Staub, Kevin E.; Winkelmann, Rainer

Working Paper

Robust estimation of zero-inflated count models

Working Paper, No. 0908

Provided in Cooperation with:
Socioeconomic Institute (SOI), University of Zurich

Suggested Citation: Staub, Kevin E.; Winkelmann, Rainer (2009) : Robust estimation of zero-inflated count models, Working Paper, No. 0908, University of Zurich, Socioeconomic Institute, Zurich

This Version is available at:
https://hdl.handle.net/10419/76158

University of Zurich

Socioeconomic Institute
Sozialökonomisches Institut

Working Paper No. 0908

Robust estimation of zero-inflated count models

Kevin E. Staub, Rainer Winkelmann

June 2009

Socioeconomic Institute
University of Zurich

Working Paper No. 0908

Robust estimation of zero-inflated count models

June 2009

Author's address: Kevin E. Staub
E-mail: staub@sts.uzh.ch

Rainer Winkelmann
E-mail: winkelmann@sts.uzh.ch

Publisher

Sozialökonomisches Institut
Bibliothek (Working Paper)
Rämistrasse 71
CH-8006 Zürich
Phone: +41-44-634 21 37
Fax: +41-44-634 49 82
URL: www.soi.uzh.ch
E-mail: soilib@soi.uzh.ch

Robust estimation of zero-inflated count models

Kevin E. Staub and Rainer Winkelmann *

University of Zurich, Socioeconomic Institute

June 26, 2009

Abstract

Applications of zero-inflated count data models have proliferated in empirical economic research. There is a downside to this development, as zero-inflated Poisson or zero-inflated Negative Binomial Maximum Likelihood estimators are not robust to misspecification. In contrast, simple Poisson regression provides consistent parameter estimates even in the presence of excess zeros. The advantages of the Poisson approach are illustrated in a series of Monte Carlo simulations.

JEL Classification: C12, C25

Keywords: excess zeros, Poisson, overdispersion, Negative Binomial regression.

*✉ Chair for Statistics and Empirical Economic Research, Socioeconomic Institute, University of Zurich, Zuerichbergstr. 14, CH-8032 Zurich , Switzerland, ☎ +41 44 634 2312, email: staub@sts.uzh.ch, winkelmann@sts.uzh.ch.

1 The problem of "excess zeros" and zero-inflated models

The Poisson regression model (PRM) is the benchmark model for regressions with count dependent variables. The so-called problem of "excess zeros", however, plagues a majority of count data applications in the social sciences, as the proportion of observations with zero counts in the sample is often much larger than that predicted by the PRM. The conventional wisdom of the pertinent literature is that with "excess zeros", Poisson regression should be abandoned in favor of modified count data models which are capable of taking into account the extra zeros explicitly.

By far the most popular of these models are zero-inflated (ZI) count models[1] (Mullahy, 1986, Lambert, 1992) such as the zero-inflated Poisson (ZIP) and zero-inflated Negative Binomial (ZINB) models. In their simplest form, these models are specified as having a probability function (pf)

$$f(y) = \begin{cases} \pi + (1-\pi)f^*(0) & \text{for } y = 0 \\ (1-\pi)f^*(y) & \text{if } y = 1, 2, 3, ... \end{cases} \tag{1}$$

where y is a count-valued random variable. The function $f^*(\cdot)$ is a standard count pf, and $\pi \in [0, 1]$ is a zero-inflation parameter which allows for any fraction of zeros. If $\pi = 0$, the ZI pf $f(\cdot)$ reduces to $f^*(\cdot)$. The two most common choices for $f^*(\cdot)$ are Poisson,

$$f^{\mathrm{P}}(y; \lambda) = \frac{\exp(-\lambda)\lambda^y}{y!} \ , \ \lambda > 0$$

and Negative Binomial,

$$f^{\mathrm{NB}}(y; \lambda) = \frac{\Gamma(\gamma + y)}{\Gamma(\gamma)\Gamma(y+1)} \left(\frac{\gamma}{\lambda + \gamma}\right)^{\gamma} \left(\frac{\lambda}{\lambda + \gamma}\right)^{y} \ , \lambda > 0 \, , \, \gamma > 0$$

Both models' expectation is equal to λ, which also gives the variance in the Poisson case. The variance in the Negative Binomial model is $\lambda + \gamma^{-1}\lambda^2$. In a regression context, it is customary to specify the mean parameter λ as a function of a vector of explanatory variables, x, the usual choice being the exponential function

$$\lambda = \exp(\alpha + x'\beta) \tag{2}$$

where β is a parameter vector conformable to x, and α is a constant element. Estimation of ZIP and ZINB proceeds by Maximum Likelihood (ML). If the equations (1) and (2) as well as $f^*(\cdot)$ are

[1] Alternative terminology includes with-zeros or zero-altered count models.

correctly specified, ML theory ensures that these estimators will be consistent and asymptotically efficient[2].

Zero-inflated models for count data are used extensively in many areas of current empirical economic research. Without any claims of being exhaustive, a list of recent applications using ZI count models includes job interviews (List, 2001), work absences (Campolieti, 2002), job changes (Heitmueller, 2004), lateness (Clark et al., 2005), patent applications (Stephan et al., 2007), cigarette consumption (Sheu et al., 2004), theatre attendance (Ateca-Amestoy, 2008), biking trips (Zahran et al., 2008) and firm FDI (Ho et al., 2009).

2 Re-enter Poisson

The conditional expectation function of the zero-inflated count data model defined by (1) and (2) is given by

$$\mathrm{E}(y|x) = (1-\pi)\lambda = \exp(\tilde{\alpha} + x'\beta) \tag{3}$$

The only difference to the CEF of the parent model (2) is a shifted constant $\tilde{\alpha} = \alpha + \ln(1-\pi)$. This suggests that $\tilde{\alpha}$ and β can be estimated consistently by any moment based estimator, for example non-linear least squares.

