Published in final edited form as:

Model Assist Stat Appl. 2013 ; 8(2): 143–150. doi:10.3233/MAS-130263.

A Bayesian model for censored positive count data in evaluating breast cancer progression

Hung-Wen Yeh^{1,2,†}, Yu Jiang¹, Lili Garrard^{1,3}, Yang Lei¹, and Byron Gajewski^{1,3}

¹Department of Biostatistics, The University of Kansas Medical Center, Kansas City, Kansas 66160

²The University of Kansas Cancer Center, Kansas City, Kansas 66160

³School of Nursing, The University of Kansas Medical Center, Kansas City, KS 66160

Abstract

Basic science researchers transplant human cancer tissues from patients with ductal carcinoma *in situ* (DCIS) to animals and observe the progression of the disease. Successful transplants show invasion of human tissues across mammary ducts in animal fat pads and cause DCIS-like lesions in one or more ducts. In this work, we consider data from a recent publication of breast cancer research where positive counts of affected ducts may be subject to censoring. We fit the data with zero-truncated Poisson (ZTP) models with an informative prior of gamma. Due to the zero-truncation and right censoring, posterior distributions may not be conventional gamma and are estimated through Markov chain Monte Carlo and grid approximation. For each of the two cell lines, we fit a model with group-specific parameters for DCIS subtypes classified by the cell surface biomarkers, and another model with a homogeneous parameter across groups. Models are compared by the Deviance Information Criterion (DIC). For the chosen prior parameter values, Bayes estimates are comparative to the maximum likelihood estimates, and the DIC favors the simpler model in both cell lines.

Keywords

count data; zero-truncated Poisson; right censoring; Markov Chain Monte Carlo; grid approximation; Deviance Information Criterion

1. Introduction

Development of breast cancer is a multi-stage process involving progression from non-invasive to invasive lesions, but not all patients with non-invasive lesions progress to invasive lesions. To understand the progression mechanism and heterogeneous progression potential, researchers often rely on animal models. For instance, Behbod et al. [1] investigates the progression from ductal carcinoma *in situ* (DCIS), a common form of non-invasive breast cancer. They obtain two DCIS cell lines (DCIS.COM and SUM-225), culture and classify the cells into several subtypes according to cell surface biomarkers, inject each subtype of cells to mammary glands of mice, and observe the disease development. However, not all human cells grow in animals, but once grown, the injected human cancer cells invade neighboring mammary ducts. A larger count of affected ducts indicates more aggressive progression. Assuming that zero counts are due to lack of progenitor cells and

[†]Corresponding author hyeh@kumc.edu 5028 K Robinson, Mail Stop 1026, 3901 Rainbow Blvd., Kansas City, KS 66160 Tel: 1-913-588-0531 Fax: 1-913-588-0252 .

that positive counts are due to invasion of injected human tissues, Yeh et al. [7] evaluates the means of Poisson distributions truncated at zero (we refer to “truncated mean” thereafter) for positive counts of affected ducts, and makes comparisons across subtypes within each cell line using the zero-truncated Poisson (ZTP) model allowing right-censored observations due to some censored observations (investigators recorded counts beyond 7 as “7+”). The analysis uses frequentist approach, and encounters some computational issues in simulation due to small sample sizes and censoring (e.g. simulation runs that generate all censored observations in any group need to be excluded, and bimodal distributions in some parameters).

In the present work, the data are analyzed by Bayesian modeling. Gelman et al. discusses general principles of modeling censored and truncated data [2]. Although a lot of literature on censored and/or truncated count data lies in the frequentist framework, literature on Bayesian modeling emerges in recent years. Howlader and Balasooriya [3] focus on count data and develop closed forms for Bayes estimates of the Poisson parameter in standard Poisson and ZTP models using informative gamma priors. Scollnik [5] applies Markov chain Monte Carlo (MCMC) to model positive count data following Poisson distributions. Vergne et al. [6] considers over-dispersion and heterogeneity due to covariates, links covariates to the parameters of Poisson and negative binomial models through log-linear models, and makes inference with MCMC. Puza et al. [4] further treats covariates as random and incorporates the covariate parameters with Poisson and regression parameters in Bayesian modeling. Among the aforementioned literature, the class of zero count is not observable (and thus truncated) so inference often extends to estimating the size of zero counts or the “total” sample size (the unobserved zero counts plus the number of positive counts).

