Statistical Methods in Medical Research
http://smm.sagepub.com/

Conditional decomposition diagnostics for regression analysis of zero-inflated and left-censored data

Yan Yang and Douglas G Simpson

Stat Methods Med Res 2012 21: 393 originally published online 10 November 2010
DOI: 10.1177/0962280210387525

The online version of this article can be found at:
http://smm.sagepub.com/content/21/4/393

Published by:
SAGE
http://www.sagepublications.com

Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
21(4) 393–408
© The Author(s) 2010
Reprints and permissions:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/0962280210387525
smm.sagepub.com
SAGE

Conditional decomposition diagnostics for regression analysis of zero-inflated and left-censored data

Yan Yang[1] and Douglas G Simpson[2]

Abstract

Health and safety studies that entail both incidence and magnitude of effects produce *semi-continuous* outcomes, in which the response is either zero or a continuous positive value. Zero-inflated left-censored models typically employ latent mixture constructions to allow different covariate processes to impact the incidence versus the magnitude. Assessment of the model, however, requires a focus on the observable characteristics. We employ a conditional decomposition approach, in which the model assessment is partitioned into two observable components: the adequacy of the marginal probability model for the boundary value and the adequacy of the conditional model for values strictly above the boundary. A conditional likelihood decomposition facilitates the statistical assessment. For corresponding residual and graphical analysis, the conditional mean and quantile functions for events above the boundary and the marginal probabilities of boundary events are investigated. Large sample standard errors for these quantities are derived for enhanced graphical assessment, and simulation is conducted to investigate the finite-sample behaviour. The methods are illustrated with data from two health-related safety studies. In each case, the conditional assessments identify the source for lack of fit of the previously considered model and thus lead to an improved model.

Keywords

binary regression, conditional likelihood, conditional quantile, left-truncated data, mixed response model, zero-inflated Tobit

1 Introduction

Health and safety studies that entail both incidence and magnitude of effects produce *semi-continuous* outcomes that are either zeroes or continuous positive values. Understanding incidence of health effects focuses the investigation on the binary partition of the response into zero (no effect) or positive (observed effect), whereas understanding the magnitude of occurred effects focuses the

[1] School of Mathematical and Statistical Sciences, Arizona State University, Tempe, AZ 85287, USA
[2] Department of Statistics, University of Illinois, Champaign, IL 61820, USA

Corresponding author:
Yan Yang, School of Mathematical and Statistical Sciences, Arizona State University, Tempe, AZ 85287, USA
Email: yy@math.asu.edu

analysis on the conditional distribution of the positive effects given that they have occurred. Different covariate processes may impact the incidence versus the magnitude. Unlike purely left-censored models such as the left-censored normal (i.e. Tobit) regression model[1], zero-inflated (ZI) left-censored models such as the ZI Tobit and left-censored logistic (CL) models[2,3] allow the flexibility to jointly model incidence and magnitude effects while allowing for differing covariate models between the two components. This is achieved via the mixture of a non-negative response distribution such as the Tobit or CL with an additional point mass at zero.

A number of recent articles have discussed about the applications of bounded-response models for health studies. For example, Moulton and Halsey[4] considered the immune response to measles vaccine. Taylor *et al.*[5] studied the level of benzene exposure in the petroleum refining industry. Simpson *et al.*[6] modelled the magnitude of lung hemorrhage due to focused exposure of clinical diagnostic ultrasound. Chai and Bailey[7] analysed the coronary artery calcification scores from an atherogenesis study. Applications also abound in medical and ecological research where the data are *discrete* with excessive zeroes, such as wildlife abundance,[8] dental caries status,[9–11] adenoma recurrence,[12] and alcohol or cigarette consumption.[13,14]

For ZI count data, the overall fit of competing models may be compared by plotting differences between the observed and estimated probability masses against non-negative integer values assumed by the response.[15] In related work, score tests have been proposed for testing a Poisson model against a zero-inflated Poisson (ZIP) model,[16] testing a ZIP model against a ZI negative binomial (NB) model,[17] detecting overdispersion in a ZIP model,[18] and testing a ZIP model against general smooth alternatives.[19] Recently, Xie *et al.*[20] considered local influence analysis for ZI generalised Poisson mixed models.

Despite considerable work on fitting various cases of bounded-response models, there is a lack of general methods for assessing the adequacy of these models. Traditional scatter plots superimposed with expectation-based or median-based fitted values may provide only limited information for assessing a model when the data are bounded (Figure 1(a)). The goal of this article is to develop useful diagnostic methods for ZI left-censored and count models. Although these flexible incidence/magnitude models are built on latent mixture models, assessment of the model requires a focus on the observable rather than latent features. To this end, we develop a conditional decomposition approach to assessing the model by partitioning the overall assessment into two *observable* components: (1) the adequacy of the marginal probability model for the boundary value and (2) the adequacy of the conditional model for values strictly above the boundary. We employ a conditional likelihood decomposition into the marginal likelihood for boundary events and the conditional likelihood for magnitudes of positive responses. For corresponding residual and graphical analysis, we investigate the general and model-based conditional mean and quantiles for events above the boundary and marginal probabilities of zeroes. Large sample standard errors (SEs) of these quantities are derived for enhanced graphical assessment.

A motivating data example is provided by the ultrasound safety study of O'Brien *et al.*[21] in which the adverse response was either absent (true zero or below a detection threshold) or an ultrasound-induced lesion whose size was measured. Due to the designed low to moderate exposure mimicking clinical ultrasound, 77.5% of zero responses were observed. The proportions of zeroes at increasing exposure were 1.00, 1.00, 1.00, 0.90, 0.90, 0.70, 0.67, 0.50 and 0.30, so the observed median responses were identically zero except for the two highest exposure groups. Consequently, a plot that compares fitted medians from competing models is not very informative. Further analysis of the data is presented in Section 5.

Section 2 presents the general decomposition of a bounded-response likelihood and defines model-independent conditional mean and quantiles. After a brief review on bounded-response models, we provide the model-based likelihood decomposition, means and quantiles.

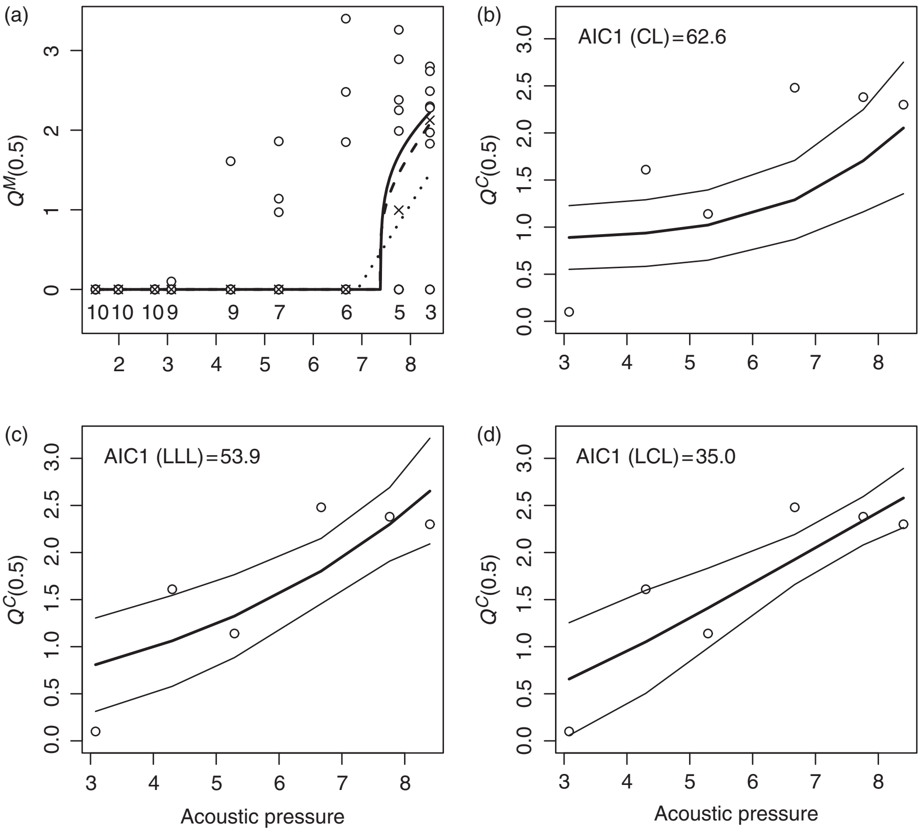

Figure 1. Ultrasound safety study: (a) Estimated marginal medians $Q^m(0.5)$ from the left-censored logistic model (CL: dotted line), logit/log-logistic model (LLL: dashed line), and logit/left-censored logistic model (LCL: solid line). The open circles are raw depth measures. The number symbols located slightly below the $y = 0$ line indicate how often a zero response occurred given exposure. The 'x' symbols denote observed marginal medians. (b)–(d): Pointwise 95% confidence intervals for conditional medians $Q^c(0.5)$ from the CL, LLL and LCL models, respectively. The open circles denote observed conditional medians.