More relevant, however, estimation based on the PRM is consistent as well. This is the case because the Poisson distribution is a member of the linear exponential family (LEF), which is the class of distributions with pf of the form

$$f^{\mathrm{LEF}}(y|\mu_x) = \exp\{a(\mu_x) + b(y) + c(\mu_x)y\}, \quad \text{where } \mu_x = \mu(x;\beta) = \mathrm{E}(y|x),$$

for $a(\mu_x) = -\mu_x$, $b(y) = -\ln(y!)$ and $c(\mu_x) = \ln(\mu_x)$. LEFs have the property that

$$\frac{\partial \log f(y|x)}{\partial \beta} = (y - \mu_x)h(x) \tag{4}$$

where $h(x) = dc(\mu_x)/d\mu_x x$. 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 (4) at the true density is zero, even though the model is misspecified, since the CEF residual $y - \mathrm{E}(y|x)$ is

[2]For more detail on these models and their estimation, see Cameron and Trivedi (1998) or Winkelmann (2008).

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 as long as the CEF is correctly specified and (first-order) identified. This holds regardless of misspecification of higher conditional moment functions but requires using a LEF distribution such as the Poisson for constructing the likelihood function. An adjustment to the covariance matrix is necessary. Gourieroux, Monfort and Trognon (1984a) refer to estimators with this property as pseudo-maximum likelihood (PML) estimators.

While Poisson regression of the zero inflated model cannot provide separate estimates for α and π, this is of secondary importance in most applied work, since knowledge of α and π is not needed in order to estimate the CEF and the semi-elasticities of the CEF with respect to some regressor x_k. The former is (3) and the latter is given by

$$\frac{\partial \mathrm{E}(y|x)/\mathrm{E}(y|x)}{\partial x_k} = \beta_k,$$

the element from β corresponding to the kth regressor. Thus, PRM estimates both these quantities consistently. Even though the data are zero-inflated, a simple PRM suffices to obtain valid estimates of the objects of interest if it is the case that the CEF is exponential as in (2).

There are other estimators that could be used to estimate β consistently based on an exponential CEF specification, such as nonlinear least squares (NLS) and various moment-based estimators. Without further assumptions, however, it is unclear how these could improve upon PRM, since PRM's first order conditions are plain orthogonality conditions between residuals and regressors. First order conditions of the other estimators either coincide with PRM's or are weighted versions of PRM's, as in the case of NLS. In the context of robust estimation, the choice of weights is arbitrary and therefore unlikely to enhance estimation[3].

Unlike PRM, the ZIP and ZINB models are not LEF members. Indeed, the log-probability function of a ZIP variable is

$$\ln f^{ZIP}(y;\lambda,\pi) =$$

[3]Furthermore, owing to its simple FOC, optimization of PRM is computationally very stable. This is not always the case with other estimators. For instance, the objective function of NLS in this context is not concave, as noted in Gourieroux, Monfort and Trognon (1984b).

$$\mathbb{1}(y=0)[\ln(\pi+(1-\pi)\exp(-\lambda))]+(1-\mathbb{1}(y=0))[\ln(1-\pi)-\lambda+y\ln\lambda-\ln(y!)]$$

which cannot be written as $a(\mu)+b(y)+c(\mu)y$ with $\mu=(1-\pi)\lambda$, as there is no way of isolating an additive component that is linear in y –i.e. $c(\mu)y$– due to the (nonlinearity of the) indicator function $\mathbb{1}(y=0)$. An analog argument holds for the ZINB log-probability and, in fact, for any ZI model generated according to (1). An additional result from Gourieroux, Monfort and Trognon (1984a) states that all PML estimators are LEF members. Consequently, ZIP and ZINB are not PML estimators and misspecification of higher conditional moments will in general lead to asymptotic bias in these models. Thus, the cure might be worse than the disease.

Following this line of reasoning, one might wonder whether there are any reasons *not* to use PRM in a ZI model context. There are some. First, one might want to predict probabilities of certain events or elasticities of such probabilities to specific regressors. Using the PRM estimates with the Poisson pf for this is inappropriate. Second, PRM can be less efficient than the ZI estimator if the zero-inflated model is correctly specified. Finally, the applicability of large sample robustness results to small samples is open for discussion. The small sample properties of the two approaches for estimating model with extra zeros are assessed in the following Monte Carlo study.

3 Monte Carlo evidence

The basic design of the experiment is as follows. We generate data by drawing $n=100,1000$ observations from the scalar random variable $x\sim N(0,1)$, and specify values for the parameters $(\alpha,\beta)=(0.5,1)$ and $\pi=0.1,0.5,0.9$. For the pf $f^*(\cdot)$, the Poisson probability function is chosen. Thus, the data generating model is ZIP with $\lambda=\exp(0.5+x)$ and zero-inflation ranging from 10% to 90%. The data are fitted to PRM and ZIP models, and the procedure is replicated 10,000 times. We call this data generating process DGP1, results for which are printed in Table 1[4].

— Table 1 about here —

[4]The Monte Carlo study was programmed in STATA/MP 10.1; program code and full output are available on request. Appendix A contains an overview of the entire Monte Carlo design.

Considering the results with $n = 1000$ first, it is evident that Poisson and ZIP estimators are both consistent estimates of the semi-elasticity β. As ZIP correctly specifies the DGP, it estimates α consistently, while the corresponding estimate of PRM are close to the true $\tilde{\alpha}$ which for $\pi = (0.1, 0.5, 0.9)$ corresponds to 0.3946, -0.1931 and -1.8026. To compare efficiency of the estimators, the numbers in parentheses in the table provide the standard deviation of $\hat{\alpha}$ and $\hat{\beta}$ computed over the 10'000 replications. The efficiency gains of using the ZIP estimator are substantial. The standard deviation of the estimator more than halves when passing from PRM to the ZIP estimate of β.

With 100 observations, both estimators still do fairly well. Larger biases and high imprecision become apparent only in the presence of very large fractions of zeros ($\pi = 0.9$). The efficiency gains of ZIP in relation to PRM melt away as the sample size shrinks.

