HEALTH ECONOMICS
Health Econ. **22**: 673–686 (2013)
Published online 24 May 2012 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/hec.2844

CONSISTENT ESTIMATION OF ZERO-INFLATED COUNT MODELS

KEVIN E. STAUB[a] and RAINER WINKELMANN[a,b,c,*]

[a] *University of Zurich, Zürich, Switzerland*
[b] *CESifo, Munich, Germany*
[c] *IZA, Bonn, Germany*

ABSTRACT

Applications of zero-inflated count data models have proliferated in health economics. However, zero-inflated Poisson or zero-inflated negative binomial maximum likelihood estimators are not robust to misspecification. This article proposes Poisson quasi-likelihood estimators as an alternative. These estimators are consistent in the presence of excess zeros without having to specify the full distribution. The advantages of the Poisson quasi-likelihood approach are illustrated in a series of Monte Carlo simulations and in an application to the demand for health services. Copyright © 2012 John Wiley & Sons, Ltd.

Received 8 August 2011; Revised 23 March 2012; Accepted 24 April 2012

KEY WORDS: excess zeros; Poisson; logit; unobserved heterogeneity; misspecification

1. INTRODUCTION

The so-called problem of "excess zeros" plagues a majority of count data applications in health economics and other social sciences. The proportion of observations with zero counts in the sample is often much larger than that predicted by standard count models, be it the Poisson or the negative binomial model.[1] The most common response is to think of the data-generating process in terms of a discrete mixture of a count random variable and a degenerate random variable with unit probability mass at zero. For instance, consider the demand for health services as measured by the number of physician visits. A person might have had zero physician visits during a given period because (i) she is a follower of alternative medicines and never visits a doctor or because (ii) she visits doctors in principle but by chance did not do so during the observed period. Type i zeros—sometimes called *structural* or *strategic*—results from a binary process, whereas type ii zeros—sometimes called *incidental*—are a realization of a count process to which only the "population at risk" is subjected.

These models allowing for two separate types of zeros are known as zero-inflated count models (Mullahy, 1986, Lambert, 1992), the most prominent ones being the zero-inflated Poisson (ZIP) and zero-inflated negative binomial (ZINB) models. In health economics, they have been used, among others, for the number of physician visits (Yen *et al*., 2001; Sarma and Simpson, 2006; Sari, 2009; Pizer and Prentice, 2011), the number of pharmacy visits (Chang and Trivedi, 2003), the number of prescriptions (Street *et al*., 1999), the number of occupational injuries (Campolieti, 2002), and the number of cigarettes smoked (Sheu *et al*., 2004; Bauer *et al*., 2007).

There are two ways to estimate the parameters of zero-inflated count data models. The standard way, pursued by all of the cited literature, is based on full maximum likelihood (ML) estimation. The alternative

*Correspondence to: Department of Economics, University of Zurich, Zürichbergstr. 14, CH-8032, Zürich, Switzerland. E-mail: rainer.winkelmann@econ.uzh.ch

[1] A formal definition is given in Section 2.1.

is to focus on the first moment and either embed it in a linear exponential family distribution and estimate the parameters by quasi-ML or use a minimum distance type estimator such as nonlinear least squares (NLS) or generalized methods of moments (GMM). The purpose of this article is to discuss the implementation of this alternative approach in detail, including its strengths and weaknesses. Specifically, we propose a Poisson quasi-likelihood (PQL) estimator that is robust to misspecification, as it estimates the regression parameters consistently regardless of the true distribution of the counts. A series of Monte Carlo experiments and an application show that PQL estimation is a promising alternative to ML estimation in moderate and large samples, avoiding sizeable biases that can potentially affect ML estimators.

The next section reviews models for zero-inflated count data. The ML and the quasi-likelihood estimation of zero-inflated models are discussed in Section 3. Section 4 presents Monte Carlo simulation results comparing the PQL estimator with the ML estimators. Section 5 illustrates the PQL estimator with logit zero inflation in an application modeling the frequency of doctor visits. Section 6 concludes the article.

2. ECONOMETRIC MODELS

2.1. Zero-inflated count data models

Zero-inflated count data models have probability function

$$f(y) = \begin{cases} \pi + (1-\pi)g(0) & \text{for } \; y = 0 \\ (1-\pi)g(y) & \text{for } \; y = 1, 2, 3, \ldots \end{cases} \tag{1}$$

where y is a count-valued random variable, $\pi \in [0, 1]$ is a zero-inflation parameter (the probability of a strategic zero), and $g(\cdot)$ is the probability function of the parent count model. Excess zeros, or zero inflation, occur by definition whenever $\pi > 0$. The mean of the zero-inflated count data model is

$$\mathrm{E}(y) = \sum_{k=1}^{\infty} (1-\pi)g(k) = (1-\pi)\mathrm{E}_g(y) \tag{2}$$

where $\mathrm{E}_g(y)$ denotes the mean of the parent distribution. A fully parametric zero-inflated count data model is obtained once the probability function of the parent count model is specified. For example, the ZIP model is obtained for

$$g(y;\lambda) = \frac{\exp(-\lambda)\lambda^y}{y!}, \lambda > 0 \tag{3}$$

with mean $\mathrm{E}_g(y) = \lambda$ and $\mathrm{E}(y) = (1-\pi)\lambda$. The main alternative to the ZIP model is the zero-inflated negative binomial model, which has the same mean as the ZIP but overdispersion in the count part of the model. Both λ and π can be parameterized in terms of exogenous explanatory variables. The standard assumptions are that

$$\lambda(x) = \mathrm{E}_g(y|x) = \exp\left(x'\beta\right) \tag{4}$$

and

$$\pi(z) = \frac{\exp\left(z'\delta\right)}{1+\exp(z'\delta)} \tag{5}$$

where z can be identical to x, overlap with x, or be completely distinct from x. The conditional expectation function (CEF) of the corresponding zero-inflated count data model is given by

$$\mathrm{E}(y|x,z) = (1-\pi(z))\lambda(x) = \frac{\exp\left(x'\beta\right)}{1+\exp(z'\delta)} \tag{6}$$

Importantly, this is the CEF of any zero-inflated count data model, not only the ZIP model, as long as Equations (4) and (5) hold. In addition, Equation (6) is fairly general, in the sense that it can capture several departures from standard count data models other than zero inflation. For example, Winkelmann and Zimmermann (1993) derived an expression such as Equation (6) from a model of underreporting (in their model of "binomial thinning," y is Poisson distributed with double indexed mean). Kim and Lee (2011) obtained Equation (6) from a hurdle-at-zero model, where $1-\pi(z)$ is the probability of having a positive count, and the CEF condition on passing the hurdle is modeled as an exponential function: $\mathrm{E}(y|x, y>0)=\exp(x'\beta)$.