In Section 3, large sample confidence intervals with delta method SEs are developed for the model-based means, quantiles and probabilities of boundary events. Section 4 investigates the finite-sample behaviour of these confidence intervals through a simulation study. The proposed conditional decomposition approach is illustrated with data from an ultrasound safety study in laboratory animals and with data from a measles vaccine study in infants in Section 5. Section 6 summarises our main findings.

2 Likelihood decomposition, means and quantiles of lower bounded data

To model the incidence and magnitude of a health outcome, let the response be a random variable such that

$$\Pr(Y \geq L) = 1 \quad \text{and} \quad 0 < \Pr(Y = L) < 1. \tag{1}$$

The incidence of the health effect is reflected by whether or not $[Y > L]$, while the magnitude is reflected in the value of Y, where Y can take on either discrete or semi-continuous values and L is a lower (detection) threshold that does not necessarily equal zero. Define $v = I(y > L)$, where y is a realisation of Y and $I(A)$ is the indicator function for event A. Let $p(\cdot)$ be the frequency or density function for $y > L$, then the likelihood contribution from an individual observation y is given by $\mathcal{L}(y) = \Pr(Y = L)^{1-v} \cdot p(y)^v$, which can be written as the product of the marginal likelihood for the incidence of the boundary event and the conditional likelihood for the magnitude of the event given that it has occurred: $\mathcal{L}(y) = \{\Pr(Y = L)^{1-v}\Pr(Y > L)^v\} \cdot \{p(y)/\Pr(Y > L)\}^v$.

Denote by $\mu^c = E(Y \mid Y > L)$ the conditional mean for values greater than L and by $\mu^m = E(Y)$ the regular or marginal mean for all values. In health effects studies, the conditional mean represents the typical size of an event conditional on its occurrence, while the marginal mean describes the average overall size of an event regardless of occurrence. When Y is semi-continuous, we can also define the conditional $(100q)$-th quantile $Q^c(q)$ of Y, given $Y > L$, by $\Pr\,(Y > Q^c(q) \mid Y > L) = 1 - q$. Further denote the regular or marginal $(100q)$-th quantile for all values by $Q^m(q)$, where $\Pr(Y > Q^m(q)) = 1 - q$ if $\Pr(Y = L) < q$ and $Q^m(q) = L$ otherwise. Quantiles of common interest include the quartiles and, if the tail behaviour of a response distribution is under check, the 5th and 95th percentiles. Due to the truncation of the conditional quantities at the lower boundary, we have the following inequalities for any outcome variable that satisfies (1), which are proven in the Appendix:

$$\mu^c > \mu^m \quad \text{and} \quad Q^c(q) > Q^m(q).$$

We review next left-inflated mixture (LIM) models and other bounded-response models that can be used to model health effects via a unified mixture construction. Early work on these models for semi-continuous data can be found in the econometrics literature,[22] where the bounded responses are termed as limited dependent variables.

2.1 LIM models and hurdle models

Consider response variables Y_i $(i = 1, \ldots, N)$ that satisfy (1) and follow mixture distributions each composed of the point mass at L and a general distribution F defined on $[L, \infty)$. The mixture density of the ith response has the form $(1 - \pi_i)\delta_L(y_i) + \pi_i f(y_i)$, where $0 < \pi_i \leq 1$ is the mixing weight, $\delta_L(u) = 1$ if $u = L$ and equals zero otherwise, and $f(\cdot)$ is the frequency or density function of F. Let W_i be the mixture-component indicator variables such that $\pi_i = \Pr(W_i = 1)$ and let Z_i be random variables generated from F. Then the marginal responses can be expressed as

$$Y_i \overset{\mathrm{d}}{=} (1 - W_i)L + W_i Z_i, \tag{2}$$

where 'd' denotes in distribution, and W_i and Z_i are statistically independent. For regression analysis define $\mu_i = E\left(Z_i^*\right)$. If Z_i is a censored or truncated variable defined on $[L, \infty)$, then Z_i^* is an uncensored or untruncated variable associated with Z_i. Otherwise $Z_i^* = Z_i$. Let $f^*(\cdot)$ and $F^*(\cdot)$ denote the frequency or density function and the cumulative distribution function for Z_i^*. The dependence of the responses on covariates is expressed through

$$\begin{aligned} h_1(\pi_i) &= \eta_i = \boldsymbol{g}_i^T \boldsymbol{\gamma} \\ h_2(\mu_i) &= \delta_i = \boldsymbol{x}_i^T \boldsymbol{\beta}, \end{aligned} \tag{3}$$

where h_1 and h_2 are known link functions, η_i and δ_i linear predictors, $\boldsymbol{g}_i$ and $\boldsymbol{x}_i$ vectors of covariates that may contain common variables, and $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ parameter vectors.

The class of LIM models includes a wide range of latent mixture models for discrete or semi-continuous data bounded on the left,[3] such as the ZIP and ZI Tobit models. In a LIM model, the general distribution F is discrete (e.g. Poisson) or left-censored (e.g. Tobit) and can produce observations equal to the lower boundary. Thus, the W_i are partially observable and the log-likelihood function is in the form of

$$l_{\mathrm{LIM}}(\boldsymbol{\theta}; \boldsymbol{y}) = \sum_{i=1}^{N} \left[(1 - v_i) \log\left\{1 - \pi_i + \pi_i \Pr\left(Z_i^* \le L\right)\right\} + v_i \log\left\{\pi_i f^*(y_i)\right\}\right],$$

where $\boldsymbol{\theta}$ is the vector containing all regression and scale parameters, $\boldsymbol{y} = (y_1, \ldots, y_N)^{\mathrm{T}}$ and $v_i = I(y_i > L)$. An expectation-maximization (EM) or quasi-Newton algorithm can be used to maximise the likelihood. A simplified case of LIM models is the standard discrete or left-censored models when the point mass at L is not needed (i.e. $\pi_i = 1$ for all i).

Unlike a LIM model in which the distribution status of a boundary value cannot be identified, a hurdle model assumes that boundary values can only come from the point mass at L. Thus, a hurdle model consists of a binary model for whether or not the response is above the boundary (hurdle) and a conditional model (e.g. a truncated-at-L Poisson or normal model) for values strictly above the boundary.[23] The log-likelihood function is given by

$$l_{\mathrm{H}}(\boldsymbol{\theta}; \boldsymbol{y}) = \sum_{i=1}^{N} \left\{(1 - v_i)\log(1 - \pi_i) + v_i \log \pi_i\right\} + \sum_{i=1}^{N} v_i \log \frac{f^*(y_i)}{\Pr\left(Z_i^* > L\right)},$$

which is a sum of the marginal likelihood for boundary events and the conditional likelihood for responses above the boundary. These two likelihood components, with distinct sets of parameters, can be maximised separately for maximising $l_{\mathrm{H}}(\boldsymbol{\theta}; \boldsymbol{y})$. For non-negative semi-continuous responses ($L = 0$), the conditional model may alternatively assume an untruncated distribution with positive support such as the log-normal.[24,25] The resulting model is usually referred to as a two-part model.

