University of Massachusetts Amherst

Count Data Regression Analysis: Concepts, Overdispersion Detection, Zero-inflation Identification, and Applications with R

Item Type	article;article
Authors	Fávero, Luiz Paulo;Souza, Rafael de Freitas;Belfiore, Patrícia;Corrêa, Hamilton Luiz;Haddad, Michel F. C.
DOI	https://doi.org/10.7275/44nn-cj68
Download date	2024-11-21 18:33:38
Link to Item	https://hdl.handle.net/20.500.14394/39711

Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Copyright is retained by the first or sole author, who grants right of first publication to *Practical Assessment, Research & Evaluation*. Permission is granted to distribute this article for nonprofit, educational purposes if it is copied in its entirety and the journal is credited. PARE has the right to authorize third party reproduction of this article in print, electronic and database forms.

Volume 26 Number 13, June 2021 ISSN 1531-7714

Count Data Regression Analysis: Concepts, Overdispersion Detection, Zero-inflation Identification, and Applications with R

Luiz Paulo Fávero, *University of São Paulo*
Rafael de Freitas Souza, *University of São Paulo*
Patrícia Belfiore, *Federal University of ABC*
Hamilton Luiz Corrêa, *University of São Paulo*
Michel F. C. Haddad, *University of Cambridge*

In this paper is proposed a straightforward model selection approach that indicates the most suitable count regression model based on relevant data characteristics. The proposed selection approach includes four of the most popular count regression models (i.e. Poisson, negative binomial, and respective zero-inflated frameworks). Moreover, it addresses two of the most relevant problems commonly found in real-world count datasets, namely overdispersion and zero-inflation. The entire selection approach may be performed using the programming language R, being all commands used throughout the paper available for practical purposes. It is worth mentioning that counting regression models are still not widespread within the social sciences.

Introduction

Poisson regression models are commonly applied in count data analysis, where the behavior of the dependent variable $Y(Y \in \mathbb{N}_0)$ may be explained by predictor variables X_{ki}, which may be either metric or dummy variables – i.e. $E(Y|X_{ki})$. Such models are based on the strict premise that there is equality between the mean and variance of the dependent variable, conditional to the predictor variables. Despite the fact that Poisson regression models consist of a suitable framework for numerous analysis involving count data, its simplicity engenders problems to many real-word applications (Cameron and Trivedi, 2013, Zeviani et al., 2014, Fávero and Belfiore, 2019).

The problem of overdispersion is frequently found in real-world count datasets, consisting of the case in which the variance of the dependent variable - conditional to the predictor variables - is statistically greater than the corresponding mean (Payne et al., 2017, Dupuy, 2018). The detection of such a problem is relevant for an appropriate choice between the Poisson or negative binomial (NB hereinafter) regression model (Blackburn, 2015). Overdispersion is commonly caused either due to heterogeneity in sample values, presence of outliers, correlated variables, omission of relevant predictors or zero inflation (Payne et al., 2018).

The adoption of Poisson regression models in the presence of overdispersion produces similar undesirable consequences compared to the absence of homoscedasticity in a linear regression setting (Cameron and Trivedi, 2013). This leads to bias in the estimated coefficients, with consequent inefficiency and inconsistency of the modeling process (Breslow, 1990, Hilbe, 2011, Smith and Faddy, 2016). Aiming at tackling such frequent problems in count data, a test for detecting overdispersion (CT test hereinafter) within count data is proposed by Cameron and Trivedi (1990).

As part of the Generalized Linear Models, the count data regression models are used for cases where the

phenomenon under study presents itself in the form of a quantitative variable, however with only discrete and non-negative values, as we have already discussed. However, it is common that some variables with count data present an excessive amount of zeros, which can cause that the estimated parameters, when estimating the traditional Poisson and NB regression models, be biased since they cannot capture the exacerbated presence of null counts. In these situations, the zero-inflated (ZI) regression models can be used, in the presence of overdispersion or not.

The ZI regression models, according to Lambert (1992), are considered a combination between a model for count data and a model for binary data, since they are used to investigate the reasons that lead to a determined number of occurrences (count) for a phenomenon, as well as lead (or not) to the actual occurrence of this phenomenon, independently of the amount of observed count.

The actual definition regarding the existence or not of an excessive amount of zeros in the dependent variable is prepared by means of a specific test, known as the Vuong test (1989), which will represent the first output to be analyzed when estimating the zero-inflated regression models.

The contribution of the present paper is the proposition of a straightforward model selection approach to select the most suitable count data regression framework given distinct data characteristics. The proposed selection approach adapts the rationale in Perumean-Chaney et al. (2013) to be performed using the programming language R, covering two of the most relevant problems frequently found in real-world count datasets, namely overdispersion and zero-inflation. Moreover, it is provided the respective R commands to perform each stage of the proposed model selection approach.

Thus, this study intends to be a tutorial article, which target audience consists of researchers who are interested in regression models applied to count data but frequently are not sure about which model framework would be preferrable. To address such a practical need observed among analysts and researchers, particularly within social scientists, this paper then has two main objectives. Firstly, it is presented a clear exposition on count data models, emphasizing the Poisson, NB, and ZI frameworks. Secondly, it is provided a step-by-step introductory guide on how such count models may be properly understood, adapted, refined, and executed through the R language.

Following this introduction, this paper is divided into six sections. Section two presents the literature review. Section three introduces our proposed model selection approach. Section four details the data and respective exploratory data analysis. Section five demonstrates the proposed approach through four empirical cases, interpreting the respective outputs, and comparing the log-likelihood and fitted values of each model. Section six concludes. All R commands used throughout the paper are provided in the Appendix.

Literature Review

According to Wooldridge (2010), a general regression model applied to count data may be described through equation (1):

$$\ln(\hat{Y}_i) = \ln(\lambda_i) = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki}\,, \quad j = 1, 2, 3, \ldots \qquad (1)$$

where λ_i represents the expected number of occurrences or incidence rate ratio of the phenomenon under study for a given exposure (i.e. a fixed interval of time in which a particular number of events is registered), α is the intercept, the coefficients estimated for each predictor variable X_j are represented by β_j, and i represents each observation in the sample.

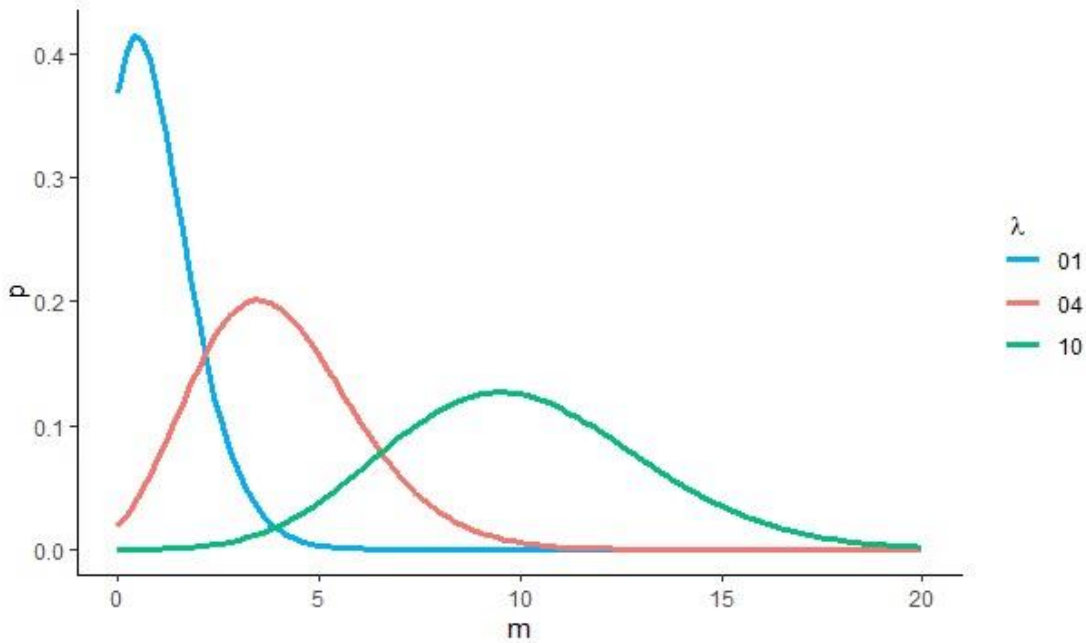
Poisson Regression Model

According to Cameron and Trivedi (2013), in general, Poisson regression models are suitable to cases in which the distribution of the occurrence of a particular phenomenon under study follows a Poisson distribution, as shown in Figure 1.