The scenario of a correctly specified model is quite unlikely in practice. More realistically, certain features of the model are invalid, raising concerns of potentially large biases. To illustrate this, we generate data from a second process, DGP2, in which we introduce an additional random error in λ that is ignored in the estimation procedures. This error can be thought of as an omitted variable that affects the mean of the count but is unobserved to the econometrician. It is well known that such unobserved heterogeneity will induce overdispersion in the Poisson part of the model. As the presence of unobserved heterogeneity is ubiquitous in empirical economic work it is particularly interesting to investigate its effects on these models.

Thus, let $\lambda = \exp(0.5 + x + v)$ with $v \sim N(-0.5, 1)$. The expectation of v is set to -0.5 to have $\mathrm{E}[e^v] = 1$. This specification leads to a quadratic variance or, equivalently, linear overdispersion function at the level of the count distribution $f^*(\cdot)$ conditional on observables, in the sense that

$$\mathrm{Var}(y^* | \, x) = \mathrm{E}(y^*|x) + \omega[\mathrm{E}(y^*|x)]^2, \quad \text{or} \quad \frac{\mathrm{Var}(y^* | \, x)}{\mathrm{E}(y^*|x)} = 1 + \omega \mathrm{E}(y^*|x)$$

where $\omega \approx 1.7$ and y^* denotes the count variable with pf $f^*(\cdot)$. To estimate this model we now also include the ZINB estimator. This is a potentially good choice since the ZINB assumes a quadratic variance function for the count part $f^*(\cdot)$. Therefore, ZINB specifies DGP2 correctly in the first two moments. ZINB is not quite the right model, however, as higher moments are

misspecified[5]. The resulting bias may be small in finite samples, however, an issue we want to explore in the Monte Carlo experiments. We know that the Poisson estimator is robust to this kind of misspecification. While the negative binomial model (without zero inflation) is a LEF as well, for a given dispersion parameter γ, it is not if γ is estimated, and we refrain from using it. Table 2 thus compares results for the ZIP, the ZINB and the Poisson models under DGP2.

— Table 2 about here —

The ZIP estimates both for α and for β are subject to serious bias in all cases reported in Table 2. The PRM estimates, on the other hand, are practically unaffected by the change in the DGP. The new DGP does make itself noticed in the larger standard deviations of the PRM estimates, though. Not surprisingly, ZINB estimates β closely. However, the misspecification of moments higher than the variance has a perceivable effect on the estimates of α, which are clearly inconsistent.

Next, we investigate estimation of two models with overdispersed $f^*(y|x)$, as before, but where this overdispersion is a constant or hyperbolic rather than a linear function of the mean. Under the new DGPs, both ZIP and ZINB will only specify the CEF correctly, while misspecifying the conditional variance function. To produce a DGP with constant overdispersion in $f^*(y)$, DGP3, the unobserved heterogeneity term v is drawn from a normal distribution with mean $\mu = -0.5\sigma^2$ and variance $\sigma^2 = \ln(\exp(-0.5 - x) + 1)$, resulting in $\mathrm{Var}(y^*|x) = 2\mathrm{E}(y^*|x)$. To obtain a DGP with hyperbolic overdispersion, DGP4, the error variance is set to $\sigma^2 = \ln(2\exp(-2(0.5-x)) + 1)$ (see Appendix A for details). Results corresponding to DGP3 and DGP4 are displayed in Tables 3 and 4.

— Tables 3 and 4 about here —

Having a look at Table 3, ZIP estimation again yields estimators that are not consistent for the true values of α and β, as to be expected. At 1000 observations, while the ZINB estimates are not as highly biased as ZIP's, the best performance is delivered by PRM. The same is true for

[5]DGP2 is a zero-inflated Poisson-log-normal model. To simulate the ZINB model, the distribution of $\log v$ would have to be Gamma.

the case with 100 observations, except for the DGP with $\pi = 0.9$, where ZINB's bias is smaller, although ZINB displays unusually high standard deviations in this DGP. A similar impression is obtained by considering Table 4. PRM consistently outperforms its competitors, and in the DGPs with low zero-inflation its efficiency is in the same order as the ZI estimators'.

These experiments demonstrate the robustness of the PRM estimator of semi-elasticities in zero-inflated, finite samples, and the biases that can arise when using its two most common ZI competitors. Next, we want to analyze whether this result is extendable to another popular class of ZI models, namely models with non-constant zero-inflation.

4 Non-constant zero-inflation

Model (1)-(2) is often generalized to allow for non-constant zero-inflation. To do so, π is specified as a function of covariates, for example as a probit or logit model. These models allow to distinguish between determinants of so-called 'structural' zeros that are due to the binary model and 'incidental' zeros stemming from the count distribution part of the model. For instance, considering job mobility, a person might not have changed job in a given time period because she is not looking for a new one (structural zero) or because despite searching she has not found another job (incidental zero). Here, we limit ourselves to logit-type zero-inflation as it is more widely represented in the existing literature. The principal issue applies, however, to any parametric models for binary variables. Under the logit assumption

$$\pi \equiv \pi(z) = \frac{\exp(z'\delta)}{1+\exp(z'\delta)}, \quad \text{and} \quad 1-\pi = \frac{1}{1+\exp(z'\delta)} \tag{5}$$

where z is a vector of covariates (possibly including a constant) determining the zero-inflation process, and δ is a conformable parameter vector. When z is correlated with x, the CEF of the new model is no longer given by (3). In this case, all of the three previous estimators are inconsistent. The CEF of the model with logit-type ZI is

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

The conventional way in which the literature has opted to estimate these models is by modifying the constant ZI models' log-likelihood function to accommodate the function of the logit

model. Thus, ZIP and ZINB estimators for this model are obtained by maximizing the corresponding log-likelihood functions with respect to $\theta = (\alpha, \beta, \delta)$ for ZIP and with respect to (θ, γ) for ZINB. As before, if the assumed model is equal to the underlying DGP, these estimators are consistent and asymptotically efficient. Under mispecification, inconsistency arises and one might again be interested in more robust estimators. By virtue of its PML property, PRM will work again, as it requires only the correct specification of the CEF which in this case is given by (6). The resulting model is formally identical with the Poisson-logit or Pogit model for underreported counts discussed by Winkelmann and Zimmermann (1993). In general, some constraints on the relationship between x and z need to hold for θ to be identified (see Papadopoulos and Santos Silva, 2008), a sufficient condition being the availability of an element in z that is excluded from x.

