Advance Access publication June 12, 2012

Political Analysis (2012) 20:387–399
doi:10.1093/pan/mps018

Compound Poisson–Gamma Regression Models for Dollar Outcomes That Are Sometimes Zero

Benjamin E. Lauderdale

London School of Economics, Methodology Institute, Columbia House,
Houghton Street, London, WC2A 2AE, UK
e-mail: b.e.lauderdale@lse.ac.uk

Edited by Jonathan N. Katz

Political scientists often study dollar-denominated outcomes that are zero for some observations. These zeros can arise because the data-generating process is granular: The observed outcome results from aggregation of a small number of discrete projects or grants, each of varying dollar size. This article describes the use of a compound distribution in which each observed outcome is the sum of a Poisson–distributed number of gamma distributed quantities, a special case of the Tweedie distribution. Regression models based on this distribution estimate loglinear marginal effects without either the ad hoc treatment of zeros necessary to use a log-dependent variable regression or the change in quantity of interest necessary to use a tobit or selection model. The compound Poisson–gamma regression is compared with commonly applied approaches in an application to data on high-speed rail grants from the United States federal government to the states, and against simulated data from several data-generating processes.

1 Introduction

Political scientists frequently fit regression models to nonnegative data denominated in units of dollars or other currencies, but in which there are at least some zero outcomes. Such data are particularly prevalent in studies of the distribution of political contributions, domestic spending, and foreign aid. Researchers studying these topics have adopted a variety of data analysis strategies. Some studies employ linear regression analyses on raw or per-capita dollars (e.g., Maizels and Nissanke 1984). Other studies employ linear regression models for log dollars, which requires a "fix" to deal with the fact that log 0 is undefined. Several such fixes are employed, including dropping observations with zero outcomes, defining an arbitrary log-scale level for the zero outcomes (e.g., Kuziemko and Werker 2006; Dollar and Levin 2006; Younas 2008), or using a tobit model that treats zeros as censored below some cutoff (e.g., Alesina and Dollar 2000; Fleck and Kilby 2001; Alesina and Weder 2002; Berthelemy and Tichit 2004). Another group of studies uses selection models (Heckman 1976, 1979) that separately model the fact of and the magnitude of nonzero outcomes (e.g., McGillivray and Oczkowski 1992; Balla et al. 2002; Neumayer 2003; Berthelemy 2006; Fleck and Kilby 2006; de Mesquita and Smith 2007).

Some of these studies explicitly check whether their substantive conclusions are stable across several of these specifications. Such checks are prudent, because all of these approaches have potential problems. Regression on dollar amounts (whether total or per-capita) imposes a linear model for the aggregation of marginal effects that is often implausible, as well as ignoring the zero bound on the outcome. Dropping zero dollar observations in a log-dependent variable regression changes the population being studied to nonzero outcomes. Placing zeros at (or below via tobit) a specific point on the log dollar scale creates outliers at arbitrarily determined points on the outcome scale. Tobit and selection models change the population of interest in a potentially undesirable

Author's note: The author thanks Kosuke Imai, Dustin Tingley, Teppei Yamamoto, and two anonymous reviewers for comments on earlier versions of this article. Supplementary materials for this article are available on the *Political Analysis* Web site.

way: estimating marginal effects over a counterfactual population of nonzero outcomes that includes the observations that were actually observed as zeros. None of these "fixes" allow scholars to estimate log-scale additive (loglinear) regression models for the entire set of *actual* observations.

This article describes an alternative approach to analyzing this kind of data, based on the idea that observed outcomes result from the additive aggregation of a nonnegative integer number of dollar amounts, each of varying dollar size. Additive aggregation of discrete allocation decisions (e.g., projects, grants, donations) is often a substantively plausible model, and is one that can predict the exact zeros that are often observed. Such processes yield *compound* distributions that are continuous, positive, and have a point mass at zero. Zeros are assumed to arise because some units receive no projects, but for units that do receive at least one project, the observation is the cumulative dollar magnitude of all projects rather than the project count.

This article focuses on the use of compound Poisson–gamma processes to generate this granular structure. Building on pre-existing research on compound Poisson–gamma distributions (see the review in Smyth 1996), Jorgensen and de Souza (1994) proposed the use of a generalized linear model (GLM) with a Tweedie distributed outcome to analyze this kind of data. Tweedie distributions are three-parameter exponential family distributions that can take the form of more familiar exponential family distributions—including the normal, Poisson, and gamma—for particular values of an index parameter (Tweedie 1984). Applied research has mostly employed the Tweedie in the range of index parameter values corresponding to the compound Poisson–gamma. This approach has been applied by scholars studying actuarial data (e.g., Jorgensen and de Souza 1994), meteorological data (e.g., Dunn 2004), and fisheries data (e.g., Shono 2008). In all these cases, certain kinds of outcomes—total insurance payouts consisting of multiple claims, total precipitation accumulated from multiple storms, total catch consisting of multiple fish—are naturally viewed as the cumulative size of a discrete number of objects, each drawn from a size distribution.

The compound Poisson–gamma regression has a particularly useful feature for descriptive research: The coefficients have the same loglinear interpretation for the expected outcome as coefficients from a log-dependent variable regression. In Section 2, I describe how this property arises from compounding separate Poisson and gamma regressions for the number and size of the discrete number of projects. Under a particular restriction on the relative scaling of the number and size of projects as a function of covariates, the compound Poisson–gamma can be fit using standard GLM fitting procedures (Jorgensen and de Souza 1994). For many applications, this approach to estimation is adequate, and is very quickly executed in exploratory analysis. However, the restriction that yields this version of the model is not required to maintain the convenient interpretation of coefficients as loglinear for $E[y_i]$. Models that allow for separate scaling of project number and size can be estimated either by the saddlepoint approximation procedure for the double Tweedie GLM described by Smyth and Jorgensen (2002) or by directly maximizing the Tweedie likelihood using recent advances in computation of the Tweedie density function (Dunn and Smyth 2008). This article uses the latter approach, which facilitates computation of robust or model-based standard errors and profile likelihood confidence intervals. A supplementary appendix describes existing R packages, as well as an R package from the author, that include code for fitting these models.

