AUSTRIAN JOURNAL OF STATISTICS
Volume 40 (2011), Number 1 & 2, 85-92

Models for Underreporting: A Bernoulli Sampling Approach for Reported Counts

Gerhard Neubauer[1], Gordana Djuraš[1] and Herwig Friedl[2]
[1]Joanneum Research, Graz Austria
[2]Graz University of Technology, Austria

Abstract: Underreporting in register systems can be analyzed using a binomial approach, where both the size and the probability parameter have to be estimated. Parameter estimation fails when overdispersion is present. Extensions of the binomial model are derived by randomizing the parameters, i.e. considering mixed models. Among these models are the beta-binomial, which results from allowing for a random reporting probability; the negative-binomial, that is the marginal when the size parameter is randomized; and the beta-Poisson model, where both binomial parameters are considered random. Likelihood based estimation is developed and inference issues are discussed. Finally the method is applied to data from the Austrian crime register.

Keywords: Deficient Register System, Mixed Distributions, Regression, Nonnested Testing, Crime Data.

1 Introduction

Underreporting is a problem in data collection that occurs, when the counting of some event is for some reason incomplete. Any reporting or counting system is prone to such errors in recording. The reasons may be quite different in the various fields of application like public health, criminology, actuarial science or production. In public health we have reporting systems for infectious diseases like HIV or chronic diseases like diabetes, and recording failures may occur as result of diagnostic errors or patients avoiding diagnosis. Crimes associated with shame are likely not to be reported to the police, just as theft of low value goods. The same holds for traffic accidents with minor damage. Insurances are faced with an unknown number of total claims as some claims are made with a delay, that may be as long as five years. An example from industrial production is the number of products that are broken within a certain period, typically the warranty period. To know this number is important for quality management. Only the number of returned products is known, but the total number includes also those goods that are not returned by customers. In all cases reporting systems give lower counts than the actual number of events. Therefore, underreporting is a widespread phenomenon and the estimation of the total number of cases is of particular interest.

As a consequence of underreporting μ, the mean of the observed counts is smaller than the true mean λ. Using a binomial model the mean of the observed counts is $\mu = \lambda\pi$, with both π, the reporting probability and λ, the total number of cases to be estimated.

Neubauer and Friedl (2006) addressed this problem by simultaneous estimation of both binomial parameters. They showed that a binomial and a beta-binomial regression model are suited for a wide range of applications. However, both models fail, if the sample variance of the observed counts is considerably larger than the sample mean. With the Poisson model this phenomenon is known as overdispersion.

Neubauer and Djuraš (2008, 2009) proposed models that allow for more overdispersion than the beta-binomial does. One is a generalized Poisson regression model, and the second is a beta-Poisson regression model.

In the following sections we give an overview of models for underreporting, introduce a regression technique for these models, discuss estimation and inference and finally present results from application to real data.

2 Models Based on a Bernoulli Sampling Scheme

Let y_t, $t = 1, \dots, T$, be a sample of counts of some register system reported over time. We start by assuming that for each time t there is the same unknown number λ of events that actually happened. The Bernoulli sampling model makes the assumption that for each event a random mechanism decides whether it is reported or not, i.e. there is some probability π for reporting an event. Hence we have random Bernoulli variables

$$R_i = \begin{cases} 1 & \text{if the event is reported} \\ 0 & \text{otherwise} \end{cases} \qquad i = 1, \dots, \lambda$$

such that

$$Y_t = \sum_{i=1}^{\lambda} R_i \sim \text{binomial}(\lambda, \pi)$$

and $\text{E}(Y_t) = \mu = \lambda\pi$ is the mean model and the mean-variance relation is characterized by $\text{var}(Y_t) = \mu(1-\pi) = \mu\phi$, $0 \le \phi \le 1$.

A more realistic model at hand results, if $\text{E}(Y_t) = \mu_t = \lambda_t\pi$ is allowed with $\lambda_t(\beta) = \exp(x_t'\beta)$ and $\pi(\alpha) = \exp(\alpha)/[1+\exp(\alpha)]$, $\alpha \in \mathbb{R}$. Here x_t is a d-vector of known regressors and β denotes the corresponding vector of unknown parameters. The likelihood contribution of the t-th observation is now

$$L(\alpha, \beta | y_t, x_t) = \binom{\lambda_t(\beta)}{y_t} \pi(\alpha)^{y_t} (1-\pi(\alpha))^{\lambda_t(\beta)-y_t} . \tag{1}$$

For real data $\text{var}(Y_t) \le \mu_t$ is often too restrictive and hence mixed models are considered as alternatives to the binomial model. Allowing for larger variability becomes possible by treating parameters as random variables. The counts now have a conditional binomial distribution.