The following Monte Carlo results illustrate estimation of these three models in the generalized setting with nonconstant zero-inflation. To accommodate zero-inflation of the logit type, DGP1 is modified by specifying $\pi = \exp(\delta_0 + \delta_1 z)/(1 + \exp(\delta_0 + \delta_1 z))$. The (scalar) variable z is jointly drawn from a bivariate standard normal distribution with x, and the correlation is set to 50%; i.e., $(x, z) \sim BVN(0, 0, 1, 1, 0.5)$. The percentage of zeros in the logit part of the model is thus determined by the vector (δ_0, δ_1). We fix $\delta_1 = 1$ and let $\delta_0 = -2.564, 0, 2.564$ which on average produces datasets where 10%, 50% and 90%, respectively, of the observations are zeros stemming from the binary process. Results for this DGP – DGP5 – are reported in Table 5.

— Table 5 about here —

The results for this DGP are broadly similar to the case with constant zero-inflation. Both ZIP and PRM provide virtually unbiased estimates of β, except for the case of both small samples and many zeros. The efficiency gains of ZIP are comparable to the ones discussed in DGP1, i.e. a reduction of the standard deviation of around 50% in large samples and of less in smaller ones. Estimation of α, on the other hand, turns out to be often problematic for PRM, resulting in biased and imprecise estimates.

As a next step, we introduce normally distributed unobserved heterogeneity to generate DGP6, the logit-type zero-inflation version of DGP2. Additionally to ZIP and PRM, the model is es-

timated by a correspondig ZINB, which specifies correctly the expectation and variance of the standard count variable. Table 6 depicts the corresponding means and standard deviations.

— Table 6 about here —

As before, inconsistency of ZIP is reflected in substantial biases in all reported mean estimates of Table 6, most of which are of the order of -15% to -20%. Estimation of the model with ZINB yields good results as would be expected. PRM's estimates, however, perform equally well in this setting. The efficiency advantage of ZINB is about 50% in most of the cases for β. Both models exhibit difficulties to estimate the constant, unbiased estimation of which appears to be possible only in large samples with moderate amount of zero-inflation.

To complete the simulation results, Tables 7 and 8 show the results of the logit-type zero-inflation equivalents of DGP3 and DGP4, the case of the count variable displaying constant or hyperbolic overdispersion due to an unobserved variable. Estimation of α is again difficult. While PRM does provide some acceptable results with 1'000 observations, the ZINB estimates of α are much more precise. In terms of bias in the more interesting estimate $\hat{\beta}$, on the other hand, PRM clearly outperforms its two competitors.

— Tables 7 and 8 about here —

As a whole, the results from the experiments in this section show that Poisson regression with an appropriately modified mean function is able to estimate the semi-elasticities in zero-inflated count models robustly compared to ZIP and ZINB.

5 Conclusions

In applied practice, by far the main quantities of interest in count models are the conditional expectation function and its semi-elasticities with respect to some regressors. Positive evidence of this can be found, for instance, in the fact that all the applications cited in Section 1 without exception limited the discussion of their estimation results to them. In this paper a case was made for the use of Poisson regression to estimate these quantities, regardless of the presence of "excess zeros".

If zero-inflation is given either by a constant factor or by a binary stochastic process unrelated to the count random variable, simple estimation of the standard Poisson regression model (PRM) yields consistent estimates of the semi-elasticities of the mean with respect to the independent variables. Otherwise, a modification of the mean function is needed. In both cases, however, estimation of the parameters needed to estimate the conditional expectation and semi-elasticities is straightforward, as was illustrated in a set of Monte Carlo experiments.

The advantage of using PRM over zero-inflated count models 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, strong a priori information about the DGP being absent, PRM may be the better choice for estimating ZI models compared to ZI estimators.

References

Ateca-Amestoy Victoria (2008), Determining heterogeneous behavior for theater attendance, *Journal of Cultural Economics*, 32, 127-151.

Cameron, A. Colin and Pravin K. Trivedi (1998), *Regression Analysis of Count Data*, Cambridge, MA: Cambridge University Press.

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

Clark Ken, Simon A. Peters and Mark Tomlinson (2005), The determinants of lateness: Evidence from British workers, *Scottish Journal of Political Economy*, 52(2), 282-304.

Gourieroux Christian, Alain Monfort and Alain Trognon (1984a), Pseudo Maximum Likelihood Methods: Theory, *Econometrica*, 52, 681-700.

Gourieroux Christian, Alain Monfort and Alain Trognon (1984b), Pseudo Maximum Likelihood Methods: Application to Poisson models, *Econometrica*, 52, 701-721.

Heitmueller Axel (2004), Job mobility in Britain: Are the Scots different? Evidence from the BHBS, *Scottish Journal of Political Economy*, 51(3), 329-358.

Ho Woon-Yee, Peiming Wang and Joseph D. Alba (2009), Merger and acquisition FDI, relative wealth and relative access to bank credit: Evidence from a bivariate zero-inflated count model, *International Review of Economics and Finance*, 18, 26-30.

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

List, John A. (2001), Determinants of securing academic interviews after tenure denial: evidence from a zero-inflated Poisson model, *Applied Economics*, 33, 1423-1431.

Papadopoulos Georgios and Joao M.C. Santos Silva (2008), Identification Issues in Models for Underreported Counts, University of Essex, Discussion Paper No. 657.