In Section 3, I compare the compound Poisson–gamma regression model to other commonly applied models in analyzing the distribution of High-Speed Rail projects in the United States in 2010 and 2011.[1] The commonly applied strategies of placing zeros at (or below via tobit) an arbitrary threshold such as log $\$1 = 0$ can perform particularly poorly. By comparison, dropping zeros, using a tobit with a cutoff just below the nonzero data or a selection model all perform reasonably, though they estimate a different quantity of interest and are less efficient than the compound Poisson–gamma in this example. In Section 4, I present results of a limited set of Monte Carlo simulations to show that the compound Poisson–gamma regression estimator can recover correct marginal effects even when the true error distribution deviates substantially from the assumed compound Poisson–gamma. In Section 5, I conclude by discussing the appropriate

[1] Benjamin E. Lauderdale, "Replication Data for Compound Poisson–Gamma Regression Models for Dollar Outcomes That Are Sometimes Zero," http://hdl.handle.net/1902.1/17924, IQSS Dataverse Network.

scope of substantive application for these methods, particularly focusing on the comparison to tobit and selection models. I argue that researchers have often adopted the latter models for bad reasons, ignoring the fact that they involve estimating effects within counterfactual populations that are difficult to identify and not always of substantive interest. Thus, even when the distributional assumptions of the compound Poisson–gamma model are violated, the fact that the model estimates loglinear marginal associations of covariates with the *actual* outcome may make it the most appropriate model for some applications.

2 Specification and Estimation

To derive the compound Poisson–gamma regression model, we begin by assuming that observed dollar outcomes arise from additive aggregation of a nonnegative integer number of projects, each of which has a positive magnitude in dollars. We further assume that the number of projects n_i for unit i is Poisson distributed, that each project's dollar size z_{ij} within unit i is gamma distributed, and that the number and size of the projects are independent.[2] These assumptions yield a compound Poisson–gamma distribution for the observable outcome y_i that is specified by the following equations:

$$p(n_i|\lambda_p) = \frac{\lambda_p^{n_i} e^{-\lambda_p}}{n_i!}, \lambda_p > 0 \tag{1}$$

$$p(z_{ij}|\lambda_g, \nu) = \frac{1}{z_{ij}\Gamma(\nu)} \left(\frac{\nu z_{ij}}{\lambda_g}\right)^{\nu} e^{-\frac{\nu z_{ij}}{\lambda_g}}, \lambda_g > 0, \nu > 0 \tag{2}$$

$$y_i = \begin{cases} \sum_{j=1}^{n_i} z_{ij} & \text{if } n_i > 0 \\ 0 & \text{if } n_i = 0. \end{cases} \tag{3}$$

Because of the independence of the Poisson and gamma processes described by Equations 1 and 2, the expected level of spending is simply $E[y_i] = E[n_i]\ E[z_{ij}] = \lambda_p \lambda_g$. This compound distribution is a special case of the Tweedie distribution. The standard parameterization of the Tweedie distribution is in terms of the expected value μ, plus a "dispersion" parameter ϕ and an "index" or "power" parameter ζ (see Tweedie 1984; Jorgensen and de Souza 1994). This parameterization can be derived from the Poisson and gamma parameterizations given above by the following change of variables: $\mu = \lambda_p \lambda_g$, $\phi = \frac{\lambda_p{}^{1-\zeta} \lambda_g{}^{2-\zeta}}{2-\zeta}$, and $\zeta = \frac{\nu+2}{\nu+1}$, where the parameter domains are $\mu \geq 0$, $\phi \geq 0$, and $1 < \zeta \leq 2$.[3] The variance of the compound Poisson–gamma distribution is $Var[y_i] = \phi \mu^{\zeta}$.[4]

When estimated separately, link functions for Poisson and gamma regressions can be specified such that both yield loglinear marginal associations of the expected outcome with covariates.[5] If we parameterize each of these component processes in a particular way, we can aggregate them into a compound Poisson–gamma regression with the same coefficient interpretation. Where x_i is a vector of covariates for unit i and β and γ are vectors of coefficients,

$$\lambda_p = e^{\frac{x_i(\beta-\gamma)}{2}} \tag{4}$$

$$\lambda_g = e^{\frac{x_i(\beta+\gamma)}{2}} \tag{5}$$

[2] Section 4 explores the performance of the MLE estimator based on these distributions in several situations where the true data-generating process deviates from the assumptions.

[3] Other values of the Tweedie index parameter ζ yield probability densities, but they are not compound Poisson-gamma distributions, and are consequently not of interest here.

[4] Note that what is substantively important is the granularity of the compound process, not the particular choice of the Poisson and gamma distributions. The compound Poisson–gamma turns out to have attractive properties for estimation, but with modern computing other possibilities can be explored. In particular, the one-parameter Poisson distribution is somewhat restrictive. The author has done some exploratory analysis with a negative binomial, but this makes computation far more difficult and is unlikely to make much difference in most applications. The gamma distribution is of less concern, since it is a two-parameter distribution.

[5] Loglinear marginal associations result from the canonical link for the Poisson, but not the gamma.

$$\mu = \lambda_p \lambda_g = e^{x_i \beta} \tag{6}$$

$$\phi = \frac{e^{\frac{1-\zeta}{2}x_i(\beta-\gamma)} \cdot e^{\frac{2-\zeta}{2}x_i(\beta+\gamma)}}{2-\zeta}. \tag{7}$$

Thus, the β coefficients describe the log-scale marginal associations of the covariates x_i with the expected dollar outcomes, while the γ coefficients describe how covariates vary in the extent to which they predict the size versus the number of the discrete grants or projects. The implicit marginal associations with size are $(\beta+\gamma)/2$, while those with number are $(\beta-\gamma)/2$.

Since $E[y_i]$ does not depend on γ, it is possible to estimate the model with the restriction that $\gamma = 0$ for all but the constant term in x_i.[6] This yields the GLM model of Jorgensen and de Souza (1994), which assumes common scaling for both size and number as a function of covariates. I will refer to this restriction hereafter as the "single equation model" and the unrestricted model as the "double equation model," following the distinction between single and double GLMs (Smyth and Jorgensen 2002). Both models assume $E[y_i] = e^{x_i\beta}$, but the single equation model is more restrictive with respect to the error distribution around that level. Regardless of whether the single equation or double equation model is estimated, the β coefficients are the key quantities of interest because they are the marginal associations with the mean dollar outcome. Because only the β coefficients predict $E[y_i]$, they can be more precisely estimated than the γ coefficients, which only describe variation in higher moments.