2.1 Randomization of Parameter π

For $Y_t|P \sim \text{binomial}(\lambda, p)$ and $P \sim \text{beta}(\gamma, \delta)$ we obtain the beta-binomial as marginal distribution of Y_t, with $\text{var}(Y_t) = \mu(1-\pi)(\lambda+\gamma+\delta)/(1+\gamma+\delta) = \mu\phi$, $\phi > 0$. We use the reparametrization $\theta = \gamma + \delta$ and $\pi = \gamma/\theta$ and in the regression model we have

$\lambda_t(\beta) = \exp(x_t'\beta)$ and $\pi(\alpha) = \exp(\alpha)/[1 + \exp(\alpha)]$, as before. The profile likelihood contribution of the t-th observation is now

$$L(\alpha, \beta|y_t, x_t, \theta) = \binom{\lambda_t(\beta)}{y_t} \frac{\mathcal{B}(y_t + \pi(\alpha)\theta, \lambda_t(\beta) - y_t + (1 - \pi(\alpha))\theta)}{\mathcal{B}(\pi(\alpha)\theta, (1 - \pi(\alpha))\theta)}, \tag{2}$$

where $\mathcal{B}(\cdot)$ is the beta function and $\gamma(\alpha) = \pi(\alpha)\theta$ and $\delta(\alpha) = (1-\pi(\alpha))\theta$. The estimation algorithm cycles between ML estimation of α and β given θ, and the method of moments estimation of θ given α and β.

2.2 Randomization of Parameter λ

Assuming $Y_t|L \sim$ binomial(l, π) and $L \sim$ Poisson(λ) we obtain marginally that $Y_t \sim$ Poisson$(\lambda\pi)$ with var$(Y_t) = \mu$, i.e. $\phi = 1$. For this model the decomposition of the mean is in general not identified. Winkelmann (2000) considers a special regression model where $\lambda_t(\beta) = \exp(x_t'\beta)$, $\pi_t(\alpha) = \exp(z_t'\alpha)/[1+\exp(z_t'\alpha)]$ and x_t and z_t are disjoint sets of regressors. This so-called Pogit model is identified and the mean decomposition gives the desired results.

Allowing for randomness in λ we state a conditional Poisson model as $L_t|K \sim$ Poisson$(k\lambda_t)$. Using $K \sim$ Gamma(ω, ω) in addition, we obtain a negative-binomial marginal distribution for Y_t with parameters ω, the expected number of unreported cases, and π, the reporting probability. The mean-variance relation is now var$(Y_t) = \mu + \mu^2/\omega = \mu\phi$, $1 \leq \phi$. Now we use the regression model $\omega_t(\beta) = \exp(x_t'\beta)$ and $\pi(\alpha) = \exp(\alpha)/[1+\exp(\alpha)]$, and the likelihood contribution is

$$L(\alpha, \beta|y_t, x_t) = \binom{\omega_t(\beta) + y_t - 1}{y_t} \pi(\alpha)^{y_t} (1 - \pi(\alpha))^{\omega_t(\beta)}. \tag{3}$$

Several distributions have the Poisson as special case, and one of them is the generalized Poisson (gP) distribution (Consul, 1989), denoted as $Y \sim$ gP(θ, τ). Neubauer, Djuraš, and Friedl (2009) showed in simulations that for a wide range of parameters the gP is equivalent to the binomial and to the negative-binomial distribution. The parameter τ tunes the type of distribution. For $\tau = 0$ we get the Poisson distribution with mean θ, for $0 < \tau \leq 1$ we have a equivalent negative-binomial and for $\tau < 0$ a equivalent (positive) binomial distribution. In the following we suggest an interpretation of the gP parameters that allows to use the gP distribution for the estimation of underreporting. The first two moments of the gP distribution are given as E$(Y) = \theta(1 - \tau)^{-1}$ and var$(Y) = \theta(1 - \tau)^{-3}$. For the positive and negative-binomial distribution we have E$(Y) = \lambda\pi$ and var$(Y) = \lambda\pi(1 - \pi)$ or var$(Y) = \lambda\pi(1 - \pi)^{-1}$. Equating the respective moments of the distributions and solving for λ and π we obtain $\pi = 1 - (1 - \tau)^{2s}$ and $\lambda = \theta\pi^{-1}(1 - \pi)^{-s/2}$, where $s = $ sign(τ).