Sheu Mei-Ling, Teh-Wei Hu, Theodore E. Keeler, Michael Ong and Hai-Yen Sung (2004), The effect of major cigarette price change on smoking behavior in California: a zero-inflated negative binomial model, *Health Economics*, 13, 721-791.

Stephan Paula E., Shiferaw Gurmu, Albert J. Sumell and Grant Black (2007), Who's patenting in the university? Evidence from the survey of doctorate recipients, *Economics of Innovation and New Technology*, 16(2), 71-99.

Winkelmann, Rainer and Klaus F. Zimmermann (1993), Poisson Logistic Regression, University of Munich, Working Paper No. 93-18.

Winkelmann, Rainer (2008), *Econometric Analysis of Count Data*, fifth edition, Berlin: Springer.

Zahran Sammy, Samuel D. Brody, Praveen Maghelal, Andrew Prelog and Michael Lacy (2008), Cycling and walking: Explaining the spatial distribution of healthy modes of transportation in the United States, *Transportation Research Part D*, 13, 462-470.

Table 1: Mean estimated α and β with 10'000 replications (DGP1)

Estimator		Mean $\hat{\alpha}$			Mean $\hat{\beta}$		
		$\pi = 0.1$	$\pi = 0.5$	$\pi = 0.9$	$\pi = 0.1$	$\pi = 0.5$	$\pi = 0.9$
ZIP	n=100	0.4932	0.488	0.3310	1.0012	1.0032	1.0591
		(0.1025)	(0.1515)	(1.0254)	(0.0757)	(0.1155)	(1.7935)
	n=1000	0.4995	0.4987	0.4935	1.0000	1.0003	1.0016
		(0.0317)	(0.0457)	(0.1058)	(0.0216)	(0.0310)	(0.0759)
PRM	n=100	0.3879	-0.2147	-1.9651	0.9986	0.9822	0.8624
		(0.1026)	(0.1837)	(0.6137)	(0.0977)	(0.2207)	(0.612)
	n=1000	0.3941	-0.1947	-1.8159	0.9999	0.9975	0.9804
		(0.0328)	(0.0628)	(0.1709)	(0.0310)	(0.0743)	(0.2056)

Notes: Standard deviations in parenthesis.

Table 2: Mean estimated α and β with 10'000 replications (DGP2)

Estimator		Mean $\hat{\alpha}$			Mean $\hat{\beta}$		
		$\pi = 0.1$	$\pi = 0.5$	$\pi = 0.9$	$\pi = 0.1$	$\pi = 0.5$	$\pi = 0.9$
ZINB	n=100	0.4211	0.2782	-0.1787	1.0001	0.9993	1.0966
		(0.1797)	(0.3573)	(3.9084)	(0.1791)	(0.2692)	(4.3761)
	n=1000	0.3969	0.2695	0.2103	0.9993	1.0026	0.9988
		(0.0542)	(0.1422)	(0.4476)	(0.0567)	(0.0823)	(0.1966)
ZIP	n=100	0.7722	0.7688	0.3106	0.8578	0.8329	0.9090
		(0.2505)	(0.3337)	(3.9447)	(0.2464)	(0.3273)	(4.5169)
	n=1000	0.7873	0.8081	0.8073	0.8837	0.8653	0.8345
		(0.0939)	(0.1238)	(0.251)	(0.1034)	(0.1326)	(0.2483)
PRM	n=100	0.3717	-0.2444	-2.0574	0.9743	0.9437	0.7947
		(0.2091)	(0.3097)	(0.8443)	(0.2344)	(0.3397)	(0.7120)
	n=1000	0.3911	-0.1988	-1.8264	0.9962	0.9898	0.9498
		(0.0747)	(0.1091)	(0.2626)	(0.0955)	(0.1347)	(0.2890)

Notes: Standard deviations in parenthesis.

Table 3: Mean estimated α and β with 10'000 replications (DGP3)

Estimator		Mean $\hat{\alpha}$			Mean $\hat{\beta}$		
		$\pi = 0.1$	$\pi = 0.5$	$\pi = 0.9$	$\pi = 0.1$	$\pi = 0.5$	$\pi = 0.9$
ZINB	n=100	0.5085	0.5255	0.2769	0.9908	0.9685	1.0165
		(0.1522)	(0.2259)	(1.1205)	(0.1215)	(0.1803)	(1.3769)
	n=1000	0.5215	0.5568	0.5606	0.9838	0.9587	0.9500
		(0.0472)	(0.068)	(0.1589)	(0.0364)	(0.0524)	(0.1255)
ZIP	n=100	0.5957	0.6202	0.4356	0.9490	0.9236	0.9463
		(0.159)	(0.2213)	(1.0558)	(0.1148)	(0.1694)	(1.4011)
	n=1000	0.6029	0.6354	0.6453	0.9505	0.9301	0.9161
		(0.0494)	(0.0690)	(0.1576)	(0.0334)	(0.0483)	(0.1173)
PRM	n=100	0.3848	-0.2210	-1.9941	0.9989	0.9810	0.8648
		(0.1377)	(0.2248)	(0.6865)	(0.1201)	(0.2398)	(0.6435)
	n=1000	0.3937	-0.1959	-1.8171	1.0001	0.9981	0.9786
		(0.0436)	(0.0726)	(0.1918)	(0.0370)	(0.079)	(0.2177)

Notes: Standard deviations in parenthesis.