The DCIS study in the present work differs from the published literature in at least two aspects. First, the “zero counts” are observed (see Table 1 in [1]), but they are caused by a different process (lack of progenitor cells) and need to be separated from positive counts (caused by the progression of DCIS tissues). In other words, inference on the size of zero counts is not relevant. Secondly, larger counts are censored. As a result, both zero-truncation and right censoring are considered in the likelihood, and the conjugate nature of gamma distributions may not be guaranteed. Furthermore, although the heterogeneity in progression potential across DCIS subtypes is incorporated into models through group-specific Poisson parameters, progression potential is evaluated by the truncated means rather than the Poisson parameters. The goal of this work is to examine and solve at least some computational issues seen in the aforementioned frequentist approach.

2. Methods

2.1 The Model and the posterior distribution

Let Y_{ij} be a random variable for the number of ducts in the i^{th} mouse fat pad affected by successful transplants of the j^{th} subtype of human cells, $i = 1, 2, \dots, n_j$ and $j = 1, 2, \dots, p + 1$. If we assume that observations in group j are generated by a Poisson distribution with a group-specific parameter λ_j , the probability mass function is

$$Pr(Y_{ij}=y_{ij}|Y_{ij}>0) = \frac{\lambda_j^{y_{ij}}}{(e^{\lambda_j} - 1)y_{ij}!}, \quad y_{ij}=1, 2, 3, \dots$$

which gives a truncated mean $\mu_j = \lambda_j / (1 - e^{-\lambda_j})$. Note the non-linear, monotonic relationship between λ_j and μ_j , and that $\lim_{\lambda_j \rightarrow 0} \mu_j = 1$ and $\lim_{\lambda_j \rightarrow \infty} \mu_j = \lambda_j$. Consider that the observations may be censored at the threshold h , the likelihood function is given as

$$p(\mathbf{y}|\boldsymbol{\lambda}) = \prod_{j=1}^{p+1} \prod_{i=1}^{n_j} \left[\frac{\lambda_j^{y_{ij}}}{y_{ij}! (e^{\lambda_j} - 1)} \right]^{1-c_{ij}} \left[1 - \sum_{y_{ij}=1}^{h-1} \frac{\lambda_j^{y_{ij}}}{y_{ij}! (e^{\lambda_j} - 1)} \right]^{c_{ij}} \quad (1)$$

where $c_{ij} = 1$ if the observation y_{ij} is censored (greater than or equal to the threshold h) and 0 otherwise.

In Bayesian modeling, we apply a Gamma prior for $\boldsymbol{\lambda}$ [3] with prior parameters k and θ , then the posterior distribution can be expressed as

$$f(\boldsymbol{\lambda}|\mathbf{y}, \mathbf{c}, k, \theta) \propto \prod_{j=1}^{p+1} \prod_{i=1}^{n_j} \left(\frac{1}{\Gamma(k)\theta^k} \lambda_j^{k-1} e^{-\lambda_j/\theta} \right) \left[\frac{\lambda_j^{y_{ij}}}{y_{ij}! (e^{\lambda_j} - 1)} \right]^{1-c_{ij}} \left[1 - \sum_{y_{ij}=1}^{h-1} \frac{\lambda_j^{y_{ij}}}{y_{ij}! (e^{\lambda_j} - 1)} \right]^{c_{ij}} \quad (2)$$

Often, the Poisson parameters can be linked to covariates which may contribute heterogeneity through a log-linear model. In this study, the covariates are the dummy variables indicating group membership, so we define beta regression parameters as

$$\beta_0 = \log \lambda_{p+1}; \quad \beta_j = \log \lambda_j - \log \lambda_{p+1} \quad (3)$$

Because of the one-to-one relationship between λ_j and μ_j , $\beta_j = 0$ implies equal means between group j and the reference group $p+1$, that is, $\lambda_j = \lambda_{p+1}$ and $\mu_j = \mu_{p+1}$.

