Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Estimating Marginal Effects with Zero-inflated Models: A Tutorial with R package mzim

Chendong Li

cliattx@tamu.edu

Oi-Man Kwok
Timothy Lawrence

Research Article

Keywords:

Posted Date: December 2nd, 2025

DOI: https://doi.org/10.21203/rs.3.rs-8224986/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Estimating Marginal Effects with Zero-inflated Models: A Tutorial with R package *mzim*

Chendong Li[1], Oi-Man Kwok[1], Timothy Lawrence[1]

[1]Department of Educational Psychology, Texas A&M University, College Station, TX, USA

Author Emails

Chendong Li: cliattx@tamu.edu

Oi-Man Kwok: omkwok@tamu.edu

Timothy Lawrence: tlawrence3@tamu.edu

Corresponding Author

Chendong Li

Harrington Tower, College Station, TX 77843-4225

Phone: +1 979 326 2250

Abstract

Count data in the psychological and health sciences are often characterized by an excess of zero values, a feature known as zero-inflation. While traditional zero-inflated models, such as the Zero-Inflated Poisson (ZIP) and Zero-Inflated Negative Binomial (ZINB), were developed to handle such data, they present challenges for applied researchers. Standard count models can produce biased estimates, and the dual-parameter output of traditional zero-inflated models provides conditional effects for a latent at-risk subpopulation, which complicates interpretation and often fails to directly answer research questions about the entire population. To address these limitations, marginalized zero-inflated (mZI) models directly estimate the population-averaged effect, yielding a single, interpretable coefficient for each predictor's overall effect. However, the adoption of mZI models has been hindered by the lack of an accessible software package. The current study has two objectives: first, it provides a tutorial on the theory, estimation, and interpretation of marginalized zero-inflated models. Second, it introduces *mzim*, a new R package designed to make both marginalized zero-inflated Poisson (mZIP) and Negative Binomial (mZINB) models readily accessible. Using an empirical example on self-reported youth abuse experiences, we demonstrate a complete workflow with the *mzim* package and compare the results from the mZINB model to traditional approaches, highlighting the practical benefits of the marginalized framework for applied researchers.

Overview

In psychological and medical research, count data—representing the number of times an event occurs—are frequently encountered (Agresti, 2002). Examples include the number of school absences in a year or chemotherapy sessions attended by cancer patients (Zorc et al., 2013; Chrischilles et al., 2009). Often, particularly when studying negative or infrequent behaviors in general populations, this data is characterized by a high number of zero-count observations, a phenomenon known as zero-inflation (ZI). Studies on substance use, violent behaviors, or risky sexual practices commonly yield such data, which is typically highly skewed (e.g., Bandyopadhyay et al., 2011; Swartout et al., 2015; Hu et al., 2011). These excess zeros surpass what would be expected if the data followed standard count distributions, reflecting a mixture of individuals who do not engage in the behavior and those who do, but may not have during the study period.

When confronted with zero-inflated count data, researchers have several analytical options, each with distinct advantages and limitations. One common approach for handling non-normal data is to use maximum likelihood estimation with robust standard errors (MLR), as implemented in software like Mplus (Muthén & Muthén, 2017). This method uses a sandwich estimator to compute standard errors that are consistent even under violations of assumptions like heteroscedasticity. Its primary limitation, however, is that while MLR can correct for general non-normality, it is not designed to address the specific two-part data generating process underlying zero-inflation. It fails to account for the mixture of structural and sampling zeros, and thus may not adequately model the phenomenon of interest.

Another approach is to ignore the zero-inflation and apply standard count models like Poisson or Negative Binomial (NB) regression. However, this misspecification leads to

significant statistical issues. Numerous simulation studies have documented that applying standard count models to zero-inflated data can produce biased parameters, and in some cases, an incorrect estimation of the effect direction (Lambert, 1992). Furthermore, the unaccounted-for variance from the excess zeros leads to underestimated standard errors, which results in an inflated Type I error rate (Campbell, 2021).

To more accurately analyze data with an excess of zeros, researchers developed specialized models that directly account for the zero-inflation process. The most widely used approaches are the Zero-Inflated Poisson (ZIP) model (Lambert, 1992) and the Zero-Inflated Negative Binomial (ZINB) model (Greene, 1994). These models explicitly accommodate the dual nature of zero-inflated data. The ZIP and ZINB models are two-part mixture models that conceptualize the data as arising from two distinct processes. The models combine a binary logistic model with a standard count model. The logistic component estimates the probability that an observation is a structural zero—originating from a subpopulation that has no chance of experiencing the event. The count component (Poisson for ZIP, Negative Binomial for ZINB) models the counts for the remaining subpopulation, which is considered at-risk of experiencing the event. For this at-risk group, a zero count is still possible and is referred to as a sampling zero. The ZINB model is an extension of the ZIP that adds a dispersion parameter to the count component, making it more flexible for data where the variance is greater than the mean (overdispersion).

These models are built on the assumption that zeros in the data arise from two distinct sources: structural zeros from individuals who will always be zero, and sampling zeros from individuals who could have a positive count but did not during the study period (He et al., 2014). For instance, in a study of cigarette use, zero counts include both participants who have never

smoked (structural zeros) and smokers who simply did not smoke during the observation window (sampling zeros) (Pittman et al., 2022). Thus, the parameters from ZIP and ZINB models are interpreted within a latent class framework, distinguishing between the at-risk sampling zero subpopulation and not-at-risk structural zero subpopulations (Loeys et al., 2012). The coefficients from the logistic part describe how predictors relate to the odds of being in the structural zero group. The coefficients from the count part describe how predictors affect the outcome only for those within the at-risk group.