The full likelihood function for the compound Poisson–gamma regression is most easily written in terms of the Tweedie distribution with a parameter restriction. Where the Tweedie distribution function is given by $\mathcal{TW}(\mu, \phi, \zeta)$, the likelihood for the observed dollar outcomes y_i is

$$\mathcal{L}(\beta, \gamma, \zeta | y, x) = \prod_{i=1}^{N} \mathcal{TW}\left(e^{x_i\beta}, \frac{e^{\frac{1-\zeta}{2}x_i(\beta-\gamma)} \cdot e^{\frac{2-\zeta}{2}x_i(\beta+\gamma)}}{2-\zeta}, \zeta\right) \tag{8}$$

$$1 < \zeta \leq 2. \tag{9}$$

A range of standard estimation approaches can be used in estimating this model, especially in its single equation form. Jorgensen and de Souza (1994) estimate the single equation model as a GLM using iteratively reweighted least squares, Smyth and Jorgensen (2002) estimate the double equation model as a double GLM (Smyth 1989), and Swan (2006) estimates the single equation model using generalized estimating equations to allow for clustered errors. More recently, Dunn and Smyth (2008) introduced numerical methods for quickly calculating the Tweedie density using Fourier inversion of the characteristic functions, facilitating estimation by methods that require direct calculation of the likelihood function. The R package "eliot," maintained by the author of this article, includes a function CPGregression that estimates the single and double equation models described above by MLE, using general purpose optimizers, returning model-based and robust Wald standard errors as well as profile likelihood confidence intervals. For further details on fitting these models, see the supplementary appendix to this article.

3 Application: US High-Speed Rail Project Distribution

Among the many projects funded by the 2009 American Recovery and Reinvestment Act (ARRA), \$8 billion was allocated to high-speed rail projects. In response to this allocation, the governors of 34 states applied for projects totaling \$55 billion. On 28 January 2010, just under \$8 billion was allocated to 28 of these proposed construction projects, a general upgrade program on the Northeast Corridor (\$112 million) and a set of small planning projects across a range of states.[7] In the subsequent 16 months, the list of active projects changed in composition for several reasons.[8]

[6] Fixing the γ coefficient on the constant term to zero would imply that the average number and average dollar size of projects in dollars are numerically equal, which would only be a sensible assumption in an extremely peculiar situation.

[7] http://www.whitehouse.gov/files/documents/100128_1400-HSRAwards-Summary_FRA%20Revisions.pdf

[8] The data on 2011 projects were downloaded during June 2011 from http://www.fra.dot.gov/rpd/HSIPR/462.shtml.

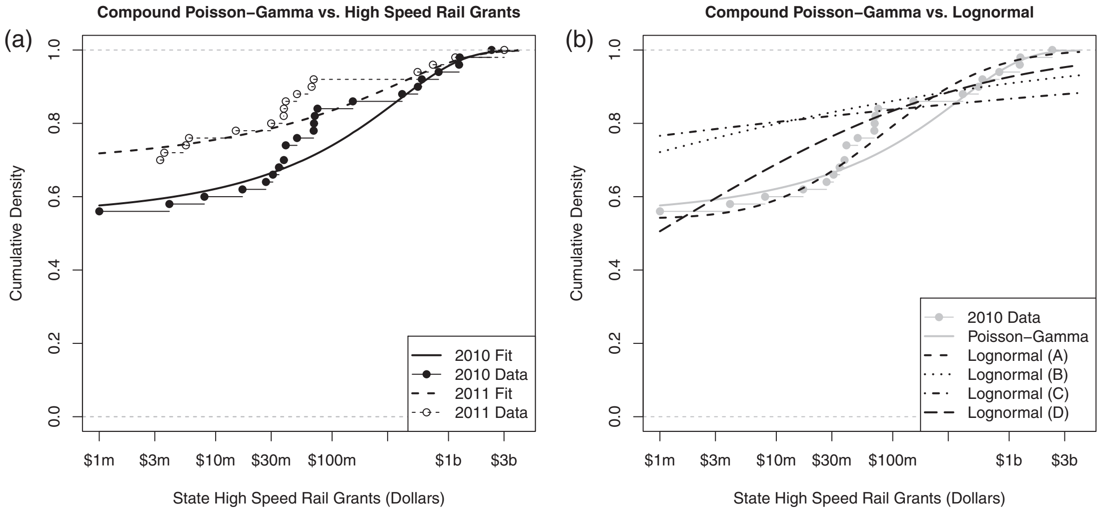

Fig. 1 Comparison of state-level high-speed rail grant data to compound Poisson–gamma fit using 2010 and 2011 data (**a**) and comparison of 2011 fit and data to four estimation strategies using a lognormal distribution (**b**). The four strategies are (A) dropping zeros, (B) replacing log 0 with log \$1 = 0, (C) treating log 0 as censored at log \$1 = 0, and (D) treating log 0 as censored at \$1 below the minimum nonzero outcome.

A further \$2.5 billion was allocated for high-speed rail by Congress in the FY 2010 budget. New Republican governors in Wisconsin, Ohio, and Florida rejected projects in their states that had been requested by previous officeholders.[9] The money from these rejected projects was subsequently reallocated to high-speed rail projects in other states.

Since planning projects had not been specified as of January 2010, I exclude them from subsequent analysis.[10] Thus, the January 2010 data consist of thirty-one construction projects across twenty-three states and the June 2011 data consist of thirty-six construction projects across sixteen states. Researchers typically analyze data aggregated over relevant political geographies, and the compound Poisson–gamma regression is designed for that kind of application. The fact that we can observe the aggregation of state-level rail grant receipts from individual project data indicates that the motivating story for the model fits the data-generating process in this case; however, the compound Poisson–gamma regression analysis uses the data after aggregation to the state level, which is the kind of data more typically available to researchers.