2.2 Estimation and model selection

Due to the truncation and censoring, the posterior distribution may not be a Gamma and is normalized by the grid approximation. Specifically, 20,000 equally spaced grid points are used to partition the un-normalized posterior distribution in the range of 0 to 20. The un-normalized posterior density is evaluated at each grid point except for the first 99 points which have “density” values $< 10^{-65}$ (they are indicated as “NA” in R) and summed over all grids to estimate the un-normalized area. The normalized posterior distribution is then obtained by dividing the un-normalized posterior distribution by the estimated un-normalized area. Ten thousand samples are then drawn from the normalized posterior distribution, with the first 1,000 samples discarded as the burn-in period. Finally, the posterior distributions of the truncated means are formed through the non-linear, monotone relationship with the Poisson parameters $\mu_j = \lambda_j / (1 - e^{-\lambda_j})$. Because positive counts (1) were censored at the threshold 7, the value $(1+7)/2 = 4$ is chosen as the prior mean. However, little is known about the prior variance so we aim at a vague prior with large variance. Thus, we use prior parameter values of $k = 1/4$ and $\theta = 16$ which give $E(\lambda_j) = 4$ and $\text{Var}(\lambda_j) = 64$ for all groups in both cell lines in this study.

When the log-linear model is used, the regression coefficients are often the parameters of interest and their posterior distributions are usually modeled with normal distributions due to their conjugate priors [6]. In the current work, we consider small λ 's in some groups and relax from the normal priors. Instead, posterior distributions of the beta coefficients are obtained from those of the Poisson parameters using Equation (3).

To examine whether heterogeneity exists in DCIS progression potential across subtypes, we also fit a null model by assuming an equal λ parameter for all groups from a given cell line.

The null models are then compared to the alternative models that allow group-specific λ parameters by the Deviance Information Criterion (DIC). Here, the deviance is $D(\lambda) = -2\log(p(y|\lambda))$ where $p(y|\lambda)$ is from Formula (1), and $DIC = p_D + \bar{D}$ where $\bar{D} = E_\lambda [D(\lambda)]$ is approximated by $D(\lambda)$ evaluated at 9,000 values of λ , and $p_D = \bar{D} - D(\bar{\lambda})$ where $\bar{\lambda}$ is the average of the 9,000 values of λ .

The analysis is performed in R version 2.14.1 for each cell line.

3. Results

In the DCIS data, 4.7% (2/43) and 40% (4/10) observations are censored for the DCIS.COM and the SUM-225 cell lines, respectively. Table 1 shows the Bayes estimates of the truncated means (μ) under the alternative and the null models for both cell lines. As a contrast, the maximum likelihood estimates (MLE) from [7] are also included. With the specified prior parameter values, the posterior means (3.13, 2.27, 2.04, 3.77, 1.44, 3.60, 3.73, 1.73, and 2.61 ducts per fat pads for DCIS.COM; 6.75 and 5.90 for SUM-225 subtypes) are nearly identical to the MLE (3.11, 2.25, 2.00, 3.75, 1.33, 3.60, 3.71, 1.67, and 2.59 for DCIS.COM; 6.75 and 5.90 for SUM-225 subtypes), and so as the posterior standard deviations and the standard errors of MLE. The effective numbers of parameters p_D are close to the number of groups ($p_D = 8.55$ vs. 0.98 for alternative and null models fitted to DCIS.COM data; $p_D = 1.98$ vs. 1.01 for SUM-25), and the DIC suggests no obvious heterogeneity in truncated means across groups for both cell lines (DIC = 165.0 vs. 164.0 for DCIS.COM and DIC = 41.6 vs. 39.9 for SUM-225). Figure 1 and the left panel of Figure 3 demonstrate the posterior distributions of truncated means for the two cell lines. The label of “mu. j ” refers to the truncated mean of group j .

