© Health Research and Educational Trust
DOI: 10.1111/1475-6773.12055
METHODS BRIEF

Beta-Binomial Regression and Bimodal Utilization

Chuan-Fen Liu, James F. Burgess Jr., Willard G. Manning, and Matthew L. Maciejewski

Objective. To illustrate how the analysis of bimodal U-shaped distributed utilization can be modeled with beta-binomial regression, which is rarely used in health services research.

Data Sources/Study Setting. Veterans Affairs (VA) administrative data and Medicare claims in 2001–2004 for 11,123 Medicare-eligible VA primary care users in 2000.

Study Design. We compared means and distributions of VA reliance (the proportion of all VA/Medicare primary care visits occurring in VA) predicted from beta-binomial, binomial, and ordinary least-squares (OLS) models.

Principal Findings. Beta-binomial model fits the bimodal distribution of VA reliance better than binomial and OLS models due to the nondependence on normality and the greater flexibility in shape parameters.

Conclusions. Increased awareness of beta-binomial regression may help analysts apply appropriate methods to outcomes with bimodal or U-shaped distributions.

Key Words. Beta-binomial, bimodal, veterans, primary care, Medicare, utilization

There are a number of situations in which health services researchers can encounter data with a bimodal (U-shaped) distribution. When such a situation arises, it may be beneficial to adopt an analytic approach that explicitly incorporates the bimodal nature of the distribution. One useful tool for this is beta-binomial regression, which has been rarely used in health services research but has the flexibility to fit various distributions, including bimodal data. This article describes some of the key issues that arise with beta-binomial regression and illustrates the use of this technique in a real-world application.

One situation in which a bimodal distribution could arise is when patients have two alternative choices of health care provider, and data measure the share of times patients use one of the providers. A recent study showed that reliance on the Department of Veterans Affairs (VA) health care system, that is, the proportion of primary care received in the VA, among

Medicare-eligible veterans had a bimodal distribution and residuals had a nonnormal distribution (Liu et al. 2011). This VA reliance measure was not only bimodal but also followed a U-shaped distribution, because most veterans tended to be either mostly or entirely reliant on VA, or hardly or not at all reliant on VA; while there was a significant proportion remaining using both VA and Medicare services. For example, from the study by Liu and colleagues, 40% of veterans had a VA reliance measure for primary care greater than 0.8; 25% had a VA reliance measure less than 0.2 in 2004; and 35% of veterans had a VA reliance measure in between 0.2 and 0.8 (Liu et al. 2011). Beta-binomial regression is a rarely used model in health services research that has the flexibility to fit various distributions, including bimodal and U-shaped distributions.

The purpose of this article was to illustrate how to model bimodally distributed utilization using a beta-binomial regression. First, we describe the basic properties of the beta-binomial distribution. Second, we describe the process that can be employed to estimate a beta-binomial regression using the statistical model developed by Guimaraes, using a fixed effect negative binomial model in Stata (xtnbreg) (Guimaraes 2005). Finally, we compare the predicted mean and predicted distributions from beta-binomial, ordinary least squares (OLS), and binomial regressions to illustrate how the different models fit the bimodally distributed VA reliance outcome, showing the superiority of the beta-binomial approach in this example.

METHODS

Sample, Data Sources, and VA Reliance Outcome

The study sample included 11,123 Medicare-eligible veterans who used VA primary care services in 2000, which is a subset of a sample from a prior study that assessed reliance on VA primary care from 2001 to 2004 (Liu et al. 2011). Data sources included VA administrative databases, Medicare claims, the Area Resource File, and 2000 Census data to proxy beneficiary characteristics

Address correspondence to Chuan-Fen Liu, Ph.D., M.P.H., is with the Northwest Center for Outcomes Research in Older Adults at the VA Puget Sound Health Care System, 1660 S. Columbian Way, Seattle, WA 98108; e-mail: Chuan-Fen.Liu@va.gov. James F. Burgess, Jr., Ph.D., is with the Center for Organization, Leadership & Management Research at the VA Boston Healthcare System, Boston, MA. Willard G. Manning, Ph.D., is with the University of Chicago, Chicago, IL. Matthew L. Maciejewski, Ph.D., is with the Center for Health Services Research in Primary Care at the Durham VA Medical Center, Durham, NC.

by zip code. The outcome variable was reliance on VA primary care defined as the proportion of all VA or Medicare primary care visits that occurred in VA in 2004. The algorithm of classifying primary care visits in VA and Medication was described in detail elsewhere (Burgess et al. 2011).