For regression we use $\theta_t(\beta) = \exp(x_t'\beta)$ and $\tau(\alpha) = 1 - \exp(-\alpha)$ to ensure $\mu_t = \theta_t(1 - \tau)^{-1} > 0$ and the likelihood contribution is given as

$$L(\alpha, \beta|y_t, x_t) = \frac{\theta_t(\beta)[\theta_t(\beta) + y\tau(\alpha)]^{y-1} \exp[-(\theta_t(\beta) + y\tau(\alpha))]}{y!}. \tag{4}$$

The parameter α is a real number that indicates Poisson overdispersion for $0 < \alpha$ and Poisson underdispersion $\alpha < 0$. Testing $\alpha = 0$ is therefore a possibility to identify near Poisson data, or in other words to test for Poisson over- or underdispersion.

2.3 Randomization of Both Binomial Parameters

Another possibility of randomizing the Poisson model is $Y_t|P \sim \text{Poisson}(\lambda P)$ and $P \sim \text{beta}(\gamma, \delta)$ which gives the marginal beta-Poisson distribution, that we write as $Y_t \sim \text{beta-Poisson}(\lambda, \gamma, \delta)$. For this model we have $\text{E}(Y_t) = \mu = \lambda\pi$, $\text{var}(Y_t) = \mu\phi$ with $\pi = \gamma/(\gamma + \delta)$ and $1 \leq \phi = 1 + \lambda(1 - \pi)/(1 + \gamma + \delta)$. The beta-Poisson distribution is also known as "Type H_1" distribution, and is usually treated in the context of contagious distributions (Johnson, Kemp, and Kotz, 2005).

As for the beta-binomial model we use the reparametrization $\theta = \gamma + \delta$ and $\pi = \gamma/\theta$, and in the regression model we have $\lambda_t(\beta) = \exp(x_t'\beta)$ and $\pi(\alpha) = \exp(\alpha)/[1+\exp(\alpha)]$, as before. The likelihood contribution of the t-th observation is now

$$L(\alpha, \beta|y_t, x_t, \theta) = \frac{\lambda_t(\beta)^{y_t}}{y_t!} \frac{\mathcal{B}(y_t + \pi(\alpha)\theta, (1 - \pi(\alpha))\theta)}{\mathcal{B}(\pi(\alpha)\theta, (1 - \pi(\alpha))\theta)} {}_1F_1[y_t + \pi(\alpha)\theta; y_t + \theta; -\lambda_t(\beta)] , \tag{5}$$

where ${}_1F_1[\cdot]$ denotes the confluent hypergeometric function.

3 Estimation and Inference

For all models maximum likelihood (ML) is used for estimating the parameters α and β in the mean $\mu_t = \lambda_t(\beta)\pi(\alpha)$. For the beta-Poisson model ML is applied to an approximated likelihood in which ${}_1F_1$ is replaced by a Laplace approximation based on its integral representation. The roots of the score equations $\partial\ell/\partial\alpha = 0$ and $\partial\ell/\partial\beta = 0$ are found by the Newton-Raphson algorithm, where $\ell = \log \prod_t L(\Omega_M|y_t, x_t)$ denotes the log-likelihood, and Ω_M is the parameter vector of some model M. For the beta-mixture models the estimation algorithm cycles between ML estimation of α and β given θ, and the method of moments estimation of θ given α and β.

