The Stata Journal (2013)
13, Number 4, pp. 810–835

Testing for zero inflation in count models: Bias correction for the Vuong test

Bruce A. Desmarais
University of Massachusetts–Amherst
Amherst, MA
desmarais@polsci.umass.edu

Jeffrey J. Harden
University of Colorado–Boulder
Boulder, CO
jeffrey.harden@colorado.edu

Abstract. The proportion of zeros in event-count processes may be inflated by an additional mechanism by which zeros are created. This has given rise to statistical models that accommodate zero inflation; these are available in Stata through the `zip` and `zinb` commands. The Vuong (1989, *Econometrica* 57: 307–333) test is regularly used to determine whether estimating a zero-inflation component is appropriate or whether a single-equation count model should be used. The use of the Vuong test in this case is complicated by the fact that zero-inflated models involve the estimation of several more parameters than the single-equation models. Although Vuong (1989, *Econometrica* 57: 307–333) suggested corrections to the test statistic to address the comparison of models with different numbers of parameters, Stata does not implement any such correction. The result is that the Vuong test used by Stata is biased toward supporting the model with a zero-inflation component, even when no zero inflation exists in the generative process. We provide new Stata commands for computing the Vuong statistic with corrections based on the Akaike and Bayesian (Schwarz) information criteria. In an extensive Monte Carlo study, we illustrate the bias inherent in using the uncorrected Vuong test, and we examine the relative merits of the Akaike and Schwarz corrections. Then, in an empirical example from international relations research, we show that errors in selecting an event-count model can have clear implications for substantive conclusions.

Keywords: st0319, zipcv, zinbcv, count models, Poisson, zero-inflated Poisson, negative binomial, zero-inflated negative binomial, Vuong test, AIC, BIC, zip, zinb

1 Introduction

In formulating statistical models, there is an inherent tension between reducing the data to a parsimonious and comprehensible summary and specifying a model that adequately captures the complexities in real data (Achen 2005). This balancing act is apparent in the modeling of event-count data with a seemingly disproportionate number of zeros. One way that this overabundance could arise is that the presence of zeros is inflated by an additional process besides the one that influences the counts that are greater than zero. Regression models for zero-inflated counts offer the benefit of accommodating multiple theories regarding the presence of zeros (Lambert 1992).

Underlying the choice between conventional-count regression and zero-inflated modeling is the common tension between overfitting and successfully explaining empirical

features of the data. A striking trait of many event-count datasets is the sheer proportion of zeros in the dependent variables. The field of international relations, which focuses to a great degree on events of an extreme and rare nature, is one in which this trait is highly prevalent. Data from international conflict and terrorism provide illustrative examples: 89% zeros in Clare (2007), 91% in Kisangani and Pickering (2007), and 97% in Neumayer and Plümper (2011). This characteristic raises an important theoretical and empirical question: Is there a process that inflates the probability of a zero case?

Much is at stake in the answer to this question. A "yes" amounts to more than just the addition of an explanatory variable—an entire process, in the form of another equation and several more parameters to estimate, is added to the model. Such an addition may be warranted if there is strong theoretical reason to expect two processes. For instance, Clare (2007) presents a theoretical differentiation of international dispute initiation and escalation. Data-driven, inductive assessments of the presence of multiple processes can be performed by formally testing whether the added complexity of the zero-inflated model improves significantly upon the fit of the standard count model. The validity of this test is critical. A false negative—choosing the standard model when the zero-inflated model should be used—directs attention away from a separate and striking component of the data-generating process (DGP). A false positive—incorrectly choosing the zero-inflated model—causes the erroneous complication of the model through the addition of an entire equation to the specification.

With this tension in mind, researchers commonly use the Vuong test (Vuong 1989) to determine whether the zero-inflated model fits the data statistically significantly better than count regression with a single equation (see Vogus and Welbourne [2003]; Anthony [2005]; Mondak and Sanders [2005]; Clare [2007]; Lee et al. [2007]; Zandersen, Termansen, and Jensen [2007]; Tiwari et al. [2009]; Nielsen et al. [2010]; Cavrini et al. [2012]; and Zhang et al. [2012]). In this article, we show that there are problems with the implementation of this test in Stata. In particular, Vuong (1989) demonstrates that bias ensues from comparing models with different parameters and suggests using an information criterion adjustment to correct this bias. The built-in Stata commands for zero-inflated count models, `zip` (zero-inflated Poisson (ZIP) regression) and `zinb` (zero-inflated negative binomial (ZINB) regression), do not implement a correction to the Vuong test statistic to account for the added parameters in the zero-inflated model. The result of having no such correction is that Stata's computation of the Vuong test statistic is strongly biased in favor of the more complex model with a zero-inflation component, even when there is no zero inflation in the true DGP.

We address this problem here by providing new Stata commands, `zipcv` and `zinbcv`, which operate exactly like `zip` and `zinb` but add computations of the Vuong test with two different corrections suggested by Vuong (1989)—one based on the Akaike information criterion (AIC) (Akaike 1974) and one based on the Bayesian (Schwarz) information criterion (BIC) (Schwarz 1978). We show that these commands allow applied researchers to properly use the Vuong (1989) test to decide between standard and zero-inflated count models.

After reviewing zero-inflated models and the details of the commands, we illustrate the use of `zinbcv` with an example from research on international disputes. Clare (2007) shows evidence from a ZINB model that redemocratizing countries with long legacies of past democratic regimes is more likely to initiate international disputes, while those with long legacies of past authoritarian regimes follow more cautious foreign policy. We show that support for these assertions is conditional on the use of the ZINB model; under a standard negative binomial (NB) model, the length of the previous democratic regime exerts a small and statistically nonsignificant effect on the expected number of disputes initiated. Moreover, while the uncorrected Vuong test statistic from `zinb` selects the ZINB model ($p < 0.05$), the test statistic with the BIC correction in `zinbcv` selects the NB model ($p < 0.05$).

2 Zero-inflated count models

The class of zero-inflated count regression models first proposed by Lambert (1992) as the ZIP model, is a mixture between a generalized linear model (GLM) for the dichotomous outcome that a count Y is equal to zero (such as logit, with covariates $\mathbf{z}$ and coefficients $\boldsymbol{\gamma}$) and a conventional event-count GLM (such as a Poisson or NB regression with covariates $\mathbf{x}$ and coefficients $\boldsymbol{\beta}$). The likelihood of a single observation is given by the following equation (Long 1997, 244),

$$l(y|\mathbf{x}, \mathbf{z}, \boldsymbol{\beta}, \boldsymbol{\gamma}) = P(\mathbf{z}'\boldsymbol{\gamma})I(y = 0) + \{1 - P(\mathbf{z}'\boldsymbol{\gamma})\}\, f(y|\mathbf{x}'\boldsymbol{\beta})$$

where P is the cumulative distribution function used to specify the dichotomous outcome that $y > 0$, and f is the probability mass function corresponding to the chosen count model (for example, the Poisson distribution). Using the log link to parameterize f, we obtain the mean of y_i:

$$\mu_i = \{1 - P(\mathbf{z}_i'\boldsymbol{\gamma})\} \exp(\mathbf{x}_i'\boldsymbol{\beta})$$

There are several important properties of this model to note. First, the probability of a zero is governed by both the dichotomous and count equations in the model. Specifically,

$$\Pr(y_i = 0) = P(\mathbf{z}_i'\boldsymbol{\gamma}) + \{1 - P(\mathbf{z}_i'\boldsymbol{\gamma})\} f(0|\mathbf{x}_i'\boldsymbol{\beta})$$

This is different from the popular hurdle models, first proposed by Mullahy (1986), in which the probability of a 0 is completely determined by a dichotomous GLM and the distribution of counts above 0 is governed by a count distribution truncated from below at 1. Second, the count regression f is not "nested" in the zero-inflated model, because the model does not reduce to f when $\boldsymbol{\gamma} = \mathbf{0}$, in which case the probability of a 0 is inflated by 0.50. The main implication stemming from these properties is that it is necessary to compare the zero-inflated model with a simple count model using a test for nonnested models. The conventional likelihood-ratio test, Wald test, or Lagrange multiplier test cannot be used (Long 1997).