In Figure 1, the term p represents the likelihood of an observation i occurring for a particular exposure m, as detailed in equation (2):

$$p(Y_i = m) = \frac{\exp(-\lambda_i)\lambda_i^m}{m!}, \; m = (0, 1, 2, 3, \ldots.). \qquad (2)$$

Figure 1. Data visualisation of the Poisson distribution



It is postulated in Fávero et al. (2020) and (Klakattawi et al., 2018) that Poisson regression models assume the existence of equidispersion within the variable of interest - i.e. $\mu_i = E(Y_i) = Var(Y_i) = \lambda$, as detailed in equations (3) and (4):

$$E(Y_i) = \sum_{m=0}^{\infty} m \frac{\exp(-\lambda)\lambda^m}{m!} = \lambda \sum_{m=1}^{\infty} \frac{\exp(-\lambda)\lambda^{m-1}}{(m-1)!} = \lambda \quad (3)$$

$$Var(Y_i) = \sum_{m=0}^{\infty} \frac{\exp(-\lambda)\lambda^m}{m!}(m-\lambda)^2 = \sum_{m=0}^{\infty} \frac{\exp(-\lambda)\lambda^m}{m!}(m^2 - 2m\lambda + \lambda^2) = \lambda^2 \sum_{m=2}^{\infty} \frac{\exp(-\lambda)\lambda^{m-2}}{(m-2)!} + \lambda \sum_{m=1}^{\infty} \frac{\exp(-\lambda)\lambda^{m-1}}{(m-1)!} - \lambda^2 = \lambda \quad (4)$$

The coefficients of a Poisson regression model are estimated by the following likelihood function (Taddy, 2015):

$$L = \prod_{i=1}^{n} \frac{\exp(-\lambda_i)\lambda_i^{Y_i}}{Y_i!} \quad (5)$$

From which the logarithm of the likelihood function is derived as follows:

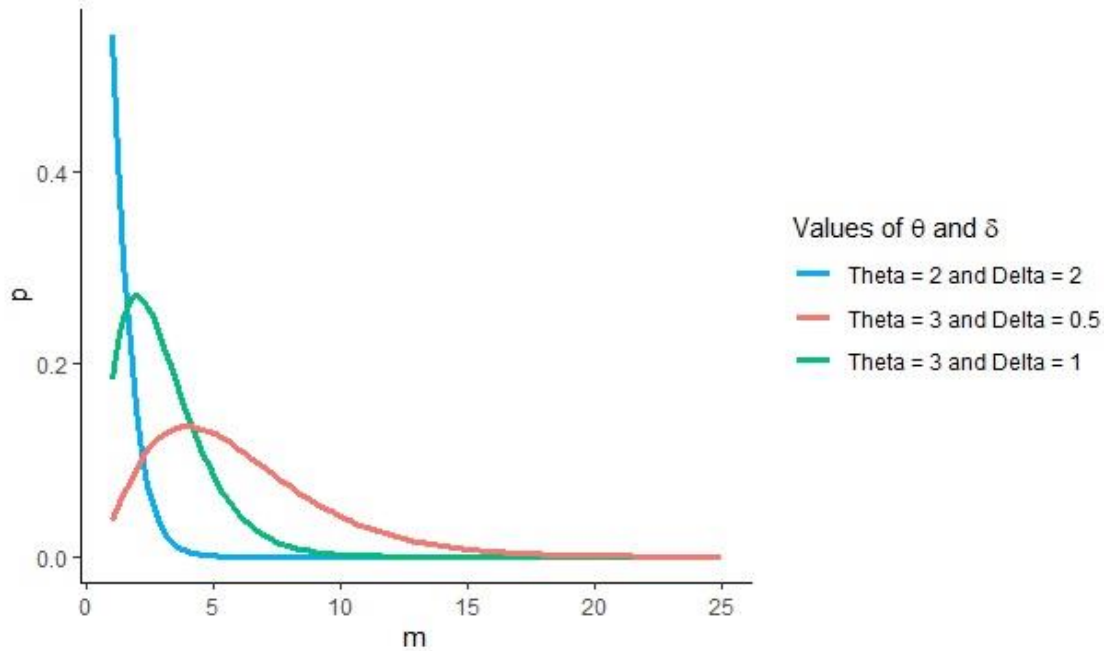
$$LL = \sum_{i=1}^{n} [-\lambda_i + Y_i \ln(\lambda_i) - \ln(Y_i!)] \quad (6)$$

Equation (6) is then iterated k times up to convergence to a maximum value (Fávero et al., 2018) – i.e. $LL = \sum_{i=1}^{n} [-\lambda_i + Y_i \ln(\lambda_i) - \ln(Y_i!)] = \max$.

Negative Binomial (NB) Regression Model

The estimation of an NB regression model is intrinsically related to the existence of overdispersion in the count dataset (Payne et al., 2018), which distribution is shown in Figure 2.

Figure 2. Data visualisation of the NB distribution



In Figure 2, p and m refer to the same terms mentioned in Figure 1, θ is the shape parameter being $\theta > 0$, and δ is the rate parameter being $\delta > 0$. The likelihood of a random portion out of the total number of occurrences of the dependent variable Y_i of i observations in exposure m may calculated through equation (7):

$$p(Y_i = m) = \binom{m+\theta-1}{\theta-1}\left(\frac{\theta}{u_i+\theta}\right)^{\theta}\left(\frac{u_i}{u_i+\theta}\right)^{m},\ m = 0,1,2,3,\ldots \quad (7)$$

where u_i represents the mean, with $u = E(Y) < Var(Y)$. Thus, the estimation of an NB regression model assumes the presence of overdispersion in the dependent variable conditional to the predictor variables (Hilbe, 2014) - i.e. $u_i = E(Y_i) < Var(Y_i)$. The mean and variance are shown in equations (8) and (9), respectively:

$$E(Y_i) = u \quad (8)$$

$$Var(Y_i) = u + \phi u^2 \quad (9)$$

$$\phi = \theta^{-1} \quad (10)$$

$$u_i = \exp(\alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki}) \quad (11)$$

As discussed in Cameron and Trivedi (1990), the parameter ϕ in equation (9) represents overdispersion within the count data. Moreover, it is postulated in Fávero et al. (2020) that cases where $\phi \to 0$, equidispersion would then be detected, indicating that a Poisson regression model would be suitable. According to the same authors, for the case in which ϕ is statistically greater than zero, then overdispersion would effectively occur, suggesting the application of an NB regression model instead.

Although the Poisson model is the most frequently used regression model for count data analysis, by definition, its distribution contains a single free parameter λ. This prevents the variance to be fitted to the mean (Payne et al., 2018), as previously explored for the case of $\lambda = Var(Y)$ in equations (3) and (4). Hence, in the presence of overdispersion, an NB regression model may provide a better fit for the count data, in which the mean of the Poisson distribution may be used as a random variable that follows a gamma distribution with an additional free parameter ϕ (Cameron and Trivedi, 1990, Fávero et al., 2020).

Thus, once overdispersion is properly tested and confirmed, then NB models are estimated via the likelihood criterion specified in equation (12), as discussed in Cameron and Trivedi (2010):

$$LL = \sum_{i=1}^{n}\left[Y_i \ln\left(\frac{\phi u_i}{1+\phi u_i}\right) - \frac{\ln(1+\phi u_i)}{\phi}\ln\Gamma(Y_i + \phi^{-1}) - \ln\Gamma(Y_i+1) - \ln\Gamma(\phi^{-1})\right] \quad (12)$$

which should be iterated up to reaching the maximum value (Hilbe, 2014) - i.e. $LL = \sum_{i=1}^{n}\left[Y_i \ln\left(\frac{\phi u_i}{1+\phi u_i}\right) - \frac{\ln(1+\phi u_i)}{\phi}\ln\Gamma(Y_i + \phi^{-1}) - \ln\Gamma(Y_i+1) - \ln\Gamma(\phi^{-1})\right] = \max.$

Zero-inflated (ZI) Regression Models

Although count models, such as Poisson and NB regression, may generate robust estimation for count data, these models may not be the most suitable frameworks in the case there is an excessive amount of zero-valued observations in the dependent variable (Perumean-Chaney et al., 2013).

According to Lambert (1992), ZI regression models are regarded as a mixture between a model for count data and a model for binary data. These models are applied to identify the reasons a particular quantity of counts occurs, regardless of the number of observed counts.

Two types of ZI models are typically considered. The first type consists of a zero-inflated Poisson (ZIP hereinafter) regression model, estimated from the combination of a Bernoulli with a Poisson distribution. The second type refers to the zero-inflated negative binomial (ZINB hereinafter) model, estimated from the combination of a Bernoulli with a Poisson-gamma distribution.

Zero-inflated Poisson (ZIP) Regression Model. The probability $\boldsymbol{p(Y_i = 0)}$, reflecting a zero count of a particular observation $\boldsymbol{i = 1, 2, \ldots, n}$ – being $\boldsymbol{n}$ the sample size, is calculated considering the sum of a dichotomic (i.e. binary) component. Analogously, the term $\boldsymbol{p_{logit_I}}$ is defined as the probability of any count not materializing exclusively due to such a dichotomic component.

Moreover, the probability $\boldsymbol{p(Y_i = m)}$ that a particular count $\boldsymbol{m = 1, 2, 3, \ldots}$ effectively occurs, is calculated by multiplying the Poisson distribution by $\boldsymbol{(1 - p_{logit_i})}$. Therefore, the general equation of the estimated probability for each observation i of an event occurring in a dichotomous manner is formulated in equation (13):

$$\boldsymbol{p_i = \frac{1}{1+exp[-(\gamma+v_1X_{1i}+v_2X_{2i}+\cdots+v_kX_{ki})]}} \quad (13)$$

Equation (13) is routinely used to calculated the probability of an event – commonly denoted by the value of one, and the probability of a non-event – commonly denoted by the value of zero. However, specifically for the ZI cases, equation (13) should consider the existence of zeros as an event and, conversely, values different from zero should be understood as a respective nonevent.