Before including covariates, it makes sense to check how well the compound Poisson–gamma fits the raw data, and to compare that fit to the lognormal distributions that are typically assumed. As Fig. 1a shows, the fit of the compound Poisson–gamma to the data is quite good (see Table 1 for the parameter estimates from this null model). In both the years 2010 and 2011, there were a few too many states receiving grant totals in the range from \$30 to \$100 million, and too few receiving \$100 million to \$300 million, relative to the compound Poisson–gamma distribution. However, these are not statistically significant differences,[11] and are minor deviations compared with the errors implicitly made by some of the most commonly used approaches for dealing with this kind of data.

[9] Republican governor Chris Christie of New Jersey rejected a tunnel project under the Hudson River in late 2010 that had support from the federal government; however, that project is not in the data set because the funding for that project predated the high-speed rail initiative begun with the ARRA.

[10] In the analysis, the general upgrade program on the Northeast Corridor is allocated among states according to the projects that were later specified from this line item. These were upgrades to the Portal Bridge in New Jersey (\$38.5 million), the B&P Tunnel in Maryland (\$60 million), and the Baltimore-Washington International Airport Station in Maryland (\$9.4 million).

[11] The confidence bands on the empirical cumulative density encompass the fitted compound Poisson–gamma, but were omitted from the plot for the sake of visual clarity.

Table 1 Compound Poisson–gamma regression model coefficients and confidence intervals for US high-speed rail projects in January 2010 and June 2011

Parameter	*Year*	*Null model*	*Single equation*	*Double equation*
		MLE (95% CI)	*MLE (95% CI)*	*MLE (95% CI)*
β_0	2010	19.89 (19.19 to 20.66)	16.45 (15.94 to 16.94)	17.61 (17.18 to 18.03)
	2011	20.51 (19.69 to 21.46)	16.82 (16.19 to 17.43)	17.30 (16.70 to 17.88)
$\beta_{\log(pop)}$	2010		1.67 (1.26 to 2.06)	1.43 (1.14 to 1.71)
	2011		1.49 (1.02 to 1.95)	1.32 (0.91 to 1.72)
β_{pres08}	2010		−0.15 (−0.20 to −0.10)	−0.05 (−0.09 to −0.01)
	2011		−0.13 (−0.18 to −0.07)	−0.10 (−0.15 to −0.04)
β_{gov11}	2010		0.45 (−0.32 to 1.18)	0.03 (−0.75 to 0.77)
	2011		−1.87 (−3.25 to −0.60)	−2.21 (−3.33 to −1.13)
γ_0	2010	18.88 (18.24 to 19.49)	18.27 (17.52 to 19.09)	19.67 (18.90 to 20.54)
	2011	18.57 (17.73 to 19.35)	18.62 (17.80 to 19.56)	20.39 (19.56 to 21.33)
$\gamma_{\log(pop)}$	2010			−0.90 (−1.63 to −0.10)
	2011			−0.47 (−1.24 to 0.36)
γ_{pres08}	2010			0.17 (0.09 to 0.24)
	2011			0.12 (0.03 to 0.19)
γ_{gov11}	2010			−0.11 (−1.19 to 1.14)
	2011			−1.82 (−3.11 to −0.28)
ζ	2010	1.72 (1.60 to 1.82)	1.70 (1.56 to 1.82)	1.74 (1.61 to 1.84)
	2011	1.76 (1.64 to 1.86)	1.65 (1.49 to 1.80)	1.64 (1.48 to 1.79)
$log(\mathcal{L}/\mathcal{L}_0)$	2010	0	21.04	23.65
	2011	0	17.08	18.84

Note. 95% intervals are model-based profile confidence intervals.

Figure 1b shows corresponding lognormal fits to the 2010 data (A) after dropping zeros, (B) after replacing log 0 with log \$1 = 0, (C) treating log 0 as censored at log \$1 = 0, and (D) treating log 0 as censored at \$1 below the minimum nonzero outcome. Strategies (B) and (C) perform extremely poorly because there are no observed grants between \$1 and \$1m. Assuming a lognormal distribution when over half the data are at (or censored below) \$1 implies that the density of points in that range should be substantially higher than in the equivalently large log-scale range between \$1m and \$1t, which in fact contains all of the nonzero data. With such a large misspecification of the underlying data-generating process, replacing log 0 with log \$1 = 0 or using the standard tobit with a truncation point at 0 leads to misleading inferences about marginal covariate associations because it is the deviations from these null models that the covariates in a regression model seek to predict. Strategy (D), which corresponds to a tobit in which the truncation point is set just below the smallest nonzero outcome, is a far better fit to these data, and is likely to be superior in other applications as well because it does not place the latent quantities that generated the zeros at values far less than the observed nonzero data points. However, it is strategy (A), dropping the zeros, that yields the best fit between the nonzero data and the lognormal distribution.[12] Young and Young (1975) provide a theoretical argument for why dropping zeros is better than placing them at an arbitrary value, a result that is almost universally ignored in applied political science research. Two equation selection models do not exactly correspond to any of these strategies for fitting a lognormal to data with zeros, but they are closest to the strategies of dropping the zeros or to using a tobit with a censoring point just below the lowest observed nonzero outcome.

[12] Note that the baseline for the plotted lognormal cdf is offset from zero to correspond to the fact that it is fit to only the nonzero observations.

When we shift to analyses with covariates, these differences in the degrees to which the underlying null models fit the data become very important. Three state-level covariates are considered: "log(*pop*)," the mean-centered natural log of the state's total population from the 2010 United States Census (range: −1.9 to 2.3); "*pres*08," the deviations from 50% in the Republican two-party presidential vote share from the 2008 general election (range: −23.0 to 16.6); and "*gov*11," the party of the state governor in 2011 (coded Republican = 1, Democratic = 0).[13] These three variables are not intended to be exhaustive (Achen 2002), and their coefficients should certainly not be interpreted causally on the basis of the regression results. However, as basic measures of demand for transportation spending, general political alignment, and the specific political inclination to accept grants for high-speed rail projects, these variables are associated with a great deal of the variation in the observed award totals by state. Moreover, we know already that there is a causal effect of having a Republican governor in 2011 on spending in 2011 (but not in 2010), so an appropriate regression method for these data should find evidence of that effect.