Beta-Binomial Model

The beta-binomial model is a combined model of the beta and binomial distributions. The binomial distribution is a discrete probability distribution arising when the probability of success (p) in each of a fixed or known number of Bernoulli trials (n) is either unknown or random. In the case of VA reliance, the probability of success is the probability that a veteran has a VA primary care visit given the number of all VA/Medicare primary care visits in a year. The beta distribution is a family of continuous probability distributions defined on the interval $(0, 1)$ parameterized by two positive shape parameters, typically denoted by a and b . These shape parameters provide a tremendous amount of flexibility to model different empirical shapes over the $(0, 1)$ interval. The beta-binomial distribution is used to model the number of successes in n binomial trials when the probability of success is p with a beta distribution with parameters a and b . Further details of the distribution and estimation assumptions can be found in Guimaraes (2005).

Figure 1: Distributions of Beta-Binomial Model

The beta-binomial is flexible in fitting different shapes depending on the values of two shape parameters, a and b (Figure 1). The beta-binomial distribution is U-shaped if both a and b are less than 1; however, other values of a and b can generate shapes that are monotonically rising toward either end or are flat. The beta-binomial is a uniform distribution if both a and b are equal to 1. The beta-binomial approximates the binomial distribution if a and b are large (>1). For application purposes in these $(0, 1)$ interval bounded problems, the beta-binomial approach is most useful in this case where the parameters follow the U-shape but are bounded by 0 and 1 inclusive of the zero and one.

We assessed the appropriateness using the beta-binomial model in our application based on the following steps. First, we examined the distribution of the VA reliance outcome and residuals, which appeared to be bimodal and U-shaped. Second, we examined the model fit using the conventional statistical approaches, including the binomial model and OLS model. We found that the binomial and OLS regressions fit the distribution poorly because the predictions of VA reliance were concentrated around the mean where we had relatively few observations in the data. Furthermore, the residuals from these two models were not normally distributed. Finally, we considered a beta-binomial model, because it is most useful in the case of extreme over-dispersion that cannot be addressed with a negative binomial model, such as U-shaped distributions with finite support bounded between 0 and 1. We fitted the beta-binomial model to confirm the U-shaped distribution of the VA reliance measure.

Beta-binomial models may also be appropriate for outcomes with distributions that are either monotonically rising or falling or are flat. The parameters a and b of the beta-binomial model can be chosen to provide flexibility to handle many possible situations in health services research that have this “probability” nature of constraining between 0 and 1, and are more diffuse than the over-dispersion capabilities of the negative binomial distribution (Morris and Lock 2009).

Process of Estimating Beta-Binomial Models

We estimated the beta-binomial regression on VA reliance in Stata, based on a fixed effect negative binomial approach (xtnbreg), developed by Guimaraes (Guimaraes 2005). This section describes the three steps for data construction and regression modeling (see Appendix SA2 for more description and Stata code).

First, one must structure the data in a particular way to estimate the parameters of the beta-binomial model using the xtnbreg command. For each

individual, there are two records per year. The first record indicates the number of visits (*pc_enctr*) that occurred in a Medicare outpatient primary care clinic (*ilocation* = 0), while the second record indicates the number of visits (*pc_enctr*) that occurred in a VA outpatient primary care clinic (*ilocation* = 1).

Second, one must estimate the shape parameters, a and b , from a beta-binomial regression without including any covariates once the data are properly structured. In this simple model, the dependent variable is the number of visits (*pc_enctr*) and the independent variable includes “*ilocation*.” In this example, the shape parameters for the VA reliance distribution were $a = 0.517$ and $b = 0.305$. Both a and b were less than 1, which indicates that the distribution is U-shaped. From this regression, we also predicted that the unadjusted mean VA reliance was 0.629, suggesting that 62.9% of total primary care visits occurred in VA.

Third, we reestimate the beta-binomial with covariates, using an approach first pioneered by Heckman and Willis (1977) and further details provided in Guimaraes (2005) (Heckman and Willis 1977; Guimaraes 2005). To reestimate the model, we need to construct the interaction terms of covariates and the variable indicating the location of visits which occurred in VA (*ilocation*). Then, we reestimate the beta-binomial regression with the covariates including *ilocation* and the interaction terms between *ilocation* and covariates categories. The shape parameters (a and b) and the mean VA reliance can be estimated for the model with covariates as well. The coefficients can be interpreted as incidence rate ratios by exponentiating the coefficients of the interaction terms, in cases where that is of interest.