Despite their statistical elegance, traditional ZIP and ZINB models present significant challenges for applied researchers. The latent class interpretation often mismatches the research objective. Researchers frequently want to assess the marginal effect of a predictor across the entire sample, not just a latent subgroup (Albert et al., 2014). For example, in an intervention study, the key question is often the average effect of the program on the total population, rather than the effect on the unobserved at-risk subgroup.

The dual-parameter output from ZIP and ZINB complicates analysis and communication. The coefficients from the count part do not represent the overall effect on the population mean. While methods exist to calculate this marginal effect post-estimation (e.g., using the delta method or bootstrapping), these techniques can be computationally intensive and difficult for applied researchers to implement correctly (Albert et al., 2014; Long et al., 2014).

The core assumption of two separated subpopulations (structural vs. at-risk) is strong and often theoretically untestable due to the latent nature of the class membership. In applied research, the distinction between these zero types is frequently ambiguous, with no empirical method to differentiate them at the individual level. Consequently, imposing this rigid latent class framework may fail to accurately reflect the true data-generating process. An alternative

framework, the hurdle model, conceptualizes this process differently. Instead of assuming two types of zeros, hurdle models treat all zeros as a single group and model the data in two stages: a binary model determines whether an observation has a zero or a positive count, followed by a separate truncated count model for the positive values only. This approach avoids the distinction between structural and sampling zeros entirely. However, a detailed comparison of hurdle models and their own set of interpretations is beyond the scope of this paper.

To overcome these challenges, marginalized zero-inflated models (mZIP and mZINB) were developed as an accessible approach for estimating marginal effects (Long et al., 2014; Preisser et al., 2017). Instead of modeling two separate latent processes, these models are structured to directly model the population-averaged (marginal) mean count while still accounting for the excess zeros. The marginalized ZIP (mZIP) model directly links the overall mean count to covariates, allowing for straightforward inference on overall exposure effects. The interpretation becomes analogous to a standard Poisson regression but is corrected for the zero-inflation. The marginalized ZINB (mZINB) model extends this framework by replacing the Poisson component with a Negative Binomial distribution, allowing it to handle both excess zeros and overdispersion simultaneously.

The marginalized approach directly resolves the primary challenges posed by traditional zero-inflated models. First, by directly modeling the marginal mean, the coefficients from an mZIP or mZINB model represent the effect of a predictor on the average outcome across the entire population. This provides a single, easy-to-interpret parameter that directly answers the most common type of research question about overall effects.

Second, marginalized models provide parameter estimates that have the same straightforward interpretation as those from standard Poisson or NB regression (i.e., log-

incidence density ratios). This eliminates the need for complex post-estimation calculations to derive marginal effects, making the results more accessible and less prone to misinterpretation.

Lastly, another conceptual advantage of the marginalized approach is how it addresses the strong, and often untestable, assumption regarding the origin of zero counts. While the marginalized model retains a logistic component similar to its traditional counterpart, its function is reconceptualized. In a traditional ZIP or ZINB model, the logistic component is central to the interpretation, defining latent classes (at-risk vs. not-at-risk) that are necessary to understanding the count parameters. In contrast, the primary role of the logistic component in a marginalized model is statistical rather than interpretive. It functions as a mathematical mechanism to adjust for the distributional properties of the data—namely, the excess zeros—thereby enabling an unbiased and accurate estimation of the marginal effect. Consequently, the marginalized framework implies that while the dual-source nature of zeros (structural vs. sampling) may exist as the underlying data-generating mechanism, this latent separation becomes subordinate to the estimation of the overall mean. The logic of the mZI approach is not that it assumes the absence of the existence of structural zeros. However, it structures the inference in a way that distinguishing between the two sources is no longer required to interpret the overall exposure effect. By focusing the count component on the marginal mean, the separation of subpopulations, even if theoretically present, becomes irrelevant for answering the research question regarding the population-averaged effect.

The implementation of traditional ZIP and ZINB models for applied researchers has been facilitated by several statistical software packages, including the *pscl* (Zeileis et al., 2008) and *glmmTMB* (McGillycuddy et al., 2025) packages in R, the *zip* and *zinb* commands in Stata (StataCorp, 2023), and PROC GENMOD in SAS (SAS Institute Inc., 2022). In contrast, the

implementation of the marginalized versions has been more limited, creating a gap between methodological development and practical application. In the R software, for instance, the mZIP model can be estimated using packages such as mcount and mzipmed. However, a limitation is the absence of a single, unified package that offers estimation for both the mZIP model and its overdispersed counterpart, the mZINB model. Accessible implementations for marginalized zero-inflated models are not readily available in other widely used statistical programs like SAS or Stata. This presents a barrier for applied researchers wishing to adopt these more interpretable models.

The current study has two objectives. First, this paper serves as a tutorial on the formulation, estimation, and implementation of mZIP and mZINB models. We present the underlying statistical theory of these models. Second, we introduce the *mzim* (Author, 2025), a newly developed R package designed to make them readily accessible to applied researchers. The major advantage of this package is that it can compute the marginal estimates based on the entire sample instead of a subsample as the traditional ZIP and ZINB models when data contain sampling zeros or the mix of both structural and sampling zeros. To demonstrate the complete workflow, from model specification to the interpretation of results, we analyze an empirical dataset using the *mzim* package. The paper will be arranged as follows: we will detail the derivations for both the regular ZIP and ZINB models and their marginalized version. Then, we will introduce the dataset we are using to illustrate the proposed methods. Next, we will present the workflow of using mZIP/mZINB models in the empirical dataset. Lastly, we will compare and interpret the results from both analyses.

