A BETA BINOMIAL MODEL FOR VICTIMIZATION

by

Mark J. Schervish

Technical Report No. 273
Department of Statistics
Carnegie-Mellon University
Pittsburgh, PA 15213
1 June 1983

SUMMARY

This report describes attempts to develop models for victimization that incorporate two features. First, occurrences of victimizations for given housing units should be dependent over time, while those from housing unit to housing unit can be independent (conditional on the parameters of the model.) Secondly, there should be an explicit way of using victimization information on a given housing unit to help impute missing observations for that housing unit. The first of these goals can be met by introducing a beta-binomial model for the number of months in which a given housing unit is victimized. The second goal has not yet been successfully met within the framework of the beta-binomial model.

1. INTRODUCTION

The National Crime Survey (NCS) is a survey of housing units nationwide. The residents of each surveyed housing unit are asked to describe any incidents in which they were victims of criminal activity. One goal of the analysis of the NCS data is to provide yearly estimates of the proportion of housing units which are crime-free (not victimized). One problem is that not every housing unit in the NCS sample during a given year, is sampled for the entire year. Also, because some housing units are victimized more than others, we take the following approach. We assume that the number of months of victimization for a given housing unit has binomial distribution conditional on a parameter p, and then that p has a beta distribution with parameters α and β. This allows the victimizations for a given housing unit to be correlated with each other, but does not force the same correlation between victimizations in different housing units. A housing unit with no missing data, which was in the sample for n months and had k months of victimization would contribute a factor of

$$\frac{\Gamma(\alpha+\beta)\Gamma(\alpha+k)\Gamma(\beta+n-k)}{\Gamma(\alpha)\Gamma(\beta)\Gamma(\alpha+\beta+n)} \tag{1.1}$$

to the likelihood function of (α, β). In section 2, some numerical results are given using 1% samples of the NCS data from 1975, 1976, and 1977. These results all assume that data is missing at random.

The goal of modelling non-response within the beta-binomial model proven more difficult to meet. The approach followed is to let each missing observation be treated as part victimization and part non-victimization. So, a housing unit in the sample for n months with k months of victimization and m months of missing data would be treated as if it had k+mx months of victimization and k+m(1-x) months of non-victimization. Three different methods for defining x are described in section 4. None of them has proven satisfactory.

2. A BETA-BINOMIAL MODEL

Equation (1.1) gives the likelihood function for one housing unit in the sample for n months with k months of victimization. The sufficient statistics, then, would be the numbers of housing units with n months in the sample and k months of victimization for all n from 1 to 12 and all k from 0 to n. These statistics have been calculated for each of the three years 1975, 1976, and 1977 for two of the 10 1% subsamples of NCS data.

Recall that the beta-binomial model can be summarized by assuming that the number of months of victimization for a given housing unit with n months of data is binomial bin(n; p) conditional on p, and p has a beta Be(α, β) distribution. The probability that a given housing

unit is crime-free in a given month under this model is the expected value of p

$$E_p = \beta/(\alpha+\beta), \tag{2.1}$$

and the probability that a given housing unit is crime free for s months is

$$\theta_s = \Pi_{i=1}^{s}\{(\beta+i-1)/(\alpha+\beta+i-1)\}. \tag{2.2}$$

We will denote θ_{12} simply θ, which is the probability that a given housing unit will be crime free for a whole year. We can measure the variation in p from housing unit to housing unit by its variance

$$V_p = \alpha\beta/\{(\alpha+\beta)^2(\alpha+\beta+1)\}.$$

Because the 1% samples were large (approximately 900 housing units each), the method of maximum likelihood was used to estimate α and β, and thereby provide estimates of E_p, θ, and V_p. The posterior means of the parameters should be approximately the MLE's. The results are summarized in Tables 1 and 2. The calculations were done using double precision on a VAX 11/780 using IMSL subroutine ZXMIN.