Table 4: Mean estimated α and β with 10'000 replications (DGP4)

Estimator		Mean $\hat{\alpha}$			Mean $\hat{\beta}$		
		$\pi = 0.1$	$\pi = 0.5$	$\pi = 0.9$	$\pi = 0.1$	$\pi = 0.5$	$\pi = 0.9$
ZINB	n=100	0.5195	0.5488	0.2721	0.9803	0.9449	1.0638
		(0.1672)	(0.2443)	(1.8518)	(0.1431)	(0.2022)	(3.4095)
	n=1000	0.5360	0.5954	0.5988	0.9745	0.9259	0.9118
		(0.0530)	(0.0758)	(0.1847)	(0.0533)	(0.0676)	(0.1485)
ZIP	n=100	0.6068	0.6415	0.4031	0.9385	0.9060	1.0024
		(0.1941)	(0.2608)	(1.8226)	(0.1393)	(0.2007)	(3.4145)
	n=1000	0.6145	0.6596	0.6732	0.9321	0.9112	0.8929
		(0.0636)	(0.0854)	(0.1906)	(0.0450)	(0.0581)	(0.1406)
PRM	n=100	0.3825	-0.2252	-2.0043	1.0015	0.9866	0.8888
		(0.1535)	(0.2404)	(0.7348)	(0.1258)	(0.2438)	(0.6470)
	n=1000	0.3939	-0.1968	-1.8202	0.9922	0.9985	0.9809
		(0.0508)	(0.0795)	(0.2043)	(0.0539)	(0.0807)	(0.2211)

Notes: Standard deviations in parenthesis.

Table 5: Mean estimated α and β with 10'000 replications (DGP5)

Estimator		Mean $\hat{\alpha}$			Mean $\hat{\beta}$		
		$E(\pi) = 0.1$	$E(\pi) = 0.5$	$E(\pi) = 0.9$	$E(\pi) = 0.1$	$E(\pi) = 0.5$	$E(\pi) = 0.9$
ZIP	n=100	0.4963	0.4834	0.3323	0.9981	1.0024	1.0764
		(0.1014)	(0.1516)	(0.6299)	(0.0795)	(0.1325)	(0.9923)
	n=1000	0.4993	0.4989	0.4871	1.0003	1.0002	1.0044
		(0.0311)	(0.0459)	(0.1142)	(0.0221)	(0.0357)	(0.1047)
PRM	n=100	1.0734	0.5052	0.0848	0.9763	0.9628	0.9230
		(1.8325)	(1.1331)	(2.6202)	(0.1265)	(0.2469)	(0.9151)
	n=1000	0.5573	0.5506	-0.0829	0.9925	0.9939	0.9609
		(0.2757)	(0.4554)	(1.0294)	(0.0449)	(0.0873)	(0.2282)

Notes: Standard deviations in parenthesis.

Table 6: Mean estimated α and β with 10'000 replications (DGP6)

Estimator		Mean $\hat{\alpha}$			Mean $\hat{\beta}$		
		$E(\pi) = 0.1$	$E(\pi) = 0.5$	$E(\pi) = 0.9$	$E(\pi) = 0.1$	$E(\pi) = 0.5$	$E(\pi) = 0.9$
ZINB	n=100	0.4269	0.3504	0.0938	0.9382	0.9580	1.0741
		(0.1793)	(0.3196)	(1.0229)	(0.1834)	(0.2779)	(1.7519)
	n=1000	0.4051	0.3365	0.1014	0.9515	0.9865	0.9960
		(0.0561)	(0.1011)	(0.5016)	(0.0604)	(0.0852)	(0.2186)
ZIP	n=100	0.7886	0.7838	0.5094	0.8212	0.8006	0.8974
		(0.239)	(0.3002)	(1.0762)	(0.2436)	(0.3152)	(2.0353)
	n=1000	0.8081	0.8199	0.8159	0.8624	0.8520	0.8128
		(0.0859)	(0.1041)	(0.2148)	(0.0952)	(0.1263)	(0.254)
PRM	n=100	0.9289	0.5276	0.2113	0.9378	0.9400	0.9362
		(1.7203)	(1.3644)	(2.8565)	(0.2597)	(0.3640)	(1.0876)
	n=1000	0.5665	0.4997	-0.3083	0.9827	0.9837	0.9446
		(0.3198)	(0.5595)	(1.1770)	(0.0976)	(0.1448)	(0.3068)

Notes: Standard deviations in parenthesis.

Table 7: Mean estimated α and β with 10'000 replications (DGP7)

Estimator		Mean $\hat{\alpha}$			Mean $\hat{\beta}$		
		$E(\pi) = 0.1$	$E(\pi) = 0.5$	$E(\pi) = 0.9$	$E(\pi) = 0.1$	$E(\pi) = 0.5$	$E(\pi) = 0.9$
ZINB	n=100	0.4877	0.4811	0.1445	0.9615	0.9681	1.0652
		(0.1551)	(0.2628)	(0.8796)	(0.1309)	(0.2117)	(1.0670)
	n=1000	0.5002	0.5350	0.4440	0.9820	0.9631	0.9535
		(0.0585)	(0.0702)	(0.3783)	(0.0420)	(0.0598)	(0.1665)
ZIP	n=100	0.6031	0.6313	0.4302	0.9381	0.9091	0.9633
		(0.1621)	(0.2248)	(0.7876)	(0.1236)	(0.2024)	(1.2085)
	n=1000	0.6086	0.6483	0.6688	0.9465	0.9173	0.8822
		(0.0519)	(0.0698)	(0.1633)	(0.0359)	(0.0553)	(0.1547)
PRM	n=100	0.9898	0.5215	0.1659	0.9721	0.9689	0.9702
		(1.7354)	(1.1552)	(2.7091)	(0.1537)	(0.2726)	(1.8048)
	n=1000	0.5640	0.5566	-0.1296	0.9914	0.9951	0.9705
		(0.3045)	(0.5087)	(1.0932)	(0.0514)	(0.0944)	(0.2476)

Notes: Standard deviations in parenthesis.