Statistical Model

ZIP and ZINB models

The traditional ZIP models for handling zero-inflated count data were first introduced by Lambert (1992) assuming that the count variable Y_i for the ith observation arises from two distinct processes: (1) zero-inflation process with probability ψ_i - the observation is an excess zero, representing a structural zero that cannot take on positive count values and (2) count process with probability $1 - \psi_i$. The count process $g(y_i|\theta_i)$ follows a standard Poisson distribution with mean $\theta_i = \mu_i$, allowing for both zero and positive counts. Thus, the probability distribution of Y_i in the ZIP model is defined as follows

$$P(Y_i = 0) = \psi_i + (1 - \psi_i)e^{-\mu_i} \tag{1}$$

$$P(Y_i = y_i) = (1 - \psi_i)g(y_i|\theta_i), y_i \in \mathbb{Z}^+$$

When the count process exhibits overdispersion, meaning the sample variance is greater than the sample mean, it violates the Poisson distribution assumption. The count process will instead follow a NB distribution with $\theta_i = (\mu_i; \phi)$ where ϕ is the overdispersion parameter, which is referred as ZINB (Greene, 1994). This allows the variance to be specified as a quadratic function of the mean in the count process: $Var(y_i) = \mu_i + \phi\mu_i^2$.

To assess the impact of covariates on the count distribution in a ZIP or ZINB model, two processes could be explicitly expressed as a function of covariates. The most natural choice to model the probability of excess zeros is to use a logistic regression model,

$$\text{logit}(\psi_i) = Z_i'\lambda \tag{2}$$

where Z_i is a vector of covariates influencing the structural zero process and λ is a vector of parameters associated with Z_i. Other link functions (e.g., probit link) can be used as well, but will not be considered here. Each element λ_j represents the log-odds change in the probability of an observation being a structural zero for a one-unit increase in the jth covariate in Z_i. A positive

λ_j indicates that as the jth covariate in Z_i, the odds of the observation being a structural zero increases.

For the count process, the impact of covariates excluding the excessive zeros can be modeled through Poisson regression or NB regression below

$$\log(\mu_i) = X_i'\beta \tag{3}$$

where X_i is a vector of covariates influencing the count process and β is a vector of parameters associated with X_i. Each element β_j represents the log change in the expected count μ_i for a one-unit increase in the jth covariate in X_i, among observations not in the excess zero group. Exponentiating β yields the incidence rate ratio (IRR), indicating the multiplicative effect on the expected count. In practice, models often have specification either $Z_i = X_i$ or that Z_i consists of a subset of the covariates X_i in the count process (Preisser et al., 2016)

It is important to note that in the traditional ZIP and ZINB models, researchers should be cautious when interpreting the parameters λ and β separately within their respective processes. Specifically, each element λ_j in the zero-inflation component represents the log-odds ratio of a one-unit increase in the jth covariate in Z_i on the probability of an observation being an excess zero. Similarly, each element β_j in the count component represents the log-incidence rate ratio (L-IRR) of a one-unit increase in the jth covariate in X_i on the mean count μ_i among individuals not in the excess zero group. However, researchers are often more interested in the overall effect of predictors on the entire population mean of the outcome rather than effects confined to latent subpopulations. This presents a challenge in the traditional ZIP and ZINB models because there is no simple summary of the effect of a one-unit increase in a predictor on the overall population

mean $\nu_i = E(Y_i)$, which is frequently the primary focus of investigation. The marginal mean of the outcome ν_i is given by

$$E(Y) = \nu_i = (1 - \psi_i)\mu_i = \frac{e^{X_i'\beta}}{1 + e^{Z_i'\lambda}} \tag{4}$$

indicating that the overall mean depends on both the zero-inflation parameters λ and the count parameters β. This dependency between predictors and the marginal mean is complex and nonlinear, making it difficult to directly interpret the impact of a one-unit increase in a predictor on the overall outcome within the traditional ZIP model.

Marginalized ZI models

To address the challenges of interpreting the overall effect of predictors on the entire population mean within the traditional ZIP and ZINB models, the mZIP and mZINB models were developed (Long et al., 2014; Preisser et al., 2016). These marginalized models use similar framework that modifies the traditional approaches by directly modeling the marginal mean of the outcome variable, providing a more straightforward interpretation of predictor effects on the overall population.