It is known that asymptotic normality of the ML estimates holds for members of the one-parameter linear exponential family. Even our simple binomial model is not a member of this family. Hence it is theoretically unclear if the desirable property of normality for parameter estimates holds for our models. From different simulation studies we have empirical evidence that the parameter estimates $\hat{\alpha}$ and $\hat{\beta}$ are approximately normally distributed, if the reporting probability is not too small. This is a reasonable finding, as one of the regularity conditions for asymptotic normality of ML estimates is the identifiability of the parameters. For $\pi \to 0$ we approach the Poisson limit in all models and for the Poisson model the decomposition $\mu = \lambda\pi$ is not identified. Hence we assume asymptotic normality for the parameter estimates and obtain pointwise confidence intervals for the derived quantities $\hat{\lambda}_t$ and $\hat{\pi}$ by the Delta Method.

For variable selection within one model the usual criteria like the t-statistic are available and of course the likelihood ratio test (LRT) principle can be used to discriminate between nested models, i.e. for models of the same distribution.

To support the decision between our models the usual likelihood approaches are not applicable and therefore non-nested testing techniques must be applied. The classical Cox test for non-nested models addresses no more than two alternatives. A simple strategy for more than two model alternatives is given by Allcroft and Glasbey (2003). It is a simulation-based approach and its main advantage lies in the fact that it does not need the estimated parameters from the simulated data. Hence, this inferential procedure is very fast and also easy to implement. To compare a set of models $\mathbf{M} = (M_k)$, $k = 1, \ldots, K$, the procedure consists of the following steps:

1. For observed data $\mathbf{y} = (y_t)$, $t = 1, \ldots, T$, estimate parameters under all models by optimizing a goodness-of-fit criterion. Denote the estimates by $\hat{\theta}_k$ and the observed values of the criterion as $\mathbf{c} = (c_k(\mathbf{y}))$.
2. Simulate a number of data from each estimated model, denoted by $\mathbf{y}^{(s)}(\hat{\theta}_k)$, $s = 1, \ldots, S$, (e.g. $S = 100$) and obtain all criteria values for each data set; thus we have S matrices $\mathbf{C}^{(s)}$ of dimension $K \times K$.
3. Obtain $\bar{\mathbf{C}}$, the mean of the S matrices, and compare $\mathbf{c}$ to the k-th column $\bar{\mathbf{c}}_k$ by assuming multivariate normality for the mean column. A measure for the difference between the observed and simulation based vector of criteria is the Mahalanobis distance $D_k = (\mathbf{c} - \bar{\mathbf{c}}_k)' V_k^{-1} (\mathbf{c} - \bar{\mathbf{c}}_k)$, where V_k is the sample variance matrix of the criteria. If the k-th model is correct then $D_k \sim \chi^2_K$.

An even simpler but only heuristic strategy to choose between non-nested models is to use the Bayesian Information Criterion (BIC), as for instance Burnham and Anderson (1998) recommend.

4 Application to Austrian Crime Data

The real data examples are taken from the Austrian online crime register SIMO. For each of 132 regions in Austria we have weekly counts of different crime categories since 2004. The models introduced above were applied to data of larger regions and crime categories bicycle theft and shop lifting.

The regression element used to model the mean consists of three components: T_t a smooth trend function, S_t a seasonality function and C_t a component for calender effects. Hence λ_t or ω_t is modelled as $\exp(\beta_0 + T_t + S_t + C_t)$. Tables 1 and 2 give the values of the log-likelihood, the Pearson statistic, the BIC, the Mahalanobis distance D, its p-value and the estimate $\hat{\pi}$. The estimation algorithm does not converge for the binomial model and thus no results are available. All other models converge and show estimates for ϕ that are larger than 1. Therefore, the binomial model is not adequate for the data.

Considering the BIC for the bicycle theft data as criterion for model selection, we find the gP distribution — or equivalently the negative-binomial distribution — appropriate for the data, and here the reporting probability is estimated as $\hat{\pi} = 0.62$. This is supported by the non-nested test, where only for the negative-binomial model the test statistic D is not significant. Figure 1 shows the bicycle theft data and the estimation results from the negative-binomial model, the estimated mean and the estimated total number with its confidence intervals. The shape of the functions is dominated by the seasonality term, which shows an increase of bicycle thefts during the warmer period of the year. The abrupt changes in the level of the functions are due to calender effects expressed as dummy variables. One of them covers the summer holiday period and the other is an additional effect for the winter season.