2.2 Model-based likelihood decomposition

Besides computational advantages, the natural decomposition of the hurdle likelihood into the marginal and conditional likelihoods also helps assess the adequacy of the marginal probability model for the occurrence of an event and the adequacy of the conditional model for the size of the event given that it has occurred. To see this, denote by l_{H0} the marginal likelihood component of the hurdle likelihood and by l_{H1} the conditional likelihood component such that $l_{\mathrm{H}}(\boldsymbol{\theta}; \boldsymbol{y}) = l_{\mathrm{H0}} + l_{\mathrm{H1}}$. The LIM likelihood can be decomposed similarly following the general decomposition of a bounded-response likelihood $\mathcal{L}(y)$ into $l_{\mathrm{LIM}}(\boldsymbol{\theta}; \boldsymbol{y}) = l_{\mathrm{LIM0}} + l_{\mathrm{LIM1}}$, where

- $l_{\mathrm{LIM0}} = \sum_{i=1}^{N} \left[(1 - v_i)\log\left\{1 - \pi_i + \pi_i \Pr\left(Z_i^* \le L\right)\right\} + v_i \log\left\{\pi_i \Pr\left(Z_i^* > L\right)\right\}\right]$, and
- $l_{\mathrm{LIM1}} = \sum_{i=1}^{N} v_i \log\left\{f^*(y_i) / \Pr\left(Z_i^* > L\right)\right\}$.

The Akaike's information criterion (AIC) of a hurdle model can then be written as $\mathrm{AIC_H} = \mathrm{AIC_{H0}} + \mathrm{AIC_{H1}}$, where $\mathrm{AIC_{H0}} = -2l_{\mathrm{H0}} + 2k_{\mathrm{H0}}$, $\mathrm{AIC_{H1}} = -2l_{\mathrm{H1}} + 2k_{\mathrm{H1}}$, $k_{\mathrm{H0}} = k_0 = \dim(\boldsymbol{\gamma})$ and $k_{\mathrm{H1}} = k_1 = \dim(\boldsymbol{\beta}) + 1$ if a scale parameter is estimated in the model and $k_{\mathrm{H1}} = k_1 = \dim(\boldsymbol{\beta})$

otherwise. Likewise, the AIC of a LIM model can be expressed as $\mathrm{AIC}_{\mathrm{LIM}} = \mathrm{AIC}_{\mathrm{LIM0}} + \mathrm{AIC}_{\mathrm{LIM1}}$, where $\mathrm{AIC}_{\mathrm{LIM0}} = -2l_{\mathrm{LIM0}} + 2k_{\mathrm{LIM0}}$, $\mathrm{AIC}_{\mathrm{LIM1}} = -2l_{\mathrm{LIM1}} + 2k_{\mathrm{LIM1}}$, $k_{\mathrm{LIM0}} = k_0 + (1-\zeta)k_1$, $k_{\mathrm{LIM1}} = \zeta k_1$, and $0 < \zeta \le 1$. The midpoint values of the range intervals for k_{LIM0} and k_{LIM1} with $\zeta = 0.5$ provide rough point summaries for these penalty terms.

2.3 Model-based conditional and marginal means

The estimated means given the level of covariates from an incidence/magnitude model may be compared with the sample means for assessment of the model. Under Equations (2) and (3), the conditional mean $\mu^c(\boldsymbol{\theta}) = E(Y \mid Y > L, \boldsymbol{g}, \boldsymbol{x})$ and marginal mean $\mu^m(\boldsymbol{\theta}) = E(Y \mid \boldsymbol{g}, \boldsymbol{x})$ based on a LIM model are respectively given by

$$\mu^c_{\mathrm{LIM}}(\boldsymbol{\theta}) = E(Z^* \cdot I(Z^* > L) \mid \boldsymbol{x}) / \Pr(Z^* > L \mid \boldsymbol{x}), \tag{4}$$

$$\mu^m_{\mathrm{LIM}}(\boldsymbol{\theta}) = \left\{1 - \pi \cdot \Pr(Z^* > L \mid \boldsymbol{x})\right\} L + \pi \cdot E(Z^* \cdot I(Z^* > L) \mid \boldsymbol{x}) \tag{5}$$

due to independence between W and Z, where $\pi = h_1^{-1}(\boldsymbol{g}^T\boldsymbol{\gamma})$. For hurdle models, the conditional mean $\mu^c_{\mathrm{H}}(\boldsymbol{\theta})$ has the same form as $\mu^c_{\mathrm{LIM}}(\boldsymbol{\theta})$, and the marginal mean is $\mu^m_{\mathrm{H}}(\boldsymbol{\theta}) = (1-\pi)L + \pi\mu^c_{\mathrm{H}}$. The model-based conditional and marginal variances can be found in a similar manner. The estimated moments are obtained by plugging in $\hat{\boldsymbol{\theta}}$, the maximum likelihood estimator of $\boldsymbol{\theta}$.

Table 1 summarises conditional and marginal means as well as marginal variances of the ZI Poisson, NB, binomial, Tobit and CL models. Note that the conditional mean is greater than the marginal mean in each model. In addition, unlike standard count models, ZI count models allow overdispersion due to excessive zeroes, as indicated by the marginal variances. For instance, a NB model has mean μ and variance $\mu + \mu^2/\kappa$, where κ^{-1} is sometimes referred to as a dispersion parameter.[26] In contrast, a ZI negative binomial model has mean $\pi\mu$ and variance $\pi\mu + (\pi\mu)^2/\kappa + \pi\mu^2(1-\pi)(1 + 1/\kappa) > \pi\mu + (\pi\mu)^2/\kappa$ when $\pi < 1$. For other left-censored LIM models, the conditional and marginal moments are computed through numerical integration.

2.4 Model-based conditional and marginal quantiles

For semi-continuous health outcomes the model-based quantiles offer a more flexible means to assessing a model. Suppose for the model under consideration that F^* is a continuous

Table 1. Conditional means, marginal means and marginal variances of the ZI Poisson, NB, binomial, Tobit and CL models. The dependence of π and μ on covariates is suppressed

F^*	$f^*(y)$	$\mu^c(\boldsymbol{\theta})$	$\mu^m(\boldsymbol{\theta})$	$\mathrm{var}^m(\boldsymbol{\theta})$
Poisson	$\exp(-\mu)\mu^y/y!$	$\mu/\{1-\exp(-\mu)\}$	$\pi\mu$	$\pi\mu + \pi\mu^2(1-\pi)$
NB	$\Gamma(\kappa + y)/\{\Gamma(\kappa)\ y!\} \cdot \mu^y\kappa^\kappa/(\mu+\kappa)^{\kappa+y}$	$\mu/\{1 - (\frac{\kappa}{\mu+\kappa})^\kappa\}$	$\pi\mu$	$\pi\mu + (\pi\mu)^2/\kappa + \pi\mu^2(1-\pi)(1 + 1/\kappa)$
Binomial	$\binom{n}{y}\mu^y(1-\mu)^{n-y}$	$n\mu/\{1-(1-\mu)^n\}$	$\pi n\mu$	$n(\pi\mu)(1-\pi\mu) + \pi n(1-\pi)(n-1)\mu^2$
Normal[a]	$1/(\sqrt{2\pi}\sigma) \cdot \exp\{-(y-\mu)^2/(2\sigma^2)\}$	$\nu/\Phi(\lambda)$	$\pi\nu$	$\pi\{\nu(\mu - \pi\nu) + \sigma^2\Phi(\lambda)\}$
Logistic[b]	$\exp(\frac{y-\mu}{\sigma}) \Big/ \left\{\sigma\left(1 + \exp(\frac{y-\mu}{\sigma})\right)^2\right\}$	$\omega(1 + \exp(-\lambda))$	$\pi\omega$	–

[a] For the ZI Tobit model, $\lambda = \mu/\sigma$ and $\nu = \mu\Phi(\lambda) + \sigma\phi(\lambda)$.