Table 8: Mean estimated α and β with 10'000 replications (DGP8)

Estimator		Mean $\hat{\alpha}$			Mean $\hat{\beta}$		
		$E(\pi) = 0.1$	$E(\pi) = 0.5$	$E(\pi) = 0.9$	$E(\pi) = 0.1$	$E(\pi) = 0.5$	$E(\pi) = 0.9$
ZINB	n=100	0.5025	0.4923	0.1578	0.9592	0.9547	1.0833
		(0.1774)	(0.2822)	(1.1012)	(0.1579)	(0.2466)	(1.8613)
	n=1000	0.5159	0.5531	0.3966	0.9744	0.9316	0.9159
		(0.0602)	(0.0825)	(0.4843)	(0.0528)	(0.0816)	(0.2052)
ZIP	n=100	0.6171	0.6619	0.4586	0.9266	0.8801	0.9617
		(0.2002)	(0.2716)	(1.0513)	(0.1533)	(0.2504)	(2.0508)
	n=1000	0.6266	0.6873	0.7265	0.9321	0.8849	0.8259
		(0.0663)	(0.0932)	(0.2098)	(0.0441)	(0.0753)	(0.2103)
PRM	n=100	0.9999	0.5121	0.1495	0.9770	0.9783	1.0341
		(1.7644)	(1.0828)	(2.7692)	(0.1622)	(0.2803)	(2.1014)
	n=1000	0.5639	0.5628	-0.1397	0.9921	0.9958	0.9757
		(0.3159)	(0.5363)	(1.1115)	(0.0537)	(0.0979)	(0.2544)

Notes: Standard deviations in parenthesis.

Appendix A Overview of the Monte Carlo design

Eight different data generating processes (DGPs) have been examined. They all are special cases of the following specification:

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

$$y^* \sim \text{Poisson}(\lambda), \quad \lambda = \exp(\alpha + \beta x + v), \quad (x, z) \sim \text{BVN}(0, 0, 1, 1, 0.5), \quad \pi = \frac{\exp(\delta_0 + \delta_1 z)}{1 + \exp(\delta_0 + \delta_1 z)}$$

where $\alpha = 0.5$, $\beta = 1$, $\delta_1 = 1$ and $v \sim N(\mu, \sigma^2)$. In DGP1 and DGP5, we let $\mu = \sigma^2 = 0$, i.e. there is no unobserved heterogeneity. With $\sigma^2 > 0$ the variance function for y^* is given by

$$\begin{aligned} \text{Var}(y^* | \, x) &= \text{E}_v[\text{Var}(y^*|x, v)] + \text{Var}_v[\text{E}(y^*|x, v)] \\ &= \exp(\alpha + x'\beta)\text{E}_v[e^v] + \exp(\alpha + x'\beta)^2 \text{Var}_v[e^v] \\ &= \exp(\alpha + x'\beta)e^{\mu + \frac{1}{2}\sigma^2} + \exp(\alpha + x'\beta)^2 (e^{\sigma^2} - 1)e^{\mu + \frac{1}{2}\sigma^2} \\ &= \text{E}(y^* | \, x) + \text{E}(y^* | \, x)^2 (e^{\sigma^2} - 1)e^{-\mu - \frac{1}{2}\sigma^2} \end{aligned}$$

The mean and variance of v are specified as

$$\begin{aligned} \mu &= -0.5\,\sigma^2 \\ \sigma^2 &= \ln\{1 + c \exp\left[(k-1)(\alpha + \beta x)\right]\} \end{aligned}$$

The parameter k controls the nonlinearity of the variance function, while c is a free overdispersion parameter. Thus, setting $k = 1$ results in a quadratic variance function

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

In DGP2 and DGP6, we set $c = e - 1$ which incidentally implies $v \sim N(-0.5, 1)$.

Linear variance functions as in DGP3 and DGP7 are obtained with $k = 0$, so that $\text{Var}(y^* | \, x) = (1 + c)\text{E}(y^* | \, x)$. In these DGP we set c equal to 1.

Last, hyperbolic overdispersion functions $\text{Var}(y^* | \, x)/\text{E}(y^* | \, x) = 1 + c/\text{E}(y^* | \, x)$ are produced by setting $k = -1$. In DGP4 and DGP8, parameter c was set to 2.

The different values for π (DGP1-3) or $\text{E}(\pi)$ (DGP4-6) were obtained setting specific values for δ_0. Thus, when z is constant and equal to zero, $\delta_0 = -2.1972, 0, 2.1972$ yields $\pi = 0.1, 0.5, 0.9$; and with $z \sim N(0, 1)$, setting $\delta_0 = -2.564, 0, 2.564$ gives $\text{E}(\pi)$ equal to the same values as before.

Table 9 summarizes the different DGPs.

Table 9: Data generating processes of Monte Carlo simulations

	$\mathrm{Var}(y^*\mid x)/\mathrm{E}(y^*\mid x)$			
Zero inflation type	$=1$ (Equidispersion)	$=1+\mathrm{E}(y^*\mid x)$ (Lin. Overdisp.)	$=2$ (Const. Overdisp.)	$=1+2/\mathrm{E}(y^*\mid x)$ (Hyperbol. Overdisp.)
Constant zero inflation $z=0$	DGP1	DGP2	DGP3	DGP4
Logit-type zero inflation $z\sim\mathrm{N}(0,1)$	DGP5	DGP6	DGP7	DGP8

Working Papers of the Socioeconomic Institute at the University of Zurich

The Working Papers of the Socioeconomic Institute can be downloaded from http://www.soi.uzh.ch/research/wp_en.html