The Bayes and ML estimates of beta coefficients in log-linear models are given in Table 2. Note that a zero value indicates no difference in Poisson parameters from the reference group. The Bayes and ML estimates are in general similar but deviate in some coefficients. Posterior distributions of these coefficient parameters are pictorially displayed in Figures 2 (DCIS.COM) and 3 (SUM-225, right panel), which suggest the beta coefficients may not follow normal distributions. The most skewed distributions are coefficients for the groups with the smallest truncated means (posterior means 1.44 and 1.73 for groups 5 and 8, respectively, from DCIS.COM). This finding confirms that using normal priors for beta coefficients may not be proper.

4. Discussion

For positive count data modeled by a zero-truncated Poisson distribution, the truncated mean and the Poisson parameter have a non-linear monotone relationship and they converge to each other as their values increase. As shown in [7], the parameters of interests in evaluation of progression potential are the truncated means rather than the Poisson parameters.

In this work, we consider the small sample sizes of the positive count data with right censoring, and perform Bayesian ZTP modeling using an informative prior. In contrast to the frequentist approach that results in bimodal distributions in some beta parameters (Figure 1(b) in [7]) which cause problems in inference, the posterior distributions are uni-modal for all parameters in both cell lines. In other words, inference can only be made on the means in frequentist CZTP model but not on the beta parameters, but inference in the Bayesian model is valid for both means and betas. The pre-selected prior parameters $k = 1/4$ and $\theta = 16$ imply $E(\lambda) = 4$ and $\text{Var}(\lambda) = 64$. Because the posterior mean is the weighted average of the prior mean and MLE, the posterior means of the truncated mean parameters

(μ_j) are expected to be greater than the MLE (all less than 4) in DCIS.COM and less than the MLE in SUM-225 (both greater than 4), but the differences are small or indiscernible due to the vague prior. The DIC favors the simpler (null) model for both cell lines, which is in agreement with the likelihood ratio test for MLE (LRT statistic $G^2(8) = 1.42$, $p = 0.077$ and $G^2(1) = 0.2$, $p = 0.65$ for DCIS.COM and SUM-225, respectively [7]). However, different models may lead to different conclusion. For example, if the DCIS.COM groups are collapsed into 5 groups (MUC-1^{med}, MUC-1^{lo}, CD24^{med}, CD24^{lo}, and random sort), the DIC favors this collapsed model (DIC = 158.39, see Tables 1 and 2 for Bayes estimates, and Figures 4 and 5 for posterior distributions).

The choice of k and θ values is arbitrary as long as they lead to a large variance such that the prior distribution is flat (i.e. a vague prior). Other prior parameter values with the same mean but different large variance values also lead to similar results (not shown).

Acknowledgments

HY was partly supported by the University of Kansas Cancer Center. HY and BJG were supported by a CTSA grant from NCRR and NCATS awarded to the University of Kansas Medical Center for Frontiers: The Heartland Institute for Clinical and Translational Research # UL1TR000001 (formerly #UL1RR033179). The contents are solely the responsibility of the authors and do not necessarily represent the official views of the NIH, NCRR, or NCATS.