In the mZIP model, we begin by specifying the zero-inflation process similarly to the traditional ZIP model. The probability of observation being an excess zero is modeled using a logistic regression as before in Equation (2) and solving for ψ_i using the inverse logit function, we obtain

$$\psi_i = \frac{1}{1 + e^{Z_i'\lambda}} \tag{5}$$

Next, instead of modeling the count process conditional on the outcome not being in the latent class of excess zeros, the mZIP model directly models the marginal mean ν_i of the outcome for the entire population using a log-linear relationship

$$\log(\nu_i) = X_i'\alpha \tag{6}$$

where $\nu_i = E(Y_i)$ is the expected count for the ith observation, X_i is a vector of covariates associating with the overall mean, and α is a vector of parameters associated with X_i. Exponentiating both sides yields

$$\nu_i = e^{X_i'\alpha} \tag{7}$$

This formulation allows the parameters α to be interpreted similarly to those in standard Poisson regression, providing a direct interpretation of the effect of predictors on the marginal mean ν_i. With previously given in Equation (4) $\nu_i = (1 - \psi_i)\mu_i$, with the known functional form of ν_i, we could solve for μ_i, the mean of the count process in the traditional zero-inflated Poisson and substituting the expressions for ν_i in Equation (7) and ψ_i in Equation (5) into equation for μ_i, we get

$$\mu_i = \frac{\nu_i}{1 - \psi_i} = \frac{e^{X_i'\alpha}}{1 - (\frac{e^{Z_i'\lambda}}{1 + e^{Z_i'\lambda}})} = e^{X_i'\alpha}(1 + e^{Z_i'\lambda}) \tag{8}$$

In order to use the ZIP model likelihood framework, we redefine $\mu_i = e^{\delta_i}$ and take the natural logarithm of both sides. Here, δ_i is not necessarily a linear function of model parameters, highlighting the interpretational challenges inherent in traditional ZIP models

$$\delta_i = \log \mu_i = X_i'\alpha + \log(1 + e^{Z_i'\lambda}) \tag{9}$$

By substituting Equation (5) and Equation (6) into Equation (9), the likelihood of the mZIP model for (γ, α) is

$$L(\gamma, \alpha|y) = \prod_{y_i} \left(1 + e^{Z_i'\lambda}\right)^{-1} \prod_{y_i=0} (e^{Z_i'\lambda} + e^{-(1+\exp(Z_i'\lambda))\exp(X_i'\alpha)}) \times \prod_{y_i>0} [e^{-(1+\exp(Z_i'\lambda))\exp(X_i'\alpha)} \left(1 + e^{Z_i'\lambda}\right)^{y_i} e^{X_i'\alpha y_i}/(y_i!)] \quad (10)$$

For datasets where the variance of the count data is greater than the mean, the mZIP model can be extended to the mZINB model. The mZINB model provides additional overdispersion parameter by using a NB distribution for the count process, while keeping the population-averaged interpretation of the marginalized framework. The mZINB model retains the same direct modeling approach for the zero-inflation probability ψ_i, and the marginal mean ν_i, as defined previously. The probability of an excess zero ψ_i is specified using the logistic model in Equation (5), and the marginal mean ν_i is modeled with the log-linear relationship shown in Equations (6) and (7). The relationship used to solve for the mean of the count process μ_i remains the same in Equation (8) but the distribution assumed for these counts is NB rather than Poisson. The core of the mZINB model is its use of the Negative Binomial (NB) distribution, whose probability mass function is detailed in the Appendix. The NB distribution is defined by its expected mean $E(y_i) = \mu_i$ and a variance function of $Var(y_i) = \mu_i + \alpha\mu_i^2$. The overdispersion parameter α allows the variance to exceed the mean. The likelihood for the mZINB model is therefore constructed by substituting this NB probability function for the Poisson in the marginalized framework in Equation (10). The resulting log-likelihood function is shown in the Appendix.

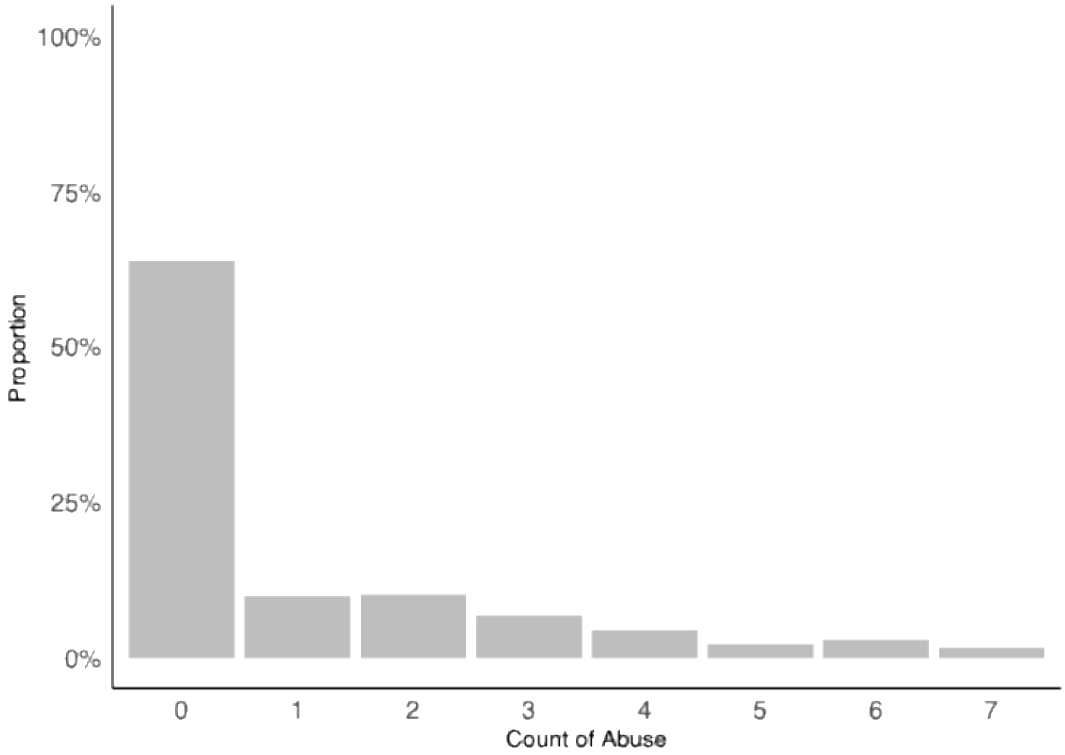
For the mZIP and mZINB, to facilitate their application, we have developed *mzim*, an R package that uses the quasi-Newton BGFS algorithm to perform the estimation. The function generates robust starting values from a standard ZIP or ZINB fit. The package provides both model-based standard errors derived from the Hessian matrix and robust sandwich estimators for variance estimation. In the following session, we will use an empirical example to detail the use of the package.