[b] For the ZICL model, $\lambda = \mu/\sigma$ and $\omega = \sigma\log(1 + \exp(\lambda))$; the marginal variance involves an infinite series expansion and is omitted.

distribution from a location-scale family with location $= \mu$ and scale $= \sigma$. Examples include the normal, logistic, t and extreme value (EV) distributions. By log transformation, this class also includes the log-normal, log-logistic and Weibull distributions. Let F_s^* denote the standardised distribution of F^* such that $\frac{Z^* - \mu}{\sigma} \sim F_s^*$ and let $f_s^*(\cdot)$ be the associated density function. Under Equations (2) and (3), the conditional and marginal $(100q)$-th quantiles of a LIM model are in the respective forms of

$$Q^c_{\mathrm{LIM}}(q, \boldsymbol{\theta}) = \mu + \sigma \cdot F_s^{*-1}\big(q + (1 - q) \cdot F_s^*(c_1)\big), \tag{6}$$

$$Q^m_{\mathrm{LIM}}(q, \boldsymbol{\theta}) = \begin{cases} \mu + \sigma \cdot F_s^{*-1}(1 - (1 - q)/\pi) & \text{if } 1 - \pi + \pi \cdot F_s^*(c_1) < q \\ L & \text{otherwise,} \end{cases} \tag{7}$$

where $c_1 = \frac{L - \mu}{\sigma}$, $\mu = h_2^{-1}(\boldsymbol{x}^T \boldsymbol{\beta})$ and $F_s^{*-1}(\cdot)$ is the inverse distribution function. For hurdle models, the conditional quantile $Q^c_{\mathrm{H}}(q, \boldsymbol{\theta})$ has the same form as $Q^c_{\mathrm{LIM}}(q, \boldsymbol{\theta})$, and the marginal quantile $Q^m_{\mathrm{H}}(q, \boldsymbol{\theta}) = \mu + \sigma \cdot F_s^{*-1}(1 - \frac{1-q}{\pi} \cdot \{1 - F_s^*(c_1)\})$ if $1 - \pi < q$ and equals L otherwise. The estimated quantiles are obtained by plugging in $\hat{\boldsymbol{\theta}}$.

3 Large sample confidence intervals

For enhanced graphical assessment of an incidence/magnitude model, we provide in this section the delta method confidence intervals for model-based means, quantiles as well as probabilities of the boundary event, $P^L(\boldsymbol{\theta}) = \Pr(Y = L \mid \boldsymbol{g}, \boldsymbol{x})$. Suppose that $\psi(\boldsymbol{\theta})$ is a real-valued, non-linear scalar function of $\boldsymbol{\theta}$ from a bounded-response model. Here we consider $\psi(\boldsymbol{\theta}) = \mu^c(\boldsymbol{\theta})$, $\mu^m(\boldsymbol{\theta})$, $Q^c(q, \boldsymbol{\theta})$, $Q^m(q, \boldsymbol{\theta})$ or $P^L(\boldsymbol{\theta})$. Let $\psi(\hat{\boldsymbol{\theta}})$ denote the estimator of $\psi(\boldsymbol{\theta})$. By the multivariate delta method through a first-order Taylor series expansion[27] at $\boldsymbol{\theta}$, the variance of $\psi(\hat{\boldsymbol{\theta}})$ is estimated by

$$\widehat{\mathrm{var}}\,(\psi(\hat{\boldsymbol{\theta}})) = \nabla\psi(\hat{\boldsymbol{\theta}})^T \cdot \widehat{\mathrm{var}}\,(\hat{\boldsymbol{\theta}}) \cdot \nabla\psi(\hat{\boldsymbol{\theta}}),$$

where $\nabla\psi(\hat{\boldsymbol{\theta}}) = (\partial/\partial\boldsymbol{\theta})\psi(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}$, and $\widehat{\mathrm{var}}\,(\hat{\boldsymbol{\theta}})$ is the estimated covariance matrix of $\hat{\boldsymbol{\theta}}$. For a LIM model, $\widehat{\mathrm{var}}\,(\hat{\boldsymbol{\theta}})$ can be obtained by the method of Louis[28] in the context of an EM algorithm or from the approximate Hessian at convergence of a quasi-Newton method.[3] For a hurdle model, $\mathrm{var}\,(\hat{\boldsymbol{\theta}})$ is a two-block diagonal matrix combining the two covariance matrices from the binary model and the conditional model. Let $z_{1-\alpha/2}$ denote the $100(1 - \alpha/2)$-th percentile of the standard normal distribution. A large sample $1 - \alpha$ confidence interval for $\psi(\boldsymbol{\theta})$ is given by

$$\psi(\hat{\boldsymbol{\theta}}) \pm z_{1-\alpha/2} \cdot \{\widehat{\mathrm{var}}\,(\psi(\hat{\boldsymbol{\theta}}))\}^{1/2}, \tag{8}$$

which may need to be modified to reflect boundary constraints: the lower confidence limits for means and quantiles should be bounded below by L, while intervals for probabilities of boundary events should be contained in [0, 1].

Next we find the gradient vectors $\nabla\psi(\boldsymbol{\theta})$ for left-censored LIM models. In case of discrete data omit from the vectors the components involving the scale parameter σ.

3.1 Confidence intervals for means

Taking the derivatives of the conditional mean (4) and marginal mean (5) with respect to $\boldsymbol{\theta} = (\boldsymbol{\gamma}^T, \boldsymbol{\beta}^T, \tau)^T$, where $\tau = \log(\sigma)$, gives

$$\nabla\mu^c_{\mathrm{LIM}}(\boldsymbol{\theta}) = \left(\mathbf{0}^T,\ \frac{\partial\mu^c_{\mathrm{LIM}}(\boldsymbol{\theta})}{\partial\mu}\cdot\frac{\partial\mu}{\partial\delta}\cdot\boldsymbol{x}^T,\ \frac{\partial\mu^c_{\mathrm{LIM}}(\boldsymbol{\theta})}{\partial\sigma}\cdot\sigma\right)^T,$$

$$\nabla\mu^m_{\mathrm{LIM}}(\boldsymbol{\theta}) = \left(\frac{\partial\mu^m_{\mathrm{LIM}}(\boldsymbol{\theta})}{\partial\pi}\cdot\frac{\partial\pi}{\partial\eta}\cdot\boldsymbol{g}^T,\ \frac{\partial\mu^m_{\mathrm{LIM}}(\boldsymbol{\theta})}{\partial\mu}\cdot\frac{\partial\mu}{\partial\delta}\cdot\boldsymbol{x}^T,\ \frac{\partial\mu^m_{\mathrm{LIM}}(\boldsymbol{\theta})}{\partial\sigma}\cdot\sigma\right)^T.$$

Computing $\nabla\mu^c_{\mathrm{LIM}}(\boldsymbol{\theta})$ and $\nabla\mu^m_{\mathrm{LIM}}(\boldsymbol{\theta})$ is straightforward for the left-censored and discrete LIM models included in Table 1. For other left-censored LIM models, no extra numerical integration is needed for computing the gradient vectors than required by the means themselves; a sketch of the proof is provided in the Appendix. These results also apply to hurdle models.

3.2 Confidence intervals for quantiles

Taking the derivatives of the conditional quantile (6) and marginal quantile (7) with respect to elements of $\boldsymbol{\theta}$ gives

$$\nabla Q^c_{\mathrm{LIM}}(q,\boldsymbol{\theta}) = \left(\mathbf{0}^T, \left\{1-(1-q)\cdot\frac{f^*_s(c_1)}{f^*_s(c_2)}\right\}\boldsymbol{x}^T, \left\{c_2 - c_1(1-q)\cdot\frac{f^*_s(c_1)}{f^*_s(c_2)}\right\}\sigma\right)^T,$$

$$\nabla Q^m_{\mathrm{LIM}}(q,\boldsymbol{\theta}) = \left(\frac{\sigma(1-q)}{\pi^2 f^*_s(c_3)}\cdot\frac{\partial\pi}{\partial\eta}\cdot\boldsymbol{g}^T, \boldsymbol{x}^T, c_3\sigma\right)^T \quad \text{if } 1-\pi+\pi\cdot F^*_s(c_1) < q$$

and $\nabla Q^m_{\mathrm{LIM}}(q,\boldsymbol{\theta}) = \mathbf{0}$ otherwise, where $c_2 = \left\{Q^c_{\mathrm{LIM}}(q,\boldsymbol{\theta}) - \mu\right\}/\sigma$ and $c_3 = \left\{Q^m_{\mathrm{LIM}}(q,\boldsymbol{\theta}) - \mu\right\}/\sigma$.