2.2. Parameters of interest

In health economic applications such as the ones cited in Section 1, the key objects of interest are predictions of the CEF as well as its derivatives. The derivative with respect to a variable w, which is an element of both vectors x and z, is given by

$$\frac{\partial \mathrm{E}(y|x,z)}{\partial w}=\frac{\exp\left(x'\beta\right)}{1+\exp(z'\gamma)}\beta_w-\frac{\exp\left(x'\beta\right)\exp\left(z'\gamma\right)}{\left(1+\exp(z'\gamma)\right)^2}\delta_w$$

where β_w and δ_w are the elements in the vectors β and δ corresponding to w. Hence, we obtain a relatively simple expression for the semi-elasticity, $[\partial\,\mathrm{E}(y|x,z)/\mathrm{E}(y|x,z)]/\partial\,w=\beta_w-\pi(z)\delta_w$.

Under the maintained assumption of a zero-inflated data-generating process, one can deduce two further quantities, namely, the effect of a regressor on the CEF of the parent model and the effect of a regressor on the probability of an excess zero. More specifically, the parameters β and δ provide the semi-elasticities of the parent model and the changes in the log-odds of strategic zeros, respectively,

$$\frac{\partial \mathrm{E}_g(y|x)/\mathrm{E}_g(y|x)}{\partial x}=\beta_x \qquad \frac{\partial log[\pi(z)/(1-\pi(z))]}{\partial z}=\delta_z$$

The estimation of these parameters of interest in general does not require the specification of a full parametric distribution because they are identified from the first moment of the model alone.

3. ESTIMATION

3.1. Maximum likelihood estimation

The log-likelihood function of the ZIP model for a sample of n independent observation tuples (y_i, x_i, z_i) is

$$\begin{aligned}\ln l^{\mathrm{ZIP}}=\sum_{i=1}^{n}&\mathbb{1}(y_i=0)\ln[\exp(z_i\delta)+\exp(-\exp(x_i\beta))]\\&+\mathbb{1}(y_i>0)[-\exp(x_i\beta)+y_ix_i\beta]-\ln(1+\exp(z_i\delta))\end{aligned} \tag{7}$$

Because the model has a finite mixture structure, the maximization of the log-likelihood function can use the EM algorithm, although direct maximization using Newton–Raphson is possible as well. Alternative estimation algorithms are discussed by Hall and Cheng (2010). If the model is correctly specified, ML theory ensures that these estimators are consistent and asymptotically efficient, provided they exist (Cameron and Trivedi, 1998; Winkelmann, 2008).

3.2. Moment-based estimation

The parameters β and δ can also be estimated directly from the conditional moment restriction (6). Such an approach is in principle preferable, because it makes fewer assumptions regarding the data-generating process than ML estimation. These additional assumptions, if violated, will invalidate ML inference but not moment-based inference. Moment-based estimators are thus more robust.

DOI: 10.1002/hec

Identification based on Equation (6) raises two issues. First, if z has a constant only, we obtain a model with constant zero inflation. In this case, the CEF of the zero-inflated model is given by

$$\mathrm{E}(y|x) = (1-\pi)\lambda(x) = \exp\left(\ln(1-\pi) + x'\beta\right) \tag{8}$$

and it is not possible to separately identify π and the constant in the parent model. Hence, the share of strategic zeros is not identified. However, most applied work focuses anyway on semi-elasticities (overall and in the parent model) and CEF predictions, and all of those can be obtained from Equation (8).

Second, assume $x = z$, that is, all variables enter the zero-inflation part as well as the parent process. In this case, two parameter vectors lead to the same CEF (see Papadopoulos and Santos Silva, 2008):

$$\mathrm{E}(y|x,z) = \frac{\exp\left(x'\beta_1\right)}{1+\exp(x'\delta_1)} = \frac{\exp\left(x'\beta_2\right)}{1+\exp(x'\delta_2)}$$

for $\beta_2 = \beta_1 + \delta_1$ and $\delta_2 = -\delta_1$. Thus, the estimation problem has two solutions. In practice, this identification problem can be overcome if the sign of at least one element in δ is known. Third, in the case where x and z differ, identification is achieved if there is at least one variable in z that is not included in x.

To implement moment-based estimators in such a just-identified case, several approaches are possible. We suggest to embed the CEF into a standard Poisson model, an application of quasi-likelihood estimation, which leads to consistent estimates and, as we will show in Monte Carlo simulations, has also good finite sample properties. Alternatively, one can solve the sample analogues of unconditional moment restrictions implied by Equation (6), such as orthogonality between the CEF errors and the functions of the regressors:

$$\mathrm{E}\left[y_i - \frac{\exp(x_i'\beta)}{1+\exp(z_i'\delta)}\right]h_1(x_i,z_i) = 0 \qquad \text{and} \qquad \mathrm{E}\left[y_i - \frac{\exp(x_i'\beta)}{1+\exp(z_i'\delta)}\right]h_2(x_i,z_i) = 0,$$

where the functions $h_1(x_i, z_i)$ and $h_2(x_i, z_i)$ are instruments. In general, unweighted orthogonality conditions between regressors and CEF errors are insufficient to identify β and δ. For instance, if $h_1(x, z) = x$, $h_2(x, z) = z$, and $x = z$, there are fewer moment conditions than parameters, and the model is underidentified.