Analysis

We compared the predicted mean and distribution of VA reliance from the beta-binomial model with binomial and OLS models. OLS is a linear regression model that minimizes the sum of squared vertical distances between the observed responses and the responses predicted by the linear approximation, and provides unbiased and efficient estimators if residuals are normally distributed. We used the binomial regression to estimate the number of success, that is, number of primary care visits in VA, in a given number of trials, that is, the number of total primary care visits in both VA and Medicare, using the Stata command “*binreg*” with a log link, a generalized linear model with extensions to the binomial family.

All beta-binomial, OLS, and binomial regression analyses adjusted for patient characteristics in the baseline year (2000) included age, gender, race, marital status, Medicaid status, VA copayment status, distance to the closest VA facility, the original reason for Medicare eligibility, comorbidity burden using the Diagnostic Cost Group (DCG), the number of nonfederal primary care physicians, and median income in the zip code level.

Analyses were conducted using *Stata 11* (StataCorp 2009). Human subjects' approvals for these analyses were obtained from the Boston, Durham, and Seattle VA Medical Centers.

RESULTS

The mean age of the study sample was 69.7, 30% were originally eligible for Medicare due to disability, a majority were male (97%) and white (91%), and 70% were married (Table 1). Only 5% were dually eligible for Medicare and Medicaid, and the majority of these (79%) were exempted from VA copayments. The average DCG risk score (0.85) indicated a comorbidity burden below that of the average Medicare beneficiary. On

Table 1: Descriptive Statistics of Medicare-Eligible VA Primary Care Users

<i>Patient Characteristic</i>	
Age (mean, SD)	69.7 (9.4)
<55 (%)	9.2
55–64 (%)	8.1
65–74 (%)	49.0
≥ 75 (%)	33.7
Female (%)	2.8
Married (%)	69.5
White (%)	90.7
Disability as the original reason for Medicare eligibility (%)	29.9
Medicaid (%)	5.0
Copay status for VA care	
Free care due to disability (%)	33.7
Free care due to income (%)	44.9
Required to pay copayments (%)	21.4
Diagnostic cost group (DCG) [†]	0.85 (0.59)
Distance to VA, [‡] miles (mean, SD)	19.3 (27.8)
Per cap income in zip code/10,000 (mean, SD)	1.994 (0.721)
Number of non-federal primary care physicians per 100,000 population in county (mean,SD)	38 (18)

[†]Measured at baseline in 2000 including diagnoses from both VA and Medicare.
[‡]Distance to the closest VA facility, including community clinics and medical centers.

average, the closest VA facility was 19.3 miles from the patient's residence. The average number of primary care visits in VA and Medicare was 4.2 visits in 2004 and an average of 63% of all primary care visits occurred in VA (i.e., VA reliance was 0.63).

We predicted mean VA reliance from the OLS, binomial, and beta-binomial models after adjusting for covariates (see note in Figure 2). The mean VA reliance from the OLS model was exactly the same as the unadjusted VA reliance (0.63), as expected. In contrast, the mean VA reliance from the binomial model (0.54) was well below the mean of adjusted and OLS means. The mean VA reliance from the beta-binomial model (0.71) was above those from the OLS and binomial regressions.

We then examined the distribution of predicted VA reliance in 2004 from the three models and compared these distributions with unadjusted VA reliance. The unadjusted distribution of VA reliance was U-shaped (Figure 2), because Medicare-eligible veterans were most commonly either entirely reliant on VA (100% because all visits in VA) or completely not reliant (0%). When compared with a normal distribution with skewness of 0 and kurtosis of 3, this graph indicates that the unadjusted VA reliance is nonnormal with skewness of -0.847 and kurtosis of 1.541 . The distribution of predicted VA reliance shows that the beta-binomial model generated predictions that

Figure 2: Distribution of VA Reliance for Primary Care across Models

tracked closely with the U-shaped distribution of unadjusted VA reliance. However, the binomial and OLS regressions fit the distribution poorly because predictions were concentrated around the mean, and residuals were not normally distributed, as expected.