Table 1: Results from the Bicycle Theft Data

Distribution	$\log L$	Pearson	BIC	D	$p(D)$	$\hat{\pi}$
Negative-binomial	-728.91	206.49	1521.63	1.92	0.59	0.61
Generalized Poisson	-728.57	204.79	1520.97	—	—	0.62
Beta-binomial	-732.87	204.79	1529.57	9.59	0.02	0.32
Beta-Poisson	-735.39	197.97	1534.60	9.85	0.02	0.63

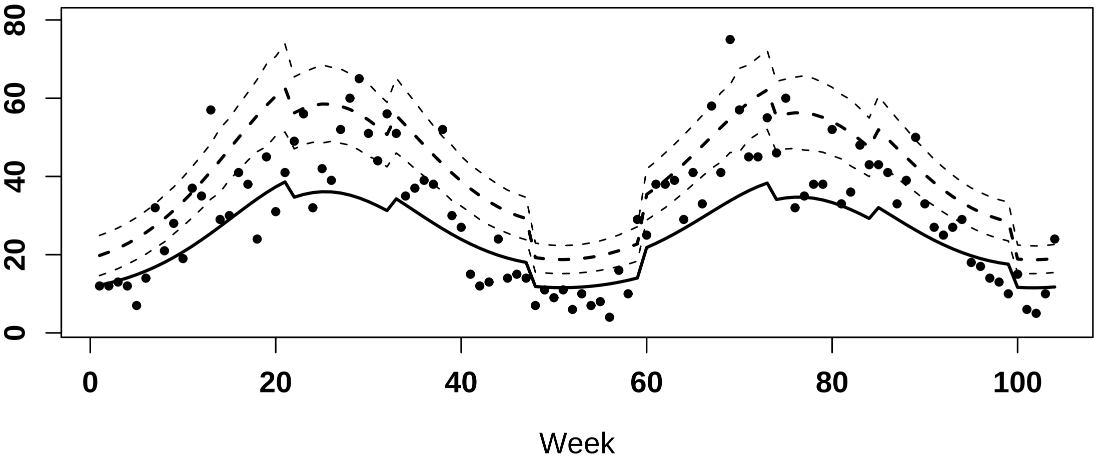

Figure 1: Results from the negative-binomial model: Bicycle theft data (points), estimated mean (solid), estimated total number of thefts with confidence interval (dashed).

Using the BIC for model selection with the shop lifting data we choose the negative-binomial model with an estimated reporting probability of $\hat{\pi} = 0.63$. Considering the test statistic D we find all models equally appropriate for the data. This is quite disappointing, as we do not know which of the estimated reporting probabilities to consider as appropriate for the data situation. In Figure 2 the estimates $\hat{\lambda}_t$ are plotted for all models as functions over the data. The functions are here also dominated by a seasonality term, with an increase during the colder period of the year.

Table 2: Results from the Shop Lifting Data

Distribution	$\log L$	Pearson	BIC	D	$p(D)$	$\hat{\pi}$
Negative-binomial	-802.31	202.70	1657.79	2.04	0.56	0.63
Generalized Poisson	-802.44	202.57	1658.06	—	—	0.64
Beta-binomial	-803.19	194.13	1659.56	4.71	0.19	0.42
Beta-Poisson	-807.19	193.02	1667.57	6.54	0.09	0.74