2.1 The Vuong test

The Vuong test is designed to compare two models (g_1 and g_2) fit to the same data by maximum likelihood. Specifically, it tests the null hypothesis that the two models fit the data equally well. The models need not be nested, nor does one of the models need to represent the correct specification. The specific metric of model fit is the Kullback–Leibler divergence (KLD) (Kullback and Leibler 1951) from the true model that generated the data (g_t). The KLD is a measure of the distance between two probability distributions, which is the basis of many measures used for model comparison and selection, including the AIC (Akaike 1974), the Takeuchi information criterion (Konishi and Kitagawa 1996), the generalized information criterion (Konishi and Kitagawa 1996), and the cross-validated log likelihood (Smyth 2000). The KLD between models g and g_t is denoted $D_{\mathrm{KL}}(g_t||g)$. The null hypothesis of the Vuong test is

$$H_0 \colon D_{\mathrm{KL}}(g_t||g_1) = D_{\mathrm{KL}}(g_t||g_2)$$

The formula for $D_{\mathrm{KL}}(g_t||g)$, where g_t and g are both models for nonnegative integers (for example, counts), is defined as

$$\begin{aligned} D_{\mathrm{KL}}(g_t||g) &= \sum_{y=0}^{\infty} \ln\left\{\frac{g_t(y)}{g(y)}\right\} g_t(y) \\ &= \sum_{y=0}^{\infty} \ln\left\{g_t(y)\right\} g_t(y) - \sum_{y=0}^{\infty} \ln\left\{g(y)\right\} g_t(y) \\ &= E_{g_t}[\ln\{g_t(y)\}] - E_{g_t}[\ln\{g(y)\}] \end{aligned}$$

From this, the null hypothesis, H_0, can be written as

$$\begin{aligned} H_0 \colon D_{\mathrm{KL}}(g_t||g_1) - D_{\mathrm{KL}}(g_t||g_2) &= 0 \\ (E_{g_t}[\ln\{g_t(y)\}] - E_{g_t}[\ln\{g_1(y)\}]) - (E_{g_t}[\ln\{g_t(y)\}] - E_{g_t}[\ln\{g_2(y)\}]) &= 0 \\ E_{g_t}[\ln\{g_1(y)\}] - E_{g_t}[\ln\{g_2(y)\}] &= 0 \quad (1) \end{aligned}$$

(1) is the difference in the expected values of the log likelihoods of g_1 and g_2 when their parameters are estimated on data generated from g_t. Importantly, $E_{g_t}[\ln\{g_1(y)\}]$ and $E_{g_t}[\ln\{g_2(y)\}]$ are not formulated under the assumption that the same sample is used to estimate the parameters and evaluate the likelihoods. For a sample size N, the Vuong test is a difference of means test (that is, a paired z test) applied to the N individual log-likelihood contributions (of the N observations) to g_1 and g_2. In the context of testing for zero inflation, the Vuong test is a test for whether the mean observation-wise difference between the log-likelihood contribution to the zero-inflation model and the contribution to the standard count model is, on average, greater than zero. Let $\widetilde{\boldsymbol{\beta}}$ be the estimate of $\boldsymbol{\beta}$ when the zero-inflation component is not included in the model, $\widehat{\boldsymbol{\beta}}$ the estimate of $\boldsymbol{\beta}$ in the zero-inflation model, and $\widehat{\boldsymbol{\gamma}}$ the estimate of $\boldsymbol{\gamma}$. Let $\mathbf{dl}$ be a vector of length N, such that the ith element is the ith individual log-likelihood difference

$$dl_i = \ln\left\{l\left(y_i|\mathbf{x_i}, \mathbf{z_i}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}}\right)\right\} - \ln\left\{f\left(y_i|\mathbf{x_i}'\widetilde{\boldsymbol{\beta}}\right)\right\}$$

The Vuong test statistic is

$$\text{Vuong} = \left(s_{dl}\sqrt{n}\right)^{-1} \sum_{i=1}^{n} dl_i$$

where s_{dl} is the standard deviation of **dl**. Because $1/N \sum_{i=1}^{n} dl_i$ is a *consistent* estimator of the quantity in (1), under H_0, the Vuong test statistic is asymptotically normally distributed by the central limit theorem (Vuong 1989).

The estimated log likelihood is a consistent estimator of the KLD, which establishes the consistency and asymptotic normality of the Vuong test statistic. However, the estimated log likelihood is a biased estimator of the KLD, a result that motivated the derivation of the AIC (Akaike 1974) and numerous other model fit statistics. This means that the Vuong test statistic is a biased estimator of the differences in the average fit of the count model and zero-inflated count model. The bias in the estimated log likelihood as an estimator of the KLD arises from the fact that the same data are used to estimate both the parameters of the model (that is, coefficients and standard errors) and the average value of the log likelihood. This "double dipping" produces a positive bias in the in-sample log likelihood as an estimator of the KLD (Konishi and Kitagawa 1996). Intuitively, this bias arises because some of the random noise from the sample gets treated as nonrandom signal when estimating the KLD with the log likelihood.

It is generally intractable to derive the value of this bias in a finite sample, so model selection criteria use asymptotic corrections. For example, the AIC uses the correction p (the number of estimated parameters), which is equal to the asymptotic bias, given that g_t is nested in the fit model. The bias is accentuated when g_1 and g_2 have a different number of parameters (p_1 and p_2, respectively), as is the case when comparing single-equation and zero-inflated count models. Vuong (1989) suggests adding an average difference in a selection-criterion-based correction factor to each dl_i to correct the bias.

For instance, if the correction factor is based on the AIC, the corrected difference in log likelihoods is given by

$$dl_i^c = dl_i + \frac{p_2 - p_1}{N} \tag{2}$$

Vuong (1989) also provides the BIC correction as

$$dl_i^c = dl_i + (p_2 - p_1)\frac{\ln(N)}{2N} \tag{3}$$

Another possibility would be to use an out-of-sample approach to computing the individual log-likelihood contributions, such as through leave-one-out cross-validation (for example, Smyth [2000]). Rendering the training and testing data independent of one another removes the optimistic bias of in-sample measures. We tested such an approach in the analysis described below and found minimal differences between it and the asymptotic corrections in (2) (AIC) and especially (3) (BIC). Not surprisingly, these differences were particularly small as N increased. Because the iterative nature of leave-one-out cross-validation produces considerable computational costs, we elected not to include it in the corrections to the Vuong test we examine here.

3 Monte Carlo simulations

Having outlined the basic premise of zero-inflated models and the Vuong (1989) test, we next examine via simulation the consequences of failing to correct the Vuong statistic when comparing standard and zero-inflation count models. The Stata commands `zip` and `zinb` offer the option of reporting a Vuong test statistic, which many researchers use (for example, Vogus and Welbourne [2003]; Anthony [2005]; Mondak and Sanders [2005]; Clare [2007]; Zandersen, Termansen, and Jensen [2007]). However, Stata's documentation for the Vuong statistic does not mention which adjustment (AIC or BIC) is used. Stata's technical support informed us that current versions of Stata do not include any adjustment. We then verified this by inspection of `zip.ado` and `zinb.ado`.[1]

We study the performance of the Vuong test in selecting between ZIP and Poisson models and ZINB and NB models. In the simulation study, we examine the consequences of two important dimensions for the performance of the uncorrected and corrected tests: first, the sample size, and second, the number of covariates in the inflation component of the model, both in generating the dependent variable and fitting the zero-inflated models. We use Stata's example dataset `fish.dta` to parameterize the simulation study. Approximately 57% of the 250 observations in this dataset have a value of 0 on the dependent variable. To define parameters for the data simulated in the Monte Carlo study, we first fit zero-inflated models with `count` as the dependent variable and standardized versions of `nofish`, `livebait`, `camper`, `persons`, and `child` as independent variables in both the count and inflation components. The linear predictors in the count components of the models in the Poisson- and NB-based simulations, respectively, are

$$\begin{aligned}\mathbf{x}'\boldsymbol{\beta} = 0.734 &- 0.384\texttt{nofish} + 0.376\texttt{livebait} + 0.264\texttt{camper} \\ &+ 0.940\texttt{persons} - 1.004\texttt{child}\end{aligned}$$

and

$$\begin{aligned}\mathbf{x}'\boldsymbol{\beta} = 0.515 &- 0.127\texttt{nofish} + 0.504\texttt{livebait} + 0.106\texttt{camper} \\ &+ 1.131\texttt{persons} - 1.013\texttt{child}\end{aligned}$$

In the following equations, the inflation components contain a number of terms equal to the number of covariates included in the respective condition in the simulation study. The formulas are given below for the Poisson and negative binomial simulations, respectively.[2]

$$\begin{aligned}\mathbf{z}'\boldsymbol{\gamma} = -0.157 &+ 1.73\texttt{child} - 0.669\texttt{persons} - 0.443\texttt{camper} - 0.176\texttt{livebait} \\ &- 0.638\texttt{nofish}\end{aligned}$$

1. Specifically, we verified this with `zip.ado` version 1.6.11 (6/6/2011) and `zinb.ado` version 1.7.12 (4/19/2012). Both of these were the current files as of 29 April 2013.
2. In the appendix, we present a replication of this Monte Carlo study using slightly different parameterizations and `medpar.dta`.