3.3 Confidence intervals for probabilities of boundary events

To evaluate the fit for values at the boundary, consider the probabilities of boundary events: $P^L_{\mathrm{LIM}}(\boldsymbol{\theta}) = 1 - \pi\cdot\Pr(Z^* > L \mid \boldsymbol{x})$ and $P^L_{\mathrm{H}}(\boldsymbol{\theta}) = 1 - \pi$. The gradient vectors are obtained by taking their derivatives with respect to elements of $\boldsymbol{\theta}$. A related tool that helps evaluate the deviation of an estimated probability from the observed is the Pearson residual:

$$(\text{observed} - \text{estimated})/\sqrt{\left\{\text{estimated} \times (1 - \text{estimated})/\text{ group size}\right\}}.$$

When the design of a study includes replicated observations at combinations of covariate levels and the group sizes are moderate, a Pearson residual with an absolute value greater than 2 provides evidence of lack of fit.[26]

4 Simulation

A simulation study is conducted to investigate the finite-sample behaviour of the delta method confidence intervals. The ZIP model and zero-inflated left-censored logistic (ZICL) model were evaluated. For each model, a binary predictor x was included in both sub-models at respective sample sizes $N = 40$, 100 and 200. The parameters took the values $\boldsymbol{\gamma} = \boldsymbol{\beta} = (0.5,\ 1/3)^T$ in the ZIP model and $\boldsymbol{\gamma} = (0.5,\ 1/3)^T$, $\boldsymbol{\beta} = (0.7,\ 0.4)^T$ and $\sigma = 0.5$ in the ZICL model. With these choices, there were 50% (37%) of zero responses and around 20% (10%) of these zeroes were from the Poisson or

Table 2. Mean estimates, Monte Carlo SEs and mean delta method SEs for (1) probabilities of zeroes and conditional means from the ZIP model and (2) conditional quartiles from the ZICL model. The sample sizes are 40, 100 and 200

			Mean			Monte Carlo SE (mean SE)		
$\psi(\boldsymbol{\theta})$	x	True value	40	100	200	40	100	200
ZIP								
$P^L(\boldsymbol{\theta})$	0	0.497	0.496	0.498	0.497	0.110 (0.108)	0.070 (0.070)	0.050 (0.050)
	1	0.373	0.373	0.372	0.373	0.106 (0.105)	0.069 (0.068)	0.048 (0.048)
$\mu^c(\boldsymbol{\theta})$	0	2.041	2.043	2.039	2.042	0.354 (0.349)	0.226 (0.222)	0.158 (0.157)
	1	2.557	2.562	2.556	2.559	0.391 (0.390)	0.248 (0.246)	0.173 (0.174)
ZICL								
$Q^c(0.25, \boldsymbol{\theta})$	0	0.494	0.538	0.508	0.502	0.185 (0.172)	0.117 (0.112)	0.082 (0.080)
	1	0.734	0.778	0.748	0.742	0.203 (0.192)	0.130 (0.127)	0.092 (0.091)
$Q^c(0.50, \boldsymbol{\theta})$	0	0.900	0.920	0.905	0.903	0.225 (0.213)	0.147 (0.141)	0.104 (0.102)
	1	1.200	1.212	1.202	1.202	0.221 (0.210)	0.142 (0.138)	0.099 (0.099)
$Q^c(0.75, \boldsymbol{\theta})$	0	1.391	1.378	1.383	1.388	0.268 (0.254)	0.174 (0.168)	0.122 (0.120)
	1	1.718	1.695	1.707	1.714	0.255 (0.243)	0.163 (0.160)	0.114 (0.114)

CL given $x=0$ ($x=1$). 10,000 runs were carried out for each simulation. ZI mixture models for the simulated trials were estimated by the quasi-Newton method implemented in the `optim` function in R.[29]

Probabilities of zeroes and conditional means of the ZIP model were assessed, excluding the two runs at $N=40$ that had non-positive definite observed information due to large estimates for γ.[3] For the ZICL model, the conditional quartiles were evaluated, excluding the 26 runs at $N=40$ and five runs at $N=100$ in which the observed information matrix was not positive definite. Table 2 reports the true values of these quantities (obtained by plugging in the true values of the parameters), mean estimates, Monte Carlo SEs and mean delta method SEs. The mean estimates are generally unbiased. The simulated and mean SEs are fairly close to each other.

Table 3 summarises error and coverage rates of 95% delta method confidence intervals for the quantities under evaluation. The lower (upper) error rate is the percent of confidence intervals located entirely above (below) the true value of a quantity. When the sample size is small, the intervals exhibit undercoverage; the lower and upper error rates appear asymmetric. As N increases, the coverage rate generally approaches the nominal level 0.95, and the error rates are reasonably balanced.

Marginal means of the ZIP model and marginal quartiles and other quantiles of the ZICL model were also assessed in other results not shown. Confidence intervals for marginal means perform similarly to those for conditional means. Marginal quantiles occasionally have biased point estimates and low-coverage intervals when the true quantile is close to the boundary value and the sample size is small. This is because the interval has zero length when the estimated marginal quantile equals the boundary value. The performance improves as N increases or when the true marginal quantile stays away from the boundary.

5 Data examples

The proposed conditional decomposition approach to assessing and comparing LIM models and other bounded-response models is illustrated with data from an ultrasound safety study and with

Table 3. Rates of lower error, coverage and upper error of 95% delta method confidence intervals for (1) probabilities of zeroes and conditional means from the ZIP model and (2) conditional quartiles from the ZICL model. The sample sizes are 40, 100 and 200

		$x=0$			$x=1$		
$\psi(\theta)$	N	Lower	Coverage	Upper	Lower	Coverage	Upper
ZIP							
$P^L(\theta)$	40	0.053	0.922	0.025	0.031	0.941	0.028
	100	0.028	0.938	0.034	0.023	0.941	0.036
	200	0.025	0.945	0.031	0.016	0.950	0.034
$\mu^c(\theta)$	40	0.005	0.920	0.075	0.011	0.933	0.056
	100	0.011	0.932	0.057	0.014	0.943	0.043
	200	0.014	0.944	0.042	0.017	0.949	0.034
ZICL							
$Q^c(0.25, \theta)$	40	0.038	0.894	0.067	0.051	0.909	0.041
	100	0.026	0.925	0.049	0.031	0.938	0.031
	200	0.024	0.939	0.037	0.030	0.944	0.025
$Q^c(0.50, \theta)$	40	0.036	0.906	0.058	0.036	0.919	0.045
	100	0.026	0.930	0.044	0.028	0.937	0.035
	200	0.027	0.940	0.034	0.027	0.948	0.025
$Q^c(0.75, \theta)$	40	0.019	0.912	0.068	0.015	0.917	0.068
	100	0.018	0.932	0.050	0.016	0.936	0.049
	200	0.021	0.941	0.038	0.017	0.945	0.038

data from a measles vaccine study. In each case the conditional assessments identify the source for lack of fit of the previously considered models and lead to an improved model. The R code implementation and data for the two examples are available at http://math.asu.edu/˜yy/MA.html. Hurdle models can also be estimated by other software including Stata.[30]

5.1 An ultrasound safety study

Our motivating data are from a randomised, blind ultrasound safety study in laboratory rabbits.[3,21] The adverse response was the depth of an ultrasound-induced lesion (in mm) on the lung surface. A zero response was recorded if no lesion was observed. The primary predictor was acoustic pressure (in MPa) for which 10 or 9 animals were exposed at each of nine exposure groups. Both lungs of an animal were exposed. For demonstration, we consider here only the first exposed lungs. Of these, $69/89 = 77.5\%$ were zero.

