Bayesian Multiple Imputation of Zero Inflated Count Data

Chin-Fang Weng
chin.fang.weng@census.gov

U.S. Census Bureau, 4600 Silver Hill Road, Washington, D.C. 20233-1912

Abstract

In government survey applications, zero-inflated count data often arise, sometimes with item nonresponse. We consider the problem of imputing missing counts. We assume that the observations/items are missing at random. We also assume the zero-inflated data is a mixture distribution: one component from a distribution degenerate at zero and one from a Poisson distribution. Both components may depend on covariates, which are always observed. We formulate a model for bivariate zero inflated count data and propose a Bayesian imputation scheme for imputing missing items by assigning priors to unknown regression parameters. Using the predictive distribution of missing items given observed items, one can impute each missing item as a random draw from the predictive distribution. We use Markov Chain Monte Carlo to generate imputed values of missing items. Multiple imputations are computed by running the chain long enough to produce multiple realizations of the missing item values. To obtain (nearly) independent draws, the chain must be thinned. We will illustrate its potential with a simulation study and with an analysis of part-time Employment data from the Annual Survey of Public Employment & Payroll (ASPEP).

Key Words: Nonresponse; Mixture distribution; Missing; Poisson; Regression

Disclaimer: This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress. Any views expressed on statistical, methodological, technical, or operational issues are those of the authors and not necessarily those of the U.S. Census Bureau.

1 Introduction

Nonresponse is a common problem in most large-scale surveys. In order for users to make a valid and efficient inference using some statistical procedure, imputing missing entries may be necessary. For some imputation methodologies, the imputation is done deterministically, providing neither adjustment for parameter estimation nor information of uncertainty of the imputed values, as though the imputed values are true values (Schafer, 1997; Gelman & Hill, 2006). Thus, variance of individual imputed values also is a main research focus of this study.

Multiple imputation (MI), developed by Rubin (1987), is a common approach to handle missing data issues. The MI technique involves three steps: impute m complete data sets, analyze each of the m completed data sets, and use ANOVA-like methods to integrate the m analyses into a final result. MI adjusts the standard errors of parameters due to the uncertainty of missing value by incorporating the variation from multiple sets of imputed data, given that the imputation model and the analysis model are the same or similar.

For public-use data sets, it is always unclear what analyses the ultimate users will conduct, so the imputation of public-use data focuses on the imputation model instead of the analysis model. Thus, the imputer should include all variables contained in the original data set in the imputation model. Also the imputer should not create or add new variables in the imputation model to avoid redefining the data set (Rubin, 1996; Schafer & Graham, 2002). It is possible to lose precision when including unimportant predictors, for example in sparse data situations. Nevertheless, it is a relatively small price to pay for the general validity of analyses of the multiply imputed data base (Rubin, 1996).

An imputation model contains two major outcomes: model parameters and imputed values. Since this is public-use data, special attention is given to imputed values, although the performance of these two components is highly correlated. Moreover, good imputed values are a byproduct of a good imputation model. Thus, modeling is the focus of this study.

In the Bayesian framework, the future or missing observation, y', can be estimated using the predictive distribution. Since the quantity y' can be considered as an additional parameter to be estimated, it can be generated using Markov Chain Monte Carlo (MCMC) methods from the conditional posterior distribution. During the iteration process, thousands of sets of imputed data are generated. Once the Markov Chain has stabilized, multiple imputations can be generated after thinning the chains (Gelman and Hill, 2006). The standard deviations of imputed values can be used to compute credible intervals.

In Section 2 the model is discussed, followed by a simulation study in Section 3. Section 4 presents a real data analysis. Section 5 presents conclusions and topics for future study.

2 Bivariate Zero Inflated Poisson (ZIP) Model

One frequently sees that a set of data contains an excess of zeros relative to standard distributions. Such zero inflated data appear in many fields, such as rainfall measurement, counts of numbers of seals, or counts of numbers of industrial defects. Some zeros are sampling zeros; for example, seals may swim under the sea instead of staying on the seashore. Some zeros are structural zeros, for example, male responses to questions about pregnancy. Many researchers have studied this problem and have developed various zero-inflated models in response. This study concerns imputation of zero-inflated count data using Bayesian modeling.