Table 1 shows the null, single, and double equation compound Poisson–gamma regression estimates for the project data from January 2010 and June 2011.[14] Since the state population variable is log transformed, the coefficient for that variable is interpreted as a power relationship between state population and spending levels. A coefficient of 1 would correspond to total spending growing proportionally to population: constant per-capita spending levels. Unlike the cases of general transportation spending formulas and earmarks (Lee 2000; Lauderdale 2008), there is no small state advantage in these high-speed rail grants. In fact, the coefficient estimates are significantly >1 in all but one specification. Holding the other variables fixed, more populous states received more money per capita.

Increasing 2008 Republican vote share is associated with decreased spending, with each additional percent of the two-party vote associated with a 5%–15% decline in spending, depending on specification and year. These are large substantive associations: A coefficient of −0.10 implies that in 2010 a typical state in which Obama received 45% of the two-party vote was awarded just 37% as much high-speed rail construction spending as was awarded to a state in which Obama received 55% of the two-party vote, all else being equal. This association is compatible with multiple causal hypotheses. Although political targeting of high-speed rail projects by the Obama administration is possible, it would be difficult to imagine that the places with the most need/demand for high-speed rail were not also the places where Obama performed best in 2008. As usual, there is no way to distinguish between these (or other) causal mechanisms with a regression analysis, but the marginal association itself can be clearly established.

The variable for having a Republican governor in 2011 does not significantly predict 2010 spending. This is a check for confounding, since the January 2010 project announcements occurred before any of the governors elected in November 2009 or 2010 could have had any impact on high-speed rail spending. In contrast, having a Republican governor in 2011 is strongly predictive of lower high-speed rail spending in 2011, with such states receiving just 11% of the spending received by states with Democratic governors in the single equation specification, other covariates equal. As noted above, we already know the causal mechanism that gives rise to this relationship: Three Republican governors rejected large quantities of high-speed rail funding at the beginning of 2011.

The single and double equation models yield very similar estimates of the marginal associations. The double equation model fits the data only slightly better than the single equation model, despite having three additional parameters. A comparison of the AIC for these two models indicates a preference for the simpler model because the extra parameters in the double equation model do little to improve model fit.

[13] Lincoln Chafee, the independent governor of Rhode Island in 2011, is coded as a Democrat. Whereas Chafee was a Republican when he served in the United States Senate, he was moderate in his voting, became an Independent after leaving office, and supported Barack Obama in the 2008 election.

[14] Model-based, rather than robust, profile confidence intervals are reported because they are more conservative in this case.

For purposes of comparison, Table 2 shows the β coefficients from commonly applied data analysis strategies that also yield loglinear marginal associations/effects: dropping zeros, replacing zeros with log $1 = 0, using a tobit model with a cutoff at zero, using a tobit model with a cutoff just below the smallest observed nonzero outcome, and using a selection model estimated by maximum likelihood.[15] Unsurprisingly, given the terrible baseline fit of the null models (B) and (C), the log dollar regression models based on replacing zeros with log $1 = 0 and using a tobit model with that value as the cutoff both perform extremely poorly. The coefficient values on the population and presidential vote variables in these models are implausibly large in magnitude, and the estimates for all variables are extremely uncertain.

The strategies of dropping zeros, using a tobit with a cutoff just below the smallest observed nonzero outcome, and using a selection model all recover plausible estimates; however, each is clearly inferior to the compound Poisson–gamma for making descriptive inferences about the marginal associations of the covariates with the state-level grant totals. All three of these models yield much wider confidence intervals than the compound Poisson–gamma. None of the three come close to statistical significance on the governor variable, in part because these strategies estimate the association with governor party only among states that received grants, rather than among all states. Moreover, if the quantity of interest is a marginal association within the entire population, including the zeros, these models do not provide the desired summary of the data.

4 Robustness to Error Distribution Misspecification

Evaluating the robustness of maximum-likelihood estimators to violations of model assumptions requires careful specification of which assumptions might be of particular concern, as well as the appropriate behavior of the estimator given the violation. In the case of the compound Poisson–gamma regression model, this article has argued that the chief benefit of using the model is that it recovers loglinear regression coefficients when zeros are present in the data. I have not argued that the compound Poisson–gamma probability distribution is going to be precisely the correct one, though in the case considered above it is closer to correct than available alternatives. Thus, if the marginal associations estimated by the regression were very dependent on the assumption of the compound Poisson–gamma error structure, the usefulness of the estimator would be greatly diminished. Therefore, the appropriate robustness check involves cases where the data-generating process deviates from the assumed compound Poisson–gamma process, but in which the expected level of the outcome remains a loglinear function of the regressors. Given this constraint on which kinds of simulation studies are worth doing, I have identified three plausible kinds of violations that change the error distribution around the expected outcome level, but do not change the scaling of the expected outcome level as a function of the regressors.

First, we can test the robustness of the single-equation model when the double-equation model is appropriate. In the following simulations, I consider cases where the covariate dependence of the model is entirely in the size rather than the number of projects. Second, we can test the robustness of the single and double equation compound Poisson–gamma models when the true data-generating process is a compound Bernoulli-lognormal process (i.e., a binary selection model with a lognormal outcome distribution for the nonzero outcomes). In the case where none of the included regressors predict selection, the regression coefficients describing the marginal associations with $E[y_i]$ describe loglinear relationships. Third, we can test the robustness of the compound Poisson–gamma models when there is dependence between the Poisson and gamma components of the generating process. Positive associations could arise if omitted variables increase both the number and size of project allocations for certain units. Negative associations could arise if a budget constraint mandates that units receiving more projects have smaller project size.

[15] The selection model is estimated by maximum likelihood without an exclusion restriction, which is generally unadvisable and especially so in small samples because it relies on the functional form of the distributions for identification. However, there is no plausible exclusion restriction given the variables under consideration. The estimated error correlation for the two equations is extremely high, so in effect the estimates are the same as those of the estimator proposed by Sartori (2003).