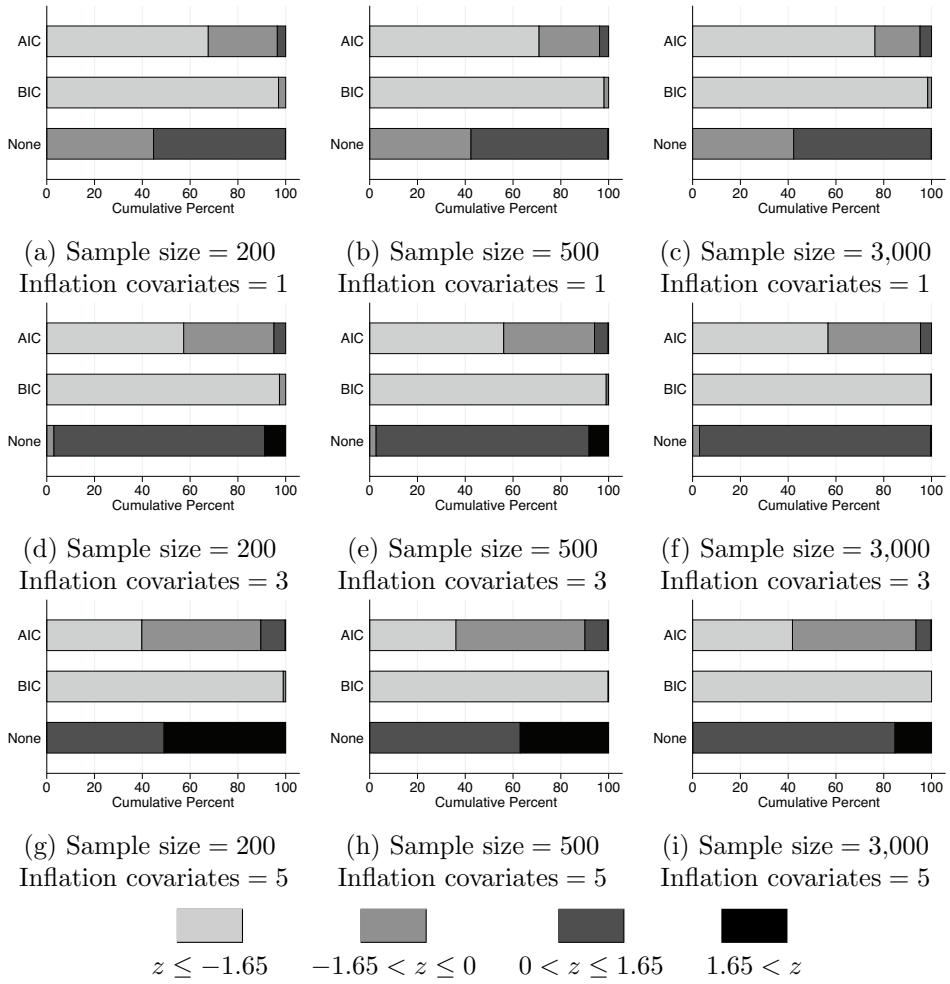
and

$$\mathbf{z}'\boldsymbol{\gamma} = -1.92 + 2.63\texttt{child} - 1.065\texttt{persons} - 1.23\texttt{camper} + 0.161\texttt{livebait} - 0.619\texttt{nofish}$$

The bias in the observed log likelihood as an estimator of the expected log likelihood, the resulting bias in the Vuong test statistic, and the bias corrections associated with AIC and BIC depend upon the sample size and the difference in the number of parameters in the two models under comparison (Konishi and Kitagawa 1996). Accordingly, the two conditions on which we focus are the sample size and the number of variables in the inflation component of the model. We examine sample sizes of 200, 500, and 3,000. In terms of the number of parameters, we vary the inflation component in two ways. First, we run simulations in which there is no zero inflation, drawing the outcomes from the Poisson and NB models, and study the performance of the three test variants when one, three, and five covariates are incorrectly included in the inflation component. We then run a second variant in which there is zero inflation and examine the performance of the tests when one, three, and five covariates are correctly included in the inflation component. Each of the 36 conditions (3 sample sizes $\times$ 3 covariate specifications $\times$ 2 inflation/no inflation $\times$ 2 distributions) is run for 1,000 iterations.[3]

Figures 1–2 present the results of the simulations under the condition of no zero inflation. The plots illustrate the results of hypothesis tests derived from the uncorrected, AIC-corrected, and BIC-corrected Vuong statistics. The graphs depict the distribution of significance test results based on the Vuong test comparing standard to zero-inflated count models with the respective correction. To demonstrate interpretation of the plots, we walk through the results conveyed in panel (a) of figure 1. Panel (a) gives the results for the simulations with a sample size of 200 and a single covariate incorrectly included in the inflation component of the ZIP model. The AIC-corrected test statistically significantly (at the 0.05 one-tailed level) selects the single-equation Poisson model around 67% of the time, supports the Poisson model (though not significantly) approximately 30% of the time, and supports the two-equation ZIP model (though not significantly) approximately 3% of the time. The BIC-corrected test statistically significantly selects the Poisson model about 97% of the time and supports the Poisson model (though not significantly) approximately 3% of the time. The Vuong test without a correction supports the Poisson model (though not significantly) approximately 45% of the time and supports the ZIP (though not significantly) approximately 55% of the time.

3. We performed all the computations presented in this section in Stata/SE 11.1 and Stata/IC 12.1.



(a) Sample size = 200 Inflation covariates = 1
(b) Sample size = 500 Inflation covariates = 1
(c) Sample size = 3,000 Inflation covariates = 1
(d) Sample size = 200 Inflation covariates = 3
(e) Sample size = 500 Inflation covariates = 3
(f) Sample size = 3,000 Inflation covariates = 3
(g) Sample size = 200 Inflation covariates = 5
(h) Sample size = 500 Inflation covariates = 5
(i) Sample size = 3,000 Inflation covariates = 5

Figure 1. Monte Carlo results with Poisson simulations. The plots depict the distribution of significance test results based on the Vuong test comparing Poisson to ZIP models with the respective correction across varying sample sizes and numbers of covariates incorrectly included in the inflation component.