Empirical Example and R Illustration

To illustrate our modeling approaches, we use data from the study, Crime during the transition to adulthood: how youth fare as they leave out-of-home care in Illinois, Iowa, and Wisconsin, 2002–2007 (Courtney & Cusick, 2010). This longitudinal research's aim was to examine how supports and services provided by these states—under varying child welfare policies—affected youth outcomes such as education, employment, housing stability, and crime. Youth who had been in care for at least one year prior to age 17 and placed due to abuse or neglect were followed over multiple interview waves (e.g., at ages 17–18, 19–20, and 21+). For the present empirical example, we focus on Wave 1 data to examine abuse experiences among 424 participants with complete information. Specifically, our outcome of interest is number of abusive events reported during Wave 1 interviews. Figure 1 shows the distribution of abusive event counts, which contains 270 zeros (64%) and 154 non-zeros (36%).

Figure 1.

Histogram of abusive event counts (N = 424).



This demonstration investigates the association between an individual's gender and school enrolment status and their self-reported number of abuse experiences at Wave 1. The primary predictors are gender (0 = female, 1 = male) and school enrolment, a dichotomous variable where a value of 1 indicates that the youth was enrolled in any form of school (high school, vocational, or college) and 0 indicates otherwise. Table 1 shows the counts of youth with zero or one or more abusive experiences broken down by their gender and school enrollment status.

Table 1.

Distribution of Participants by Abuse Experience, Gender, and School Enrollment Status (N = 424)

		Abuse Counts	
Gender	School Enrollment	Zero Events	One or More Events
Female	Enrolled	70	39
	Not Enrolled	81	38
Male	Enrolled	63	33
	Not Enrolled	56	44
	Total	270	154

Test for overdispersion

Before fitting a zero-inflated model, it is crucial to select the appropriate underlying distribution for the count process. It requires a test for overdispersion, a condition where the variance of the count data is greater than the mean, thus violating a key assumption of the Poisson distribution. The NB distribution serves as an alternative, as it includes an additional parameter to account for this excess variability, providing a better fit in such situations. To statistically evaluate the presence of overdispersion, a likelihood ratio test (LRT) is performed to compare the fit of a standard Poisson model with that of an NB model (Cameron & Trivedi, 2013). For a valid comparison, both models must be specified with the identical set of covariates. The null hypothesis (H_0) of the LRT is that no overdispersion exists, implying the Poisson model is sufficient. The alternative hypothesis (H_1) states that the data are significantly overdispersed,

indicating the NB model offers a better fit. In R, this test can be implemented using the *odTest()* function from the *pscl* package (Zeileis et al., 2008). A statistically significant result ($\chi^2 = 261.10$, p-value < 0.01) from this test with race and school enrolment as the covariates provides the justification for selecting an NB-based model, such as ZINB or mZINB, for the subsequent analysis.

Results

In order to demonstrate the application of the *mzim* package and compare the mZINB model against other common approaches, we analyze the predictors of self-reported abuse experiences using the Wave 1 data. We will compare three distinct methods: MLR, the traditional ZINB model, and the mZINB model.

First, we fit a linear regression model using MLR. Despite the outcome being count data, applied researchers might default to this approach due to its familiarity and the availability of robust standard error corrections (MLR) in statistical softwares (e.g., Mplus), which are often seen as a simple solution for violations of normality and homoscedasticity. The linear model is specified as:

$$Y_i = \delta_0 + \delta_1 Gender_i + \delta_2 School_i + \epsilon_i \quad (11)$$

where Y_i is the abusive events for individual i, $Gender_i$ is a dichotomous variable for the ith individual (1 = male, 0 = female), and $School_i$ is the ith individual's school enrollment status (1 = enrolled, 0 = not enrolled). The results in Table 2 show that being male was significantly ssociated with a decrease of 0.39 abusive events on average compared to females ($\delta_1 = -0.39$, p-value = 0.014). School enrollment was not found to be a statistically significant predictor ($\delta_2 = 0.247$, p-value = 0.256). Interpreting these coefficients presents fundamental

challenges stemming from model misspecification. First, the model imposes an additive structure on the data, which is conceptually misaligned with count outcomes where predictor effects are typically understood to be multiplicative (e.g., as incidence rate ratios). Second, the linear model fails to account for the data generating process inherent to zero-inflation.