Table 1: Estimates for Subsample #1

Item Estimated	YEAR 1973	1974	1975
Number[1]	823	948	937
α	0.726	0.499	0.736
β	18.3	12.1	17.0
E_p	0.038	0.040	0.042
θ	0.692	0.705	0.673
$V_p^{1/2}$	0.043	0.053	0.046

[1]Number of housing units in subsample #1.

An obvious question which arises is "How well does this model fit the data?" One of the major goals is to produce an estimate of the number of crime free housing units in a given year. We will address the question of how well the model fits the data in terms of the number of crime free housing units in each of the three years. Let ϕ_s stand for the probability that a housing unit with s months of data is crime free for those s months. The beta-binomial model says that $\phi_s = \theta_s$ defined in (2.2). A more general model would be to say that the values of ϕ_s are unconstrained, that is, they are twelve independent parameters. If we assume a uniform prior distribution for the vector $\phi = (\phi_1,\ldots,\phi_{12})$ over the product of

Table 2: Estimates for Subsample #10

Item Estimated	YEAR 1973	1974	1975
Number[1]	836	966	955
α	0.537	0.545	0.557
β	11.3	11.5	13.8
E_p	0.045	0.045	0.039
θ	0.675	0.675	0.703
$V_p^{1/2}$	0.058	0.057	0.049

[1]Number of housing units in subsample #10.

12 unit intervals $[0.0, 1.0]^{12}$, the posterior distribution of ϕ will be that of 12 independent beta random variables with means of

$$(r_s+1)/(n_s+2)$$

and variances of

$$(r_s+1)(n_s-r_s+1)/\{(n_s+2)^2(n_s+3)\},$$

where n_s is the number of housing units with s months of data, and r_s is the number of those housing units which were crime free.

To see how well the beta-binomial model fits the data, consider an unobserved housing unit which would have s months of data, and let Y_s be a random variable equal to 1 if the unit is crime free, and 0 if not. We will compute the posterior expected squared difference between Y_s and $\hat{\theta}_s$ for each s, and compare it to the variance of Y_s. The expected squared difference will be larger than the variance, but, if the model fits well, it will not be much larger. The expected squared difference between a random variable and a constant is the variance of the random variable plus the square of the difference between the constant and the mean of the random variable. This difference is known as the *bias* of the constant as an estimate of the random variable. As a measure of the lack of fit of the beta-binomial model, we will use the square of the bias of $\hat{\theta}_s$ as an estimator of Y_s divided by the variance of Y_s for each s. We will then average these ratios with weights proportional to n_s.

The mean of Y_s is the mean of ϕ_s, and the variance of Y_s is the variance of ϕ_s plus the expected value of $\phi_s(1-\phi_s)$. The variance is then

$$(r_s+1)(n_s-r_s+1)/(n_s+2)^2.$$

This was then divided into

$$[(r_s+1)/(n_s+2) - \hat{\theta}_s]^2,$$

multiplied by n_s, summed over all s, and divided by the total number of housing units observed. The results are in Table 3.

Table 3: Lack of Fit Measures for Beta-Binomial Model

Subsample	YEAR 1973	1974	1975
1	0.013	0.018	0.021
10	0.025	0.022	0.010

The results in Table 3 can be summarized by saying that under a model which allows the ϕ_s to be unrelated to each other, thereby allowing very close fit to the data, the expected squared error for predicting whether a future housing unit will be crime free increases only by about 2% when the beta-binomial prediction is used rather than the optimal prediction under the more general model. On the surface, a 2% increase may seem like a small amount. But it must be compared to the lack of fit of some other model to put it in perspective. For example, if the worst possible model only had a 3% increase, then 2% would look quite large. To see that 2% is a close fit, compare the lack of fit measures in Table 4 for the model which says that the data doesn't matter. Under this model, which ought to fit very poorly, the prediction for Y_s is always

Table 4: Lack of Fit Measures for Poorly Fitting Model

Subsample	YEAR 1973	1974	1975
1	0.842	1.595	2.291
10	2.321	2.464	0.492

The increases in expected squared error for this poorly fitting model range from 50% to 250%, which are much larger than the 2% increses for the beta-binomial model. So, on an absolute scale, the beta-binomial model fits well. In the next section, we will compare this model to another well-fitting model.