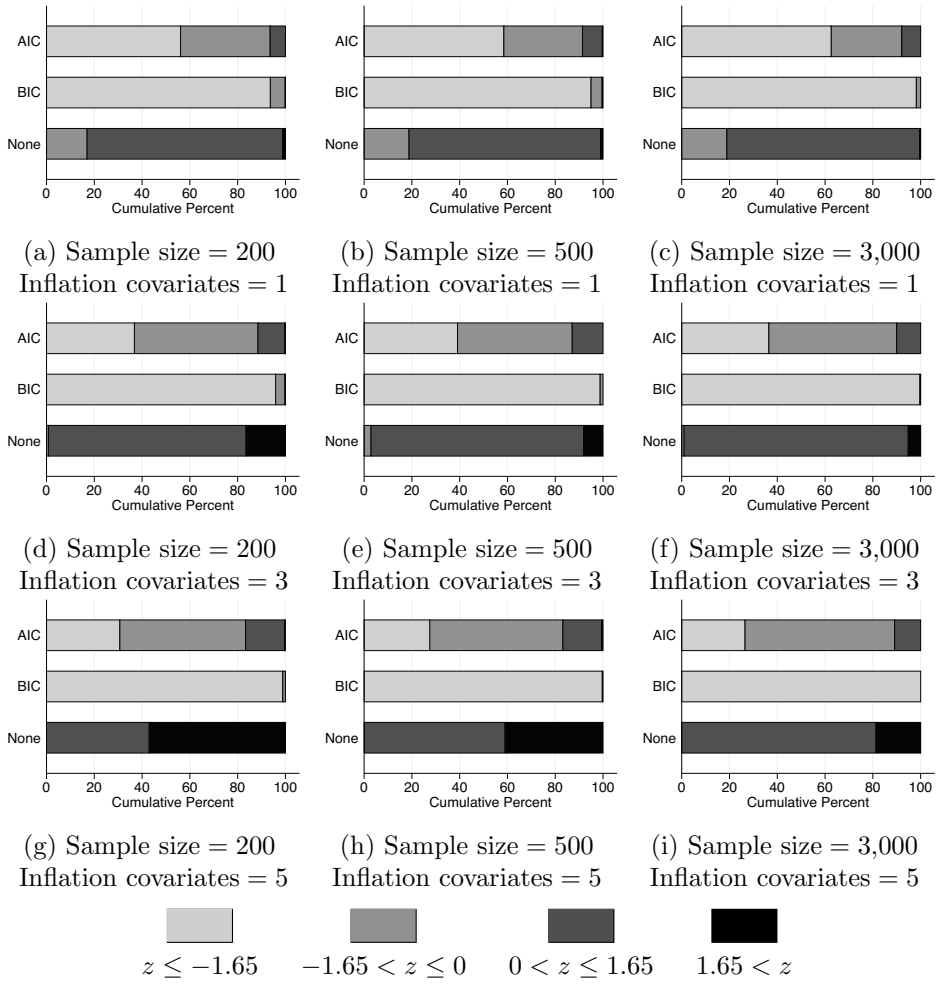


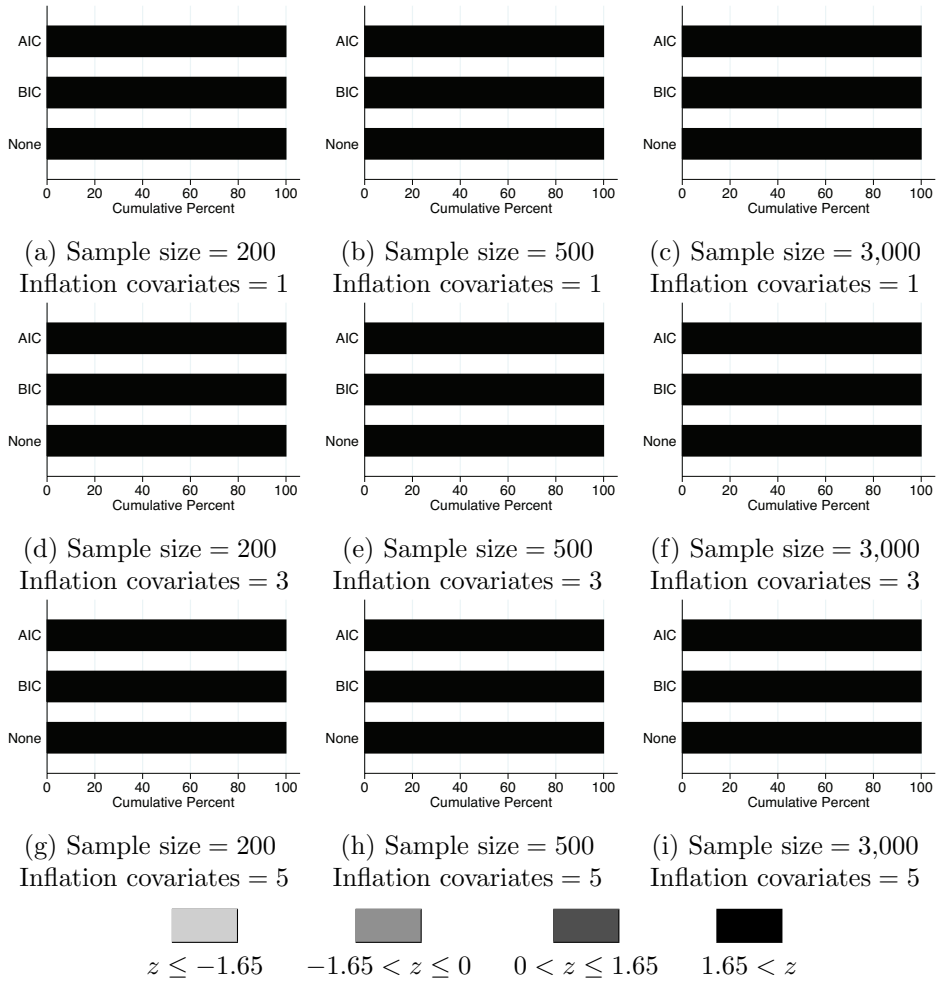
Figure 2. Monte Carlo results with NB simulations. The plots depict the distribution of significance test results based on the Vuong test comparing NB to ZINB models with the respective correction across varying sample sizes and numbers of covariates incorrectly included in the inflation component.

When there is no zero inflation in the DGP, the BIC-corrected statistic performs the best, and the uncorrected statistic performs the worst. The BIC-corrected statistic is statistically significantly negative ($p < 0.05$, one tailed)—in favor of the single-equation model—in 95–100% of the iterations. In contrast, the uncorrected Vuong statistic is positive in more than 80% of the iterations and statistically significantly in favor of the zero-inflated NB model in 5–60% of the iterations. The poor performance of the uncorrected test depends heavily on sample size and the number of covariates included in the

inflation component.[4] However, it is particularly critical to note that not once in the simulation runs without zero inflation did the uncorrected test result in a statistically significant rejection of the zero-inflated model. The AIC-corrected test performs moderately better in the no-inflation condition. In 20–60% of iterations, the zero-inflated model is statistically significantly rejected, and the single-equation model is virtually never rejected. However, the degree to which the AIC favors the single-equation count model decreases with the number of covariates incorrectly included in the inflation component.

Figures 3 and 4 present results in which zero inflation is a component of the generative process. In the Poisson-based simulations, all the tests perform equally well, nearly always rejecting the single-equation model. However, when it comes to the NB-based simulations, the uncorrected Vuong statistic performs the best in selecting the correctly specified model—nearly always statistically significantly rejecting the single equation model. The AIC-corrected test performs moderately well in the small sample ($N = 200$) conditions, significantly favoring the zero-inflated model in 40–50% of the iterations and virtually always statistically significantly selecting the zero-inflation model in the larger sample-size conditions. The performance of the BIC-corrected statistic—performing the worst among the three when ZINB is the correct model—varies substantially across the sample size and covariate conditions. The tendency for the BIC-corrected statistic to statistically significantly reject NB is inversely related to the number of covariates correctly included in the zero-inflation component and directly related to the sample size.

4. Specifically, the smaller the sample and the larger the number of covariates incorrectly included in the inflation component, the more likely the uncorrected test is to statistically significantly reject the zero-inflated model.



(a) Sample size = 200 Inflation covariates = 1
(b) Sample size = 500 Inflation covariates = 1
(c) Sample size = 3,000 Inflation covariates = 1
(d) Sample size = 200 Inflation covariates = 3
(e) Sample size = 500 Inflation covariates = 3
(f) Sample size = 3,000 Inflation covariates = 3
(g) Sample size = 200 Inflation covariates = 5
(h) Sample size = 500 Inflation covariates = 5
(i) Sample size = 3,000 Inflation covariates = 5

Figure 3. Monte Carlo results with ZIP simulations. The plots depict the distribution of significance test results based on the Vuong test comparing Poisson to ZIP models with the respective correction across varying sample sizes and numbers of covariates correctly included in the inflation component.

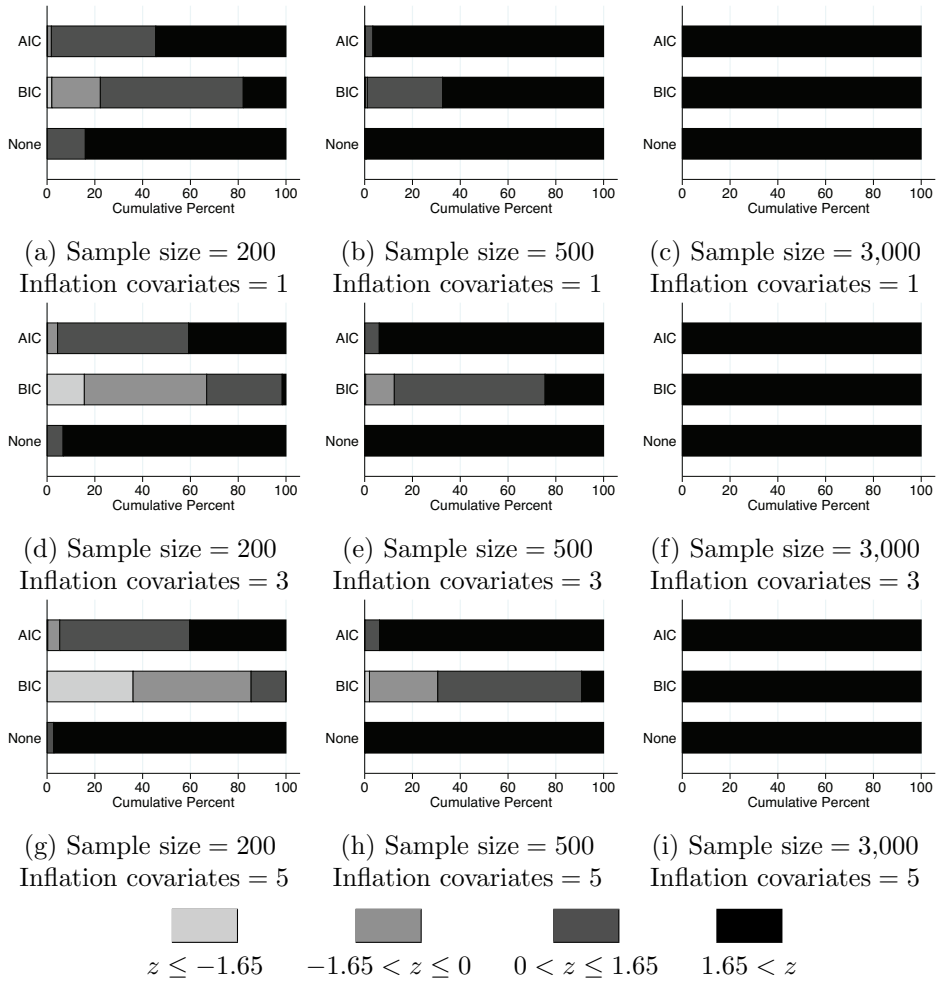


Figure 4. Monte Carlo results with ZINB simulations. The plots depict the distribution of significance test results based on the Vuong test comparing NB to ZINB models with the respective correction across varying sample sizes and numbers of covariates correctly included in the inflation component.