In a ZIP model estimation, when combining a binary logistic estimation with another counting data estimation , it is assumed that there are two processes generating zero values. One of such processes is due to the binary distribution (structural zeros) and the remaining one due to the Poisson distribution (sampling zeros).

The combination between structural zeros and sampling zeros is represented in equation (14). The former follows a Bernoulli distribution - in which zero-valued and non-zero observations are considered as an event and non-event, respectively – and the latter follows a Poisson distribution.

$$\begin{cases} \boldsymbol{p(Y_i = 0) = p_{logit_i} + (1 - p_{logit_i})exp(-\lambda_i)} \\ \boldsymbol{p(Y_i = m) = (1 - p_{logit_i})\frac{exp(-\lambda_i)\lambda_i^m}{m!}} \end{cases} \quad (14)$$

where $\boldsymbol{Y \sim ZIP(\lambda, p_{logit_I})}$, being $\boldsymbol{p_{logit_I}}$ calculated through equation (15):

$$\boldsymbol{p_{logit_i} = \frac{1}{1+exp[-(\gamma+v_1W_{1i}+vW_{2i}+\cdots+v_qW_{qi})]}} \quad (15)$$

where $\boldsymbol{W_q}$ consists of the q-th explanatory variable that originates structural zeros.

$$\boldsymbol{\lambda_i = exp(\alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki})} \quad (16)$$

It is worth noting that if $\boldsymbol{p_{logit_i} = 0}$, then the distribution of probabilities of equation (14) is clearly summarized in the Poisson distribution, including cases where $\boldsymbol{Y_i = 0}$. In other words, ZIP regression models present two zero generating processes, in which one refers to the binary distribution (structural zeros) and the other to the Poisson distribution (sampling zeros). From equation (14), it is possible then to formulate the objective function detailed in equation (17), which objective is to estimate the parameters $\boldsymbol{\alpha, \beta_1, \beta_2, \dots, \beta_k, \gamma, \nu_1, \nu_2, \dots, \nu_q}$ of a ZIP regression model.

$$\boldsymbol{LL = \sum_{Y_i=0} ln\left[p_{logit_i} + \left(1 - p_{logit_i}\right) exp(-\lambda_i)\right] + \sum_{Y_i>0}\left[ln\left(1 - p_{logit_i}\right) - \lambda_i + Y_i\, ln(\lambda_i) - ln(Y_i!)\right] = max} \quad (17)$$

Based on equations (15) and (16), one may define that while the occurrence of structural zeros is influenced by a vector of explanatory variables $\boldsymbol{W_1, W_2, \dots, W_q}$, the occurrence of a particular count $\boldsymbol{m}$ is influenced by a vector of $\boldsymbol{X_1, X_2, \dots, X_k}$. A potential application consists of inserting the same variable into two vectors to investigate if such a variable simultaneously influences the occurrence of an event and, if so, quantifying the occurrences (i.e. counts) of the phenomenon being explored.

Zero-inflated Negative Binomial (ZINB) Regression Model. The probability $p(Y_i = 0)$ of the occurrence of zero count p_{logit_i} of a particular observation $i = 1,2, \dots, n$ – being n the sample size, is also calculated based on the sum between a dichotomic with a count component. However, differently from the ZIP regression model, in the case of the ZINB regression framework, the probability $p(Y_i = m)$ of occurrence of a particular m count follows a Poisson-Gamma distribution.

Thus, the combination between structural zeros and sampling zeros is represented in equation (18). The former follows a Bernoulli distribution – in which zero-valued and non-zero observations are considered as an event and non-event, respectively – and the latter follows a Poisson-gamma distribution.

$$\begin{cases} p(Y_i = 0) = p_{logit_i} + \left(1 - p_{logit_i}\right)\left(\dfrac{1}{1 + \phi u_i}\right)^{\frac{1}{\phi}} \\ p(Y_i = m) = \left(1 - p_{logit_i}\right)\left[\dbinom{m + \phi^{-1} - 1}{\phi^{-1} - 1}\left(\dfrac{1}{1 + \phi u_i}\right)^{\frac{1}{\phi}}\right. \end{cases} \quad (18)$$

where $Y \sim ZINB\left(\phi, u, p_{logit_i}\right)$, ϕ represents the inverse of the shape parameter of a gamma distribution and, analogously to the ZIP regression model, p_{logit_i} is calculated following equations (19) and (20):

$$p_{logit_i} = \frac{1}{1 + \exp\left[-\left(\gamma + \nu_1 W_{1i} + \nu_2 W_{2i} + \cdots + \nu_q W_{qi}\right)\right]} \quad (19)$$

$$u_i = \exp(\alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki}) \quad (20)$$

Based on equation (18), the objective function detailed in equation (21) may be formulated, aiming at estimating the parameters $\phi, \alpha, \beta_1, \beta_2, \dots, \beta_k, \gamma, \nu_1, \nu_2, \dots, \nu_q$ of a ZINB regression model.

$$LL = \sum_{Y_i=0} \ln\left[p_{logit_i} + \left(1 - p_{logit_i}\right)\left(\frac{1}{1+\phi u_i}\right)^{\frac{1}{\phi}}\right] + \sum_{Y_i>0} \ln\left[\left(1 - p_{logit_i}\right) + Y_i \ln\left(\frac{\phi u_i}{1+\phi u_i}\right) - \frac{\ln(1+\phi u_i)}{\phi} + \ln\Gamma(Y_i + \phi^{-1}) - \ln\Gamma(Y_i + 1) - \ln\Gamma(\phi^{-1})\right] = \max \quad (21)$$

In the case that $p_{logit_i} = 0$ in equation (18), the distribution of probabilities is then restricted to the Poisson-gamma distribution, including cases where $Y_i = 0$. Then, the ZINB regression models also present two

zero generating processes, resulting from both the binary and Poisson-gamma distribution.

Data Diagnostic Tests

In the present study, two tests are explored in order to identify two relevant problems frequently present in real-world count data. Firstly, it is detailed a test for detecting overdispersion, following the proposition in Cameron and Trivedi (1990). Secondly, it is presented a test to confirm excessive zeros in the dependent variable, following the proposition in Vuong (1989).

Overdispersion Cameron and Triverdi (CT) Test. The CT test for overdispersion detection in count data proposed by Cameron and Trivedi (1990), where H_0 is the equidispersion given by $Var(Y|X) = E(Y|X)$, based on equation (22):

$$Var(Y|X) = E(Y|X) + \phi[E(Y|X)]^2 \quad (22)$$

It is worth noticing that equation (22) is similar to the variance function of the NB regression model in equation (10). For the test in equation (22), the significance of parameter ϕ must be verified, in which $H_0: \phi = 0$ and $H_1: \phi > 0$. For the detection of overdispersion in the count data, at a certain level of significance, it is postulated that a Poisson regression model should be estimated *a priori*.

Subsequently, an auxiliary ordinary least squares (OLS) regression model without intercept is also estimated. Its dependent variable Y^*, given by equation (23), is then computed using the fitted values of λ from the initially established Poisson regression model.

$$Y_i^* = \frac{[(Y_i - \lambda_i)^2 - Y_i]}{\lambda_i} \quad (23)$$

The auxiliary model in equation (23) sets λ as its single predictor variable, following equation (24):

$$\hat{Y}_i^* = \beta\lambda_i \quad (24)$$

Furthermore, subsequently to the estimation of the auxiliary model in equation (24), it is performed a Student's *t*-test to analyze the *p*-value of the predictor variable λ. In the cases in which $p > |t| > sig$, it is then assumed that, at a certain significance level, there is equidispersion in the data. Conversely, if $p > |t| \leq sig$ then overdispersion, at a certain significance level, is confirmed.

This overdispersion test may be performed in R programming language, using the overdisp() function of the overdisp package introduced in Freitas Souza et. al. (2020).

Excess Zeros Vuong Test. The Vuong test (Vuong, 1989) is used to check for the presence of excessive zeros within the dependent variable. This test compares two model frameworks estimated over the same data, for instance models M_1 and M_2. Its null hypothesis H_0 is that both models M_1 and M_2 fit adequately to the underlying data. Such models cannot be nested and their dependent variables must consist of count data (Desmarais and Harden, 2013).

This test adopts the Kullback-Leibler divergence (Kullback and Leibler, 1951) of the data-generating model M_t. The difference of the Kullback-Leibler divergence between model M_t and a particular model M is algebraically expressed as $D_{KL}(M_t||M)$. Hence, the H_0 of the test is formulated as $H_0: D_{KL}(M_t||M_1) = D_{KL}(M_t||M_2)$.