Zero Inflated Poisson Model

In a data set, let the discrete random variable Y_i be the *i*-th observed count, $i = 1, \dots, N$. Assume that Y_i is distributed as a mixture of two components: (1) responses which are zero with probability one (perfect state or zero state); (2) responses which follow a Poisson distribution (Poisson state). Assume that an unobserved random variable w_i indicates the state membership of the observed count, either the perfect state or the Poisson state. Note that if y_i > 0, the observed count is definitely in the Poisson state, but if y_i = 0, the subject may have been in either of the two states. This key feature makes the ZIP model different from the Poisson model. The w_i are assumed to be from a Bernoulli distribution with parameter p_i, such that $P(w_i = 1) = p_i$ and $P(w_i = 0) = 1 - p_i$. If $w_i = 1$ then $Y_i = 0$, coming from the perfect state, and if $w_i = 0$ then $Y_i = y$, y = 0, 1, 2, …, coming from a Poisson distribution. Therefore, Y_i has the ZIP distribution:

$$Y_i \sim \begin{cases} 0, & \text{with probability } p_i \\ \text{Poisson}(\lambda_i), & \text{with probability } (1 - p_i), \end{cases} \tag{1}$$

where Poisson (λ_i) is defined as $\mathrm{P}(Y_i = y_i) = \exp(-\lambda_i)\lambda_i^{y_i}/y_i!$.

Covariates can enter into the ZIP model in two places: in a logistic regression model for p_i and in a loglinear Poisson regression model for λ_i.

a) The logistic regression model is for predicting the state, either perfect state or Poisson state. The probability p_i is expressed as

$$\text{logit}(p_i) = \log\left(\frac{p_\text{i}}{1 - p_i}\right) = \boldsymbol{Z}_i{}'\boldsymbol{\beta} = \beta_0 + \beta_1 Z_i^1 + \cdots + \beta_h Z_i^h. \tag{2}$$

b) The log-linear regression for the Poisson mean is expressed as

$$\log(\lambda_i) = \boldsymbol{X}_i{}'\boldsymbol{\alpha} = \alpha_0 + \alpha_1 X_i^1 + \cdots + \alpha_k X_i^k. \tag{3}$$

Here $\boldsymbol{Z}_i$ and $\boldsymbol{X}_i$ are covariate vectors and $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are vectors of regression coefficients for the logistic regression model and the loglinear Poisson regression model, respectively. The components of $\boldsymbol{Z}_i$ and $\boldsymbol{X}_i$ could be the same or different from each other. This model was essentially proposed by Lambert (1992).

For imputation, we would like to model the Y_i using a ZIP distribution and impute directly using MCMC. However, because of limitations of the Bayesian Inference Using Gibbs Sampling **(**BUGS) program (Spiegelhalter et al., 2003), there is no function to do imputation for nonstandard distributions with count data. We sidestep this problem by creating the Bivariate ZIP model.

Bivariate Zero Inflated Poisson Model

Let (X, Y) be a pair of observed counts where X is always observed, and Y may be missing. We expect $E(Y|X = x)$ to be an increasing function of x . We propose the following model:

$$P[X = x] = P[X = x|\text{zero state}]I\{x = 0\}P(\text{zero state})$$

$$+P[X = x|\text{Poisson state}][1 - P(\text{zero state})]$$

$$= \pi I\{x = 0\} + (1 - \pi)e^{-\mu}\mu^x/x! \quad , \tag{4}$$

$$P[Y = y|X = x] = \pi I\{x = y = 0\} + (1 - \pi)e^{-(\gamma x)}(\gamma x)^y/y! \tag{5}$$

$$x, y = 0,1, \cdots$$

Through γx, the model relates a missing Y to an observed X, making imputation possible. In the bivariate ZIP model, we assume:

a) Observations stay in the same state from one year to the next.
b) Y and X have an approximate linear relationship: $Y \approx \gamma X + e$ such that on average γx equals μ_y.
c) The regression relationship may involve covariates $\boldsymbol{v}$.