Table 2 Marginal associations/effects and confidence intervals for five alternative approaches to that of Table 1

Parameter	*Year*	*(A) Drop zeros*	*(B) Add one*	*(C) Tobit (0)*	*(D) Tobit (Min)*	*Selection*
		MLE (95% CI)	*MLE (95% CI)*	*MLE (95% CI)*	*MLE (95% CI)*	*MLE (95% CI)*
β_0	2010	16.34 (14.24 to 18.44)	7.06 (4.00 to 10.12)	−2.50 (−9.96 to 4.95)	12.65 (10.57 to 14.74)	15.16 (12.21 to 18.11)
	2011	17.16 (14.35 to 19.98)	6.18 (2.81 to 9.56)	−9.90 (−22.78 to 2.99)	12.97 (10.44 to 15.5)	13.96 (13.40 to 14.53)
$\beta_{\log(pop)}$	2010	1.07 (0.17 to 1.97)	4.74 (3.01 to 6.47)	9.45 (5.70 to 13.21)	2.80 (1.77 to 3.82)	1.66 (0.37 to 2.94)
	2011	0.68 (−0.43 to 1.79)	2.96 (1.05 to 4.87)	9.11 (3.21 to 15.01)	1.95 (0.79 to 3.12)	1.50 (0.30 to 2.70)
β_{pres08}	2010	−0.14 (−0.31 to 0.02)	−0.49 (−0.70 to −0.27)	−1.15 (−1.68 to −0.62)	−0.32 (−0.47 to −0.17)	−0.19 (−0.36 to −0.01)
	2011	−0.07 (−0.30 to 0.15)	−0.36 (−0.59 to −0.12)	−1.40 (−2.32 to −0.48)	−0.26 (−0.44 to −0.08)	−0.22 (−0.34 to −0.09)
β_{gov11}	2010	0.38 (−1.35 to 2.11)	1.15 (−2.94 to 5.23)	3.07 (−4.61 to 10.75)	0.79 (−1.29 to 2.87)	0.50 (−1.16 to 2.16)
	2011	−1.07 (−3.35 to 1.21)	−1.68 (−6.19 to 2.83)	−2.17 (−14.19 to 9.86)	−1.03 (−3.39 to 1.34)	−1.13 (−3.19 to 0.93)

Maximum-likelihood estimation on the basis of such a "correlated compound Poisson–gamma" process is not feasible, but it is not difficult to generate correlated compound Poisson–gamma variates.

The following procedure uses a bivariate gaussian copula function to generate a draw from either a compound Bernoulli-lognormal or a compound Poisson–gamma in which there is correlation between the constituent distributions that make up the compound distribution.[16] First, generate one draw from a standard normal distribution. Second, convert that standard normal variate to its quantile (i.e., a uniform random variate), and convert the quantile to either the desired Bernoulli random variate via the Bernoulli cumulative distribution function or a Poisson random variate via the Poisson cumulative distribution function. The resulting draw is n_i. If $n_i = 0$, then $y_i = 0$. Third, if $n_i > 0$, generate a series of n_i draws from a univariate normal with mean and variance chosen such that each of these draws and the initial standard normal draw form a bivariate normal draw with mean vector equal to zero, variance equal to one, and covariances equal to ρ. Fourth, convert each of these n_i normal random variates into either a lognormal or a gamma random variate by converting to the normal quantile and then applying the appropriate cumulative distribution function. Fifth, for the Poisson–gamma case, compute the sum of the gamma random variates to give the correlated compound Poisson–gamma draw y_i. Although an infinite variety of other correlation structures are possible, this structure allows us to tune a single parameter $\rho \in [-1, 1]$ in order to assess the consequences of positive and negative correlations of the constituent distributions.

One way to generate simulated data sets would be to base them on the data and estimates from the high-speed rail example. In a typical application, this would be a good way to do a sensitivity analysis. For the purposes of this article, however, the large marginal associations in the high-speed rail data set make such an analysis a poor demonstration of model robustness. Consequently, I simulate data sets with far more subtle relationships. I draw X from a multivariate normal with mean vector [0, 0, 0] and a covariance matrix in which $\sigma_1 = \sigma_2 = \sigma_3 = 1$, $\rho_{12} = 0.25$, $\rho_{13} = -0.50$, and $\rho_{23} = 0$. I assume that the correct regression coefficients are $\beta = [0.25, 0, 0]$.[17] This "true" association is moderate in scale, given the range of X_1, so as to create a nontrivial test for the estimator. For the compound Poisson–gamma generating process, I set $\lambda = 1$, which yields $1/e = 36\%$ zero observations; I set the Bernoulli probability for the selection model generating process to yield the same fraction of zeros. For the compound Poisson–gamma generating process, I set the baseline size of the gamma draws when $X_1 = 0$ to \$1 million. For the compound Bernoulli-lognormal generating process, I set the baseline size of the nonzero observations when $X_1 = 0$ to \$1 million divided by $1 - 1/e$, which yields an identical scaling of the expected value. For the compound Poisson–gamma generating process, $\nu = 1$ in the gamma distribution; for the compound Bernoulli-lognormal generating process, $\sigma = 1$ in the lognormal distribution. All of these values could be varied; however, attempting to simulate model performance under all of the possibly relevant combinations of parameter values quickly exits the realm of computational feasibility.

[16] Where $\sim$ denotes a random draw from the given distribution, F_1^{-1} is either a Bernoulli or Poisson inverse cumulative distribution function, F_2^{-1} is either a lognormal or gamma inverse cumulative distribution function, and Φ is the normal cumulative distribution function, the following procedure generates the desired random variates:

$$n_i^{\mathcal{N}} \sim \mathcal{N}(0, 1) \tag{10}$$

$$n_i = F_1^{-1}\left(\Phi(n_i^{\mathcal{N}})\right) \tag{11}$$

$$z_{ij}^{\mathcal{N}} \sim \mathcal{N}(\rho n_i^{\mathcal{N}}, 1 - \rho^2) \tag{12}$$

$$z_{ij} = F_2^{-1}\left(\Phi(z_{ij}^{\mathcal{N}})\right) \tag{13}$$

$$y_i = \begin{cases} \sum_{j=1}^{n_i} z_{ij} & \text{if } n_i > 0 \\ 0 & \text{if } n_i = 0 \end{cases}. \tag{14}$$

[17] Because the X matrix has non-trivial correlation between the first covariate and the remaining two, this creates an opportunity for the estimates to misattribute the association with X_1 to X_2 or X_3.