The data were analysed using the CL model, logit/log-logistic model (LLL: a two-part model assuming the logit link for the binary model and the log-logistic distribution for the conditional model), logit/left-truncated logistic model (a hurdle model assuming the truncated-at-zero logistic distribution) and logit/left-censored logistic model (LCL: a LIM model assuming the censored-at-zero logistic distribution). The models, with corresponding likelihood and AIC decompositions for zeroes and positive responses, are summarised in Table 4. Based on the AIC values, the hurdle model and LIM model give the best fit for the data and are closely comparable.

Figure 1(a) summarises the models with estimated marginal median curves. Due to high proportions of zero responses given exposure, the observed marginal medians are identically zero

Table 4. Ultrasound safety study: Parameter estimates, SEs, and likelihood and AIC decompositions for the CL model, logit/log-logistic model, logit/left-truncated logistic model and logit/left-censored logistic model

	Censored logistic	Logit/log-logistic	Logit/truncated logistic	Logit/censored logistic
Estimate (SE)				
γ_0	–	−5.124 (1.084)	−5.124 (1.084)	−4.809 (1.182)
γ_1	–	0.695 (0.162)	0.695 (0.162)	0.652 (0.175)
β_0	−6.748 (1.670)	−0.900 (0.522)	−0.466 (0.652)	−0.597 (0.680)
β_1	0.977 (0.223)	0.223 (0.070)	0.361 (0.088)	0.378 (0.092)
$\log(\sigma)$	0.215 (0.192)	−1.488 (0.195)	−1.163 (0.192)	−1.160 (0.192)
Likelihood decomposition				
l_0	−33.45	−32.90	−32.90	−32.71
l_1	−29.81	−23.97	−15.99	−6.01
Total	−63.27	−56.87	−48.89	−48.71
AIC decomposition				
AIC_0	66.90–72.90	69.80	69.80	69.41–75.41
AIC_1	65.63–59.63	53.93	37.98	38.02–32.02
Total	132.53	123.74	107.78	107.43

except for the two highest exposure groups. As a result, the estimated marginal medians are not very informative in comparing different models. Figure 1(b)–(d) displays 95% confidence intervals for conditional medians in which the midpoint values of decomposed AIC range-intervals for positive responses are used as point summaries in the two censored logistic cases (CL and LCL). The estimated conditional median line from the censored model in Figure 1(b) indicates a systematic downward bias. Moreover, it curves the wrong way in the sense that a lesion can only get so large no matter how much acoustic pressure. The two-part model in Figure 1(c) shows some improvement over the censored model, while the LIM model in Figure 1(d) improves further and seems adequate. The decomposed AIC values agree with the plots. The hurdle medians are very similar to those from the LIM model and omitted. Plots of 95% confidence intervals for probabilities of zeroes are similar for all models considered and appear adequate (not shown). The associated Pearson residuals all have sizes smaller than or close to one.

Thus, the hurdle model and LIM model both fit the ultrasound exposure data well. The purely censored model fails to fit positive lesion sizes adequately due to the *coexistence* of sizeable portions of zeroes and large positive values at high exposure (e.g. at the highest exposure level 30% of zero responses coexisted with 70% of size measurements greater than 1.8 mm). The two-part model allows differing covariate processes for the binary model and the conditional model, but it seems sensitive to non-separability of values at and above the boundary.[3]

5.2 A measles vaccine study

Moulton and Halsey[4] analysed data from a safety and immunogenicity study of measles vaccine in infants. The response was the log-transformed antibody concentration, with $86/330 = 26.1\%$ of observations censored at the lower detection limit of log(0.1) international units. In their article, the authors employed the logit/left-censored normal model (LCN: a LIM model assuming the left-censored normal distribution) as an improvement to the left-censored normal model (CN) for

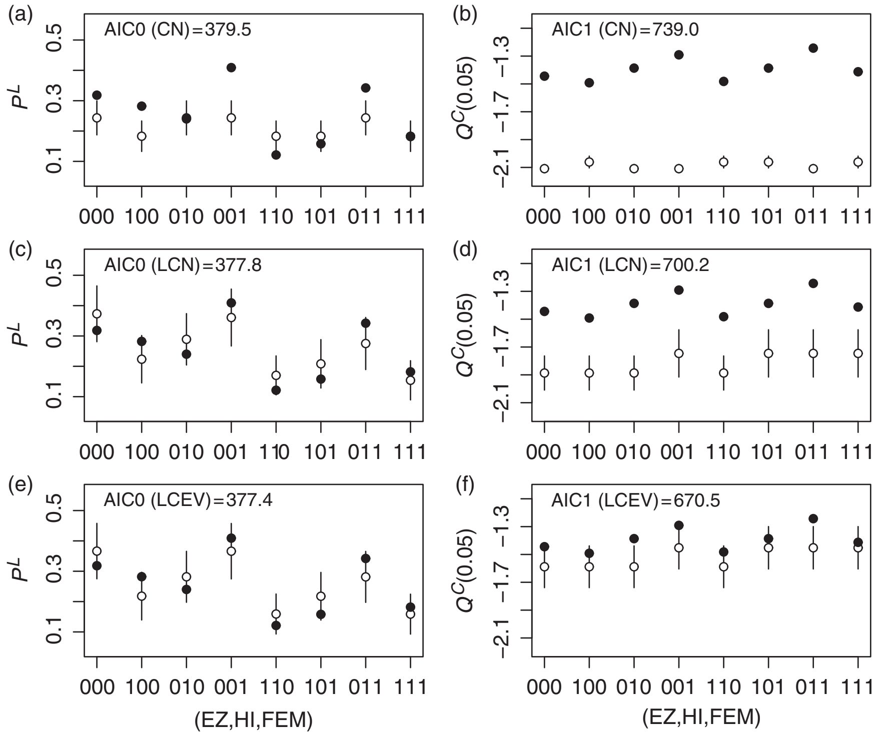

Figure 2. Measles vaccine study: Pointwise 95% confidence intervals for probabilities of boundary values P^L and conditional 5th quantiles $Q^c(0.05)$ from (a) and (b): the left-censored normal model (CN), (c) and (d): the logit/left-censored normal model (LCN), and (e) and (f): the logit/left-censored extreme value model (LCEV). The solid circles are observed quantities. The confidence intervals are represented by vertical bars centred at model-based quantities (the open circles).

investigating the effects of vaccine strain (EZ: 1-Edmonston Zagreb, 0-Schwarz), vaccine dose (HI: 1-high, 0-medium) and sex (FEM: 1-female, 0-male).

Figure 2(a)–(d) depicts 95% confidence intervals for probabilities of boundary events and conditional 5th quantiles based on the CN and LCN models. The observed proportion of boundary values given a combined level of the predictors ranges from 0.12 to 0.41. The probability intervals from the CN model in Figure 2(a) miss a few observed quantities from the upper side, while those from the LCN model look adequate. The conditional 5th quantile intervals from the CN model in Figure 2(b) are severely biased downwards. The corresponding intervals from the LCN model are less biased and have more reasonable lengths, but none of them contain the observed quantities.

The underestimation of conditional 5th quantiles by the two normal models (CN and LCN) indicates that a symmetric normal distribution is inadequate for modelling the seemingly right-skewed response distribution (Figure 3(a)). Due to a few extreme values at the upper tail, a normal distribution tends to have a large scale parameter and thus overshoots the lower tail of the data distribution. This is reflected in Figure 3(b) and (c), the scatter plots for data above the boundary enhanced by estimated conditional 5th–50th–95th quantiles from the normal models.