As we will show in the next section, the first-order conditions of the quasi-likelihood estimator imply weighted orthogonality condition between the regressors and the CEF errors, $h_1(x, z) = w_1(x, z)x$ and $h_2(x, z) = w_2(x, z)z$. In this just identified case, the optimal weights depend on higher-order moments of the model, which are left unspecified. However, the Monte Carlo study here (and related Monte Carlo results elsewhere; see Santos Silva and Tenreyro 2006, 2011) suggest that the PQL estimators performs well in general.

3.3. Quasi-ML

Quasi-ML estimation is based on distributions within the linear exponential family (LEF), whose probability function can be written as (Gourieroux *et al*., 1984a)

$$f^{\mathrm{LEF}}(y|\mu_x) = \exp\{a(\mu_x) + b(y) + c(\mu_x)y\}, \tag{9}$$

where $\mu_x = \mu(x\,;\beta) = \mathrm{E}(y|x)$ and $\mu_x = (\partial\, a(\mu_x)/\partial\,\mu_x)/(\partial\, c(\mu_x)/\partial\,\mu_x)$. LEFs have the property that the score function can be written as

$$\frac{\partial \log f(y|x)}{\partial\beta} = (y-\mu_x)h(x) \tag{10}$$

where $h(x) = (\mathrm{d}c(\mu_x)/\mathrm{d}\mu_x)(\partial\,\mu_x/\partial\,x)$. Suppose the true model is $g_0(y|x) \neq f(y|x)$ but $\mathrm{E}_0(y|x) = \mu_x$ for some value β_0. Thus, the CEF is correctly specified. In this case, the expectation of Equation (10) at the true density is zero, although the model is misspecified because the CEF residual $y - \mathrm{E}(y|x)$ is mean independent of x and thus has zero covariance with any function $h(x)$. As the empirical score converges to the expected score by the law of large numbers, the solution to the ML first-order conditions converges in probability to the true CEF parameters (see also White, 1982; Gourieroux *et al.*, 1984b).

The Poisson distribution is an LEF member with $a(\mu_x) = -\mu_x$, $b(y) = -\ln(y\,!)$ and $c(\mu_x) = ln(\mu_x)$. Therefore, although the data are zero inflated, a Poisson regression yields valid estimates of the objects of interest as long as the CEF is correctly specified. Valid standard errors require the usual White adjustment to the covariance matrix. The PQL estimator for the model with nonconstant zero inflation is obtained by maximizing

$$ql(\beta,\delta) = \sum_{i=1}^{n} y_i \ln\tilde{\lambda}(x_i, z_i, \beta, \delta) - \tilde{\lambda}(x_i, z_i, \beta, \delta) \tag{11}$$

where $\tilde{\lambda}(x_i, z_i, \beta, \delta) = \exp(x_i\beta)/(1 + \exp(z_i\delta))$. The first-order conditions are

$$\frac{\partial ql(\beta,\delta)}{\partial\beta} = \sum_{i=1}^{n}\left(y_i - \frac{\exp(x_i\beta)}{1+\exp(z_i\delta)}\right)x_i = 0$$

and

$$\frac{\partial ql(\beta,\delta)}{\partial\delta} = -\sum_{i=1}^{n}\left(y_i - \frac{\exp(x_i'\beta)}{(1+\exp(z_i'\delta))}\right)\frac{\exp(z_i'\delta)}{1+\exp(z_i'\delta)}z_i = 0$$

Solving the nonlinear first-order conditions using the Newton–Raphson or related algorithms is relatively straightforward (Stata code is provided in Appendix A). This estimator for zero-inflated count data is consistent even if the true data-generating process is not Poisson distributed—as is by definition the case with excess zeros. The gain of PQL estimation relative to the ML estimation of fully parametric zero-inflated count models is robustness to misspecification. The main cost of PQL estimation relative to the ML estimation of a correctly specified model is a loss of precision. In the next section, we explore both aspects, relative bias and efficiency loss of PQL relative to ML, for varying sample sizes using Monte Carlo simulations.

4. MONTE CARLO EVIDENCE

4.1. Simulation design

To compare the performance of the PQL estimator with its main competitors, the ZIP and the ZINB ML estimators, we create three setups. All of them are obtained from the following basic experimental design. The count dependent variable y is specified as

$$y = \begin{cases} 0 & \text{with probability } \pi(z) \\ y^* & \text{with probability } 1 - \pi(z) \end{cases}$$

where $y^* | \, x, v \sim \text{Poisson}(\lambda(x, v))$, and $\lambda(x, v)$ and $\pi(z)$ are given by

$$\lambda(x, v) = \exp(\beta_0 + \beta x + v), \quad \pi(z) = \frac{\exp(\delta_0 + \delta z)}{1 + \exp(\delta_0 + \delta z)} \tag{12}$$

with $x = q_1$ and $z = (q_1, q_2)$. The scalar regressors q_1 and q_2 follow χ^2 distributions with one degree of freedom; q_1 is rescaled to have a variance of 0.1. Thus, the data-generating process contains two regressors, one of which is excluded from the count part. The focus is on the estimation of β and $\delta = (\delta_1, \delta_2)'$, all three of which are set to 1. The parameter β_0 is set to -0.5, which ensures a low mean of the parent count process with a substantial fraction of incidental zeros (~45%). The degree of zero inflation is controlled by δ_0. All simulation experiments are run for two levels of zero inflation, 10% and 50%. These values are chosen to reflect the range of percentages of observations with $y = 0$ typically encountered in applications where zero-inflated models are used. With 10% zero inflation, the total fraction of zeros in the data is approximately 50%; with 50% zero inflation, it is

DOI: 10.1002/hec

approximately 75%. ZI models are unlikely to be of use if the proportion of zeros in the data is higher.[2] To obtain 10% zero inflation, δ_0 is set equal to -4.2; a value of $\delta_0 = -1.1$ results in 50% zero inflation.

