003** -0.076:0.070		**0.014** -0.058:0.087		**-0.008** -0.08:40.06	
β_RESTAU		**-0.376** -0.436:-0.315		**-0.090** -0.15:-0.03		**-0.372** -0.434:-0.311	
β_LINCOME		**0.241** 0.193:0.281		**0.131** 0.095:0.167		**0.237** 0.199:0.274	
β_AGE		**-0.005** -0.007:-0.003		**0.005** 0.004:0.007		**-0.005** -0.007:-0.004	
β_LPRICE		**0.061** -0.234:0.253		**0.115** -0.344:0.406		**-0.061** -0.321:0.382	
DIC	17.293	16.923	3.950	3.790	2.220	16.854	1.052

The values below each parameter in bold, separated by ":", have 95% confidence intervals.
DIC = *Deviance Information Criterion*

TABLE 3 – Estimated parameters for each model (database 2)

Parameters	Poisson Distrib.	Poisson Regression	NBD Distrib.	NBD Regression
λ_0	**8.244** 8.149:8.342			**7.134** 6.065:7.065
r			**14.000** 12.54:15.69	**7.784** 0.002:70.021
α			**1.698** 1.521:1.905	**1.091** 0.00:10.12
β_0 (intercept)		**1.963** 1.891:2.029		
β_DEPEN		**0.021** 0.013:0.028		**0.020** 0.013:0.028
β_AGE		**0.002** 0.001:0.003		**0.002** 0.001:0.003
β_SEX		**-0.024** -0.048:0.001		**-0.025** -0.048:0.001
β_SINGLE		**0.037** -0.005:0.078		**0.036** -0.005:0.078
β_MARRIED		**0.018** -0.013:0.050		**0.017** -0.015:0.049
β_DIVORC		**0.005** -0.044:0.054		**0.005** -0.043:0.052
β_INCOME		**0.028** -0.107:0.157		**0.003** -0.013:0.020
DIC	18.499	18.464	17.617	18.465

The values below each parameter printed in bold separated by ":", have 95% confidence intervals.
DIC = Deviance Information Criterion

FIGURE 1 - Histograms of observed values of the sample and values predicted by the models (database 1).