DISCUSSION

This study illustrates the application of beta-binomial regression to the analysis of a bimodal utilization measure. The beta-binomial model fits better than the OLS and binomial model, following from the nondependence on normality and the greater flexibility in shape parameters. Using the beta-binomial model produced a less biased estimate of predicted VA reliance because this model's flexibility can fit this type of bimodal distribution better than other forms of regression (Chatfield and Goodhardt 1970; Heckman and Willis 1977).

In statistical analyses, researchers generally ask research questions that are operationalized as a contrast in means between two groups with little consideration for the shape of the distribution. For outcomes with bimodal or U-shaped distributions, the distribution is important to consider if our goals are to understand the actual distribution and to ensure that we can generate unbiased regression coefficients and mean values that generate reasonable predictions. Our study shows that binomial or OLS models may match the unadjusted means but poorly estimate the entire distribution when the outcome is bimodally distributed. The extreme flexibility of the shape parameters of the beta-binomial model allows us to estimate a regression that tracks closely to the underlying distribution of residuals. Our study shows that the shape of the distribution is critical, because significant shifts occurred at both extremes of VA reliance (Liu et al. 2011).

Analysts dealing with binomial utilization or other outcomes that require estimation with bimodal or U-shaped distributions should consider this rarely used beta-binomial method, because mean effects will be more plausible and predicted values across the range of the data will track to the actual data much better. Moreover, using easily implemented software in Stata, estimating the beta-binomial distribution is straightforward (see Appendix SA2). Even without access to Stata, the backbone of the beta-binomial is simple ratios of gamma distributions that make it relatively easy to code. We illustrate the use of beta-binomial regression in this U-shaped distribution example, but beta-binomial regression is extraordinarily flexible for all types

of distributions of the outcome of interest like all beta distribution-based estimation procedures.

OLS regression is well known for its robust ability to estimate means across a wide range of distributions. In many health services research applications, we desire an ability to track the whole distribution of the data better for utilization, cost, quality, or other outcomes, especially those where continuity of care issues arise from utilization or other outcomes obtained by more than one provider. In such cases, beta-binomial regression is easy to implement, easy to graph and interpret, and can capture a wide range of distributions, including U-shaped distributions not well captured by any other method. A greater use of beta-binomial regression in such cases is recommended.

ACKNOWLEDGMENTS

Joint Acknowledgment/Disclosure Statement: This work was supported by the Office of Research and Development, Health Services Research and Development Service, Department of Veterans Affairs, project number IIR 04-292. Dr. Maciejewski was also supported by a Research Career Scientist award from the Department of Veterans Affairs (RCS 10-391). The views expressed herein are those of the authors and do not necessarily represent the views of the Department of Veterans Affairs, the University of Washington, Boston University, the University of Chicago, and Duke University.

REFERENCES

- Burgess, J. F. Jr., M. L. Maciejewski, C. L. Bryson, M. Chapko, J. C. Fortney, M. Perkins, N. D. Sharp, and C. F. Liu. 2011. "Importance of Health System Context for Evaluating Utilization Patterns across Systems." *Health Economics* 20 (2): 239–51.
- Chatfield, C., and G. J. Goodhardt. 1970. "The Beta-Binomial Model for Consumer Purchasing Behaviour." *Applied Statistics* 19 (3): 11.
- Guimaraes, P. 2005. "A Simple Approach to Fit the Beta-Binomial Model." *The Stata Journal* 5 (3): 10.
- Heckman, J. J., and R. J. Willis. 1977. "A Beta-Logistic Model for the Analysis of Sequential Labor Force Participation by Married Women." *Journal of Political Economy* 85 (1): 32.
- Liu, C. F., W. G. Manning, J. F. Burgess Jr, P. L. Hebert, C. L. Bryson, J. Fortney, M. Perkins, N. D. Sharp, and M. L. Maciejewski. 2011. "Reliance on Veterans

Affairs Outpatient Care by Medicare-Eligible Veterans.” *Medical Care* 49 (10): 911–7.

Morris, C. N., and K. F. Lock. 2009. “Unifying the Named Natural Exponential Families and Their Relatives.” *The American Statistician* 63 (3): 247–53.

StataCorp 2009. *STATA Reference Manual: Release 11*. College Station, TX.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of this article:

Appendix SA1: Author Matrix.

Appendix SA2: Process of Estimating Beta-Binomial Models.

Appendix SA3: A Simple Approach to Fit the Beta-Binomial Model.