Covariates can enter into the model in three places: in the logistic regression model for p_i , in the loglinear Poisson regression model for λ_i , and in the regression of Y on X for γ_i.

a) Logistic regression for P[zero class]

$$\text{logit}(\pi_i) = \boldsymbol{v}_i'\boldsymbol{b} \qquad \Rightarrow \pi_i = e^{\boldsymbol{v}_i'\boldsymbol{b}}/\left(1 + e^{\boldsymbol{v}_i'\boldsymbol{b}}\right) \tag{6}$$

b) Poisson regression for mean of X

$$\log(\mu_i) = \boldsymbol{v}_i'\boldsymbol{a} \qquad \Rightarrow \mu_i = e^{\boldsymbol{v}_i'\boldsymbol{a}} \tag{7}$$

c) Regression of Y on X

$$\log(\gamma_i) = \boldsymbol{v}_i'\boldsymbol{c} \qquad \Rightarrow \gamma_i = e^{\boldsymbol{v}_i'\boldsymbol{c}} \tag{8}$$

where $\boldsymbol{v}$ is a vector of covariates; $\boldsymbol{a}, \boldsymbol{b}$, and $\boldsymbol{c}$ are vectors of regression coefficients and γ is the slope in the regression of Y on X.

3 Simulation Study

3.1 Generating Simulated Data

These simulated data were generated based on a bivariate ZIP model. An R program for generating the data was written according to formulas (4) and (5) in Section 2. The total sample size was N = 3,000. One covariate, ***G***, with three levels was created. The subgroup sizes were n_1 = 1,200, n_2 = 1,500, and n_3 = 300. The missing values were randomly selected within each group: 120 (10%) from Group 1, 225 (15%) from Group 2, and 15 (5%) from Group 3. The overall correlation of x and y was 0.87.

Table 1 Simulation Values

G	*n*	Missing in *Y*	***b***	***c***	Mean (*X*)
Level 1	1,200	120 (10%)	-2.2	.05	17
Level 2	1,500	225 (15%)	-2.0	.00	15
Level 3	300	15 (5%)	-1.7	-.05	20

These simulated values are simplifications of values observed in the real data set of Section 4.

3.2 Analysis of Simulated Data

In the simulation study, we attempted to use reasonably informative priors for the unknown parameters in Equations (4) and (5). We believe that if the prior is too vague, it can lead to numerical instability causing BUGS to crash. Since we have a sample size of N = 3,000, we expect the prior effect to wash out because N is large. The prior distributions of the parameters are listed in Table 2. The precision parameters τ_b and τ_c were given Gamma (0.1, 0.1) priors.

Table 2 Parameters, Priors and Estimates Based on Simulated Data

Parameter	True	Prior choices	Estimate	SD	CV
b_1	-2.2	norm(-2, τ_b)	-2.26	.096	0.04
b_2	-2.0	norm(-2, τ_b)	-2.06	.080	0.04
b_3	-1.7	norm(-2, τ_b)	-1.77	.152	0.09
c_1	.05	norm(0, τ_c)	0.043	.0078	0.18
c_2	.00	norm(0, τ_c)	-0.006	.0074	-----
c_3	-.05	norm(0, τ_c)	-0.040	.0143	0.36

The estimated parameters, standard deviations and CVs are also presented in Table 2. All the parameter estimates were in 95% credible intervals. Parameters b_3 and c_3 have larger standard deviations than their counterparts. This is probably because of the smaller sample size of group 3. The $\boldsymbol{c}$ parameters are more difficult to estimate than the $\boldsymbol{b}$ parameters; this is because when the pair (x_i, y_i) is incomplete, no contribution to the likelihood for estimation of $\boldsymbol{c}$ parameters is provided. The true value for c_2 was set to 0 and the estimate of c_2 is close to 0, so no CV is provided for c_2 .

The 95% credible interval and sum of squared (SS) of imputation errors are used to measure the imputation performance. SS is defined as

$$SS = \sum \frac{\left(y_{\text{i,imp}} - y_{\text{i,true}}\right)^2}{\text{number of values imputed}}. \tag{9}$$

We observed SS = 15.8 and overall 96% of the true values were covered by their 95% credible intervals.