Therefore, considering that the Vuong test refers to a comparison between count data models for ZI detection, besides estimating an ZI regression model – either a ZIP or ZINB framework, it is also necessary to estimate an additional regression model for comparison purposes, consisting of either a Poisson or NB model estimation. More objectively, this test should be applied either to compare a Poisson against a ZIP model estimation or an NB against an ZINB model estimation.

This overdispersion test may be performed in R programming language, using the vuong() function of the pscl package introduced in Zeileis, Kleiber, and Jackman (2008).

Model Selection Approach

The model selection approach proposed in the present study is comprised of four widely used count regression model frameworks and two statistical tests to identify problems commonly found in real-world count datasets. The aim of this selection approach is to guide analysts in their decision on which count regression model would be the most appropriate according to

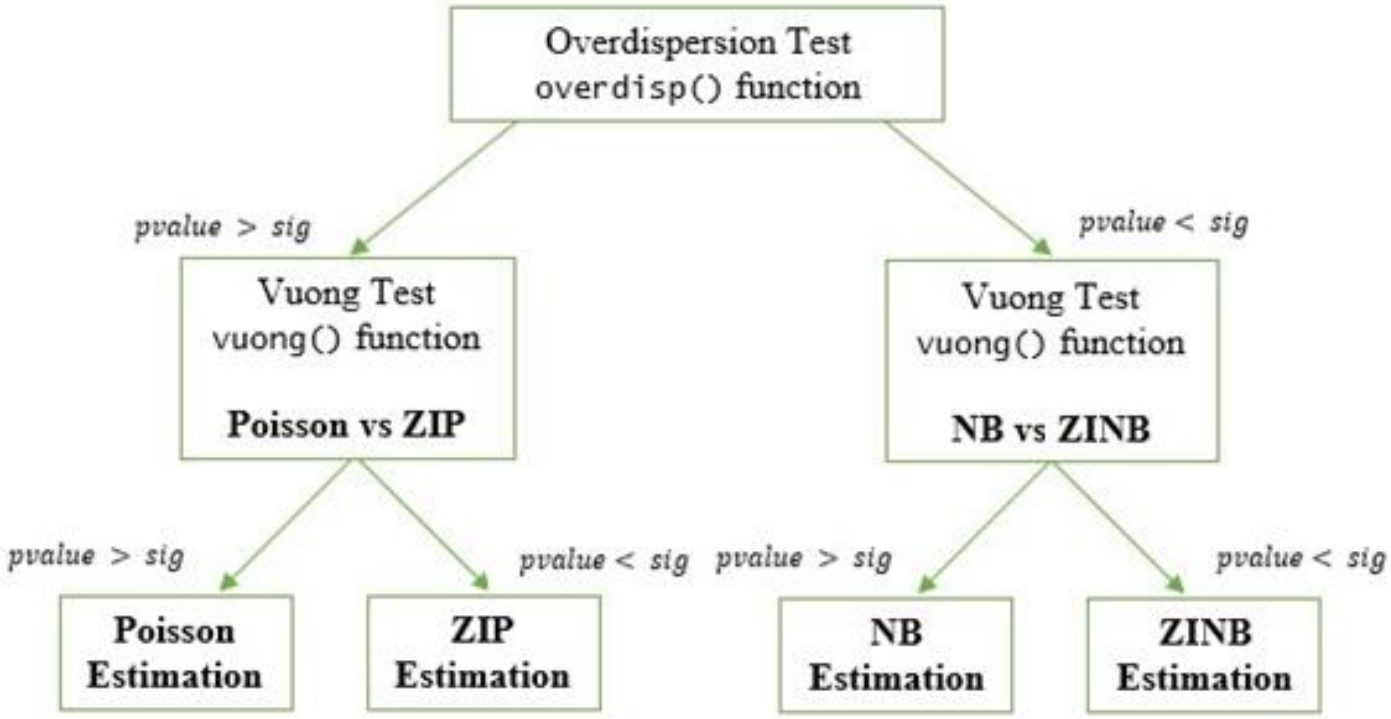
specific characteristics of the dataset. The relation between regression models for count data and the presence of overdispersion and/or excessive amount of zero-valued observations in the dependent variable are summarized in Table 1.

Thus, in cases in which there is an excessive amount of zero-valued observations in the dependent variable, the ZIP and ZINB regression models are more appropriate. Nonetheless, when there is overdispersion, then NB regression models would consist of more suitable alternatives, either its classical version or the ZI framework. As shown in Figure 3 as well, in the case that the dependent variable has both attributes of excessive zeros and overdispersion, then the ZINB regression model would be the best framework to be adopted. The proposed count regression model selection approach illustrated in Figure 3 is originally introduced in Perumean-Chaney et al. (2013) and then adapted in the present study in order to enable its execution using R programming language.

Table 1. Regression models for count data, overdispersion, and excess of zeros

Dependent variable attribute	Count data regression model			
	Poisson	NB	ZIP	ZINB
Excessive zeros	No	No	Yes	Yes
Overdispersion	No	Yes	No	Yes

Figure 3. Flowchart of the proposed count regression model selection approach, adapted from Perumean-Chaney et al. (2013).



Data

The data to illustrate the application of the proposed count regression model selection approach contain information about approvals of students in a specific national high school exam in Brazil reported in 160 public or private schools. This dataset is available as supplementary material to this paper. Moreover, all R codes used in the following empirical applications are provided in the Appendix. As reported in Table 2, the dataset contains four variables, 160 observations, and no missing values.

In Table 3 are summarized the descriptive statistics of variables *approvals*, *professors* and *hours*, as well as the frequency of occurrence of the categorical variable *public*. It is reported that almost 60% of the schools do not present any student approved in the considered national high school exam, suggesting the presence of an excessive amount of zero-valued observations in the dependent variable.

It is worth noticing that the variance of the variable *approvals* is 14 times larger than the respective mean. The histogram of the dependent variable *approvals* is shown in Figure 4.

Table 2. Name, type of data, and description of the variables in the dataset

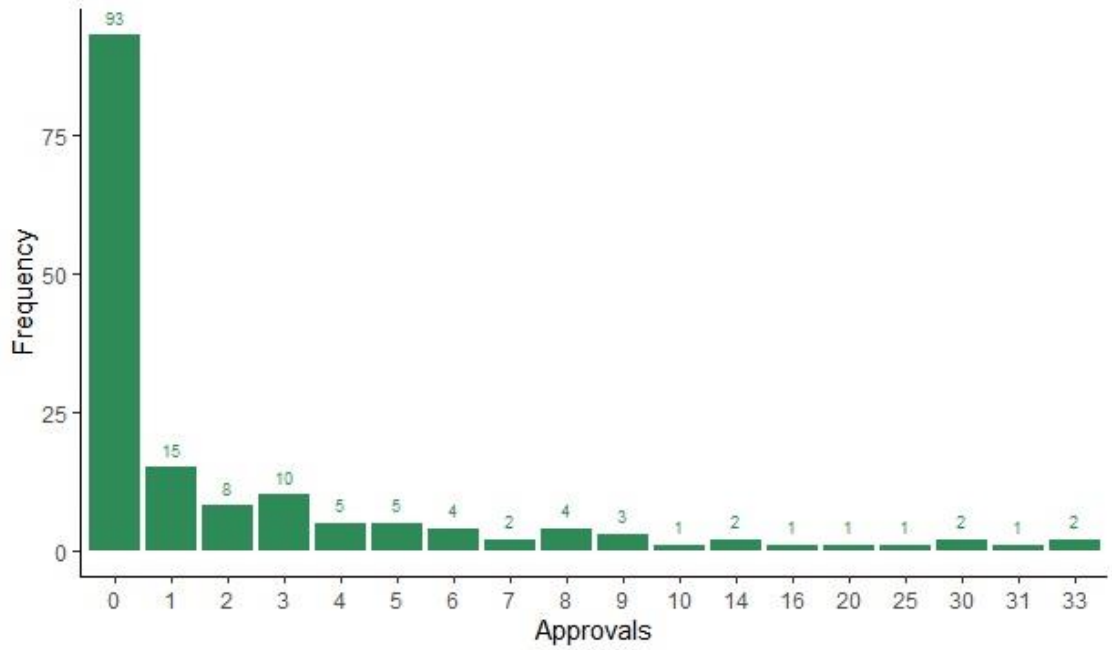
Variable name	Type of data	Variable description
Approvals	Quantitative discrete	Number of students approved per school in the national high school exam.
Professors	Quantitative discrete	Number of full-time teachers in each school.
Hours	Quantitative continuous	Average number of hours studied per week by the students from each school.
Public	Categorical dichotomous	Information about whether the school is public (i.e. 0 = no, 1 = yes).

Table 3. Descriptive statistics of the variables in the dataset

Variables			
approvals	*professors*	*hours*	*public*
Smallest: 0.00	Smallest: 7.00	Smallest: 1.00	Rel. freq. of 0: 57%
Quartile 1: 0.00	Quartile 1: 14.00	Quartile 1: 7.07	
Median: 0.00	Median: 20.00	Median: 10.24	Rel. freq. of 1: 43%
Quartile 3: 3.00	Quartile 3: 27.00	Quartile 3: 14.23	
Largest: 33.00	Largest: 29.00	Largest: 38.00	
Mean: 2.87	Mean: 18.01	Mean: 11.83	
Variance: 40.74	Variance: 59.91	Variance: 57.67	

Note. Rel. freq. stands for relative frequency.