Before concluding the discussion of lack of fit, it should be noted that there are other methods for measuring the lack of fit of the beta-binomial model. First of all, we need not have restricted attention to crime-free housing units. We could have also considered all those with exactly one month of victimization, those with exactly two months of victimization, etc. Except for s=12, the number of victimized housing units with s months of data is quite small. A large number of lack of fit measures based on such small samples would not be very useful. Secondly, we could have used more familiar lack of fit measures such as chi-suared statistics. One problem with such measures, however, is that there is no explicit alternative to compare a model to. Without any alternative in mind, it is difficult to justify any particular lack of fit

statistic.

3. COMPARISON TO AN AD HOC ESTIMATOR

Griffin (1983) considers an intuitively plausible estimator of θ which she calls the *modified ad-hoc estimator* $\hat{\theta}_1'$. The formula for this estimator is

$$\hat{\theta}_1' = \Sigma_{s=1}^{12} sr_s / \{\Sigma_{s=1}^{12} sr_s + 12\Sigma_{s=1}^{12}(n_s - r_s)\}.$$

This estimator is not based explicitly on any model, but there is a model for victimization under which it is strongly consistent (see Griffin, 1983). Under this model,

$$\theta_s = 12\theta / \{(12-s)\theta + s\}. \tag{3.1}$$

By plugging the value of $\hat{\theta}_1'$ into (3.1) for θ, we can obtain a set of estimates for θ_s which differ from those obtained from the beta-binomial model. The remarkable fact is that the two sets of estimates are nearly identical in all cases compared. For subsample #10, lack of fit measures were calculated for the modified ad hoc model in the same manner as for the beta-binomial model. For the three years 1973-5, the measures were 0.028, 0.025, and 0.009. These compare favorably with those of the beta-binomial model in Table 3 row 2. The advantage to the modified ad hoc model is that it has only one parameter rather than two to estimate. The disadvantages are that it does not allow (without further modification) estimates of the probabilities of being victimized in 1, 2, 3, etc. months, and that the formula (3.1) is not intuitively understandable.

4. ATTEMPTS TO MODEL MISSING DATA

As previously mentioned, the attempts to model missing data within the beta-binomial model consisted of treating each month of missing data as if it were part victimization and part non-victimization. A number x between 0 and 1 would be added to the number of months of victimization for every month of missing data for a given housing unit. The number 1-x would be added to the number of months of non-victimization. The likelihood function for a given housing unit with n months of observed data, m months of missing data, and k months of victimization would be

$$\frac{\Gamma(\alpha+\beta)\Gamma(\alpha+k+mx)\Gamma(\beta+n-k+m(1-x))}{\Gamma(\alpha)\Gamma(\beta)\Gamma(\alpha+\beta+n+m)} \tag{4.1}$$

Two different methods of defining x were considered. A third method of modelling non-response did not involve defining x, and is described after the first two. None of the methods proved useful for maximum likelihood estimation.

The first method of defining x was to let it be a third parameter in the model. From

the fact that the fact that the gamma function is convex for positive arguments, it follows that the minimum of (4.1), as a function of x for fixed α and β, will occur at

$$x_0 = 0.5(\beta - \alpha + m)/m. \tag{4.2}$$

The maximum will occur at one of the endpoints 0 or 1. If x_0 is not between 0 and 1, then (4.1) will be monotone in x for fixed α and β. The maximum would then occur at x = 0 if $\beta > \alpha$ and at x = 1 if $\beta < \alpha$. Typically, $\hat{\beta}$ will be much larger than $\hat{\alpha}$ so that the maximum would occur at x = 0, and not much useful information would be contained in the MLE.