Our simulation study illustrates two important points regarding the use of the Vuong test for choosing between zero-inflated and single-equation count models. First, failure to correct for the additional parameters estimated in the zero-inflation model by using the uncorrected Vuong statistic results in a substantial tendency toward erroneously rejecting the single-equation model when there is no zero inflation in the generative process. In small to moderate sample sizes with five or more covariates included in the inflation component, this tendency can exceed 40%. Second, the AIC and BIC correc-

tions exhibit their usual relative strengths. The BIC correction is better at conclusively supporting the more parsimonious single-equation model when it is appropriate, and the AIC is better at conclusively supporting the more extensively specified zero-inflated model when it is appropriate. Moreover, neither the AIC- nor the BIC-corrected tests exhibit the extreme tendency toward statistical significance in the wrong direction that is exhibited by the uncorrected statistic when there is no zero inflation in the generative process. In larger samples, the BIC-corrected test appears to exhibit an advantage in that it both performs very well at rejecting the zero-inflated model when there is no zero inflation and rejecting the single-equation model when zero inflation is present. In contrast, the AIC-corrected test does not perform well at rejecting the zero-inflation model when there is no zero inflation, even in our large-sample conditions.

4 Model selection in international relations

Having shown the problem with the uncorrected Vuong statistic with simulation and the corresponding improvements offered by the AIC or BIC correction, we now turn to their application to data from recent work using event-count models in international relations (Clare 2007).[5] Our objectives here are to demonstrate that the different implementations of the Vuong test for zero inflation can produce considerably different results in an applied setting and to provide an illustration of our `zinbcv` command.[6] In addition to illustrating the use of the new command, we show that the selection made by the test is critical to our understanding of important processes, such as conflict behavior.

4.1 The data

Clare (2007) examines the conflict behavior of democratizing regimes. He posits that redemocratizing states are more likely to initiate conflict, especially when there is a longer democratic history in the state. In contrast, he expects a stronger authoritarian legacy to correspond with less initiation of conflict. Clare's (2007) primary theoretical claim is that leaders of democratizing states face varying degrees of threat of losing power to the old authoritarian regime because of failed foreign policy. Thus democratizing states have more freedom to maneuver in foreign policy decision making when the authoritarian legacy is weak and less freedom when it is strong.

Using nation-year as the unit of analysis for the period 1950–1990, Clare (2007) models the count of disputes initiated by a state in a given year as a function of several independent variables, including indicators for the regime type (see Clare [2007, 267]), and measures of the duration of the most recent authoritarian and democratic regimes. The core test of the theory comes through the interaction of redemocratization—an indicator for a state that is in the process of democratizing—and each of these two

5. The data for this example in Stata format are publicly available at the *Journal of Peace Research* replication data archive:
http://www.prio.no/Journals/Journal/?x=2&content=replicationData#2007.

6. The `zipcv` command works in exactly the same way as `zinbcv`, so we only show the latter to conserve space.

regime duration measures. Clare (2007) expects the interaction between redemocratization and duration of the most recent authoritarian regime to produce a negative coefficient, which indicates a drop in the expected number of disputes initiated when the past authoritarian legacy is longer. In contrast, he expects redemocratization × duration of the most recent democratic regime to produce a positive coefficient, which indicates an increase in the expected number of disputes initiated when the past democratic regime is longer.

4.2 Computing the Vuong test

Clare (2007) uses the ZINB model in estimation because 89% of the 3,955 cases in the data contain a 0 on the dependent variable.[7] He also reports the uncorrected Vuong test statistic of 3.22, which corresponds to a statistically significant selection of the ZINB over the NB ($p \approx 0.001$). However, this test statistic is problematic because it does not correct for additional parameters from the inflation equation. To obtain the corrected test statistics, as well as the results of the `zinb` routine, we use `zinbcv` in the exact same way as `zinb`:

7. He includes the same set of covariates in the count and inflation equations.

```
. use clare-monadic-replication.dta
. zinbcv init_count stable_dem1 stable_aut1 redem1 redem1_pautdur1
> redem1_pdemdur1  growth prop_demsregion1 riots1,
> inflate(stable_dem1 stable_aut1 redem1 redem1_pautdur1 redem1_pdemdur1
> growth prop_demsregion1 riots1) vuong nolog
Zero-inflated negative binomial regression        Number of obs   =       3955
                                                  Nonzero obs     =        442
                                                  Zero obs        =       3513
Inflation model = logit                           LR chi2(8)      =      98.05
Log likelihood  = -1541.107                       Prob > chi2     =     0.0000
```

init_count	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
init_count						
stable_dem1	.0792226	.2257374	0.35	0.726	-.3632145	.5216598
stable_aut1	.3562161	.2275137	1.57	0.117	-.0897026	.8021349
redem1	.9875878	.7367568	1.34	0.180	-.4564291	2.431605
redem1_pautdur1	-.1189129	.0615584	-1.93	0.053	-.2395651	.0017393
redem1_pdemdur1	.0721211	.0428799	1.68	0.093	-.011922	.1561642
growth	.0000189	2.05e-06	9.22	0.000	.0000149	.000023
prop_demsregion1	-1.434321	.2373772	-6.04	0.000	-1.899572	-.9690703
riots1	.0270679	.0126341	2.14	0.032	.0023056	.0518302
_cons	-1.671941	.2222414	-7.52	0.000	-2.107526	-1.236356
inflate						
stable_dem1	-12.35688	595.2899	-0.02	0.983	-1179.104	1154.39
stable_aut1	1.260302	1.375148	0.92	0.359	-1.434939	3.955542
redem1	4.442412	5.332937	0.83	0.405	-6.009953	14.89478
redem1_pautdur1	-.5272241	.5540954	-0.95	0.341	-1.613231	.5587829
redem1_pdemdur1	.6353599	.4939444	1.29	0.198	-.3327533	1.603473
growth	-.0000544	.0000169	-3.22	0.001	-.0000876	-.0000213
prop_demsregion1	-16.1701	10.59818	-1.53	0.127	-36.94215	4.601959
riots1	-.2118948	.1103174	-1.92	0.055	-.428113	.0043234
_cons	-.4104956	1.399768	-0.29	0.769	-3.15399	2.332999
/lnalpha	-.4193847	.3544373	-1.18	0.237	-1.114069	.2752996
alpha	.6574512	.2330252			.3282207	1.316925

```
Vuong test of zinb vs. standard negative binomial:  z =    3.22  Pr>z = 0.0006
                                                                 Pr<z = 0.9994
                     with AIC (Akaike) correction:  z =    1.77  Pr>z = 0.0386
                                                                 Pr<z = 0.9614
                    with BIC (Schwarz) correction:  z =   -2.80  Pr>z = 0.9974
                                                                 Pr<z = 0.0026
. display e(vuong)
3.219827
. display e(vuongAIC)
1.7674489
. display e(vuongBIC)
-2.7950052
```

This prints all the information users are accustomed to seeing with `zinb`, but also includes the corrected versions of the Vuong statistic under the uncorrected version. Additionally, the exact values are stored in `e(vuongAIC)` and `e(vuongBIC)`. In this case, the Vuong test statistic with the AIC correction is 1.77, which still corresponds

to a statistically significant selection of the zero-inflated model ($p \approx 0.04$). However, the Vuong test statistic with the BIC correction is -2.80, which represents a significant selection of the standard NB model ($p \approx 0.003$). In short, there is considerable variation in the results from the Vuong test. However, given the results from our above simulation and the relatively large sample size, we place more weight on the BIC-corrected Vuong test and conclude that the more parsimonious NB is the appropriate model.

4.3 Implications of the results

Table 1 summarizes results from both the ZINB and NB models. Notice first that the original ZINB shows support for Clare's (2007) hypotheses. In particular, the coefficient on redemocratization×duration of the most recent authoritarian regime is negative, the coefficient on redemocratization×duration of the most recent democratic regime is positive, and both are statistically significant. This indicates that in states that are transitioning to democracy, a longer legacy of authoritarian rule contributes to a decline in the number of disputes initiated, while a longer legacy of democratic rule corresponds with an increase in disputes. Additionally, these effects are substantively meaningful. As Clare (2007, 270–271) notes, an increase of 1 year in the duration of the previous authoritarian regime corresponds to an 11% drop in the expected number of disputes, while the same increase for democratic regimes produces an 8% increase in the expected number of disputes.