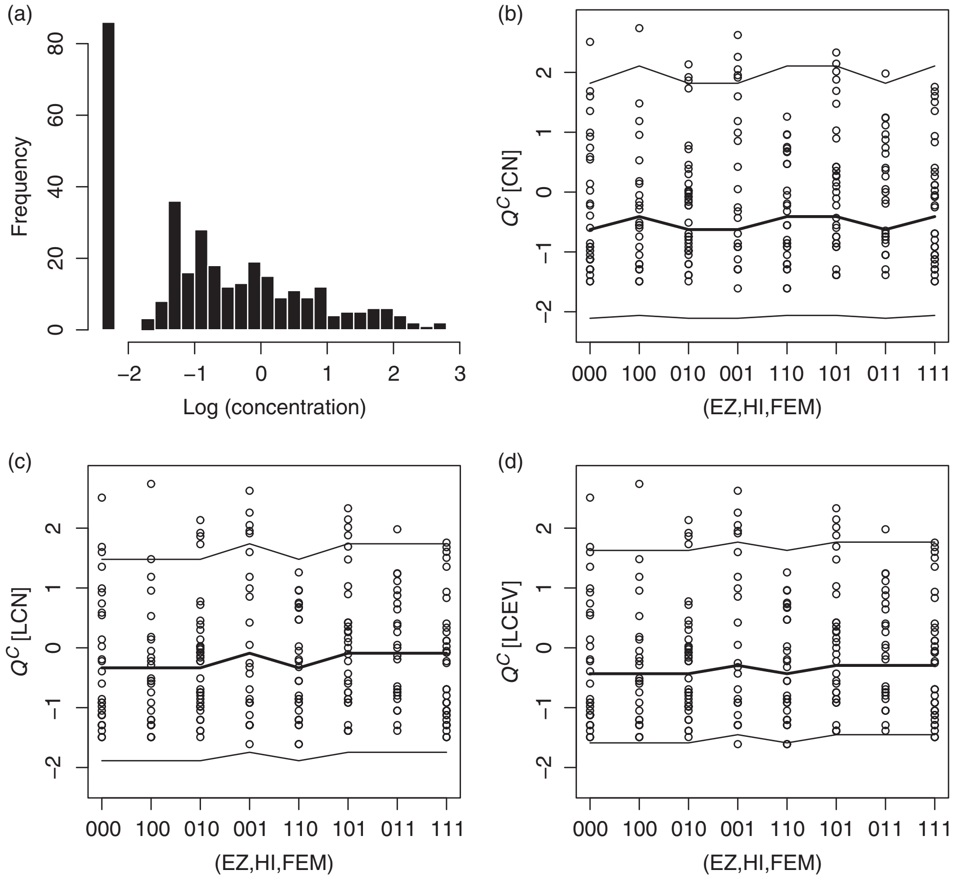

Figure 3. Measles vaccine study: (a) Frequency histogram of log-transformed antibody concentration. (b)–(d): Estimated conditional 5th quantiles (lower thin lines), conditional medians (thick lines) and conditional 95th quantiles (upper thin lines) from the left-censored normal model (CN), the logit/left-censored normal model (LCN) and the logit/left-censored extreme value model (LCEV), respectively. The open circles denote the (uncensored) log-concentration measurements above the lower boundary log(0.1).

Given the evidence of data skewness and residual skewness in the normal models, we fit the logit/left-censored extreme value model (LCEV: a LIM model assuming the left-censored EV distribution) that is more suitable for right-skewed data. Table 5 summarises the hurdle models and LIM models with an underlying normal or EV distribution. All models retain the same effects as the LCN model in Moulton and Halsey[4] for comparison. The EV hurdle and LIM models have similar parameter estimates, with smaller decomposed AIC values for observations above the boundary than the normal hurdle and LIM models. The LCN model suggests that gender had some effect on antibody concentration given that it was detectable (p-value = 0.08). The LCEV model, however, shows a weaker gender effect (p-value = 0.18), which is consistent with a more recent analysis of measles vaccine trials.[31]

Figure 2(e)–(f) displays confidence intervals for probabilities of boundary events and conditional 5th quantiles from the LCEV model. The probability intervals are similar to those from the LCN

Table 5. Measles vaccine study: Parameter estimates, SEs, likelihood decomposition, and AIC decomposition for the logit/left-truncated normal model, logit/left-truncated EV model, logit/left-censored normal model and logit/left-censored EV model

	Logit/truncated normal	Logit/truncated EV	Logit/censored normal	Logit/censored EV
Estimate (SE)				
γ_0	0.548 (0.201)	0.548 (0.201)	0.649 (0.220)	0.550 (0.201)
γ_1(EZ)	0.730 (0.263)	0.730 (0.263)	0.824 (0.304)	0.731 (0.263)
γ_2(HI)	0.389 (0.256)	0.389 (0.256)	0.422 (0.288)	0.389 (0.256)
β_0	−0.416 (0.117)	−0.724 (0.073)	−0.399 (0.112)	−0.723 (0.073)
β_1(FEM)	0.296 (0.158)	0.139 (0.103)	0.271 (0.154)	0.139 (0.103)
log(σ)	0.122 (0.059)	−0.234 (0.054)	0.119 (0.058)	−0.234 (0.054)
Likelihood decomposition				
l_0	−184.17	−184.17	−184.41	−184.18
l_1	−348.58	−333.77	−348.60	−333.77
Total	−532.76	−517.94	−533.01	−517.95
AIC decomposition				
AIC$_0$	374.35	374.35	374.82–380.82	374.35–380.35
AIC$_1$	703.16	673.54	703.20–697.20	673.54–667.54
Total	1077.51	1047.89	1078.01	1047.90

model and suggest a good fit for boundary values. Most of the conditional 5th quantile intervals contain the observed quantities, though some small downward bias still exists. Figure 3(d) presents estimated conditional 5th–50th–95th quantiles from the LCEV model; it better conforms to the spread of individual observations than the CN and LCN models.

6 Conclusions

Bounded-response models are increasingly used for modelling incidence and magnitude of health effects. In this article, we develop a novel conditional decomposition approach in which the overall assessment of an incidence/magnitude model is partitioned into the adequacy of the marginal probability model for the occurrence of a boundary event and the adequacy of the conditional model for the size of the event given that it has occurred. Likelihood and AIC decompositions are used to facilitate comparison of goodness-of-fit of different models at and above the boundary and to identify the source for lack of fit of a model. For corresponding residual and graphical analysis, we establish delta method confidence intervals for the model-based conditional means, conditional quantiles and probabilities of the boundary event. A small simulation study indicates that delta method SEs perform well in finite samples, and the confidence intervals behave reasonably for moderate to large samples.

As has been demonstrated through two health-related safety studies, the conditional approach provides insights not evident in purely marginal assessment of the model via scatter plots, marginal fitted values and the like. When a sizeable number of boundary values and values well above the boundary coexist given covariates in a data set, a purely censored regression model often fails to provide adequate fits for both the boundary event and values above the boundary. A hurdle model or a LIM model, on the other hand, allows differing covariate processes for the binary model and the conditional model, and is thus able to better fit the data. For skewed data, the tail behaviour of the conditional distribution (e.g. conditional 5th and 95th quantiles) is usually more informative in assessing a model.

The proposed conditional decomposition diagnostics may be used for assessing marginal analysis of correlated bounded responses. The working independence likelihood can be assumed for the AIC decomposition, and the sandwich formula is used to estimate the covariance matrix of parameter estimates required in computing the delta method confidence interval in Equation (8).[3] The diagnostics then address the adequacy of the marginal models for the data.

Acknowledgements

This study was supported in part by NIH grant EB002641-09 and NSF grant DRL-0909630. The authors thank two referees for careful reading and constructive comments that lead to an improved article.