In addition, the parent CEF in Equation (12) contains a random component v, which is distributed as Normal $(-0.5\sigma^2, \sigma^2)$. The parent distribution of y^* given x but marginalized over v is then the Poisson log-normal distribution (e.g. see Winkelmann, 2008) with variance function

$$\mathrm{Var}(y^*|x) = \exp(\beta_0 + \beta x) + [\exp(\beta_0 + \beta x)]^2 \left(e^{\sigma^2} - 1\right)$$

Different variance functions for y^* can be obtained by specifying σ^2 as function of x. Letting

$$\sigma^2(x) = \ln\{1 + c\exp[(k-1)(\beta_0 + \beta x)]\}$$

the parameter k controls the nonlinearity of the variance function, whereas c is a free overdispersion parameter. Such a parameterization allows to explore variance functions quite freely without abandoning the normal distribution assumption of v nor the exponential CEF of the parent process.

We use three setups. First, we set $c = 0$ and $k = 0$; it follows that $\sigma^2 = 0$ (no unobserved heterogeneity), and the data-generating process is indeed ZIP with $\lambda = \exp(-0.5 + x)$ and zero inflation of 10% or 50%. This first setup allows us to compare the efficiency loss of PQL relative to the correctly specified and thus asymptotically efficient ZIP ML estimator.

A second setup is obtained for $c = \exp(1) - 1$ and $k = 1$. It follows that $\sigma^2 = 1$, and there is quadratic overdispersion. Although both ZIP and ZINB are misspecified (because the true parent process is Poisson log-normal), we expect the ZINB to behave quite satisfactorily as the misspecification is limited to higher-order moments, not mean and variance. The ZIP model by contrast assumes equality of mean and variance and is thus unlikely to produce good results. The PQL estimator is robust to this kind of misspecification and should work well.

In our third setup, we set $c = 2$ and $k = -1$, implying a variance function with additive constant

$$\mathrm{Var}(y^*|x) = \mathrm{E}(y^*|x) + 2$$

The corresponding variance-to-mean ratio is now hyperbolic. In this case, all three estimators—ZIP, ZINB, and PQL—only specify the first moment correctly. This should not matter for PQL but lead to bias for ZIP as well as ZINB.

An all cases, two sample sizes were considered (5000 and 50,000 observations, respectively). The number of replications was 10,000 for data-generating processes with 5000 observations, and 1000 replications for those with 50,000 observations. The Monte Carlo study was programmed in STATA/MP 11.1; code is available from the authors on request.

4.2. Results

The results of the three simulation setups are displayed in Table I, which is divided into three panels, each presenting results for one of the three setups. Following the focus in the literature, we concentrate on the main parameters of interest, the semi-elasticity of the parent process β, and the change in the log-odds of strategic zeros, δ_1 and δ_2, whose true values are 1. The main entries in Table I are the mean of the QL and ML estimates $\hat{\beta}$, $\hat{\delta}$, and δ_2 over replications. The numbers in parentheses give standard deviations.

The left-hand panel titled "No overdispersion" shows the results for the first setup in which the data-generating process is a ZIP model. The first row of results is for the ZIP ML estimator on samples of 5000 observations. The ZIP estimates of $\hat{\beta}$ are very close to the true value on average, regardless of whether the degree of zero inflation is 10% or 50%. Higher degrees of zero inflation imply less information from which to estimate β, and so the standard deviation is higher with 50% zero inflation. The opposite is the case with $\delta = (\delta_1, \delta_2)$. Low degrees of zero inflation imply having to identify the effects of the variables in z on strategic zeros with little information, and so δ is estimated less precisely. PQL estimates β quite well too, although its

[2] In such cases, it may be advisable to consider estimating a binary model.

DOI: 10.1002/hec

Table I. Monte Carlo results

	No overdispersion						Quadratic overdispersion						Additive overdispersion					
	10% zero inflation			50% zero inflation			10% zero inflation			50% zero inflation			10% zero inflation			50% zero inflation		
Estimator	$\hat{\beta}$	δ_1	δ_2	$\hat{\beta}$	δ_1	δ_2	$\hat{\beta}$	δ_1	δ_2	$\hat{\beta}$	δ_1	δ_2	$\hat{\beta}$	δ_1	δ_2	$\hat{\beta}$	δ_1	δ_2
$n=5000$																		
ZIP	1.000	1.024	1.015	1.001	1.008	1.005	0.957	0.353	0.628	0.935	0.781	0.899	0.919	0.057	0.562	0.861	0.600	0.867
	(0.021)	(0.189)	(0.099)	(0.048)	(0.118)	(0.067)	(0.044)	(0.142)	(0.061)	(0.071)	(0.110)	(0.061)	(0.028)	(0.132)	(0.051)	(0.055)	(0.095)	(0.058)
ZINB							0.997	1.070	1.030	0.999	1.043	1.031	0.980	0.913	0.964	0.921	0.854	1.001
							(0.033)	(0.223)	(0.944)	(0.069)	(0.153)	(0.088)	(0.028)	(0.343)	(1.455)	(0.060)	(0.159)	(0.094)
PQL	0.970	1.248	1.216	1.050	1.117	1.087	0.917	1.249	1.250	1.024	1.093	1.094	0.960	1.259	1.234	1.078	1.140	1.088
	(0.532)	(0.941)	(0.565)	(0.591)	(0.745)	(0.202)	(1.090)	(1.521)	(0.697)	(1.024)	(1.156)	(0.215)	(0.576)	(1.320)	(0.830)	(2.187)	(2.233)	(0.215)
$n=50{,}000$																		
ZIP	1.000	1.005	1.002	1.000	1.000	1.001	0.959	0.349	0.623	0.938	0.776	0.896	0.922	0.057	0.558	0.867	0.598	0.864
	(0.006)	(0.059)	(0.031)	(0.015)	(0.037)	(0.022)	(0.017)	(0.044)	(0.019)	(0.026)	(0.033)	(0.020)	(0.010)	(0.039)	(0.016)	(0.019)	(0.030)	(0.018)
ZINB							0.997	1.048	1.042	1.002	1.032	1.024	0.982	0.887	1.027	0.923	0.837	0.988
							(0.010)	(0.067)	(0.037)	(0.022)	(0.045)	(0.027)	(0.008)	(0.090)	(0.047)	(0.019)	(0.047)	(0.027)
PQL	0.983	1.038	1.038	0.999	1.017	1.021	0.967	1.029	1.043	1.000	1.016	1.021	0.986	1.046	1.040	0.999	1.013	1.019
	(0.194)	(0.335)	(0.122)	(0.127)	(0.216)	(0.067)	(0.367)	(0.500)	(0.139)	(0.146)	(0.250)	(0.072)	(0.177)	(0.312)	(0.119)	(0.143)	(0.237)	(0.069)