0908	Robust estimation of zero-inflated count models, Kevin E. Staub, Rainer Winkelmann June 2009, 22 p.
0907	Competitive Screening in Insurance Markets with Endogenous Wealth Heterogeneity, Nick Netzer, Florian Scheuer, April 2009, 28 p.
0906	New Flight Regimes and Exposure to Aircraft Noise: Identifying Housing Price Effects Using a Ratio-of-Ratios Approach, Stefan Boes, Stephan Nüesch, April 2009, 40 p.
0905	Patents versus Subsidies – A Laboratory Experiment, Donja Darai, Jens Großer, Nadja Trhal, March 2009, 59 p.
0904	Simple tests for exogeneity of a binary explanatory variable in count data regression models, Kevin E. Staub, February 2009, 30 p.
0903	Spurious correlation in estimation of the health production function: A note, Sule Akkoyunlu, Frank R. Lichtenberg, Boriss Siliverstovs, Peter Zweifel, February 2009, 13 p.
0902	Making Sense of Non-Binding Retail-Price Recommendations, Stefan Bühler, Dennis L. Gärtner, February 2009, 30 p.
0901	Flat-of-the-Curve Medicine – A New Perspective on the Production of Health, Johnnes Schoder, Peter Zweifel, January 2009, 35 p.
0816	Relative status and satisfaction, Stefan Boes, Kevin E. Staub, Rainer Winkelmann, December 2008, 11 p.
0815	Delay and Deservingness after Winning the Lottery, Andrew J. Oswald, Rainer Winkelmann, December 2008, 29 p.
0814	Competitive Markets without Commitment, Nick Netzer, Florian Scheuer, November 2008, 65 p.
0813	Scope of Electricity Efficiency Improvement in Switzerland until 2035, Boris Krey, October 2008, 25 p.
0812	Efficient Electricity Portfolios for the United States and Switzerland: An Investor View, Boris Krey, Peter Zweifel, October 2008, 26 p.
0811	A welfare analysis of "junk" information and spam filters; Josef Falkinger, October 2008, 33 p.
0810	Why does the amount of income redistribution differ between United States and Europe? The Janus face of Switzerland; Sule Akkoyunlu, Ilja Neustadt, Peter Zweifel, September 2008, 32 p.
0809	Promoting Renewable Electricity Generation in Imperfect Markets: Price vs. Quantity Policies; Reinhard Madlener, Weiyu Gao, Ilja Neustadt, Peter Zweifel, July 2008, 34p.
0808	Is there a U-shaped Relation between Competition and Investment? Dario Sacco, July 2008, 26p.
0807	Competition and Innovation: An Experimental Investigation, May 2008, 20 p.
0806	All-Pay Auctions with Negative Prize Externalities: Theory and Experimental Evidence, May 2008, 31 p.
0805	Between Agora and Shopping Mall, Josef Falkinger, May 2008, 31 p.
0804	Provision of Public Goods in a Federalist Country: Tiebout Competition, Fiscal Equalization, and Incentives for Efficiency in Switzerland, Philippe Widmer, Peter Zweifel, April 2008, 22 p.
0803	Stochastic Expected Utility and Prospect Theory in a Horse Race: A Finite Mixture Approach, Adrian Bruhin, March 2008, 25 p.

0802 The effect of trade openness on optimal government size under endogenous firm entry, Sandra Hanslin, March 2008, 31 p.

0801 Managed Care Konzepte und Lösungsansätze – Ein internationaler Vergleich aus schweizerischer Sicht, Johannes Schoder, Peter Zweifel, February 2008, 23 p.

0719 Why Bayes Rules: A Note on Bayesian vs. Classical Inference in Regime Switching Models, Dennis Gärtner, December 2007, 8 p.

0718 Monoplistic Screening under Learning by Doing, Dennis Gärtner, December 2007, 29 p.

0717 An analysis of the Swiss vote on the use of genetically modified crops, Felix Schläpfer, November 2007, 23 p.

0716 The relation between competition and innovation – Why is it such a mess? Armin Schmutzler, November 2007, 26 p.

0715 Contingent Valuation: A New Perspective, Felix Schläpfer, November 2007, 32 p.

0714 Competition and Innovation: An Experimental Investigation, Dario Sacco, October 2007, 36p.

0713 Hedonic Adaptation to Living Standards and the Hidden Cost of Parental Income, Stefan Boes, Kevin Staub, Rainer Winkelmann, October 2007, 18p.

0712 Competitive Politics, Simplified Heuristics, and Preferences for Public Goods, Felix Schläpfer, Marcel Schmitt, Anna Roschewitz, September 2007, 40p.

0711 Self-Reinforcing Market Dominance, Daniel Halbheer, Ernst Fehr, Lorenz Goette, Armin Schmutzler, August 2007, 34p.

0710 The Role of Landscape Amenities in Regional Development: A Survey of Migration, Regional Economic and Hedonic Pricing Studies, Fabian Waltert, Felix Schläpfer, August 2007, 34p.

0709 Nonparametric Analysis of Treatment Effects in Ordered Response Models, Stefan Boes, July 2007, 42p.

0708 Rationality on the Rise: Why Relative Risk Aversion Increases with Stake Size, Helga Fehr-Duda, Adrian Bruhin, Thomas F. Epper, Renate Schubert, July 2007, 30p.

0707 I'm not fat, just too short for my weight – Family Child Care and Obesity in Germany, Philippe Mahler, May 2007, 27p.

0706 Does Globalization Create Superstars?, Hans Gersbach, Armin Schmutzler, April 2007, 23p.

0705 Risk and Rationality: Uncovering Heterogeneity in Probability Distortion, Adrian Bruhin, Helga Fehr-Duda, and Thomas F. Epper, July 2007, 29p.

0704 Count Data Models with Unobserved Heterogeneity: An Empirical Likelihood Approach, Stefan Boes, March 2007, 26p.

0703 Risk and Rationality: The Effect of Incidental Mood on Probability Weighting, Helga Fehr, Thomas Epper, Adrian Bruhin, Renate Schubert, February 2007, 27p.

0702 Happiness Functions with Preference Interdependence and Heterogeneity: The Case of Altruism within the Family, Adrian Bruhin, Rainer Winkelmann, February 2007, 20p.

0701 On the Geographic and Cultural Determinants of Bankruptcy, Stefan Buehler, Christian Kaiser, Franz Jaeger, June 2007, 35p.

0610 A Product-Market Theory of Industry-Specific Training, Hans Gersbach, Armin Schmutzler , November 2006, 28p.

0609 Entry in liberalized railway markets: The German experience, Rafael Lalive, Armin Schmutzler, April 2007, 20p.

0608 The Effects of Competition in Investment Games, Dario Sacco, Armin Schmutzler, April 2007, 22p.