Figure 4. Histogram of the variable *approvals*



Empirical Applications with R

This section contains four empirical cases that demonstrate the practical usefulness of the model selection approach proposed in this study. In such cases, the data detailed in Section 4 are estimated using Poisson, NB, ZIP, and ZINB regression models, which are executed with functions and packages of the R programming language.

Besides the algorithms used, the functional form and respective parameters of the models involved are also detailed in order to algebraically demonstrate the analysis performed behind each respective R algorithm. Furthermore, empirical applications in R of the CT test for overdispersion and Vuong test for excess zeros are also illustrated.

Following the model selection approach detailed in Section 3, the first case explores the Poisson regression model. Subsequently to the modeling stage, it is performed the CT test, which consists of a first step towards an informed decision making on which count regression framework would be the most suitable alternative. Then, the NB regression model is estimated and assessed for comparison purposes. Lastly, the ZIP and ZINB regression models are also estimated and the Vuong test is then also performed.

Case 1: Poisson Estimation and Overdispersion Test

A preliminary comparative analysis between the mean and variance of the dependent variable reported in Table 3, along with the graphical interpretation of the histogram shown in Figure 4, suggest the presence of overdispersion. However, despite of such preliminary evidences, it is not possible yet to confirm its effective presence because, in fact, these statistics need to be conditional to the respective explanatory variables.

In the R programming language, the estimation of the Poisson regression model may be performed using the glm() algorithm of the stats package, which is part of the standard installation. The functional form of the Poisson regression model, ensuing a stepwise procedure, is composed in equation (25):

$$\lambda_i = \exp(-1.0912 + 0.1279 \times hours_i - 0.4123 \times public_i) \qquad (25)$$

All predictor variables in equation (25) have parameters statistically significant at the significance level of 5%, *ceteris paribus*. It is worth mentioning that the variable *professor* is removed from this estimation not

because it is not statistically significant but instead due to the fact that is not statistically significant along with the remaining predictor variables, *ceteris paribus*.

It is also relevant to validate if the Poisson regression model is more appropriate than the NB regression framework. In other words, the analyst needs to test for the presence of overdispersion in the dependent variable conditioned to the predictor variables. Following the model selection approach detailed in Section 3, the CT test is then used.

Upon determining the fitted values λ of the regression model in equation (25), it is then possible to generate the dependent variable Y^*, as formulated in equation (23). The functional form of the auxiliary OLS model without an intercept is shown in equation (26):

$$Y_i^* = 0.0522 \times \lambda_i \qquad (26)$$

Such an auxiliary OLS model without an intercept indicates that the coefficient of the variable λ_i results in a *p*-value of $0.442 > 0.05$. Then, the H_0 of the CT test cannot be rejected. Hence, the equidispersion of the dependent variable, conditional to the predictor variables, is confirmed at the significance level of 5%.

A straightforward routine to perform this overdispersion test is using the overdisp() function. Such an R function should be performed subsequently to the application of the $n-1$ dummies procedure, if applicable depending on the data context. The ouputs of this R function to execute the CT test for overdispersion is shown in Figure 5, confirming the presence of overdispersion in this particular case.

In the light of the outputs in Figure 5, there is evidence that the Poisson would be preferable in comparison with the NB regression model. Nonetheless, it is worth mentioning that the Vuong test has not yet been performed and, therefore, such a model preference is still preliminary. In addition, to calculate the confidence intervals of the regression model in equation (25), one may use the confint() function of the stats package.

Regarding the model parameters in equation (25), the intercept α is -1.0912. The measurement of $\exp(\beta_j)$ denotes the change, on average, of the rate of the dependent variable while changing the respective predictor variable in one unit, *ceteris paribus*. For instance, the slope coefficient of variable *hours* in equation (25) is 0.1279, which yields $\exp(0.1279) = 1.1364$. Hence, the average rate of approvals of students should be multiplied by a factor of 1.1364 for every unit increase in the variable *hours*, *ceteris paribus*. Therefore, one unit increase in the variable *hours* leads to an increase, on average, of 13.64% in the rate of approvals of students in the school, *ceteris paribus*.

Similarly, one may analyze the slope coefficient of the variable *public* in equation (25). Considering that its $\beta_j = -0.4123$, then $\exp(-0.4123) = 0.6621$. Consequently, there is a decline, on average, of 33.79% in the rate of approvals in the national exam for public schools in comparison with private schools, *ceteris paribus*. In other words, the average rate of approvals in the 160 schools should be multiplied by a factor of 0.6621 as a consequence of the fact that the school is public.

In addition, to compute the respective predicted values while preserving the ranges of the predictor variables, then the predict() function of the stats package may be used.

Figure 5. Outputs of the CT test for overdispersion using the overdisp() function

```
Overdispersion Test - Cameron & Trivedi (1990)

data:  dataset
Lambda t test score: = 0.77123, p-value = 0.4417
alternative hypothesis: overdispersion if lambda p-value is less
than or equal to the stipulated significance level
```

Case 2: NB Estimation

In R, the estimation of the NB regression model may be performed through the glm.nb() algorithm of the MASS package. Although the CT test for overdispersion previously indicates that the Poisson regression model should be used in the data context explored in Case 1, in this subsection an estimation using the NB regression model is performed for pedagogical purposes. The functional form of the NB regression model is formulated in equation (27), as follows:

$$u_i = \exp\left(-2.0224 + 0.0509 \times professors_i + 0.1362 \times hours_i - 1.0914 \times public_i\right) \quad (27)$$

At the significance level of 5%, all predictor variables in the regression model in equation (28) have parameters statistically different from zero, *ceteris paribus*. As the parameter ϕ is 0.8956, then $\theta = 1.1166$ with a standard error of 0.329. The ratio between θ and its standard error results in $z > 1.96$, confirming that the parameter θ is statistically different from zero, at the confidence level of 95%.

Apparently, such results may seem counterintuitive due to the fact that the CT test indicates equidispersion in the dependent variable conditioned to the predictor variables. Nevertheless, the parameter ϕ of the NB regression model seems to be statistically different from zero. However, it is worth noticing that the Vuong test has not yet been performed and the lack of such an information may cause serious avoidable problems to the analyst.

Regarding the NB regression model parameters in equation (27), its intercept α is -2.0224 and, analogously to the model in equation (25), the measurement of $\exp(\beta_j)$ denotes the change, on average, of the rate of the dependent variable while changing the respective predictor variable in one unit, *ceteris paribus*.

The slope coefficient of variable *professors* in equation (27) is 0.0509, which yields $\exp(0.0509) = 1.0523$. Hence, the average rate of approvals of students should be multiplied by a factor of 1.0523 for every unit increase in the variable *professors*, *ceteris paribus*. Consequently, one unit increase in the variable *professors* leads to an increase, on average, of 5.23% in the rate of approvals in the national high school exam, *ceteris paribus*.

Analogously, for every unit increase in the variable *hours* the rate of approvals should be multiplied by a factor of 1.1459, yielding an increase, on average, of 14.59% in the rate of approvals in the national exam, *ceteris paribus*. Similarly, the fact that a school is public would cause a decrease, on average, of 66.43% in the rate of approvals, *ceteris paribus*.

In addition, analogously to the estimation performed in Case 1, one may compute the confidence intervals of the NB regression model using the confint() function of the MASS package as well as the predicted values, while preserving the ranges of the predictor variables, using the predict() function of the stats package.

Case 3: ZIP Estimation and Vuong Test

To illustrate the application of the Vuong test to detect the presence of excessive zeros in the dependent variable, this case explores the ZIP regression model estimation, which is also compared to the Poisson regression estimation performed in Case 1. In R, the ZIP regression model may be performed through the zeroinfl() function of the pscl package.

It is worth mentioning that there is not yet a corresponding function in R to perform a stepwise procedure to remove predictor variables whose parameters are not statistically significant in ZI regression models. In view of such a limitation, the authors then carefully select the most suitable ZIP regression model considering the variables in the dataset. The portion of structural and sampling zeros among all zeros in the dataset are shown in equations (28) and (29), respectively:

$$p_{logit_i} = \frac{1}{1+\exp[-(2.9868-0.1138\times professors_i-0.1743\times hours_i+2.9351\times public_i)]} \quad (28)$$

$$\lambda_i = \exp(-0.0808 + 0.0949 \times hours_i) \quad (29)$$

Based on equations (28) and (29), the functional form of the ZIP regression model is then formulated in equation (30):

$$\lambda_{ZIP_i} = \left\{1 - \frac{1}{1+\exp[-(2.9868-0.1138\times professors_i-0.1743\times hours_i+2.9351\times public_i)]}\right\}\times \exp(-0.0808 + 0.0949 \times hours_i) \quad (30)$$

Before discussing the parameters of equation (30), the results of the Vuong test have to be analyzed. Such test outputs generated in R are shown in Figure 6, comparing the Poisson regression model explored in Case 1 and formulated in equation (25) with the ZIP regression model formulated in equation (30). In this case, the Vuong test indicates a better adequacy of the ZIP regression model, resulting in $z = -4.8751$ and a *p*-value smaller than 0.05.