References

- [1]. Behbod F, Kittrell FS, LaMarca H, Edwards D, Kerbawy S, Heestand JC, Young E, Mukhopadhyay P, Yeh HW, Allred DC, Hu M, Polyak K, Rosen JM, Medina D. An intra-ductal human-in-mouse transplantation model mimics the subtypes of ductal carcinoma in situ. *Breast Cancer Research*. 2009; 11:R66. DOI: 10.1186/bcr1847. [PubMed: 19735549]
- [2]. Gelman, A.; Carlin, JB.; Stern, HS.; Rubin, DB. *Bayesian Data Analysis*. 2nd Edition. Chapman and Hall/CRC; Florida: 2003. p. 231-236.
- [3]. Balasooriya, H.A. Howlader and U. Bayesian estimation of the distribution function of the Poisson model. *Biometrical Journal*. 2003; 45:901–911.
- [4]. Puza BD, Johnson HL, O'Neil TJ, Barry SC. Bayesian truncated Poisson regression with application to Dutch illegal immigrant data. *Communications in Statistics – Simulation and Computation*. 2008; 37:1565–1577.
- [5]. Scollnik DPM. Inference concerning the size of the zero class from an incomplete Poisson sample. *Communications in Statistics – Theory and Methods*. 1997; 26:221–236.
- [6]. Vergne T, Calavas D, Cazeau G, Durand B, Dufour B, Grosbois V. A Bayesian zero-truncated approach for analysing capture–recapture count data from classical scrapie surveillance in France. *Preventive Veterinary Medicine*. 2012; 105:127–135. [PubMed: 22421503]
- [7]. Yeh HW, Gajewski BJ, Mukhopadhyay P, Behbod F. The zero-truncated Poisson with right censoring: an application to translational breast cancer research. *Statistics in Biopharmaceutical Research*. (in press).

Figure 1.
Posterior distributions of truncated means of infected mammary ducts for mice injected with the subtypes of DCIS.COM cells

Figure 2.
Posterior distributions of beta coefficients associated with Poisson parameters for mice injected with the subtypes of DCIS.COM cells

Figure 3. Posterior distributions of truncated means (MU) of infected mammary ducts and beta coefficients (bhat) associated with Poisson parameters for mice injected with the subtypes of SUM-225 cells

Figure 4. Posterior distributions of truncated means of infected mammary ducts for collapsed groups of mice injected with the subtypes of DCIS.COM cells

Figure 5.
Posterior distributions of beta coefficients associated with Poisson parameters for collapsed groups of mice injected with the subtypes of DCIS.COM cells

Table 1

Bayes estimates and MLE for truncated means of mammary ducts with DCIS-like lesions

Cell line	Group	Bayes estimates								MLE	SE
		2.5%	25%	50%	75%	97.5%	Mean	SD			
DCIS.COM	Alternative model ($p_D = 8.55$; DIC = 165.0)										
	1. CD49 ^{hi} /MUC-1 ^{med}	1.84	2.53	3.00	3.64	5.01	3.13	0.83	3.11	0.83	
	2. CD49 ^{hi} /MUC-1 ^{lo}	1.38	1.82	2.17	2.61	3.77	2.27	0.62	2.25	0.61	
	3. CD49 ^{med} /MUC-1 ^{lo}	1.27	1.63	1.94	2.32	3.37	2.04	0.55	2.00	0.54	
	4. CD49 ^{med} /MUC-1 ^{med}	2.27	3.11	3.66	4.32	5.80	3.77	0.91	3.75	0.92	
	5. CD49 ^{hi} /CD24 ^{med}	1.06	1.17	1.32	1.58	2.46	1.44	0.38	1.33	0.35	
	6. CD49 ^{hi} /CD24 ^{lo}	2.29	3.03	3.52	4.09	5.35	3.60	0.79	3.60	0.80	
	7. CD44 ^{hi} /CD24 ^{lo}	2.53	3.24	3.67	4.15	5.25	3.73	0.69	3.71	0.69	
	8. CD44 ^{hi} /CD24 ^{med}	1.10	1.35	1.59	1.97	3.12	1.73	0.53	1.67	0.51	
9. Random sort $\hat{\tau}$	1.82	2.27	2.56	2.90	3.65	2.61	0.47	2.59	0.47		
Collapsed model ($p_D = 4.94$; DIC = 158.39)											
DCIS.COM	1. MUC-1 ^{med}	2.38	3.00	3.40	3.83	4.77	3.44	0.62	3.44	0.62	
	2. MUC-1 ^{lo}	1.50	1.84	2.09	2.38	3.08	2.14	0.41	2.13	0.41	
	3. CD24 ^{med}	1.12	1.31	1.47	1.69	2.32	1.54	0.32	1.50	0.31	
	4. CD24 ^{lo}	2.74	3.30	3.62	3.99	4.79	3.66	0.52	3.67	0.52	
	9. Random sort $\hat{\tau}^{\#}$	1.82	2.27	2.56	2.90	3.65	2.61	0.47	2.59	0.47	
	Null model ($p_D = 0.98$; DIC = 164.0)										
	All groups combined	2.40	2.65	2.80	2.96	3.27	2.81	0.22	2.81	0.23	
	Alternative model ($p_D = 1.98$; DIC = 41.6)										
	1. CD49 ^{med} /CD24 ^{hi}	4.26	5.74	6.65	7.64	9.90	6.75	1.44	6.75	1.44	
SUM-225	2. CD49 ^{med} /MUC-1 ^{hi} $\hat{\tau}$	4.07	5.18	5.82	6.55	8.09	5.90	1.03	5.90	1.04	
	Null model ($p_D = 1.01$; DIC = 39.9)										
	Both groups combined	4.70	5.62	6.15	6.77	8.00	6.21	0.85	6.22	0.85	