Table 3 Mean coefficients for single and double equation compound Poisson–gamma regression models estimated on 1000 simulated data sets of size 100, under several deviations from the assumed data-generating process (DGP)

DGP correlation:		$\rho = -0.8$		$\rho = 0$		$\rho = 0.8$	
Estimator:		*1 Eq.*	*2 Eq.*	*1 Eq.*	*2 Eq.*	*1 Eq.*	*2 Eq.*
Poisson-gamma DGP	$\bar{\beta}_1$	0.277	0.243	0.247	0.252	0.242	0.250
	$\bar{\beta}_2$	0.003	0.003	−0.001	0.000	−0.002	0.001
	$\bar{\beta}_3$	−0.002	−0.004	−0.009	−0.001	0.001	0.002
Bernoulli-lognormal DGP	$\bar{\beta}_1$	0.280	0.251	0.254	0.249	0.275	0.246
	$\bar{\beta}_2$	0.001	−0.000	0.009	0.010	−0.009	−0.008
	$\bar{\beta}_3$	−0.003	−0.004	−0.001	0.000	−0.005	−0.005

Note. The true marginal associations for the three regressors are $\beta = [0.25, 0, 0]$.

In Table 3, I report mean coefficient estimates based on one thousand simulated data sets of size one hundred, for each of the twelve combinations of three correlation levels, two compound distributions, and two estimators (single and double equation). Both estimators are fit on the same simulated data sets within each data-generating process. For both the compound Poisson–gamma and compound Bernoulli-lognormal data-generating processes, I have tested copula correlation levels of $\rho = -0.8$, 0, and 0.8. Bias is minimal across the simulation trials for both the single and double equation estimators. If the true data-generating process has $E[y_i] = e^{x_i\beta}$, both estimators do a reasonable job of recovering β even in these cases where the error structure around that mean function is not the assumed independent compound Poisson–gamma distribution. There is some evidence that the double equation estimator is slightly more robust against deviations from the assumed error distribution, but both estimators are close enough to unbiased so as to make estimator performance a negligible concern for most social scientific applications.

The uncertain small sample properties of misspecified maximum-likelihood estimators are always a concern with models like the ones considered here. The presented simulations demonstrate that the compound Poisson–gamma estimator returns approximately unbiased regression estimates in the presence of several deviations from the assumed data-generating process, but only for a particular set of parameter values, and only when the mean level is correctly specified. Although more exhaustive simulation of performance under a wider variety of misspecifications and parameter values could be completed, these simulations do indicate that performance of the compound Poisson–gamma regression estimators is not especially dependent on the exact data-generating process. It is advisable that researchers using these methods do sensitivity analyses, conditional on the relevant theoretical expectations and parameter values for their own studies.

5 Conclusion

The “zero problem” in log-dependent variable regression is only a problem if researchers are willing to assume models that are substantively implausible. The compound Poisson–gamma model solves this problem by changing the model to fit the data, rather than the other way around. The primary value of the compound Poisson–gamma regression is that it allows scholars to generate interpretable descriptions of associations between covariates and the actual outcomes in their data. Where widely used strategies either fail to recover plausible estimates, only estimate associations among units receiving nonzero outcomes, or estimate associations in counterfactual populations, the compound Poisson–gamma yields simple characterizations of marginal relationships across the entire sample or population. Even in the case where there are very few zeros in the data, there is little downside to using the compound Poisson–gamma regression model instead of a regression on log spending. As $\zeta \to 2$, the Tweedie distribution approaches a gamma distribution and the point mass at $y_i = 0$ disappears. A gamma regression will generally behave similarly to a log-dependent variable regression.

The choice between the compound Poisson–gamma and tobit or selection models should be made in part on a case-specific evaluation. In certain applications, like the one considered here, the underlying process that generates the data is in fact the aggregation of a small integer number of projects of varying dollar size, and the compound Poisson–gamma is clearly appropriate. In other applications, it may be more substantively defensible to model the outcome using other models, though neither tobit nor selection models have the elegant loglinear coefficient interpretation of the compound Poisson–gamma. Consequently, to the extent that descriptive inferences about the marginal associations of the outcome with covariates across the whole sample/population are desired, the compound Poisson–gamma is still a useful tool even if the underlying aggregation process is not correct.

But researchers should not only look to the likely data-generating process when choosing a regression model. The quantities of interest estimated by the compound Poisson–gamma are fundamentally different than those estimated by a tobit or selection model. A canonical application of selection models in econometrics is the relationship between labor market entry decisions and expected wage levels. When actual wage income is zero because an individual is choosing not to enter the labor market, modeling her counterfactual wage is important for modeling her entry decision (and vice versa). However, in some of the applications that political scientists have applied selection models to, the corresponding counterfactual is of dubious interest. Consider the case of foreign aid distribution. Although one can certainly hypothesize about the counterfactual foreign aid that would have been received by a country, had it received foreign aid, such quantities are only interesting if one has a model of foreign aid distribution that has a binary decision process that might depend on the counterfactual aid level.[18] Some theoretical models of foreign aid distribution do in fact assume a binary decision about whether to offer aid that is dependent on the aid level that would be offered (e.g., Dudley and Montmarquette 1976), but in most applications the selection model is being used as a fix to deal with zeros, not because scholars have a theory that involves counterfactual outcomes. As a result, authors end up worrying about identification restrictions that are only relevant if they care about identifying the marginal associations with covariates *under the counterfactual where all countries receive aid.* If a researcher's goal is to describe marginal associations between *actual* foreign aid and the variables of interest, the selection model's counterfactual structure is counterproductive.