4 Analysis of Real Data

Annual Survey of Public Employment & Payroll (ASPEP)

The Annual Survey of Public Employment & Payroll (ASPEP) is a survey conducted annually that seeks to estimate the employment and payroll data for state and local governments in all states plus the District of Columbia. The ASPEP contains five type of government: counties, municipalities, townships, special districts, and school districts. There are five variables collected on the ASPEP form: Full-Time and Part-Time Payroll, Full-Time and Part-Time Employment, and Part-Time Hours. This paper focuses on Independent School Systems, Part-Time Employment (PTE), numbers of part time employees, in local government independent school systems. The design variables are State (50 states plus D.C.; 51 levels), Function (or job type; 5 levels; see Table 4 for detail), and School Level (7 levels).

Table 3 summarizes PTE observations for Independent School Systems in 2012. We see that PTE has excessive zeros and a high non-response rate. We notice that for reported data only, the proportion of zero values is 18% (4,045/22,086); while in the imputed data, the proportion of zero values is 67% (10,711/15,998).

Table 3 Zero Proportion and Non-response Rate of PTE Variable

2012 ASPEP Independent School System Part-time Employment		Reported Status		
		Imputed	Reported	Total
Observation Values	Zero	10,711	4,045	14,756 (38.7%)
	Positive Value	5,287	18,041	23,328 (61.3%)
	Total	15,998 (42.0%)	22,086 (58.0%)	38,084 (100%)

Source: Annual Survey of Public Employment & Payroll, U.S. Census Bureau

From Table 4, we see that the PTE variable is not missing completely at random (MCAR), the rates of missingness vary across the levels of the Function variable.

Table 4 2012 ASPEP Independent School Part-time Employment Variable Missing Status

2012 ASPEP Independent School System Part-time Employment		Reported Status		
		Imputed	Reported	Total
Function	012 Elementary/Secondary Education Instructional	3,227 (24%)	10,489 (76%)	13,716
	016 Higher Education – Other	3,183 (88%)	444 (12%)	3,627
	018 Higher Education – Instructional	3,183 (88%)	445 (12%)	3,628
	052 Libraries	3,177 (96%)	123 (4%)	3,300
	112 Elementary/Secondary Education – Other	3,228 (23%)	10,585 (77%)	13,813
	Total	15,998	22,086	38,084

Source: Annual Survey of Public Employment & Payroll, U.S. Census Bureau

We analyzed a subset of the ASPEP Part-time employment data, totaling 1,525 units, measured in both Year 2011 and Year 2012, respectively. These 1,525 units responded to all survey items, yielding 3,174 counts of the above Function codes. The correlation between x (PTE, Year 2011) and y (PTE, Year 2012) is 0.92. Descriptive statistics of the complete data set are presented in Table 5.

Table 5 Descriptive Statistics of Complete Data Set

Function	N	Variable	Mean	SD	Growth Rate $(\bar{y}/\bar{x})$
012 Elementary/Secondary Education - Instructional	1,373	x	228	484	0.97
		y	222	487	
016 Higher Education - Other	169	x	487	546	0.99
		y	480	514	
018 Higher Education - Instructional	205	x	498	562	1.04
		y	518	529	
052 Libraries	13	x	30	32	0.70
		y	21	21	
112 Elementary/Secondary Education - Other	1,414	x	175	477	1.03
		y	180	677	
Total	3,174				

Source: Annual Survey of Public Employment & Payroll, U.S. Census Bureau

The data contain 312 zeros in PTE2011 and 287 zeros in PTE2012. This situation is shown in Table 6 for our data.

Table 6 Observations Class Status by Year

		PTE2012		
		Zero Value	Positive Value	Total
PTE2011	Zero Value	152	160	312
	Positive Value	135	2,727	2,862
	Total	287	2,887	3,174

Source: Annual Survey of Public Employment & Payroll, U.S. Census Bureau

One set of missing data is created by randomly selecting and removing 913 observations from y, resulting in 29% missing overall. Missing at random (MAR) data, are created by randomly choosing observations from each function where the missing rate is proportional to the original data, shown in Table 4. In this data file, there are several available covariates: state (51 levels), function (5 levels), school level (7 levels), x (count variable).