References

1. Tobin J. Estimation of relationships for limited dependent variables. *Econometrica* 1958; **26**: 24–36.
2. Berk KN and Lachenbruch PA. Repeated measures with zeros. *Stat Methods Med Res* 2002; **11**: 303–316.
3. Yang Y and Simpson DG. Unified computational methods for regression analysis of zero-inflated and bound-inflated data. *Comput Stat Data Anal* 2010; **54**: 1525–1534.
4. Moulton LH and Halsey NA. A mixture model with detection limits for regression analyses of antibody response to vaccine. *Biometrics* 1995; **51**: 1570–1578.
5. Taylor DJ, Kupper LL, Rappaport SM and Lyles RH. A mixture model for occupational exposure mean testing with a limit of detection. *Biometrics* 2001; **57**: 681–688.
6. Simpson DG, Ho M, Yang Y, Zhou J, Zachary JF and O'Brien WD. Excess risk thresholds in ultrasound safety studies: Statistical methods for data on occurrence and size of lesions. *Ultrasound Med Biol* 2004; **30**: 1289–1295.
7. Chai HS and Bailey KR. Use of log-skew-normal distribution in analysis of continuous data with a discrete component at zero. *Stat Med* 2008; **27**: 3643–3655.
8. Welsh AH, Cunningham RB, Donnelly CF and Lindenmayer DB. Modelling the abundance of rare species: Statistical Models for counts with extra zeros. *Ecol Modell* 1996; **88**: 297–308.
9. Böhning D, Dietz E and Schlattmann P. The zero-inflated Poisson model and the decayed, missing and filled teeth index in dental epidemiology. *J R Stat Soc* 1999; **A162**: 195–209.
10. Dietz E and Böhning D. On estimation of the Poisson parameter in zero-modified Poisson models. *Comput Stat Data Anal* 2000; **34**: 441–459.
11. Mwalili SM, Lesaffre E and Declerck D. The zero-inflated negative binomial regression model with correction for misclassification: An example in caries research. *Stat Methods Med Res* 2008; **17**: 123–139.
12. Hsu C-H. A weighted zero-inflated Poisson model for estimation of recurrence of adenomas. *Stat Methods Med Res* 2007; **16**: 155–166.
13. Kelley ME and Anderson SJ. Zero inflation in ordinal data: Incorporating susceptibility to response through the use of a mixture model. *Stat Med* 2008; **27**: 3674–3688.
14. Wang H and Heitjan DF. Modeling heaping in self-reported cigarette counts. *Stat Med* 2008; **27**: 3789–3804.
15. Lambert D. Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics* 1992; **34**: 1–14.
16. van den Broek J. A score test for zero inflation in a Poisson distribution. *Biometrics* 1995; **51**: 738–743.
17. Ridout M, Hinde J and Demétrio CGB.. A score test for testing a zero-inflated Poisson regression model against zero-inflated negative binomial alternatives. *Biometrics* 2001; **57**: 219–223.
18. Hall DB and Berenhaut KS. Score tests for heterogeneity and overdispersion in zero-inflated Poisson and binomial regression models. *Can J Stat* 2002; **30**: 415–430.
19. Thas O and Rayner JCW. Smooth tests for the zero-inflated Poisson distribution. *Biometrics* 2005; **61**: 808–815.
20. Xie FC, Wei BC and Lin JG. Assessing influence for pharmaceutical data in zero-inflated generalized Poisson mixed models. *Stat Med* 2008; **27**: 3656–3673.
21. O'Brien WD, Yang Y, Simpson DG, et al. Threshold estimation of ultrasound-induced lung hemorrhage in adult rabbits and comparison of thresholds in mice, rats, rabbits and pigs. *Ultrasound Med Biol* 2006; **32**: 1793–1804.
22. Cragg JH. Some statistical models for limited dependent variables with application to the demand for durable goods. *Econometrica* 1971; **39**: 829–844.
23. Min Y and Agresti A. Random effect models for repeated measures of zero-inflated count data. *Stat Model* 2005; **5**: 1–19.
24. Zhou XH and Tu W. Comparison of several independent population means when their samples contain log-normal and possibly zero observations. *Biometrics* 1999; **55**: 645–651.
25. Li N, Elashoff DA, Robbins WA and Xun L. A hierarchical zero-inflated log-normal model for skewed responses. *Stat Methods Med Res* Prepublished September 24, 2008; DOI:10.1177/0962280208097372.
26. Agresti A. *Categorical data analysis*. Hoboken, NJ: Wiley-Interscience, 2002, pp.165–210.
27. Casella G and Berger RL. *Statistical inference*. Pacific Grove, CA: Duxbury/Thomson Learning, 2002.
28. Louis TA. Finding the observed information when using the EM algorithm. *J R Stat Soc* 1982; **B44**: 226–233.
29. R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org 2008.
30. McDowell A. From the help desk: hurdle models. *Stata J* 2003; **3**: 178–184.
31. Moulton LH and Halsey NA. A mixed gamma model for regression analyses of quantitative assay data. *Vaccine* 1996; **14**: 1154–1158.

Appendix

Proposition 1 *For a random variable that satisfies* (1), *the conditional mean* μ^c *is larger than the marginal mean* μ^m.

Proof To show $\mu^c > \mu^m$, note that $\mu^c = E(Y \cdot I(Y > L))/\Pr(Y > L)$ and $\mu^m = \Pr(Y = L) \cdot L + E(Y \cdot I(Y > L))$. It follows that

$$\begin{aligned}\mu^c - \mu^m &= \frac{\Pr(Y = L)}{\Pr(Y > L)} \cdot \big\{E(Y \cdot I(Y > L)) - \Pr(Y > L) \cdot L\big\}\\ &> \frac{\Pr(Y = L)}{\Pr(Y > L)} \cdot \big\{L - \Pr(Y > L) \cdot L\big\} \;\geq\; 0.\end{aligned}$$

Proposition 2 *For a semi-continuous random variable that satisfies* (1), *the conditional quantile* $Q^c(q)$ *exceeds the marginal quantile* $Q^m(q)$.

Proof When $\Pr(Y = L) \geq q$, $Q^m(q) = L < Q^c(q)$. If $\Pr(Y = L) < q$, then

$$\begin{aligned}\Pr(Y > Q^m(q)) &= \Pr(Y > Q^c(q))/\Pr(Y > L)\\ &> \Pr(Y > Q^c(q)).\end{aligned}$$

Proposition 3 *For a semi-continuous response variable that satisfies (1), the gradient vectors of the conditional mean (4) and marginal mean (5) are obtained without extra numerical integration than required by the means themselves* (*if any*).

Proof Define $g_1(\boldsymbol{\theta}) = \Pr(Z^* > L \mid \boldsymbol{x})$ and $g_2(\boldsymbol{\theta}) = E(Z^* \cdot I(Z^* > L) \mid \boldsymbol{x})$. Then the conditional mean (4) and marginal mean (5) of a LIM model can be expressed as

$$\begin{aligned}\mu^c_{\text{LIM}}(\boldsymbol{\theta}) &= g_2(\boldsymbol{\theta})/g_1(\boldsymbol{\theta}),\\ \mu^m_{\text{LIM}}(\boldsymbol{\theta}) &= \big\{1 - \pi \cdot g_1(\boldsymbol{\theta})\big\}L + \pi \cdot g_2(\boldsymbol{\theta}).\end{aligned}$$

For the rest of the proof we suppress the dependence of $\mu^c_{\text{LIM}}(\boldsymbol{\theta})$, $\mu^m_{\text{LIM}}(\boldsymbol{\theta})$, $g_1(\boldsymbol{\theta})$ and $g_2(\boldsymbol{\theta})$ on $\boldsymbol{\theta}$ for shorthand notation.

It is clear that the derivatives $(\partial/\partial t)\mu^c_{\text{LIM}}$ and $(\partial/\partial t)\mu^m_{\text{LIM}}$ involve only π, g_j and $(\partial/\partial t)g_j$ for $t = \pi$, μ or σ and $(\partial/\partial \pi)g_j = 0$ for $j = 1$, 2. Write $g_1 = 1 - F^*_s(c_1)$ and $g_2 = \mu g_1 + \sigma E\big(Z^*_0 \cdot I\big(Z^*_0 > c_1\big) \mid \boldsymbol{x}\big)$, where $c_1 = \frac{L-\mu}{\sigma}$ and $Z^*_0 = \frac{Z^*-\mu}{\sigma} \sim F^*_s$. Then

$$\begin{aligned}(\partial/\partial\mu)\, g_1 &= 1/\sigma \cdot f^*_s(c_1),\\ (\partial/\partial\sigma)\, g_1 &= c_1/\sigma \cdot f^*_s(c_1),\\ (\partial/\partial\mu)\, g_2 &= g_1 + L/\sigma \cdot f^*_s(c_1),\\ (\partial/\partial\sigma)\, g_2 &= (g_2 - \mu g_1)/\sigma + c_1 L/\sigma \cdot f^*_s(c_1).\end{aligned}$$

Thus, the only integration required for the delta method confidence intervals for μ^c_{LIM} and μ^m_{LIM} is that required for g_1 and g_2. This quantity is available explicitly for the models included in Table 1, in which case no numerical integration is necessary.