Next, we fit the traditional ZINB model shown in Equation (12)

$$logit(\pi_i) = \gamma_0 + \gamma_1 Gender_i + \gamma_2 School_i \quad (12)$$
$$\log(\mu_i) = \beta_0 + \beta_1 Gender_i + \beta_2 School_i$$

The same set of covariates is both included in the count and structural zero process in the ZINB model. Parameter estimation obtained from *zeroinfl()* function from package *pscl* in R is shown in Table 2. First, the ZI process models the likelihood of an individual belonging to the structural zero or never-abused group. In this component, gender was not a significant predictor of group membership (γ_1 = 0.046, *p*-value = .623). However, school enrollment was highly significant (γ_2 = −0.289, *p*-value = .003). The odds ratio of 0.749 (exp(−0.289)) indicates that the odds of belonging to the structural zero group were 25.1% lower for youth enrolled in school compared to those not enrolled. Second, the count process models the expected number of abuse incidents for the subpopulation of youth that already being considered at-risk. Within this group, being male was significantly associated with a 27.4% lower expected abuse count compared to females (IRR = 0.73; β_1 = −0.320, *p*-value = .024). Being enrolled in school was not statistically significantly associated with the expected abuse count (β_2 = −0.306, *p*-value = .094). Despite its statistical appropriateness, the traditional ZINB model presents the exact real-world challenges for applied researchers discussed earlier. The primary research question—What is the overall effect of gender and school enrollment on abuse for all youth?—is not directly answered. Instead, the model provides two separate answers for unobserved subgroups. This difficulty

arises from the model's foundational and non-testable assumption about the nature of the zero counts. The ZINB framework posits that the observed zeros are a mix of two types: structural zeros from youth who are truly not at risk and would never be abused, and sampling zeros from youth who are at risk but happened to not experience an abusive event during the data collection window. However, this categorization is purely theoretical. For any given youth with a zero count, it is impossible to empirically verify which group they belong to. The statistical inference of the model, which produces separate estimates for the at-risk and never-abused groups, depends on the validity of this untestable assumption.

Lastly, we fit the mZINB model using the same set of covariates, with the specification detailed in Equation (12).

$$logit(\pi_i) = \gamma_0 + \gamma_1 Gender_i + \gamma_2 School_i \quad (12)$$
$$\log(\nu_i) = \alpha_0 + \alpha_1 Gender_i + \alpha_2 School_i$$

In contrast to the traditional ZINB which models the conditional mean (μ_i) for the non-zero subgroup, the mZINB directly models the marginal mean ($\nu = E(Y_i)$) for the entire population as a function of gender and school enrollment. The logistic model for the structural zero process remains identical to that of the traditional ZINB. The R code implementing this model with the *mzim()* function in *mzim* package is shown below. Its key arguments, *count_formula* and *zi_formula*, specify the models for the marginal mean and structural zero process, respectively. Additionally, while the package computes robust sandwich estimators by default, model-based standard errors derived from the Hessian matrix can be requested by setting *robust = FALSE* in the summary function, as demonstrated. The resulting parameter estimates are shown in Table 2.

```
model_mzinb <- mzim(count_formula = abuse ~ Gender + School,
```

```
                    zi_formula = ~ Gender + School,
                    data = abuse,
                    family = "zinb")
summary(model_mzinb, robust = FALSE) # model-based S.E.
```

The zero-inflation process in the mZINB, which models the odds of belonging to the structural zero or never-abused group, is interpreted identically to the traditional ZINB. In this component, gender was not a significant predictor of group membership (γ_1 = 0.213, *p*-value = .371), but school enrollment was significant (γ_2 = −0.832, *p*-value = .007). The odds ratio is 0.435 (exp(−0.832)), indicating that the odds of being a structural zero were 43.5% lower for youth in school. In other words, individuals enrolled in school were less likely to report and experience abusive events. The key distinction between this model and the traditional ZINB lies in the marginal mean process. Unlike the traditional ZINB which estimates effects conditional on the at-risk subpopulation, this component models the overall average abuse count for the entire population. For this marginal mean, male participants were significantly associated with a 34.6% lower expected abuse count (or reported fewer abusive events) across the whole sample compared to their female counterparts (IRR = 0.65; α_1= −0.425, *p*-value = .032). School enrollment was not significantly associated with the overall average abuse count (α_2 = 0.228, *p*-value = .375). The mZINB model's results resolve the interpretive challenges posed by the preceding analyses. The advantage is that the marginal mean component provides a direct answer to the research question regarding population-level effects. For instance, the model yields a single incidence rate ratio (IRR = 0.71) for gender, representing its overall effect across the entire sample. This direct inference that was unavailable from the traditional ZINB model. Also, the role of the zero-inflation component is reconceptualized. Rather than defining interpretive latent

classes, it serves as a nuisance parameter(s) that accounts for the excess zeros, allowing the marginal effect to be estimated without bias. The validity of the main finding is therefore no longer dependent on the untestable assumption about the nature of the zeros.

Table 2.

ZINB and mZINB results for the number of self-reported abuse events.

	MLR				ZINB				mZINB			
Variable	Coefficient	Estimate	S.E.	p-value	Coefficient	Estimate	S.E.	p-value	Coefficient	Estimate	S.E.	p-value
								Structural zero process				
Intercept	δ_0	0.976	0.203	<0.001	γ_0	-0.336	0.088	< 0.001	γ_0	0.928	0.286	0.001
Gender	δ_1	-0.392	0.160	0.014	γ_1	0.046	0.095	0.623	γ_1	0.213	0.238	0.371
School Enrollment	δ_2	0.247	0.217	0.256	γ_2	-0.289	0.098	0.003	γ_2	-0.832	0.306	0.007
								Count process				
Intercept					β_0	1.261	0.177	< 0.001	α_0	-0.017	0.248	0.977
Gender					β_1	-0.320	0.142	0.024	α_1	-0.425	0.166	0.032
School enrollment					β_2	-0.306	0.183	0.094	α_2	0.228	0.257	0.375
Overdispersion					ϕ	6.740	4.158		ϕ	6.846	4.243	

Discussion

This paper has served as a tutorial on the formulation, estimation, and implementation of marginalized zero-inflated models. We have presented the underlying statistical theory for both traditional and marginalized versions of the ZIP and ZINB models and demonstrated their application using an empirical example. Another contribution of this work is the introduction of the *mzim* R package, a user-friendly tool to make mZIP and mZINB models readily accessible to applied researchers in psychology, health, and other social sciences. By providing straightforward implementation, we bridge the gap between complex statistical theory and practical data analysis.