Modeling issues

Some data features became modeling challenges:

a) Covariates

Covariate state has 51 levels. When all the covariates are used in the model, some state x function cells contain no observations or only very few observations. This creates sparse data problems.

b) Distribution

The minimum value of both variables is zero and their maximum values are about 6,500. The means are around 270. If these data had a Poisson distribution, there should have been at most a few zero values; however, there is a high percentage of zeros in the data. If all the zeros are taken out and we take a log transformation on the rest of the data, the data are still far from normally distributed. See the QQ plot in Figure 1.

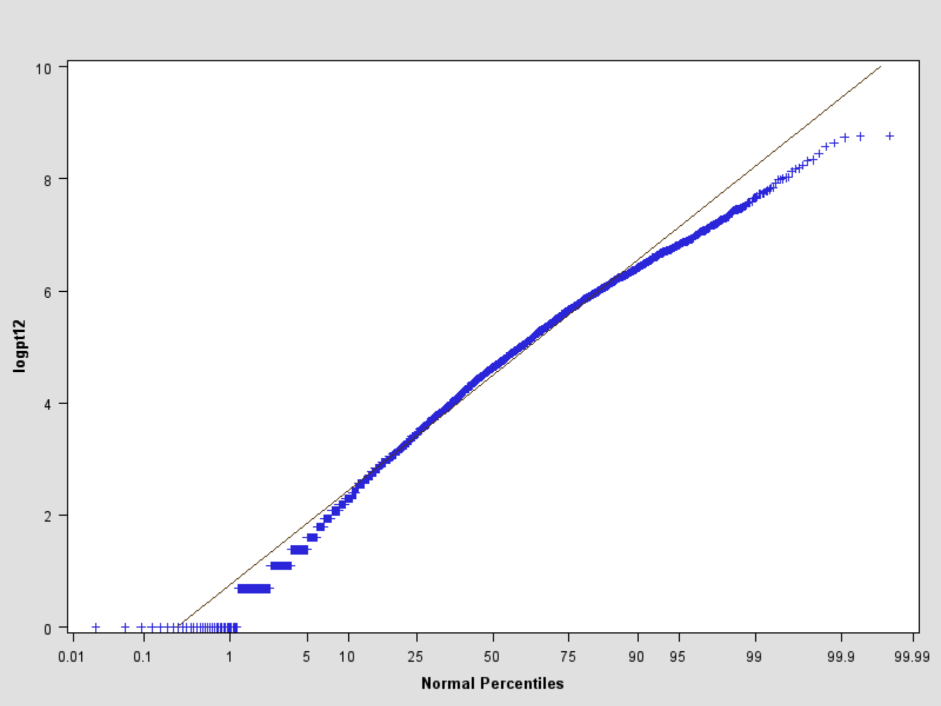

Figure 1 QQ Plot of Part-Time Employment Data in Year 2012
Source: Annual Survey of Public Employment & Payroll, U.S. Census Bureau

c) Overdispersion

From Table 5, we see that the ASPEP data set has an overdispersion problem. For example, the function 012 has sample mean 228 and standard deviation of 484. In a Poisson distribution, the mean should equal the variance.

d) Constant variance

From Figure 2, we see there is a linear relationship between Y and X . However, the variance of Y given X seems constant: the variance for small values is about the same as for large values. The model says that Var($Y|X$) should be small if X is small, large if X is large. That means the Poisson model does not describe this data well.