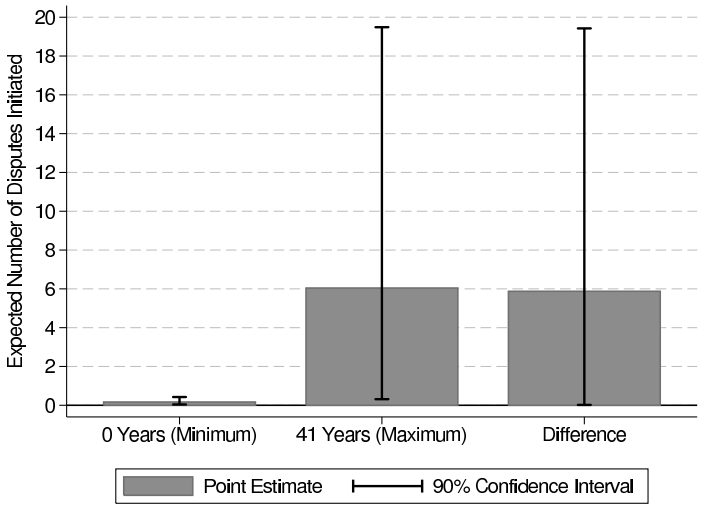
Table 1. ZINB and NB results from Clare (2007)

Variable	ZINB	NB
Stable democracy	0.08	−0.09
	(0.23)	(0.22)
Stable autocracy	0.36*	0.14
	(0.23)	(0.20)
Redemocratization	0.99	0.95*
	(0.74)	(0.43)
Redemocratization ×	−0.12*	−0.10*
Duration of most recent authoritarian regime	(0.06)	(0.04)
Redemocratization ×	0.07*	0.03
Duration of most recent democratic regime	(0.04)	(0.03)
Economic growth	1.9e−5*	2.1e−5*
	(2.1e−6)	(2.3e−6)
Riots	0.03*	0.04*
	(0.01)	(0.01)
Other democratic countries in the region	−1.43*	−0.76*
	(0.24)	(0.21)
Intercept	−1.67*	−2.02*
	(0.22)	(0.21)
N (zeros)	3,955 (3,514)	3,955 (3,514)
Vuong (uncorrected)		3.22*
Vuong (AIC)		1.77*
Vuong (BIC)		−2.80*

Note: Cell entries report coefficient estimates with standard errors in parentheses for Clare (2007) original ZINB model estimates and a replication using NB. Positive Vuong test statistic values indicate a selection of the ZINB model, and negative values indicate a selection of the NB model.
* $p < 0.05$ (one-tailed).

However, note that the coefficients on each of these interaction terms decline in magnitude in the NB model with redemocratization × duration of the most recent democratic regime dropping by more than half the value of the ZINB estimate. Furthermore, this latter coefficient is no longer statistically significant at the 0.05 level in the NB model. We assess the substantive implications of this in figure 5. Both graphs plot the expected number of disputes initiated by redemocratizing regimes on the y axis at the minimum and maximum values of duration of the most recent democratic regime (0 and 41 years, respectively). The third bar plots the difference between these two estimates. Panel (a) gives results from the ZINB model, and panel (b) shows NB results. Note that the difference is large (≈ 6 disputes) if ZINB is used but small (< 1 dispute) with the better-fitting NB. Thus at least half of the support for the original theory depends on

using the ZINB model instead of the standard NB. This is problematic in light of the fact that the BIC-corrected Vuong test clearly supports the rejection of the ZINB model.



(a) ZINB

(b) NB

Figure 5. Change in the expected number of disputes initiated by redemocratizing states from the minimum (0 years) to maximum (41 years) observed value of duration of the most recent democratic regime. The difference is large (≈ 6 disputes) if the ZINB model is used but small (< 1 dispute) with the better-fitting standard NB.

5 Conclusions

In formulating and evaluating statistical models of event-count processes, we encounter an inherent tension between developing a parsimonious summary of the data and accounting for meaningful empirical peculiarities. Because these processes are often defined on events that are relatively rare, such as dispute initiation, scholars regularly confront datasets with many zeros on the dependent variable. An important question stemming from this characteristic centers on whether some of these zeros arise because of an additional generative mechanism. If so, the proper inferential method is to fit a count model with a zero-inflation equation to account for the second process. This added complexity comes with a risk; statistically, specifying an inflation equation when one is not needed reduces the efficiency of the estimator and convolutes interpretation. Perhaps worse, theoretically, the inclusion of an additional equation in the model focuses researchers' efforts on a potentially erroneous account of the process under study.

A common response to this tension is the use of Vuong (1989) nonnested model-selection procedure, which provides a test statistic that can be used to compare standard and zero-inflated count models fit to the same data. We show there are problems with the current implementation of this test in applied research. In particular, the Vuong test executed in Stata's `zip` and `zinb` commands does not implement any correction for the added parameters estimated for the inflation equation, which leads to a test that favors the zero-inflated models, even when there is no zero inflation in the generative process. We solve this problem with the `zipcv` and `zinbcv` commands. These commands include all the functionality of the `zip` and `zinb` commands that are currently in Stata but report the uncorrected, AIC-corrected, and BIC-corrected Vuong test statistics.

Finally, in a replication analysis, we apply the findings from the simulation studies to real data. Results show that the process of selecting between competing, count models can have implications for substantive conclusions from results on international political processes. In the presence of nontrivial model dependence shown in the example, researchers need statistically sound criteria on which to make decisions. The suggestions given here provide such criteria for scholars in selecting between standard and zero-inflated count models.

6 References

Achen, C. H. 2005. Let's put garbage-can regressions and garbage-can probits where they belong. *Conflict Management and Peace Science* 22: 327–339.

Akaike, H. 1974. A new look at the statistical model identification. *IEEE Transactions on Automatic Control* 19: 716–723.

Anthony, D. 2005. Cooperation in microcredit borrowing groups: Identity, sanctions, and reciprocity in the production of collective goods. *American Sociological Review* 70: 496–515.

Cavrini, G., S. Broccoli, A. Puccini, and M. Zoli. 2012. EQ-5D as a predictor of mortality and hospitalization in elderly people. *Quality of Life Research* 21: 269–280.

Clare, J. 2007. Democratization and international conflict. *Journal of Peace Research* 44: 259–276.

Kisangani, E. F., and J. Pickering. 2007. Diverting with benevolent military force: Reducing risks and rising above strategic behavior. *International Studies Quarterly* 51: 277–299.

Konishi, S., and G. Kitagawa. 1996. Generalised information criteria in model selection. *Biometrika* 83: 875–890.

Kullback, S., and R. A. Leibler. 1951. On information and sufficiency. *Annals of Mathematical Statistics* 22: 79–86.

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

Lee, Y.-G., J.-D. Lee, Y.-I. Song, and S.-J. Lee. 2007. An in-depth empirical analysis of patent citation counts using zero-inflated count data model: The case of KIST. *Scientometrics* 70: 27–39.

Long, J. S. 1997. *Regression Models for Categorical and Limited Dependent Variables*. Thousand Oaks, CA: Sage.

Mondak, J. J., and M. S. Sanders. 2005. The complexity of tolerance and intolerance judgments: A response to Gibson. *Political Behavior* 27: 325–337.

Mullahy, J. 1986. Specification and testing of some modified count data models. *Journal of Econometrics* 33: 341–365.

Neumayer, E., and T. Plümper. 2011. Foreign terror on Americans. *Journal of Peace Research* 48: 3–17.

Nielsen, S. E., G. McDermid, G. B. Stenhouse, and M. S. Boyce. 2010. Dynamic wildlife habitat models: Seasonal foods and mortality risk predict occupancy-abundance and habitat selection in grizzly bears. *Biological Conservation* 143: 1623–1634.

Schwarz, G. 1978. Estimating the dimension of a model. *Annals of Statistics* 6: 461–464.

Smyth, P. 2000. Model selection for probabilistic clustering using cross-validated likelihood. *Statistics and Computing* 10: 63–72.

Tiwari, A., J. A. VanLeeuwen, I. R. Dohoo, G. P. Keefe, J. P. Haddad, H. M. Scott, and T. Whiting. 2009. Risk factors associated with Mycobacterium avium subspecies paratuberculosis seropositivity in Canadian dairy cows and herds. *Preventive Veterinary Medicine* 88: 32–41.

Vogus, T. J., and T. M. Welbourne. 2003. Structuring for high reliability: HR practices and mindful processes in reliability-seeking organizations. *Journal of Organizational Behavior* 24: 877–903.

Vuong, Q. H. 1989. Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica* 57: 307–333.

Zandersen, M., M. Termansen, and F. S. Jensen. 2007. Testing benefits transfer of forest recreation values over a twenty-year time horizon. *Land Economics* 83: 412–440.

Zhang, X., Y. Lei, D. Cai, and F. Liu. 2012. Predicting tree recruitment with negative binomial mixture models. *Forest Ecology and Management* 270: 209–215.

About the authors

Bruce A. Desmarais is an assistant professor in the Department of Political Science at the University of Massachusetts–Amherst and a core faculty member in the Computational Social Science Initiative at the University of Massachusetts–Amherst.

Jeffrey J. Harden is an assistant professor in the Department of Political Science at the University of Colorado–Boulder.

A Appendix: Simulations using medpar.dta

Here we replicate our Monte Carlo simulation using Stata's example dataset for the `ztnb` (zero-truncated NB) command, `medpar.dta`, to parameterize the simulation study. The settings in this replication of the simulation study differ in a few ways from the original version. First, the sample sizes differ slightly. Second, we use fewer variables in the count and inflation components such that the sets of variables in each equation are disjoint. Third, the dataset used to select parameter values has relatively fewer zeros.