Using an empirical example, the advantages of the marginalized approach becomes evident when contrasted with the other models. The linear model with MLR, while simple to implement, was misspecified. It imposed an additive structure on a multiplicative process and failed to account for the two-part data generating mechanism of zero-inflation, making its coefficient difficult to interpret meaningfully. The traditional ZI models presented the core interpretive challenge as well and it is what this paper sought to address. It produces two distinct results for each predictor: one for the odds of belonging to the structural zero group, and another for the outcome count only within the latent at-risk subpopulation. While statistically sound, this fails to directly answer the researcher's question, if it is about the overall effect of a predictor on the entire population. The interpretation is conditional on an assumption about the nature of zeros for each participant. The marginalized model, on the other hand, provides a single, interpretable coefficient for each predictor that represents the population-averaged (marginal) effect. In this framework, the zero-inflation component is reconceptualized not as a tool for

interpreting latent classes, but as a statistical adjustment (i.e., a nuisance parameter) that corrects for the data's distributional properties.

A central goal of this tutorial is to clarify for applied researchers the differences between the two frameworks. A theoretical distinction between the two modeling frameworks lies in their conceptualization of the zero values observed in the data. Traditional zero-inflated models (ZIP and ZINB) are essentially mixture models that differentiate zeros into two latent categories: structural zeros and sampling zeros. Under this framework, structural zeros are viewed as originating from a subpopulation that is immune to the event—individuals who are not at-risk and therefore have a probability of zero for the outcome. Conversely, sampling zeros are considered to arise from the count distribution itself; these represent individuals who are at-risk and capable of experiencing the event but simply did not do so during the specific observation window. Valid inference in traditional models relies on the assumption that this latent structure accurately reflects the true data-generating process.

However, the marginalized framework (mZIP and mZINB) shifts the inferential focus. The logic of using mZI models is not based on the assumption that structural zeros do not exist, but rather that distinguishing them from sampling zeros is unnecessary for marginal inference. The model statistically accounts for the excess zeros through the logistic component (i.e., acknowledging the distributional reality of the zero-inflation), but parameterizes the count component to marginalize over these latent classes. Thus, while the separation of zeros into at-risk and not-at-risk groups may theoretically exist, the marginalized estimator treats this latent structure as a nuisance characteristic rather than a primary target of inference, allowing for a direct interpretation of effects across the entire population.

The choice between the two frameworks is also not merely technical but is dictated by the research question. Traditional ZIP/ZINB models are appropriate when the goal is to understand latent heterogeneity, which is to separately model the predictors of belonging to a not-at-risk group versus the predictors of event frequency among the at-risk subpopulation. In contrast, marginalized mZIP/mZINB models are the superior choice when the research question concerns the population-averaged impact of a predictor across an entire sample, a common scenario in psychological science. Selecting the marginalized model provides a direct answer to this type of question and avoids the common errors that misinterpret the conditional coefficients from a traditional ZINB model as marginal effects (Preisser et al., 2016). Another contribution of this work is bridging the gap between this advanced statistical method and its practical application. The *mzim* R package provides a user-friendly tool to make both mZIP and mZINB models readily accessible, allowing applied researchers to focus on answering their research questions correctly rather than on complex post-estimation calculations.

In summary, this paper seeks to encourage the use of marginalized zero-inflated models. By moving beyond the restrictive latent class interpretations of traditional models, researchers can now address questions about population-level effects with greater clarity and statistical rigor. It is our hope that the theoretical explication and the practical tools provided via the *mzim* package will encourage the broader adoption of these methods, ultimately leading to more accurate inference in psychological and health research.

Declarations

Funding

The authors did not receive support from any organization for the submitted work.

Conflicts of Interests/Competing Interests

The authors have no relevant financial or non-financial interests to disclose.

Ethics Approval

This study involves the secondary analysis of de-identified data from the *Midwest Evaluation of the Adult Functioning of Former Foster Youth* study. As this study utilizes pre-existing, anonymized data, specific ethical approval for this analysis was not required.

Consent to Participate

Not applicable.

Consent for Publication

Not applicable.

Availability of Data and Materials

The code and data used in the demonstration are available at the following OSF repository.

Code availability

The mzim package developed for this tutorial is open-source after review. The full analysis code used for the empirical example in this article is available on the OSF repository.

Open Practices Statement

The analysis code and the software package mzim described in this article are available at. This study was not preregistered.