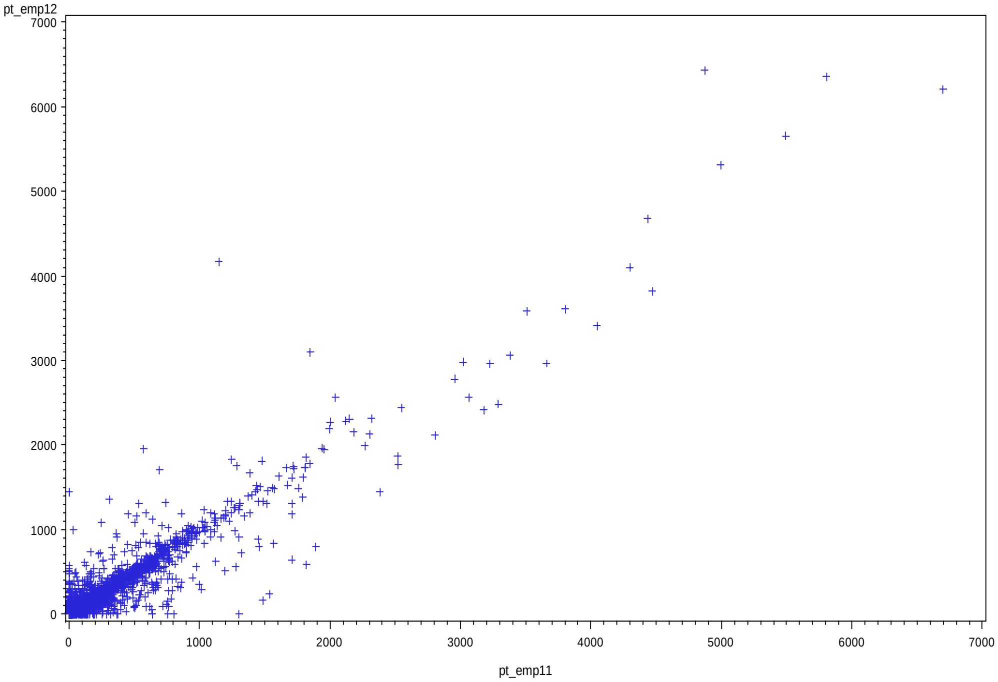

Figure 2 Scatter plot of part-time employment data in years 2011 and 2012
Source: Annual Survey of Public Employment & Payroll, U.S. Census Bureau

Results

The bivariate ZIP model is too simple to handle the complicated ASPEP PTE data. For model parameters, 4 slope estimates are outside of the 95% interval; 5th slope estimate has a wide interval. For imputed values, sum of squares of imputation errors SS = 22,859 and the proportion of true values covered by their 95% credible interval is only 52%. Clearly, the poor coverage rate of the bivariate ZIP model also suggests a lack of fit.

5 Conclusions and Future Research

We have created a bivariate ZIP model to impute year-to-year zero-inflated count data. One nice feature of this bivariate ZIP model is that it takes the Bayesian approach through MCMC and is able to provide credible intervals for imputed values. Traditional imputation methods are unable to do this.

In the simulation study, the model performed well in regression coefficient estimation; all parameter values are within their 95% credible interval. Similarly, the missing value imputation was successful; the true values fell in their 95% credible intervals in 96.8% of cases.

We applied our models to data on part-time employment in independent school districts, collected from the Annual Survey of Public Employment & Payroll (ASPEP). The ASPEP data has covariates with many levels, unknown distribution, overdispersion, and nearly constant variance. The bivariate ZIP model did not fit the ASPEP data. Only 52% of true values were covered by their 95% credible intervals.

In the future, an extension of the bivariate ZIP model will be created. We are developing a mixed effects bivariate negative binomial model, which will address the features of the ASPEP data mentioned in the previous paragraph.

Acknowledgement

The author wishes to thank Dr. Joseph Schafer for consultations on this research. The author also wishes to thank Terri Craig, Paul Villena, Andre Williams, and Franklin Winters for helpful comments and encouragement.

References

Gelman, A. and Hill, J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Model*, New York: Cambridge University Press.

Lambert, D. (1992). Zero-Inflated Poisson Regression Models with an Application to Defects in Manufacturing, *Technometrics*, **34**, 1-14.

Rubin, D, B. (1987). *Multiple Imputation for Nonresponse in Surveys*, New York: Wiley.

Rubin, D. B. (1996). Multiple Imputation After 18+ Years, *Journal of American Statistical Association*, **91,** 473-489.

Schafer, J, L. (1997). *Analysis of Incomplete Multivariate Data*, London: Chapman & Hall.

Schafer, J. L. and Graham, J. W. (2002). Missing Data: Our View of the State of the Art. *Psychological Methods*, **7**, 147-177.

Spiegelhalter, D. Thomas, A. Best, N. & Lunn, D. (2003) *WinBUGS User Manual, Version 1.4*. MRC Biostatistics Unit, Institute of Public Health and Department of Epidemiology and Public Health, Imperial College School of Medicine, UK available at http://www.mrc-bsu.cam.ac.uk/bugs.