Note. Entries are the average estimates over replications. Standard deviations in parentheses. True values: $\beta=\delta_1=\delta_2=1$; 10,000 replications for $n=5000$ and 1000 for $n=50{,}000$.

DOI: 10.1002/hec

precision is approximately one order of magnitude lower than ZIP. The Monte Carlo results suggest, however, that low levels of zero inflation can make the estimation of δ quite difficult for PQL with 5000 observations. In the corresponding entries of Table I, biases are approximately 25%. Higher levels of zero inflation visibly mitigate this problem, although biases are still approximately 10%. With 50,000 observations, the performance of PQL improves substantially. Although for β the picture is the same as with the smaller sample, PQL now also obtains satisfactory estimates of δ. However, some finite sample bias is still visible and the efficiency loss relative to ZIP is still quite large.

The middle panel ("Quadratic overdispersion") contains results obtained under the second setup where unobserved heterogeneity is causing the parent model to exhibit quadratic overdispersion. As ZINB correctly specifies the CEF and the variance function, the panel additionally includes results from this estimator. The pattern for the ZINB ML estimates on 5000 observation samples echoes the one for the previous ZIP ML: Although β is estimated quite precisely and free of bias, estimates of δ are more noisy, especially in the low zero-inflation case. Increasing the sample size to 50,000 improves ZINB's performance. For instance, the bias in δ_1, which is 7% and 4.3% for low and high zero inflation, respectively, is down to 4.8% and 3.2%. PQL's performance is quite remarkable here. Although with 5000 observations the same large biases as in the previous setup are visible, in larger sample sizes the corresponding biases are smaller than those of the ZINB (2.9% and 1.6%). The estimates for β are quite similar for PQL and ZINB. However, the standard deviations of the PQL are approximately one order of magnitude larger than those of the ZINB. The inconsistency of ZIP is reflected in substantial biases in all reported mean estimates.

In the right-hand panel, the data are drawn from a process with additive overdispersion of y^*, so that both ZIP and ZINB only specify the CEF correctly. ZIP estimation again yields estimators that are not consistent for the true value of β in any of the entries of Table I. The biases in the estimated δ are very large, ranging up to 40% for δ_1. Moreover, their persistency in the larger sample size unmasks them as asymptotic biases. The ZINB estimator does not work well either. With 50,000 observations and 10% zero inflation, the average estimate for δ_1 is biased downward by 11.3%. With 50% zero inflation, the bias for the same parameter is more than 15%, and in addition, β displays a bias of 7.7%. By contrast, the performance of PQL is much better throughout. PQL's corresponding biases are only 1.3% for δ_1 and 0.1% for β. Indeed, a look at PQL's results across the three panels shows that the presence and form of overdispersion bears no effect on its performance.

As mentioned in Section 3.2, many consistent moment-based estimators can be obtained as alternatives to PQL estimation which uses the functions $h_1(x,z)=x$ and $h_2(x,z)=\pi(z)z$ as instruments. For instance, NLS estimation minimizes

$$\sum_{i=1}^{n}\left(y_i-\frac{\exp(x_i'\beta)}{(1+\exp(z_i'\delta))}\right)^2$$

leading to $h_1(x,z)=[1-\pi(z)]\lambda(x)x$ and $h_2(x,z)=\pi(z)[1-\pi(z)]\lambda(x)z$ as instruments; that is, it weights PQL's moment conditions with the CEF, $[1-\pi(z)]\lambda(x)$. However, this estimator does not work well for the used data-generating processes. Appendix B contains additional simulation results for this estimator in samples with 50,000 and 500,000 observations. Its results compare very poorly with PQL.[3] An explanation for this is that the CEF weighting implied by NLS down weights the observations with high probability of excess zeros (low CEF), thus drastically reducing the information from which to estimate δ. Accordingly, the largest biases of the NLS estimator are for δ and in setups with few excess zeros.

4.3. Further results without exclusion restrictions

In the simulations of Table I, q_2, the regressor excluded from x, provides independent variation to the zero-inflation process, an ideal setting. In applications, exclusion restrictions may often not be justifiable. To address this

[3] In addition to mean and standard deviation, the median is reported in Appendix B because mean estimates of NLS for $n=50,000$ are severely distorted by outliers. For the other estimators discussed, mean and median are similar.