Although the discussion in this article has focused on the case of distributive spending by governments to geographic entities (e.g., foreign aid, building projects, etc.), these are not the only political science data where exact zeros appear in otherwise continuous, positive outcomes. For example, political donation data have many zeros. In such cases, there is sometimes not only a lower bound on donation ($0), but also an upper bound due to legal limits, which makes tobit models with censoring on both sides seem particularly attractive. But there is still the same problem with the interpretation of tobit models in this context, which is that they do not estimate marginal associations for actual donations. Instead, they model marginal associations for the donations that would be made in a world where arbitrarily large donations were possible.[19] Sometimes a model of the "propensity to donate" is sensible, but it is not clear that measuring marginal associations with covariates within the population of counterfactual donations is *always* preferable to violating the error distribution assumptions of the compound Poisson–gamma. In making trade-offs between estimating substantively meaningful quantities of interest and having accurate assumptions about error distributions, political scientists have perhaps been too inclined toward the latter extreme.

[18] The tobit model is a special case of this kind of model, one where donors do not bother to make donations that would have been smaller than some cutoff.

[19] If no log transformation is employed, the interpretation problem with using a tobit with a cutoff at zero is actually more severe, since the associations are then estimated for the counterfactual where negative donations are allowed. Rational political donors would prefer to donate arbitrarily large negative dollar amounts to their political opponents, so this counterfactual is hardly a basis for meaningful estimates.

References

Achen, C. H. 2002. Toward a new political methodology: Microfoundations and ART. *Annual Review of Political Science* 5:423–50.

Alesina, A., and D. Dollar. 2000. Who gives foreign aid to whom and why? *Journal of Economic Growth* 5(1):33–63.

Alesina, A., and B. Weder. 2002. Do corrupt governments receive less foreign aid? *American Economic Review* 92(4):1126–37.

Balla, S. J., E. D. Lawrence, F. Maltzman, and L. Sigelman. 2002. Partisanship, blame avoidance, and the distribution of legislative pork. *American Journal of Political Science* 46(3):515–25.

Berthelemy, J.-C. 2006. Bilateral donors' interest vs. recipients' development motives in aid allocation: Do all donors behave the same? *Review of Development Economics* 10(2):179–94.

Berthelemy, J.-C., and A. Tichit. 2004. Bilateral donors' aid allocation decisions—a three-dimensional panel analysis. *Internation Review of Economics and Finance* 13:253–74.

de Mesquita, B. B., and A. Smith. 2007. Foreign aid and policy concessions. *Journal of Conflict Resolution* 51:251–84.

Dollar, D., and V. Levin. 2006. The increasing selectivity of foreign aid, 1984–2003. *World Development* 34(12):2034–46.

Dudley, L., and C. Montmarquette. 1976. A model of the supply of bilateral foreign aid. *American Economic Review* 66(1):132–42.

Dunn, P. K. 2004. Occurrence and quantity of precipitation can be modeled simultaneously. *International Journal of Climatology* 24:1231–9.

Dunn, P. K., and G. K. Smyth. 2008. Evaluation of Tweedie exponential dispersion model densities by Fourier inversion. *Statistics and Computing* 18:73–86.

Fleck, R. K., and C. Kilby. 2001. Foreign aid and domestic politics: Voting in Congress and the allocation of USAID contracts across congressional districts. *Southern Economic Journal* 67(3):598–617.

Fleck, R. K., and C. Kilby. 2006. How do political changes influence US bilateral aid allocations? Evidence from panel data. *Review of Development Economics* 10(2):210–23.

Heckman, J. J. 1976. The common structure of statistical models of truncation, sample selection, and limited dependent variables and a simple estimator for such models. *Annals of Economic and Social Measurement* 5:475–92.

Heckman, J. J. 1979. Sample selection bias as a specification error. *Econometrica* 47(1):153–61.

Jorgensen, B., and M. C. P. de Souza. 1994. Fitting Tweedie's compound Poisson model to insurance claims data. *Scandanavian Actuarial Journal* 1:69–93.

Kuziemko, I., and E. Werker. 2006. How much is a seat on the Security Council worth? Foreign aid and bribery at the United Nations. *Journal of Political Economy* 114(5):905–30.

Lauderdale, B. E. 2008. Pass the pork: measuring legislator shares in Congress. *Political Analysis* 16:235–49.

Lee, F. E. 2000. Senate representation and coalition building in distributional politics. *American Political Science Review* 94(1):59–72.

Maizels, A., and M. K. Nissanke. 1984. Motivations for aid to developing countries. *World Development* 12(9):879–900.

McGillivray, M., and E. Oczkowski. 1992. A two-part sample selection model of British bilateral foreign aid allocation. *Allied Economics* 24(12):1311–9.

Neumayer, E. 2003. Do human rights matter in bilateral aid allocation? A quantitative analysis of 21 donor countries. *Social Science Quarterly* 84(3):650–66.

Sartori, A. E. 2003. An estimator for some binary-outcome selection models without exclusion restrictions. *Political Analysis* 11(2):111–38.

Shono, H. 2008. Application of the Tweedie distribution to zero-catch data in CPUE analysis. *Fisheries Research* 93:154–62.

Smyth, G. K. 1989. Generalized linear models with varying dispersion. *Journal of the Royal Statistical Society (Series B)* 51:47–60.

Smyth, G. K. 1996. Regression analysis of quantity data with exact zeros. In *Proceedings of the Second Australia-Japan Workshop on Stochastic Models in Engineering, Techology, and Management*, 572–80. Techology Management Center, University of Queensland.

Smyth, G. K., and B. Jorgensen. 2002. Fitting Tweedie's compound Poisson model to insurance claims data: Dispersion modeling. *Astin Bulletin* 32(1):143–57.

Swan, T. 2006. Generalized estimating equations when the response variable has a Tweedie distribution: An application for multi-site rainfall modeling. PhD thesis, University of Southern Queensland.

Tweedie, M. C. K. 1984. An index which distinguishes between some important exponential families. In *Statistics: Applications and new directions*, eds. J. K. Ghosh and J. Roy, 579–604. Calcutta: Indian Statistical Institute.

Younas, J. 2008. Motivation for bilateral aid allocation: Altruism or trade benefits. *European Journal of Political Economy* 24:661–74.

Young, K. H., and L. Y. Young. 1975. Estimation of regressions involving logarithmic transformation of zero values in the dependent variable. *American Statistician* 29(3):118–20.