Reference

Albert, J. M., Wang, W. & Nelson, S. (2014). Estimating overall exposure effects for zero-inflated regression models with application to dental caries. *Statistical Methods in Medical Research*, *23*(3), 257–278. https://doi.org/10.1177/0962280211407800

Agresti, A. (2002). *Categorical data analysis* (2nd ed.). John Wiley & Sons. https://doi.org/10.1002/0471249688

Bandyopadhyay, D., DeSantis, S. M., Korte, J. E. & Brady, K. T. (2011). Some Considerations for Excess Zeroes in Substance Abuse Research. *The American Journal of Drug and Alcohol Abuse*, *37*(5), 376–382. https://doi.org/10.3109/00952990.2011.568080

Cameron, A. C., & Trivedi, P. K. (2013). *Regression Analysis of Count Data* (Vol. 53). Cambridge University Press.

Campbell, H. (2021). The consequences of checking for zero-inflation and overdispersion in the analysis of count data. *Methods in Ecology and Evolution*, *12*(4), 665–680. https://doi.org/10.1111/2041-210x.13559

Chrischilles, E. A., Pendergast, J. F., Kahn, K. L., Wallace, R. B., Moga, D. C., Harrington, D. P., Kiefe, C. I., Weeks, J. C., West, D. W., Zafar, S. Y. & Fletcher, R. H. (2009). Adverse Events Among the Elderly Receiving Chemotherapy for Advanced Non–Small-Cell Lung Cancer. *Journal of Clinical Oncology*, *28*(4), 620–627. https://doi.org/10.1200/jco.2009.23.8485

Courtney, Mark E., and Cusick, Gretchen Ruth. Crime During the Transition to Adulthood: How Youth Fare As They Leave Out-of-Home Care in Illinois, Iowa, and Wisconsin, 2002-2007. Inter-university Consortium for Political and Social Research [distributor], 2010-12-14. https://doi.org/10.3886/ICPSR27062.v1

Greene, W. H. (1994). *Accounting for excess zeros and sample selection in Poisson and negative binomial regression models* (NYU Working Paper No. EC-94-10). New York University. https://ssrn.com/abstract=1293115

He, H., Wang, W., Crits-Christoph, P., Gallop, R., Tang, W., Chen, D.-G. D., & Tu, X. M. (2014). On the implication of structural zeros as independent variables in regression analysis: Applications to alcohol research. *Journal of Data Science*, *12*(3), 439–460. https://doi.org/10.1016/j.data.2014.07.003

Hu, M.-C., Pavlicova, M. & Nunes, E. V. (2011). Zero-Inflated and Hurdle Models of Count Data with Extra Zeros: Examples from an HIV-Risk Reduction Intervention Trial. *The American Journal of Drug and Alcohol Abuse*, *37*(5), 367–375. https://doi.org/10.3109/00952990.2011.597280

Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics, 34*(1), 1-14. https://doi.org/10.1080/00401706.1992.10485228

Loeys, T., Moerkerke, B., De Smet, O., & Buysse, A. (2012). The analysis of zero-inflated count data: Beyond zero-inflated Poisson regression. *British Journal of Mathematical and Statistical Psychology, 65*(2), 163–180. https://doi.org/10.1111/j.2044-8317.2011.02031.x

Long, D. L., Preisser, J. S., Herring, A. H. & Golin, C. E. (2014). A marginalized zero-inflated Poisson regression model with overall exposure effects. *Statistics in Medicine*, *33*(29), 5151–5165. https://doi.org/10.1002/sim.6293

McGillycuddy, M., Warton, D. I., Popovic, G., & Bolker, B. M. (2025). Parsimoniously fitting large multivariate random effects in glmmTMB. *Journal of Statistical Software*, *112*(1), 1–19. https://doi.org/10.18637/jss.v112.i01

Muthén, L. K., & Muthén, B. O. (2017). Mplus: Statistical Analysis with Latent Variables: User's Guide (Version 8). Los Angeles, CA: Authors.

R Core Team. (2024). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

SAS Institute Inc. (2022). *SAS/STAT® 16.1 user's guide*. SAS Institute Inc.

StataCorp. (2023). *Stata 18 base reference manual*. Stata Press.

Swartout, K. M., Thompson, M. P., Koss, M. P. & Su, N. (2015). What Is the Best Way to Analyze Less Frequent Forms of Violence? The Case of Sexual Aggression. *Psychology of Violence*, *5*(3), 305–313. https://doi.org/10.1037/a0038316

Pittman, B., Buta, E., Garrison, K. & Gueorguieva, R. (2022). Models for Zero-Inflated and Overdispersed Correlated Count Data: An Application to Cigarette Use. *Nicotine and Tobacco Research*, *25*(5), 996–1003. https://doi.org/10.1093/ntr/ntac253

Preisser, J. S., Das, K., Long, D. L., & Divaris, K. (2016). Marginalized zero-inflated negative binomial regression with application to dental caries. *Statistics in medicine*, *35*(10), 1722–1735. https://doi.org/10.1002/sim.6804

Preisser, J. S., Stamm, J. W., Long, D. L., & Kincade, M. E. (2012). Review and recommendations for zero-inflated count regression modeling of dental caries indices in epidemiological studies. *Caries Research*, *46*(4), 413–423. https://doi.org/10.1159/000339893

Zeileis, A., Kleiber, C., & Jackman, S. (2008). Regression models for count data in R. *Journal of Statistical Software*, *27*(8), 1–25. https://doi.org/10.18637/jss.v027.i08

Zorc, C. S., O'Reilly, A. L. R., Matone, M., Long, J., Watts, C. L. & Rubin, D. (2013). The relationship of placement experience to school absenteeism and changing schools in young,

school-aged children in foster care. *Children and Youth Services Review*, *35*(5), 826–833. https://doi.org/10.1016/j.childyouth.2013.02.006