DOI: 10.1002/hec

issue, we repeated the simulations setting $\delta_2 = 0$, that is, ridding the data-generating processes from the additional regressor. The result is a specification with one regressor that enters both parts of the model. As discussed before, the PQL estimator has now two solutions. To choose between them, we use the out-of-sample information that in the data-generating process $\delta_1 = 1 > 0$ and always select the solution where the estimated coefficient $\hat{\delta}_1 > 0$.[4]

The results are not that different from those in the previous section. With 5000 observations, the ZIP ML estimation of the ZIP model is adequate (left-hand panel, "No overdispersion"). Likewise, the ZINB ML estimation of the quadratic overdispersion process (middle panel) is comparable with that in Table I. However, at $n = 5000$, PQL estimation struggles with substantial finite sample bias and large standard deviations. As we will illustrate with the application in the next section, however, PQL estimation with such sample sizes may not be problematic if additional regressors are available: variation from more regressors can help estimating the parameters more precisely.

When passing to the results corresponding to 50,000 observations, the improvement in PQL's performance is noteworthy. In settings with 10% zero inflation, PQL's bias never exceeds 2%. In contrast, ZINB displays biases of up to 4% in the quadratic overdispersion process, and up to 17.6% in the additive overdispersion setup (right panel), ZIP performs even worse. With high zero inflation, some bias remains in PQL's estimates, but its magnitude is visibly smaller than those of misspecified ZIP and ZINB ML estimation.

To summarize, the results from the Monte Carlo experiments in this section demonstrate the robustness of the PQL estimator in zero-inflated, finite samples and the biases that can arise when using its two most common ML competitors.

Table II. Further Monte Carlo results: no exclusion restriction on regressors

	No overdispersion				Quadratic overdispersion				Additive overdispersion			
	10% zero inflation		50% zero inflation		10% zero inflation		50% zero inflation		10% zero inflation		50% zero inflation	
Estimator	$\hat{\beta}$	$\hat{\delta}_1$	$\hat{\beta}$	$\hat{\delta}_1$	$\hat{\beta}$	$\hat{\delta}_1$	$\hat{\beta}$	$\hat{\delta}_1$	$\hat{\beta}$	$\hat{\delta}_1$	$\hat{\beta}$	$\hat{a}_1$
$n = 5000$												
ZIP	0.998 (0.048)	1.010 (0.095)	0.996 (0.133)	1.005 (0.074)	1.026 (0.060)	0.643 (0.064)	0.955 (0.150)	0.882 (0.065)	0.932 (0.059)	0.485 (0.053)	0.788 (0.148)	0.770 (0.058)
ZINB					0.982 0.064	1.051 0.122	0.979 0.159	1.037 0.103	0.889 0.064	1.193 0.161	0.815 0.165	1.014 0.123
PQL	1.001 (0.279)	1.157 (0.537)	1.914 (5.054)	1.828 (2.002)	1.002 (0.312)	1.206 (0.934)	2.004 (5.876)	1.933 (2.275)	1.006 (0.293)	1.177 (0.792)	2.300 (7.712)	1.955 (2.712)
$n = 50{,}000$												
ZIP	1.000 (0.015)	1.003 (0.030)	1.001 (0.041)	1.001 (0.023)	1.026 (0.019)	0.637 (0.020)	0.962 (0.047)	0.880 (0.020)	0.932 (0.018)	0.480 (0.016)	0.795 (0.046)	0.768 (0.018)
ZINB					0.983 0.020	1.040 0.036	0.989 0.049	1.028 0.031	0.890 0.020	1.176 0.047	0.825 0.052	0.997 0.036
PQL	0.997 (0.084)	1.017 (0.071)	1.060 (0.853)	1.063 (0.170)	0.998 (0.092)	1.019 (0.078)	1.075 (0.914)	1.074 (0.182)	0.997 (0.084)	1.018 (0.072)	1.080 (0.886)	1.067 (0.183)

Note. Entries are the average estimates over replications. Standard deviations in parentheses. True values: $\beta = \delta_1 = 1$.

5. ILLUSTRATION: DEMAND FOR PHYSICIAN SERVICES

We illustrate the PQL estimation of a count model with logit zero inflation in a well-known health economics application. In particular, the goal is to estimate how health insurance affects the frequency of doctor visits. The

[4] For all estimators, we used the true values of the parameters as starting values for the optimization. The fraction of $\hat{\delta}_1$ estimated by PQL that were negative was essentially zero.

DOI: 10.1002/hec

data set is identical to the one used by Cameron and Trivedi (1986). The sample of 5190 individuals is extracted from the Australian Health Survey 1977–1978. The dependent variable is the number of consultations with a doctor or specialist in the 2-week period before the interview. The mean is 0.302, and the variance is 0.637. Further details, and a motivation of the selection of explanatory variables, are given by Cameron and Trivedi (1986) and the references quoted therein (Table II).

Regressors include demographics (sex, age, and age squared), income, various measures of health status (number of reduced activity days, general health questionnaire score, recent illness, chronic condition 1, and chronic condition 2), and three types of health insurance coverage (Levyplus, Freepoor, and Freerepat—the former representing a higher level of coverage and the latter two a basic level supplied free of charge).

Table III contains the regression results for the PQL estimator (in the first two columns) as well as for the fully parametric ZIP (in columns 3 and 4) and ZINB (in columns 5 and 6) models. In each case, the same regressors enter the logit model for zero inflation and the log-linear CEF of the parent model. As discussed earlier, this means that the PQL estimator has two solutions. To identify the correct one, we must rely on out-of-sample information: we surmise that individuals having experienced illnesses recently are more likely to be part of the population in demand for doctor visits, which implies that the sign of the parameter on "illness" is negative—that is, reducing the probability of an extra zero—and report this set of estimates as δ.[5]

There is no significant difference in the magnitude of standard errors across models—the standard errors of the ZIP are smaller than those of the ZINB and the PQL, the latter two being roughly similar. Thus, the precision of PQL estimation should be fine although there are no exclusion restrictions.

A likelihood ratio test between ZIP and ZINB clearly favors the latter.[6] Although this is an indication of the presence of unobserved heterogeneity and overdispersion, it does not mean, however, that the ZINB is the "right" model. If the overdispersion is misspecified, the ZINB estimator is inconsistent, regardless of fitting the data better than ZIP.