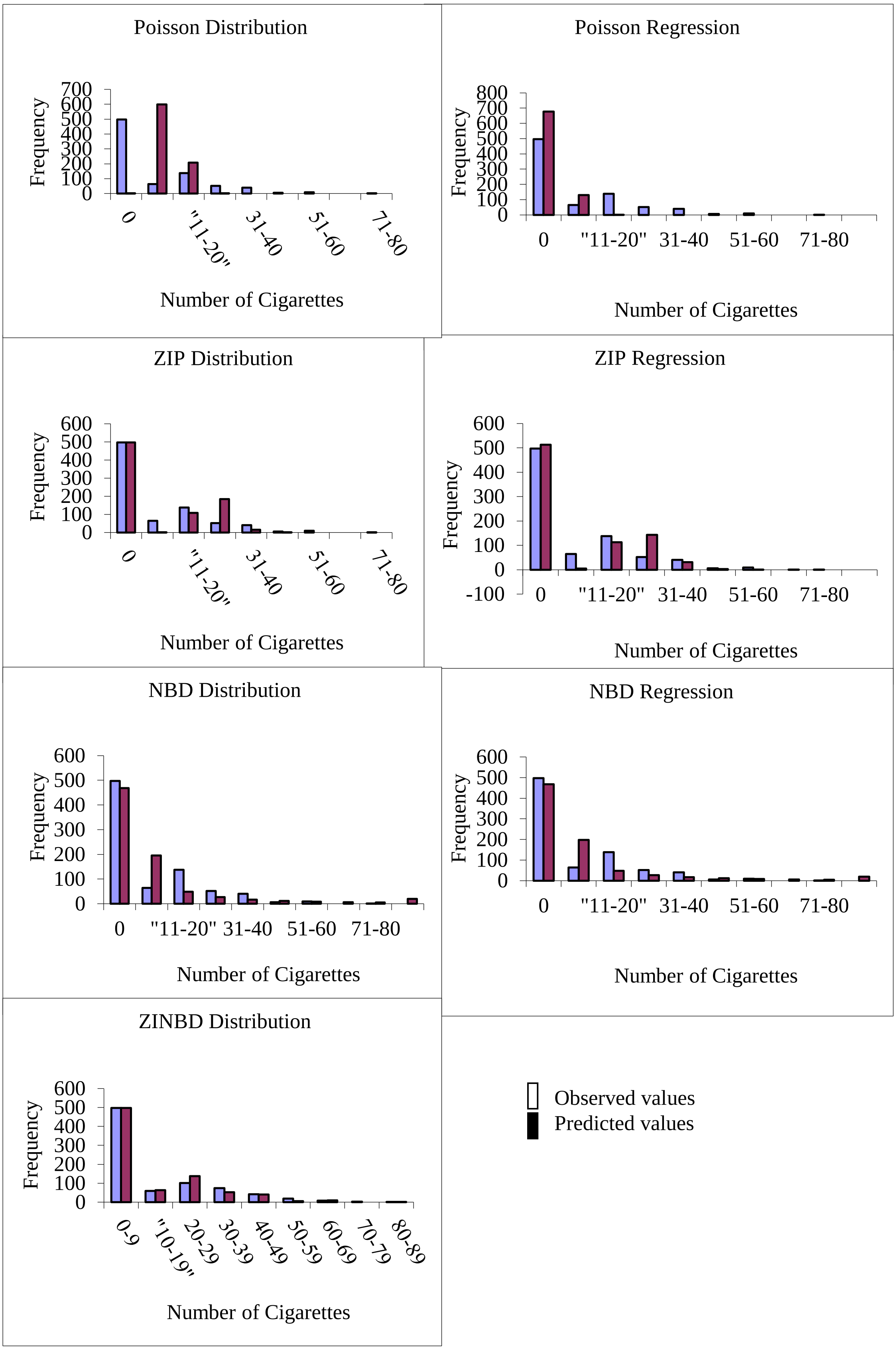
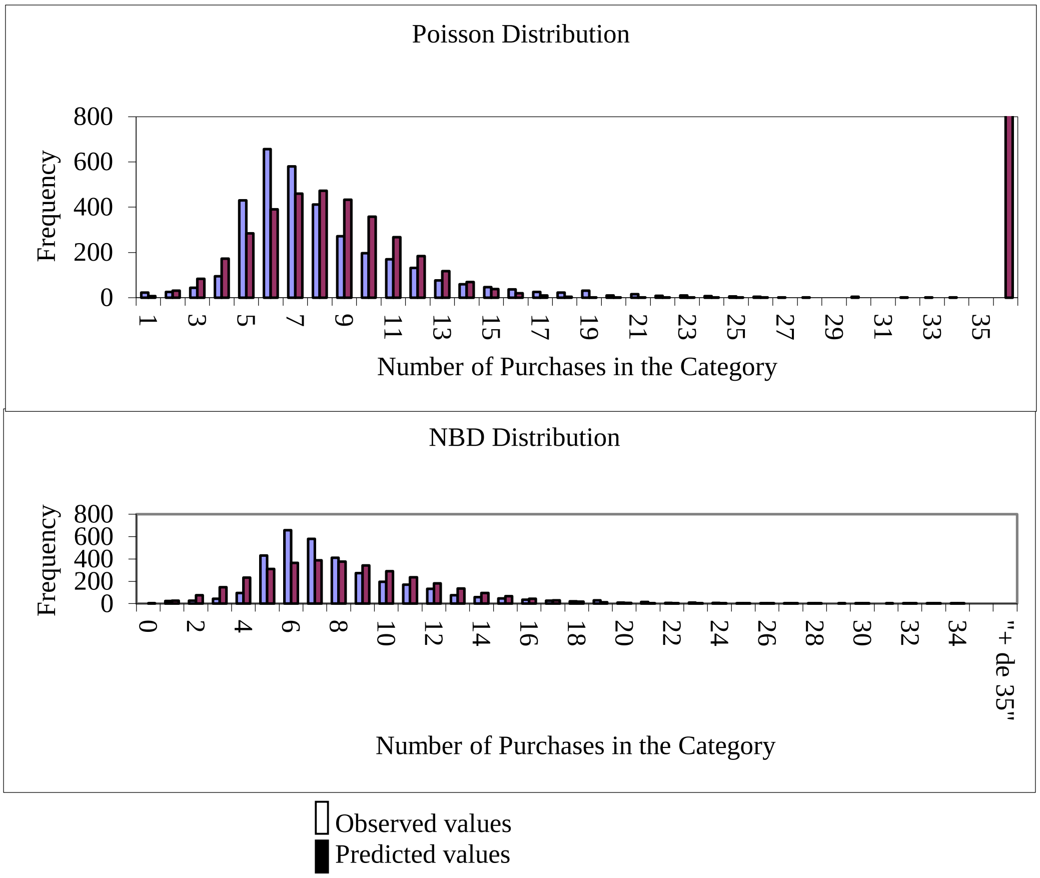


FIGURE 2 - Histograms of observed values of the sample and values predicted by the models (database 2).


Poisson Distribution
Frequency
Number of Purchases in the Category
NBD Distribution
Frequency
"+ de 35"
Number of Purchases in the Category
Observed values
Predicted values