Approximately 8% of the 1,495 observations in `medpar.dta` have a value of 1 on the dependent variable. We generate zeros by subtracting 1 from the dependent variable before proceeding with the simulation study. To define parameters in the simulation study, we first fit ZINB and ZIP models with `los` $-$ 1 as the dependent variable and standardized versions of `hmo`, `age`, and `type1` as independent variables in the count component and `died`, `white`, and `age80` as independent variables in the inflation component. The linear predictors in the count components of the models from which the outcome is simulated in the Poisson- and NB-based simulations, respectively, are

$$\mathbf{x}'\boldsymbol{\beta} = 0.249 - 0.039\texttt{hmo} - 0.016\texttt{age} - 0.168\texttt{type1}$$

and

$$\mathbf{x}'\boldsymbol{\beta} = 2.22 - 0.035\texttt{hmo} - .0013\ \texttt{age} - 0.164\texttt{type1}$$

The inflation components contain a number of terms in the following equations equal to the number of covariates included in the respective condition in the simulation study. The formulas are again given for the Poisson and NB simulations, respectively.

$$\mathbf{z}'\boldsymbol{\gamma} = -2.94 + 1.12\texttt{died} + 0.190\texttt{white} - 0.096\texttt{age80}$$

and

$$\mathbf{z}'\boldsymbol{\gamma} = -15.5 + 10.0\texttt{died} + 0.315\texttt{white} - 0.101\texttt{age80}$$

In this version of the simulation study, we examine sample sizes of 300, 700, and 2,000. In terms of the number of parameters, we vary the inflation component in two ways. First, we run the simulation without zero inflation in the DGP and study the performance of the three test variants when one, two, and three covariates are incorrectly included in the inflation component. We then run a second variant in which there is zero inflation and examine the performance of the tests when one, two, and three covariates are correctly included in the inflation component. Each of the 36 conditions (3 sample sizes $\times$ 3 covariate specifications $\times$ 2 inflation/no inflation $\times$ 2 distributions) is run for 1,000 iterations.[8]

Figures 6–7 present the results of the simulations with `medpar.dta`. As in the simulations from section 3, when there is no zero inflation in the DGP, the BIC-corrected statistic performs the best, and the uncorrected statistic performs the worst. The BIC-corrected statistic is statistically significantly negative ($p < 0.05$, one tailed)—in favor of the single-equation model—in 95–100% of the iterations. In contrast, the uncorrected Vuong statistic is positive in approximately 80% or more of the iterations. Unlike the simulations in section 3, under this design, the uncorrected test is rarely statistically significant in favor of the zero-inflated model. However, just as in the previous simulations, not once did the uncorrected test result in a statistically significant rejection of the zero-inflated model. The AIC-corrected test performs moderately better in the no-inflation condition. In approximately 45–65% of the iterations, the zero-inflated model is statistically significantly rejected, and the single-equation model is virtually never rejected. However, as we found before, the degree to which the AIC favors the single-equation count model decreases with the number of covariates incorrectly included in the inflation component.

8. We performed all the computations presented in this section in Stata/SE 11.1 and Stata/IC 12.1.

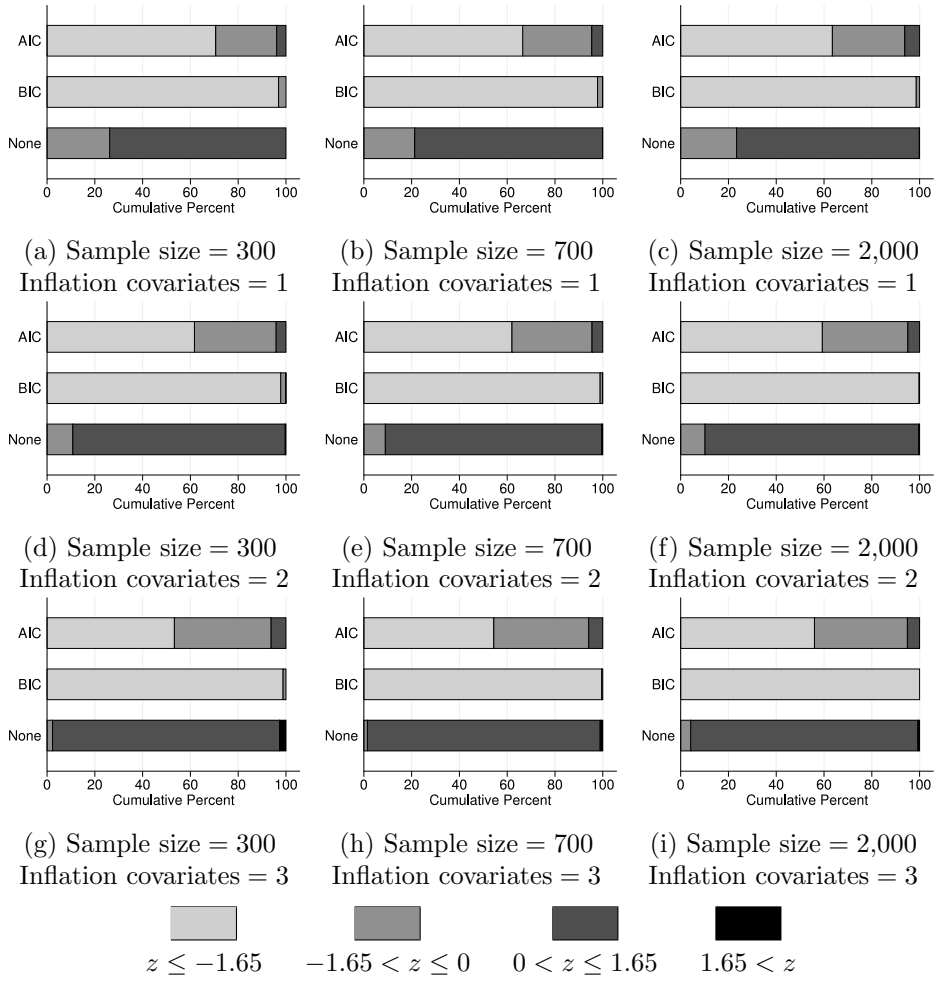


Figure 6. Monte Carlo results with Poisson simulations (`medpar.dta`). The plots depict the distribution of significance test results based on the Vuong test comparing Poisson to ZIP models with the respective correction across varying sample sizes and numbers of covariates incorrectly included in the inflation component.