The second method of defining x was to set it equal to E_p defined in (2.1). It was hoped that if those housing units with missing data had more victimizations than the others, the estimate of E_p would be increased over its estimate assuming the data missing at random. However, the same feature which drove the estimate of x above to 0, causes the estimate of E_p to be smaller under this new model unless the housing units with missing data had more months of victimization than of non-victimization. Since this was rarely the case, the MLE of E_p was of necessity smaller under this new model than under the missing-at-random-model, regardless of whether those housing units with missing data had more or fewer victimizations than the others.

As an example, take the year 1973 for subsample #1. Table 5 gives a summary of the victimization records for those housing units with missing data.

Table 5: Summary of Missing Data

No. Missing Months	No. Housing Units	Percent Victimization[1]
0	667	3.60
1	16	4.55
2	21	6.67
3	14	2.38
4	20	3.96
5	12	2.38
6	44	8.45
7	5	0
8	8	6.25
9	5	0
10	5	0
11	6	0

[1]The entry in this column on row i is the average over all housing units with i months of missing data of the percentage of months of data for which a victimization was recorded.

Note that most housing units with missing data had more months of victimization per month of data than those with complete data. However, the MLE of E_p under the second model described above was only 0.015, which is much smaller than 0.038 estimated using the missing-at-random model.

A third attempt to model missing data was to say that conditional on p, the probability that there was a victimization during a missing month was p, and the probability of no victimization was 1−p. The contribution to the likelihood of α and β for a housing unit with n months of observed data, m months of missing data, and k months of victimization would be

$$\int_{(0,1)} p^{\alpha+k-1}(1-p)^{\beta+n-k-1}[p^2+(1-p)^2]^m dp \; \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}$$

This method, unfortunately, has the same problem that the other two have. The extra factor $p^2+(1-p)^2$ is maximized at p=0 and p=1. This will cause the MLE of E_p to decrease as long as the MLE of α is much less than the MLE of β.

5. CONCLUSION

A model for victimization has been proposed and fit to data from the National Crime Survey. The model was designed so that there would be dependencies between victimizations at each housing unit. The model is related to a gamma−poisson model for repeated events described by Nelson (1982). In the gamma−poisson model, the conditional probability given γ of having x victimizations in time t would equal $e^{-\gamma t}(\gamma t)^x/x!$, while γ would have a gamma distribution $\Gamma(\alpha, \beta)$. To compare this to the beta−binomial model, assume that γ has units of victimizations per month. Then one can write

$$p = 1 - e^{-\gamma},$$

where p is the (conditional) probability of being victimized in a given month. Transforming the $\Gamma(\alpha, \beta)$ density for γ into the density of p, we obtain

$$\{\beta^\alpha/\Gamma(\alpha)\}[-\log_e(1-p)]^{\alpha-1}(1-p)^{\beta-1} \qquad (5.1)$$

for the density of p. If α is much smaller than β, then this density will be concentrated near small values of p, for which $-\log_e(1-p)$ is approximately p. Also, if α is much smaller than β, β^α is approximately $\Gamma(\alpha+\beta)/\Gamma(\beta)$. with these two approximations, (5.1) can be approximated by a beta $Be(\alpha, \beta)$ density. These two models should then produce similar results when only number of months of victimization are available, rather than the total number of victimizations and when they occurred.

6. ACKNOWLEDGEMENTS

This research was performed in part under Contract No. J−LEAA−015−79 with the Bureau of Justice Statistics, and in part under Grant 81−IJ−CX−0087 from the National Institute of Justice, both in the Office of Justice Assistance Research and Statistics, U.S. Department of Justice. Points of view and opinions stated herein are those of the author and do not represent the official position or policies of the U.S. Department of Justice.

REFERENCES

Griffin, D. (1983). *Consistency of some intuitive estimators of the prevalence of victimization.* Technical Report 271, Department of Statistics, Carnegie-Mellon University, January.

Nelson, J. (1982). *The Dirichlet-gamma-poisson model of repeated events: a multivariate description of crimal victimization in American cities.* Technical Report, Michael J. Hindelang Criminal Justice Research Center, Inc., September.

END