It is worth mentioning that Desmarais and Harden (2013) propose a correction to the Vuong test. This correction is based on the Akaike information criterion (AIC) and Bayesian information criterion (BIC) statistics. Such a correction should be implemented to eliminate potential biases that may affect the model selection decision regarding the most appropriate regression framework.

Regarding the outputs in Figure 6, while the Vuong test results in $z = -4.8751$, the AIC and BIC corrected statistics are $z = -4.6448$ and $z = -4.2908$, respectively, all yielding a *p*-value smaller than 0.05. Therefore, the results of the Vuong test with AIC and BIC correction also indicate that the ZIP regression model is the most appropriate in this context.

In equation (30), the predictor variables *professors*, *hours* and *public* correspond to the variables W_1, W_2 and W_3 in equation (15), respectively, while the variable *hours* refers to a predictor variable X_1, following equation (16). Hence, the regression model formulated in equation (30) aims to estimate if the probability of structural zeros in equation (28) is affected by the number of full-time teachers in each school included in the sample (i.e. variable *professors*), the average number of hours studied per week by the students from each school (i.e. variable *hours*), and whether the school is public or private (i.e. variable dummy *public*).

Furthermore, such a regression model also intends to predict whether the occurrence of a particular count of approvals in the national high school exam (i.e. variable *approvals*) is influenced by the average number of hours studied per week in the respective school i (i.e. variable *hours*), as shown in equation (29). Therefore, it is likely that variables used as W_q are also used as X_k (Hilbe, 2011).

Thus, in the case of an ZI estimate, the measurement of $\exp(\nu_q)$ would indicate a change, on average, of the occurrence of excess structural zeros in the dependent variable *approvals*, whereas $\exp(\beta_j)$ would explain the change, on average, in the rate of approvals of students with respect to the counting data, including sampling zeros, *ceteris paribus*.

Regarding the part of the ZIP regression model in equation (30) that calculates the occurrence of structural zeros, its intercept γ is 2.9868. For instance, as the slope coefficient of variable *professors* is $\nu_1 = -0.1138$, then $\exp(-0.1138) = 0.8924$. This means that one unit increase in the variable *professors* leads to a decrease, on average, of 10.76% in the chances of the occurrence of structural zeros, *ceteris paribus*.

Figure 6. Outputs of the Vuong test comparing the Poisson with the ZIP regression model

```
Vuong Non-Nested Hypothesis Test-Statistic:
(test-statistic is asymptotically distributed N(0,1) under the
 null that the models are indistinguishible)
------------------------------------------------------------
              Vuong z-statistic             H_A    p-value
Raw                  -4.875077 model2 > model1 5.4383e-07
AIC-corrected        -4.644827 model2 > model1 1.7018e-06
BIC-corrected        -4.290799 model2 > model1 8.9016e-06
```

In the case of the slope coefficient of variable *hours*, which is $\nu_2 = -0.1743$, then $\exp(-0.1743) = 0.8401$. Such a result means that one unit increase in the variable *hours* leads to a decrease, on average, of 15.99% in the chances of the occurrence of structural zeros in the rate of approvals in the national high school exam, *ceteris paribus*. Moreover, the variable *public* yields $\nu_3 = 2.9351$, which means that the fact that a school is public would result in an increase, on average, of 1,782.34% in the chances of the occurrence of structural zeros, *ceteris paribus*.

In summary, the larger the number of full-time professors and the average number of hours studied per week by the students in the school, the smaller is the probability of non-occurrence of approvals. On the other hand, public schools do not favor the approval of students in the national high school exam.

In addition, the intercept α of the sampling zeros is -0.0808 and the slope coefficient β_1 of variable *hours* is 0.0949, resulting in $\exp(0.0949) = 1.0995$. This indicates that one unit increase in the variable *hours* leads to an increase, on average, of 9.95% in the chances of the occurrence of approvals, *ceteris paribus*. Thus, a unit increase in the variable *hours*, besides contributing to a smaller probability of the existence of zero inflation, also influences an increase in the number of students approved in the national high school exam, *ceteris paribus*.

Similarly to Cases 1 and 2, the confidence intervals of ZIP regression model estimations may be computed using the confint() function, either for the structural or sampling zeros. Lastly, to predict values while preserving the ranges of the predictor variables, one may use the predict() function of the stats package.

Case 4: ZINB Estimation and Vuong Test

Although the CT test indicates equidispersion and the Vuong test suggests that the ZIP regression model is the most appropriate framework considering the sample data, for pedagogical purposes the ZINB regression model is estimated in this subsection.

In R, similarly to the ZIP model, the ZINB regression model estimation may be performed using the zeroinfl() function of the pscl package. Moreover, as there is not yet a corresponding function in R to perform a stepwise procedure to remove predictor variables whose parameters are not statistically significant in ZI regression models, the authors carefully select the most suitable ZINB regression model using the variables in the dataset.

Algebraically, the estimated ZINB regression model is described by equation (31), in which the parameters of all predictor variables are statistically different from zero at the significance level of 5%, *ceteris paribus*.

$$u_{ZINB_i} = \left\{1 - \frac{1}{1+\exp[-(2.9868-0.1138\times professors_i-0.1743\times hours_i+2.9351\times public_i)]}\right\} \exp(-0.0808 + 0.0949 \times hours_i) \quad (31)$$

The outputs of the Vuong test comparing the NB regression model explored in Case 2 and expressed in equation (27) with the ZINB regression model expressed in equation (31), with AIC and BIC correction, are shown in Figure 7.

The Vuong test shows a significant z statistics at the confidence level of 95%. This indicates that the ZINB is preferable to the NB regression model due to the confirmation of the presence of an excessive amount of zero-valued observations.

Regarding the part of the ZINB regression model in equation (31) that calculates the occurrence of structural zeros, the intercept γ is 2.9868. As the slope coefficient ν_1 of the variable *professors* is –0.1138, then $\exp(-0.1138) = 0.8924$. This indicates that one unit increase in the variable *professors* leads to a decrease, on average, of 10.76% in the chances of the occurrence of structural zeros, *ceteris paribus*.

As the slope coefficient ν_2 of variable *hours* is -0.1743, then $\exp(-0.1743) = 0.8401$. This means that one unit increase in the variable *hours* leads to a decrease, on average, of 15.99% in the chances of the occurrence of structural zeros in the rate of approvals, *ceteris paribus*. Moreover, regarding the variable *public*, $\nu_3 = 2.9351$ and then $\exp(2.9351) = 18.8234$. This means that the fact that a school is public would result in an increase, on average, of 1,782.34% in the chances of the occurrence of structural zeros, *ceteris paribus*.

The intercept α of the sampling zeros is -0.0808 and the slope coefficient β_1 of variable *hours* is 0.0949, resulting in $\exp(0.0949) = 1.0995$. This indicates that one unit increase in the variable *hours* leads to an

increase, on average, of 9.95% in the chances of the occurrence of approvals in the national high school exam, *ceteris paribus*.

The parameter ϕ is not statiscally different from zero, resulting in a *p*-value of 0.889. Therefore, there is further evidence of the existence of equidispersion in the dependent variable conditioned to the predictor variables. It is worth mentioning that in Case 2, when exploring the NB regression model, the parameter ϕ appears to be statistically significant. However, such a statistically significant result is, actually, a consequence of the fact that in that particular case the Vuong test is not performed and, therefore, there is no information about the presence of inflation of zeros in the dependent variable.

Moreover, the fact that ϕ is not statistically different from zero in the ZINB regression model estimation produces fitted values that causes the model to retrograde to a ZIP regression model. This is evidenced through the strong similarity of the parameters and the log-likelihood ratio (LR) test comparing both ZI regression estimations, as shown in Figure 8.

As detailed in Figure 8, the outputs $\chi^2 = 0.0001$ with one degree of freedom and *p*-value of 0.992 explicit that, at a confidence level of more than 95% (99%, for instance), there is no statistically significant differences between the ZIP and ZINB regression models, despite the fact that the ZINB regression model contains an additional parameter (i.e. ϕ). Lastly, in R, the LR test may be performed using thelrtest() function of the lmtest package.

Figure 7. Outputs of the Vuong test comparing the NB with the ZINB regression model

```
Vuong Non-Nested Hypothesis Test-Statistic:
(test-statistic is asymptotically distributed N(0,1) under the
 null that the models are indistinguishible)
------------------------------------------------------------
              Vuong z-statistic             H_A    p-value
Raw                  -4.111756 model2 > model1 1.9633e-05
AIC-corrected        -3.856389 model2 > model1 5.7537e-05
BIC-corrected        -3.463741 model2 > model1 0.00026636
```

Figure 8. Outputs of the LR test comparing the estimation of the ZIP (model 1) and ZINB regression model (model 2)

```
Likelihood ratio test

Model 1: approvals ~ hours | professors + hours + public
Model 2: approvals ~ hours | professors + hours + public
  #Df  LogLik Df Chisq Pr(>Chisq)
1   6 -210.42
2   7 -210.42  1 1e-04      0.992
```

Log-likelihood and Fitted Values Comparison

After illustrating the practical usefulness of the proposed count model selection approach in Section 3 through four cases exemplified in Section 4 - in which the ZIP regression model is selected as the most suitable framework, in Figure 9 is presented the estimated log-likelihood of each regression model included in the present study, using the logLik() function of the stats package.