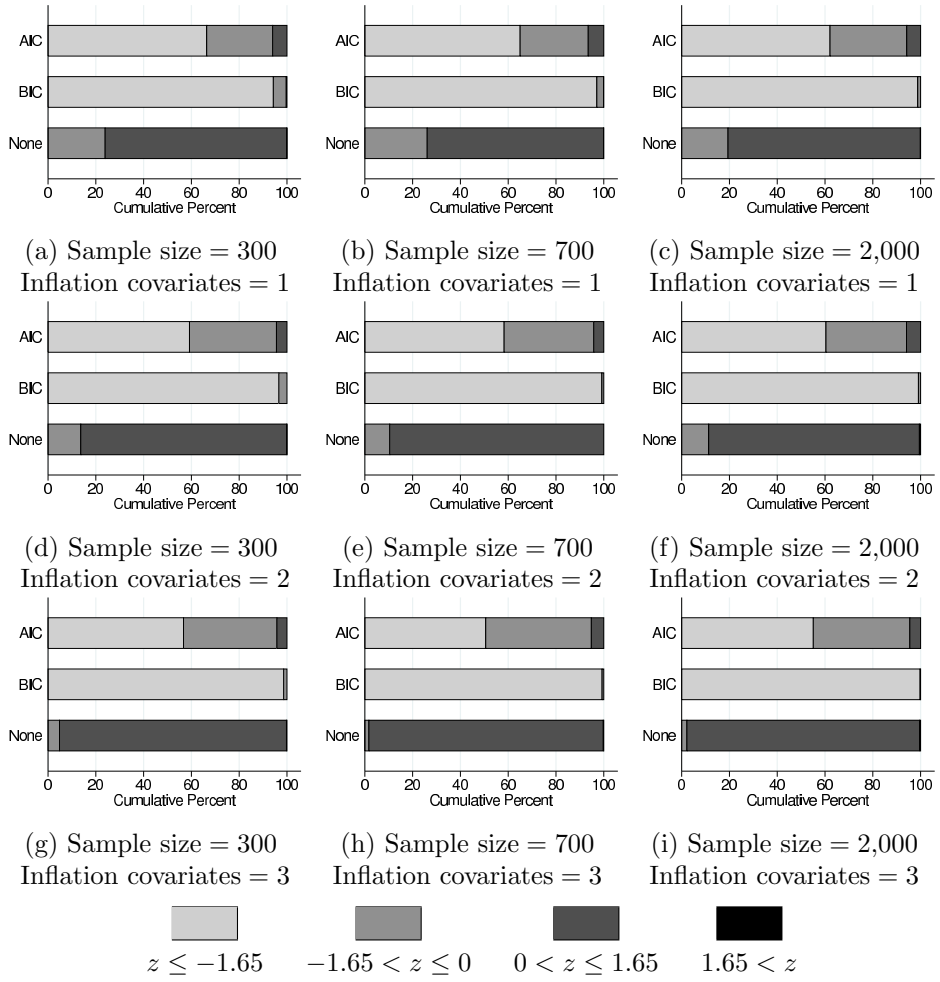
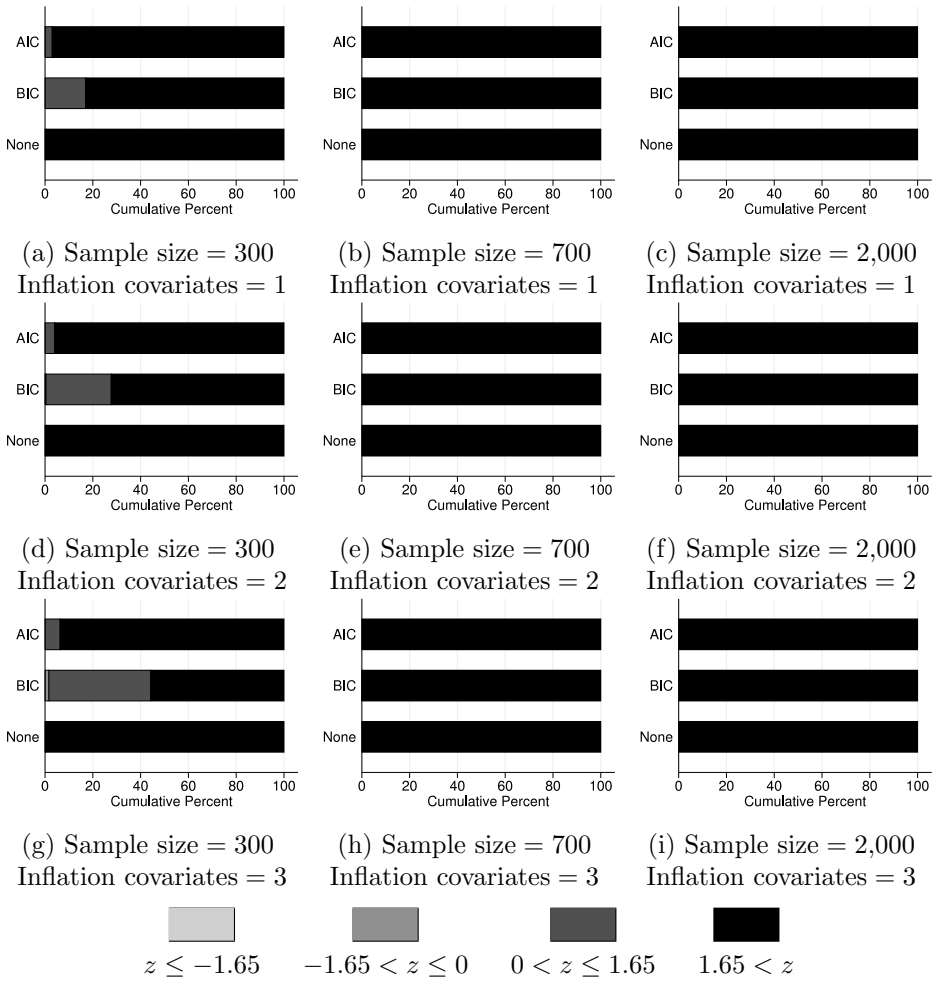


Figure 7. Monte Carlo results with NB simulations (`medpar.dta`). The plots depict the distribution of significance test results based on the Vuong test comparing NB to ZINB models with the respective correction across varying sample sizes and numbers of covariates incorrectly included in the inflation component.

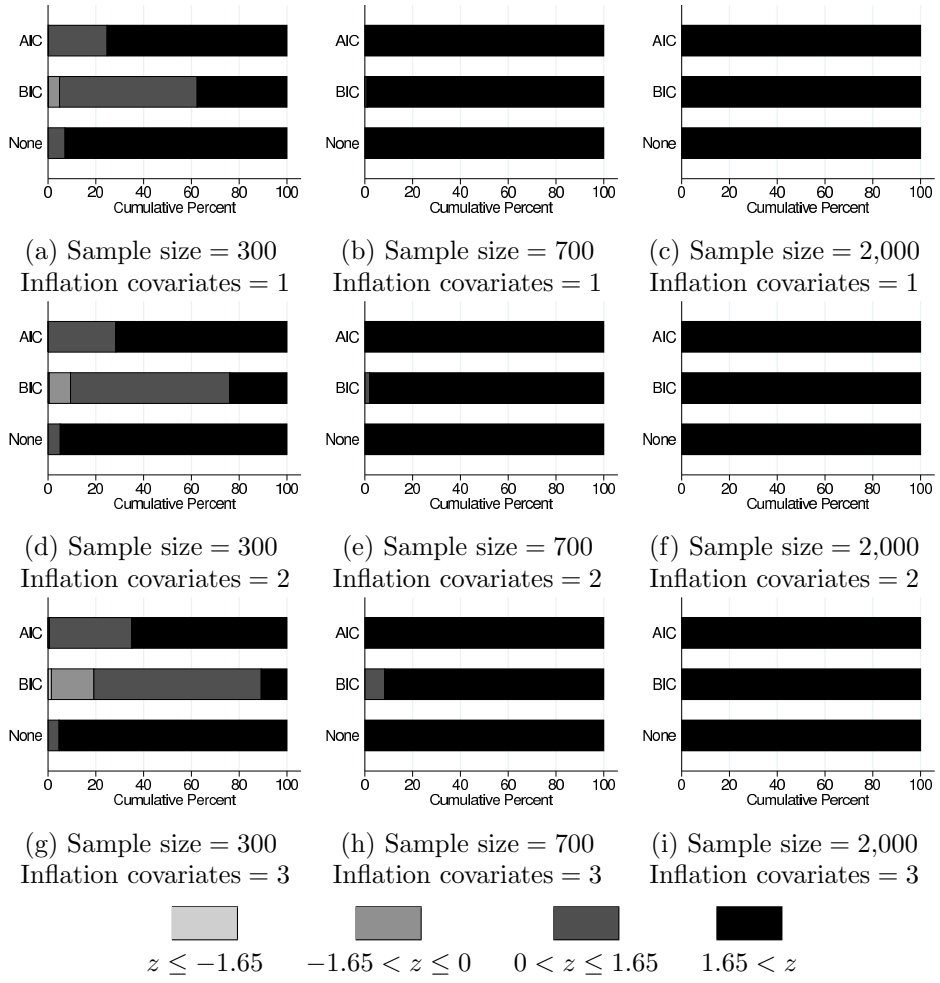
Figures 8 and 9 present results in which zero inflation is a component of the generative process. In this condition, the uncorrected Vuong statistic performs the best in selecting the correctly specified model—nearly always statistically significantly rejecting the single-equation model. The AIC-corrected test performs fairly well in the small-sample ($N = 300$) conditions, significantly favoring the zero-inflated model in 60–80% of the iterations, but virtually always statistically significantly selects the zero-inflation model in the larger sample-size conditions. The performance of the BIC-corrected statistic, which performs the worst among the three statistics when zero inflation is the correct

model, varies substantially across the sample-size and covariate conditions. As we found above, the tendency for the BIC-corrected statistic to statistically significantly reject the single-equation model is inversely related to the number of covariates correctly included in the zero-inflation component and directly related to the sample size.



(a) Sample size = 300, Inflation covariates = 1
(b) Sample size = 700, Inflation covariates = 1
(c) Sample size = 2,000, Inflation covariates = 1
(d) Sample size = 300, Inflation covariates = 2
(e) Sample size = 700, Inflation covariates = 2
(f) Sample size = 2,000, Inflation covariates = 2
(g) Sample size = 300, Inflation covariates = 3
(h) Sample size = 700, Inflation covariates = 3
(i) Sample size = 2,000, Inflation covariates = 3

Figure 8. Monte Carlo results with ZIP simulations (`medpar.dta`). The plots depict the distribution of significance test results based on the Vuong test comparing Poisson to ZIP models with the respective correction across varying sample sizes and numbers of covariates correctly included in the inflation component.



(a) Sample size = 300 Inflation covariates = 1
(b) Sample size = 700 Inflation covariates = 1
(c) Sample size = 2,000 Inflation covariates = 1
(d) Sample size = 300 Inflation covariates = 2
(e) Sample size = 700 Inflation covariates = 2
(f) Sample size = 2,000 Inflation covariates = 2
(g) Sample size = 300 Inflation covariates = 3
(h) Sample size = 700 Inflation covariates = 3
(i) Sample size = 2,000 Inflation covariates = 3

Figure 9. Monte Carlo results with ZINB simulations (`medpar.dta`). The plots depict the distribution of significance test results based on the Vuong test comparing NB to ZINB models with the respective correction across varying sample sizes and numbers of covariates correctly included in the inflation component.