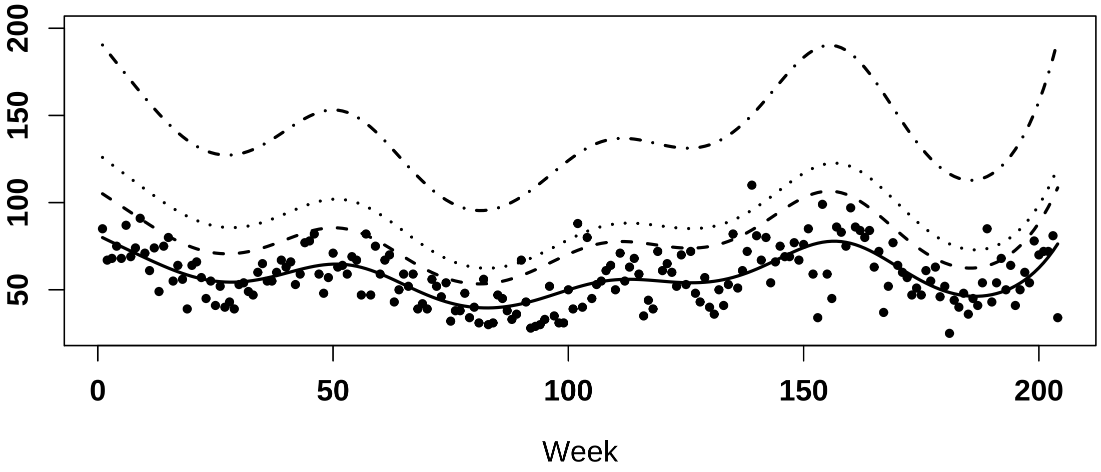

Figure 2: Shop lifting data: Estimated mean (solid) and estimated total number of crimes from the negative-binomial (dotted), the beta-binomial (dashed) and the beta-Poisson model (dash-dotted).

5 Conclusion

Reporting systems often produce counts of some event. These counts are incomplete if the system is deficient for some reason. The most prominent example are crime data, where underreporting is prevailing for various crime categories. We propose a method based on a Bernoulli sampling scheme that brings the binomial distribution as most simple model for the estimation of underreporting. Several generalizations are proposed that allow the estimation of the binomial parameters, when data show large overdispersion. The estimation relies on maximum likelihood and hence the usual inferential tools like LRT or BIC are available. The proposed method and estimation technique shows good performance in a simulation studies when $\mathrm{var}(Y) = \mu\phi$ and $\phi \neq 1$, and it is also reasonable to assume asymptotic normality of parameter estimates. Finally the method is applied to two examples of Austrian crime data, bicycle theft and shop lifting. For the bicycle data a generalized Poisson model is found to fit the data. For the shop lifting data the results are not conclusive. Therefore we use an inferential procedure to decide between models for underreporting. Usual likelihood approaches are not applicable and therefore a testing technique to distinguish between non-nested models is applied.

References

Allcroft, D. J., and Glasbey, C. A. (2003). A simulation-based method for model evaluation. *Statistical Modelling*, *3*, 1-14.

Burnham, K. P., and Anderson, D. R. (1998). *Model Selection and Inference: A Practical Information-Theoretic Approach.* New York: Springer.

Consul, P. C. (1989). *Generalized Poisson Distributions. Properties and Applications.* New York: Marcel Dekker.

Johnson, N. L., Kemp, A. W., and Kotz, S. (2005). *Univariate Discrete Distributions.* Hoboken: Wiley.

Neubauer, G., and Djuraš, G. (2008). A generalized Poisson model for underreporting. In P. Eilers (Ed.), *Proceedings of the 23rd International Workshop on Statistical Modelling. Utrecht, 7-11 July 2008* (p. 368-373).

Neubauer, G., and Djuraš, G. (2009). A beta-Poisson model for underreporting. In J. Booth (Ed.), *Proceedings of the 24th International Workshop on Statistical Modelling. Ithaca, NY, 20-24 July 2009* (p. 255-260).

Neubauer, G., Djuraš, G., and Friedl, H. (2009). *Maximum Likelihood for Size-Estimation: Some Results on Properties and Limitations.* (Tech. Rep. No. 4). Graz: Joanneum Research.

Neubauer, G., and Friedl, H. (2006). Modelling sample sizes of frequencies. In J. Hinde, J. Einbeck, and J. Newell (Eds.), *Proceedings of the 21st International Workshop on Statistical Modelling. Galway, Ireland, 3-7 July 2006* (p. 401-408).

Winkelmann, R. (2000). *Econometric Analysis of Count Data.* Berlin: Springer.

Authors' addresses:

Gerhard Neubauer and Gordana Djuraš
JOANNEUM RESEARCH
POLICIES - Centre for Economic and Innovation Research
Statistical Applications
Leonhardstraße 59
A-8010 Graz
Austria
E-mails: Gerhard.Neubauer@Joanneum.at and Gordana.Djuras@Joanneum.at
URL: www.joanneum.at/policies/sta

Herwig Friedl
Institute of Statistics
Graz University of Technology
Münzgrabenstraße 11/III
8010 Graz
Austria
E-mail: HFriedl@TUGraz.at
URL: www.statistics.tugraz.at