It is reassuring, therefore, that the parameter estimates are quite insensitive to the choice of specification in many instances, but there are exceptions. For instance, the ZINB model detects no statistically significant effect of having a chronic health condition in either part of the model. Under PQL, the second indicator has large negative and statistically significant effect on the probability of an extra zero and thus increases the expected number of visits. Inferences from PQL and ZINB also differ regarding insurance status. "Freepoor" and "Levyplus" are statistically significant in the ZINB but not so in the PQL model, suggesting some caution in interpreting these effects.

6. CONCLUDING REMARKS

The main quantities of interest in most count data applications are the CEF, changes in the probability of strategic zeros, and semi-elasticities of the parent count model with respect to some regressors. For instance, all applications cited in Section 1 without exception limited the discussion of their estimation results in the CEF and such effects. This article proposed a new approach based on PQL estimation as a way to estimate these quantities without having to specify more than the CEF, as opposed to the full distribution as is necessary with the traditional ZIP and ZINB ML estimators.

The key advantage of using PQL over ZIP and ZINB is its robustness to misspecification. Given the pervasive uncertainty about the data-generating processes in practice, using estimators for ZI models seems unwise if concerns about bias from higher-order misspecification exist. The relatively mild misspecifications of the DGP presented in the Monte Carlo experiments frequently resulted in noticeable biases, suggesting that PQL may be the better choice for estimating ZI models compared with ZI ML estimators in the absence of strong *a priori* information about the DGP. This conclusion will be the more compelling the larger the data set at hand.

[5] If we are wrong, we are erroneously reporting PQL estimates of $-\delta$ for the ZI part (and of $\beta+\delta$ for the parent process).

[6] Because the test involves a null hypothesis at the boundary at the parameter space, the likelihood ratio test statistic has a nonstandard distribution, namely, a probability mass of 0.5 at 0 and a half $\chi^2_{(1)}$ distribution for positive values, see Chernoff (1954).

Table III. Zero-inflation models for number of doctor consultations ($n = 5190$)

	PQL		ZIP		ZINB	
Variable	ZI	Parent	ZI	Parent	ZI	Parent
Sex	−0.275 (0.228)	0.003 (0.135)	−0.488 ***	−0.027 (0.072)	−0.592 (0.228)***	0.010 (0.084)
Age × 10^{-2}	8.864**	3.784*	10.496***	3.128**	10.677**	2.103
	(3.986)	(2.212)	(3.271)	(1.297)	(4.386)	(1.541)
Age squared × 10^{-4}	−10.611 (4.379)*	−3.882 (2.341)*	−13.337 (3.690)***	−3.409 (1.374)**	−13.821 (5.002)***	−2.187 (1.639)
Income	−0.269 (0.349)	−0.288 (0.203)	−0.437 (0.264)*	−0.295 (0.113)***	−0.365 (0.346)	−0.214 (0.133)
Levyplus	−0.381 (0.253)	−0.032 (0.158)	−0.433 (0.197)**	−0.034 (0.096)	−0.640 (0.264)**	−0.095 (0.114)
Freepoor	0.278 (0.830)	−0.385 (0.512)	0.308 (0.508)	−0.377 (0.239)	0.111 (0.659)	−0.481 (0.283)*
Freerepat	−0.974 (0.339)**	−0.254 (0.202)	−1.149 (0.305)***	−0.215 (0.117)*	−1.375 (0.447)***	−0.189 (0.140)
Illness	−0.345 (0.092)**	0.002 (0.045)	−0.416 (0.081)***	0.049 (0.025)**	−0.672 (0.156)***	0.052 (0.029)*
Activity days	−1.114 (0.198)**	0.047 (0.014)***	−1.256 (0.238)***	0.083 (0.006)***	−1.787 (0.653)***	0.104 (0.008)***
General health questionnaire score	−0.080 (0.043)*	0.016 (0.020)	−0.097 (0.039)**	0.018 (0.011)	−0.105 (0.056)*	0.023 (0.014)*
Chronic condition 1	−0.242 (0.262)	−0.078 (0.164)	−0.127 (0.199)	−0.013 (0.092)	−0.119 (0.279)	−0.000 (0.108)
Chronic condition 2	−0.754 (0.352)**	−0.144 (0.180)	−0.604 (0.306)**	−0.034 (0.103)	−0.489 (0.414)	0.055 (0.121)
Constant	1.452 (0.739)**	−0.618 (0.472)	0.786 (0.572)	−1.050 (0.255)***	0.622 (0.753)	−1.233 (0.296)***
α						0.578 (0.086)
Log-likelihood						−3107.6

Note. Standard errors in parentheses (robust standard errors for PQL). α indicates the overdispersion parameter of the negative binomial type II distribution.
***Statistical significance at 1% level.
**Statistical significance at 5% level.
*Statistical significance 10% level.

APPENDIX A

STATA CODE FOR THE PQL ESTIMATION OF ZERO-INFLATED COUNT MODELS

The following Stata code first loads the program for the PQL estimation of zero-inflated count models, pqlzi, and then exemplifies its use with a data set from the Stata Web site, fish.dta. The only purpose of the example is to illustrate pqlzi’s use; the particular model estimated on these data is nonsensical. Program pqlzi uses mean function $\pi\lambda$ instead of $(1-\pi)\lambda$. We often found this to have better convergence properties. It means that all the estimates from the binary part (eq 2-output) have the “wrong” sign. For example, “−1.81" should be read as “1.81.” If preferred, this can be changed by deleting the two “+‘theta2’” bits in the program.

```
clear all

** Load pqlzi program

capture program drop pqlzi
program define pqlzi
args lnf theta1 theta2
quietly replace ‘lnf’ = ///
    - exp(‘theta1’ + ‘theta2’)/(1+exp(‘theta2’)) ///
          + $ML_y1*ln(exp(‘theta1’ + ‘theta2’)/(1+exp(‘theta2’))) ///
          - lnfactorial($ML_y1)
end

** Use Stata’s example dataset

webuse fish

** Get initial values for pqlzi

poisson count persons livebait /* get initial values for count part  */
mat po = e(b)
logit count child camper       /* get initial values for binary part */
mat lo = e(b)

** Estimate pqlzi model

ml model lf pqlzi (eq1: count = persons livebait) (eq2: child camper), vce(robust)
ml init po lo, copy skip /* load initial values  */
ml maximize             /* estimate pqlzi model */

** Compare to other ZI models

zinb count persons livebait, inflate(child camper) /* compare to zinb */
zip count persons livebait, inflate(child camper)  /* compare to zip  */
```