$\hat{\tau}$ Reference group

$\hat{\tau}^*$ this group is not collapsed with another group thus remains the same as in the 9-group model

Table 2

Bayes estimates and MLE of beta coefficients for the Poisson parameters

Cell line	Parameter	Bayes estimates							MLE	SE
		2.5%	25%	50%	75%	97.5%	Mean	SD		
DCIS.COM	<u>Alternative model</u>									
	β_0 (Random sort)	0.30	0.67	0.83	0.99	1.27	0.82	0.24	0.85	0.24
	β_1 (CD49 ^{hi} /MUC-1 ^{med})	-0.63	-0.07	0.20	0.48	0.99	0.20	0.41	0.23	0.40
	β_2 (CD49 ^{hi} /MUC-1 ^{lo})	-1.27	-0.57	-0.24	0.08	0.64	-0.26	0.49	-0.20	0.47
	β_3 (CD49 ^{med} /MUC-1 ^{lo})	-1.58	-0.79	-0.43	-0.08	0.50	-0.46	0.53	-0.39	0.52
	β_4 (CD49 ^{med} /MUC-1 ^{med})	-0.30	0.20	0.44	0.68	1.16	0.43	0.37	0.44	0.36
	β_5 (CD49 ^{hi} /CD24 ^{med})	-2.97	-1.98	-1.36	-0.81	0.05	-1.40	0.80	-1.35	0.99
	β_6 (CD49 ^{hi} /CD24 ^{lo})	-0.29	0.15	0.40	0.63	1.09	0.39	0.35	0.40	0.35
DCIS.COM	<u>Collapsed model</u>									
	β_0 (Random sort)	0.30	0.66	0.83	0.99	1.27	0.82	0.25	0.85	0.24
	β_1 (MUC-1 ^{med})	-0.27	0.13	0.35	0.57	0.99	0.35	0.32	0.35	0.32
	β_2 (MUC-1 ^{lo})	-1.10	-0.56	-0.29	-0.04	0.45	-0.30	0.39	-0.29	0.39
	β_3 (CD24 ^{med})	-2.37	-1.45	-1.02	-0.63	0.02	-1.06	0.61	-0.99	0.60
	β_4 (CD24 ^{lo})	-0.12	0.24	0.43	0.63	1.03	0.43	0.29	0.42	0.29
	<u>Null model</u>									
	All groups combined	0.75	0.88	0.95	1.02	1.14	0.95	0.10	0.96	0.10
SUM-225	<u>Alternative model</u>									
	β_0 (CD49 ^{med} /MUC-1 ^{hi})	1.39	1.64	1.76	1.88	2.10	1.75	0.18	1.77	0.18
	β_1 (CD49 ^{med} /CD24 ^{hi})	-0.44	-0.06	0.13	0.32	0.68	0.13	0.28	0.14	0.28
SUM-225	<u>Null model</u>									
	Both groups combined	1.53	1.72	1.82	1.91	2.08	1.81	0.14	1.83	0.14

* Reference group