The largest log-likelihood estimations are produced by both ZI regression models, while the Poisson regression model presents the smallest log-likelihood value. The Poisson and NB regression models are 23% and 14% smaller than the ZIP regression model, respectively.

As the variable *hours* consists of the single predictor variable included in the sampling zeros part of the ZIP and ZINB regression models, then it is performed a comparison between observed values along with their respective fitted values of all four estimated regression models, as shown in Figure 10.

Figure 10 also suggests, through a data visualization approach, a better fit of the ZIP regression model estimation in comparison with the remaining model frameworks. Moreover, it is possible to notice that the ZINB regression estimation naturally retrogrades to the ZIP regression model framework in cases where $\phi \rightarrow 0$.

Conclusion

In this paper is presented a straightforward model selection approach to indicate the most suitable count data regression model, contemplating relevant data characteristics. The proposed approach adapts the rationale in Perumean-Chaney et al. (2013), covering two of the most relevant problems commonly found in real-world data count – namely, overdispersion and zero-inflation, while enabling analysts to perform the entire selection approach using the programming language R.

The comparison between the mean and variance of the dependent variable does not consist of a suitable approach for overdispersion examination. In similar fashion, assuming a particular amount of zero-valued observations is not the appropriate procedure to identify the zero-inflation characteristic in the dependent variable.

Figure 9. Log-likelihood estimations comparing all four regression model frameworks

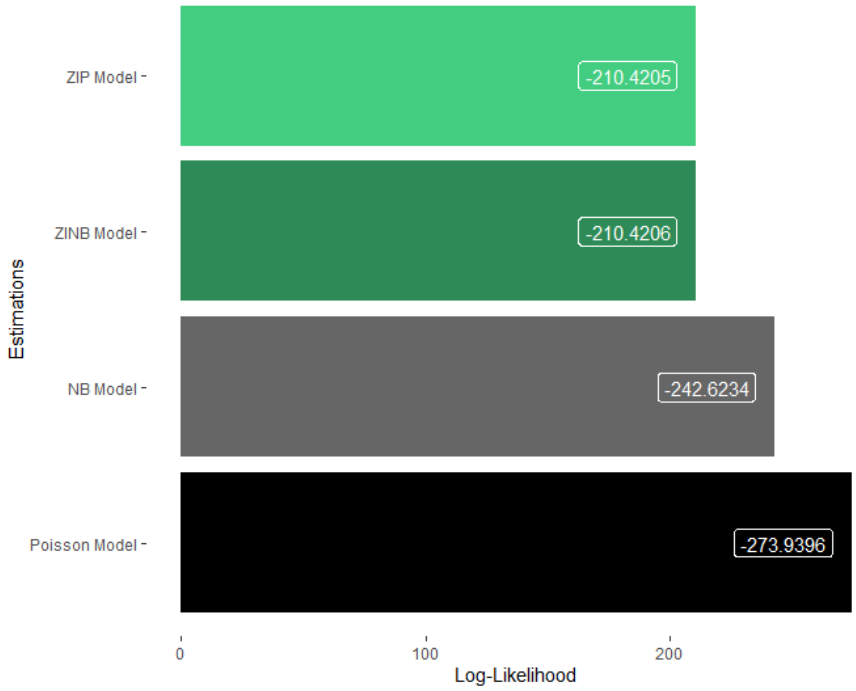
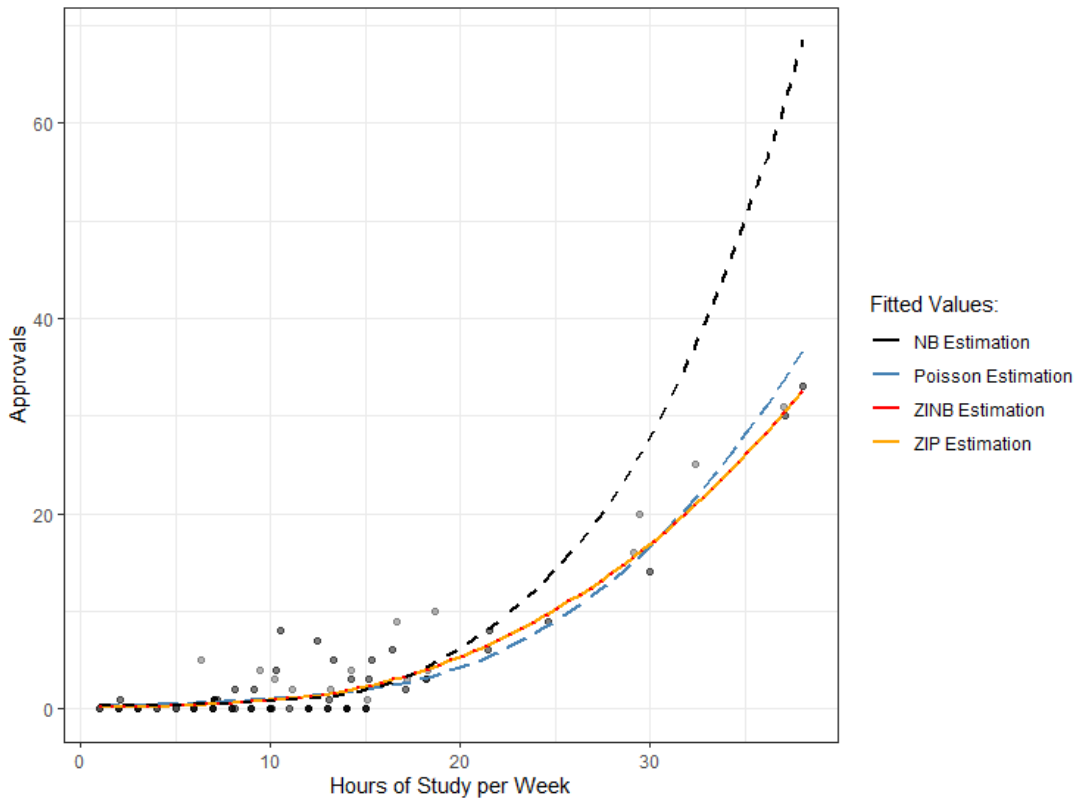


Figure 10. Comparison of fitted values of all estimated regression models, based on the variables *approvals* and *hours*



Contrastingly, the CT test for overdispersion proposed by Cameron and Trivedi (1990) consists of a reliable procedure to confirm overdispersion within the dependent variable conditional to the predictor variables. Complementarily, the Vuong test proposed by Vuong (1989) is a solid statistical tool to identify the presence of excessive amount of zero-valued observations in the dependent variable. Such tests support the decision to be made by the analyst on which count regression model should be used considering data characteristics.

In the present study the Poisson, NB, ZIP, and ZINB regression model are estimated and discussed through four empirical cases executed with R. The CT and Vuong test are also applied to the data. Considering the dataset used (i.e. information about approvals of students from 160 schools in a Brazilian national high school exam), the CT test indicates equidispersion within the dependent variable, conditioned to the predictor variables. Additionally, the outputs of the Vuong test confirm the presence of zero-inflation in the dependent variable.

Furthermore, the cases explored in this study emphasize that if the zero-inflation data characteristic is not properly detected, this may induce the analyst to incorrectly believe that there is overdispersion in the data by finding an apparently statistically significant ϕ parameter. Consequently, the respective misleading conclusion would indicate that the NB regression would be the most appropriate model whereas, in fact, the best alternative in such a data context would be the ZIP regression model instead.

It is also underscored that if a count regression estimation is performed without properly analyzing the underlying distribution of the dependent variable as well as neglecting the importance of applying statistical tests to confirm relevant data idiosyncrasies, then this potentially leads to bias in the estimated coefficients, with consequent inefficiency and inconsistency of the modeling process.

Finally, all commands in R used throughout the paper are presented in the Appendix.

References

Blackburn, M. L. (2015). The relative performance of Poisson and negative binomial regression estimators. *Oxford Bulletin of Economics and Statistics,* 77**,** 605-616.

Breslow, N. (1990). Tests of hypotheses in over dispersed Poisson regression and other quasi-likelihood models. *Journal of the American Statistical Association,* 85**,** 565-571.

Cameron, A. C., & Trivedi, P. K. (1990). Regression-based tests for overdispersion in the Poisson model. *Journal of econometrics,* 46**,** 347-364.

Cameron, A. C., & Trivedi, P. K. (2010). Microeconometrics Using Stata, Revised Edition. Stata Press. *College Station*.

Cameron, A. C., & Trivedi, P. K. (2013). *Regression analysis of count data*, Cambridge university press.