APPENDIX B

FURTHER MONTE CARLO RESULTS: NLS ESTIMATOR

	No overdispersion						Additive overdispersion					
	10% zero inflation			50% zero inflation			10% zero inflation			50% zero inflation		
	$\hat{\beta}$	δ_1	δ_2	$\hat{\beta}$	δ_1	δ_2	$\hat{\beta}$	δ_1	δ_2	$\hat{\beta}$	δ_1	δ_2
$n = 50{,}000$												
Median	0.960	1.660	1.613	0.970	1.085	1.111	0.962	1.655	1.617	0.969	1.085	1.111
Mean	1.334	21.293	28.424	1.967	1.996	1.487	1.498	21.844	28.342	2.907	2.946	1.121
SD	4.760	45.678	65.868	11.936	11.301	9.257	10.924	47.189	63.599	24.130	24.039	0.311
$n = 500{,}000$												
Median	0.969	1.178	1.223	0.977	1.030	1.044	0.971	1.170	1.232	0.977	1.029	1.043
Mean	0.976	1.175	1.225	1.076	1.086	1.018	0.977	1.173	1.221	1.070	1.081	1.020
SD	0.043	0.255	0.303	0.605	0.543	0.106	0.044	0.250	0.292	0.588	0.527	0.106

Note. 1000 replications for $n = 50{,}000$; 100 for $n = 500{,}000$. True values: $\beta = \delta_1 = \delta_2 = 1$.

ACKNOWLEDGEMENTS

The authors thank two anonymous referees for very helpful comments.

REFERENCES

Bauer T, Göhlmann S, Sinning M. 2007. Gender Differences in Smoking Behavior. *Health Economics* **16**: 895–909.

Cameron AC, Trivedi PK. 1986. Econometric models based on count data: Comparisons and applications of some estimators and tests. *Journal of Applied Econometrics* **1**: 29–53.

Cameron AC, Trivedi PK. 1998. *Regression Analysis of Count Data*. Cambridge University Press: Cambridge, MA.

Campolieti M. 2002. The recurrence of occupational injuries: Estimates from a zero-inflated count model. *Applied Economics Letters* **9**: 595–600.

Chang F-R, Trivedi PK. 2003. Economics of Self-Medication: Theory and Evidence. *Health Economics* **12**: 721–739.

Chernoff H. 1954. On the distribution of the likelihood ratio. *Annals of Mathematical Statistics* **25**: 573–578.

Gourieroux C, Monfort A, Trognon A. 1984a. Pseudo Maximum Likelihood Methods: Theory. *Econometrica* **52**: 681–700.

Gourieroux C, Monfort A, Trognon A. 1984b. Pseudo Maximum Likelihood Methods: Application to Poisson models. *Econometrica* **52**: 701–721.

Hall DB, Shen J. 2010. Robust estimation for zero-inflated Poisson regression. *Scandinavian Journal of Statistics* **37**: 237–252.

Kim Y-S, Lee M-J. 2011. Effect of informal family care on formal health care: Zero-inflated endogenous count for censored response, University of York, HEDG Working Paper 10/11.

Lambert D. 1992. Zero-inflated Poisson regression with an application to defects in manufacturing. *Technometrics* **34**: 1–14.

Mullahy J. 1986. Specification and Testing of Some Modified Count Data Models. *Journal of Econometrics* **33**: 341–365.

Papadopoulos G, Santos Silva JMC. 2008. Identification Issues in Models for Underreported Counts, University of Essex, Discussion Paper No. 657.

Pizer SD, Prentice JC. 2011. Time Is Money: Outpatient Waiting Times and Health Insurance Choices of Elderly Veterans in the United States. *Journal of Health Economics* **30**: 626–636.

Santos Silva JMC, Tenreyro S. 2006. The log of gravity. *The Review of Economics and Statistics* **88**: 641–658.

Santos Silva JMC, Tenreyro S. 2011. Further simulation evidence on the performance of the Poisson pseudo-maximum likelihood estimator. *Economics Letters* **112**: 220–222.

Sari N. 2009. Physical Inactivity and its Impact on Healthcare Utilization. *Health Economics* **18**: 885–901.

Sarma S, Simpson W. 2006. A mircroeconometric analysis of Canadian health care utilization. *Health Economics* **15**: 219–239.

Sheu M-L, Hu T-W, Keeler TE, Ong M, Sung H-Y. 2004. The effect of major cigarette price change on smoking behavior in California: a zero-inflated negative binomial model. *Health Economics* **13**: 721–791.

Street A, Jones A, Furuta A. 1999. Cost sharing and pharmaceutical utilisation and expenditure in Russia. *Journal of Health Economics* **18**: 459–472.

White H. 1982. Maximum Likelihood Estimation of Misspecified Models. *Econometrica* **50**: 1–25.

Winkelmann R. 2008. *Econometric Analysis of Count Data*, fifth edition. Springer: Berlin.

Winkelmann R, Zimmermann KF. 1993. Poisson-Logistic Regression, Department of Economics, University of Munich, Working Paper No. 93–18.

Yen ST, Tang C-H. Su S-JB. 2001. Demand for Traditional Medicine in Taiwan: A Mixed Gaussian-Poisson Model Approach. *Health Economics* **10**: 221–232.

DOI: 10.1002/hec