Desmarais, B. A. & Harden, J. J. (2013). Testing for zero inflation in count models: Bias correction for the Vuong test. *The Stata Journal,* 13**,** 810-835.

Dupuy, J.-F. (2018). *Statistical methods for overdispersed count data*, Elsevier.

Fávero, L. P., & Belfiore, P. (2019). *Data science for business and decision making*, Academic Press.

Fávero, L. P. L., Belfiore, P., Dos Santos, M. A., & Souza, R. F. (2020). Overdisp: A Stata (and Mata) Package for Direct Detection of Overdispersion in Poisson and Negative Binomial Regression Models. *Statistics, Optimization & Information Computing,* 8**,** 773-789.

Fávero, L. P. L., Dos Santos, M. A., & Serra, R. G. (2018). Cross-border branching in the Latin American banking sector. *International Journal of Bank Marketing.*

Freitas Souza, R. F., Fávero, L. P., Belfiore, P., and Correa., H. L. (2020). Package 'overdisp'.

Hilbe, J. M. (2011). *Negative binomial regression*, Cambridge University Press.

Hilbe, J. M. (2014). *Modeling count data*, Cambridge University Press.

Klakattawi, H. S., Vinciotti, V., and Yu, K. (2018). A simple and adaptive dispersion regression model for count data. *Entropy,* 20**,** 142.

Kullback, S., and Leibler, R. A. (1951). On information and sufficiency. *The 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.

Payne, E. H., Gebregziabher, M., Hardin, J. W., Ramakrishnan, V., and Egede, L. E. (2018). An empirical approach to determine a threshold for assessing overdispersion in Poisson and negative binomial models for count data. *Communications in Statistics-Simulation and Computation,* 47**,** 1722-1738.

Payne, E. H., Hardin, J. W., Egede, L. E., Ramakrishnan, V., Selassie, A., and Gebregziabher, M. (2017). Approaches for dealing with various sources of overdispersion in modeling count data: Scale adjustment versus modeling. *Statistical methods in medical research,* 26**,** 1802-1823.

Perumean-Chaney, S. E., Morgan, C., McDowall, D., and Aban, I. (2013). Zero-inflated and overdispersed: what's one to do? *Journal of Statistical Computation and Simulation,* 83**,** 1671-1683.

Smith, D. and Faddy, M. (2016). Mean and variance modeling of under-and overdispersed count data. *Journal of Statistical Software,* 69**,** 1-23.

Taddy, M. 2015. Distributed multinomial regression. *The Annals of Applied Statistics,* 9**,** 1394-1414.

Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica: Journal of the Econometric Society***,** 307-333.

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data*, MIT press.

Zeileis, A., Kleiber, C., & Jackman, S. 2008. Regression Models for Count Data in R. *Journal of Statistical Software*, 27, 1-25.

Zeviani, W. M., Ribeiro Jr., P. J., Bonat, W. H., Shimakura S. E., and Muniz, J. A. (2014). The Gamma-count distribution in the analysis of experimental underdispersed data. *Journal of Applied Statistics,* 41**,** 2616-2626.

Citation:

Fávero, L. P., Souza, R. F., Belfiore, P., Corrêa, H. L., & Haddad, M. F. C. (2019). Count Data Regression Analysis: Concepts, Overdispersion Detection, Zero-inflation Identification, and Applications with R. *Practical Assessment, Research & Evaluation*, 26(13). Available online: https://scholarworks.umass.edu/pare/vol26/iss1/13/

Corresponding Author

Luis Paulo Fávero
University of São Paulo

email: lpfavero [at] usp.br

Appendix. Model Selection Approach Program in R

Loading packages:

```
library(tidyverse)
library(MASS)
library(overdisp)
library(pscl)
library(lmtest)
library(reshape2)
```

Loading dataset:

```
load("dataset.RData")
```

Descriptive statistics of the variables in the dataset (Table 3):

```
summary(dataset)
```

Histogram of the variable *approvals* (Figure 4)

```
dataset %>%
  ggplot() +
  geom_bar(aes(x = factor(approvals)), stat="count", fill = "seagreen") +
  geom_text(aes(x = factor(approvals),label = ..count..),
            stat="count", vjust = -0.8, size = 2.5, color = "seagreen") +
  labs(x = "Approvals",
       y = "Frequency") +
  theme_classic()
```

Poisson estimation:

```
poisson_model <- glm(approvals ~ professors + hours + public,
                     data = dataset,
                     family = "poisson")
```

Stepwise procedure (Expression 25):

```
poisson_final_model <- step(object = poisson_model,
                            k = qchisq(p = 0.05,
                                       df = 1,
                                       lower.tail = FALSE),
                            trace = 0)
```

Parameters of the Poisson final model (Expression 25):

```
summary(poisson_final_model)
```

Creating a dependent variable Y* (ystar) for the auxiliary OLS model, as shown in expression (26):

```
dataset["lambda"] <- poisson_final_model$fitted.values

dataset["ystar"] <- ((dataset$approvals - dataset$lambda) ^ 2 - dataset$approvals) /
dataset$lambda
```

Estimating the auxiliary OLS Model Y*~λ, without the intercept (Expression 26):

```
aux_ols_model <- lm(formula = ystar ~ 0 + lambda,
                    data= dataset)
```

Parameters of the auxiliary OLS Model (aux_ols_model) Y*~λ, without the intercept (Expression 26):

```
summary(aux_ols_model)
```

Use of the overdisp() command (Figure 5):

```
overdisp(x = dataset,
         dependent.position = 1,
         predictor.position = 2:4)
```

Negative Binomial estimation:

```
nb_model <- glm.nb(approvals ~ professors + hours + public,
                   data = dataset)
```

Stepwise procedure (Expression 27):

```
nb_final_model <- step(object = nb_model,
                       k = qchisq(p = 0.05,
                                  df = 1,
                                  lower.tail = FALSE),
                       trace = 0)
```

Parameters of the Negative Binomial final model (Expression 27):

```
summary(nb_final_model)
```

θ Parameter:

```
nb_final_model$theta
```

ϕ Parameter ($1/\theta$):

```
1 / nb_final_model$theta
```

z Statistics of θ Parameter:

```
nb_final_model$theta / nb_final_model$SE.theta
```

ZIP Estimation (Expression 30):

```
zip_model <- zeroinfl(formula = approvals ~ hours | professors + hours + public,
                      data = dataset,
                      dist = "poisson")
```

Parameters of the ZIP model (Expression 30):

```
summary(zip_model)
```

Vuong test - poisson_final_model vs zip_model (Figure 6):

```
vuong(poisson_final_model,zip_model)
```

ZINB Estimation (Expression 31):

```
zinb_model <- zeroinfl(formula = approvals ~ hours | professors + hours + public,
                       data = dataset,
                       dist = "negbin")
```

Parameters of the ZINB model (Expression 31):

```
summary(zinb_model)
```

Vuong test - poisson_final_model vs zip_model (Figure 7):

```
vuong(nb_final_model,zinb_model)
```

LR Test (Figure 8):

```
lrtest(zip_model, zinb_model)
```

Log-likelihood and Fitted Values Comparison (Figure 9)

```
data.frame(Poisson = logLik(poisson_final_model),
           NegBin = logLik(nb_final_model),
           ZINB = logLik(zinb_model),
           ZIP = logLik(zip_model)) %>%
  rename(`Poisson Model` = 1,
         `NB Model` = 2,
         `ZIP Model` = 4,
         `ZINB Model` = 3) %>%
  melt() %>%
  ggplot(aes(x = variable, y = (abs(-value)), fill = factor(variable))) +
  geom_bar(stat = "identity") +
  geom_label(aes(label = (round(value,4))), hjust = 1.2, color = "white") +
  labs(y = "Log-Likelihood",
       x = "Estimations") +
  coord_flip() +
  scale_fill_manual("Legenda:",
                    values = c("black", "gray40", "seagreen", "seagreen3")) +
  theme(legend.title = element_blank(),
        panel.background = element_rect("white"),
        legend.position = "none")
```

Comparison of fitted values of all estimated regression models, based on the variables *approvals* and *hours* - Figure 10:

```
dataset %>%
  mutate(zip_lambda = zip_model$fitted.values,
         u = nb_final_model$fitted.values,
         zinb_u = zinb_model$fitted.values) %>%
  ggplot() +
  geom_point(aes(x = hours, y = approvals), alpha = 0.3) +
  geom_smooth(aes(x = hours, y = lambda,
                  color = "Poisson Estimation"), linetype = "longdash", se = F) +
  geom_smooth(aes(x = hours, y = zip_lambda,
                  color = "ZIP Estimation"), linetype = "solid", se = F) +
  geom_smooth(aes(x = hours, y = u, color = "NB Estimation"), linetype = "dashed", se = F) +
  geom_smooth(aes(x = hours, y = zinb_u, color = "ZINB Estimation"), linetype = "dotted", se = F)
+
  scale_color_manual("Fitted Values:",
                     values = c("black", "steelblue", "red", "orange")) +
  labs(x = "Hours of Study per Week",
       y = "Approvals") +
  theme_bw()
```