625.721 Zero-Inflation in Count Data:

Methodological Insights from Poisson-Based Regression Models

(Conditioning, Regression, and Models for Count Data)

Jeremy Hirschler

August 2025

1 Introduction

One of the fields that commonly sees count data is healthcare and hospitality. The number of missed appointments at a doctor's clinic, the number of nights a hotel guest stays within a given month, and the number of emergency room (ER) visits by patients in a year are all real-world examples of count data that are useful for people who work these jobs. However, in all these cases, it's also common to observe large amounts of zeros in the data; many people may not stay at a hotel for a given season, some individuals never miss appointments, and many patients do not visit the ER at all given a particular year.

In these cases, these excessive zeros go beyond what a common count distribution, such as the Poisson or negative binomial distributions, are able to accommodate. This type of data is called zero-inflated data. There are several ways in which zeros may occur. Looking at the example for ER visits, some zero counts can occur because individuals are generally healthy and don't require a trip to the ER. These are considered structural zeros, in which a count response for that individual is certain to be zero for that particular time period. By contrast, we may also see sampling zeros, which are typically due to random variation in the population, since it's also possible for these counts to be nonzero. Sampling zeros arise when individuals who are at risk for an ER visit did not require one over a particular timeframe, but it's also possible for these counts to be nonzero.

In a Poisson model, it's assumed that the mean and variance are equal, that is $E(Y) = Var(Y)$. With zero-inflated data however, the data is usually overdispersed due to the extra zeros, and therefore $Var(Y) > E(Y)$ in most cases. The negative binomial distribution is better able to handle overdispersion from the data, though cannot distinguish between the structural and sampling zeros since it is assumed that there is only a single underlying process generating the data. Trying to use these models for zero-inflated count data can lead to unreasonable model fitting. To address the challenges posed by these excess zeros in the count data, analysts often look towards two types of models: A Zero-Inflated Poisson (ZIP) model (Lambert 1992) [1], and a model put forth

called a hurdle model (Mullahy 1986) [2] in order to handle zero-inflated data when regular count models are unrealistic. Both models extend the Poisson model, but differ slightly in how they model the generation of the zeros and positive counts.

The ZIP model assumes that zeros come from a mixture of two processes: one that generates only zeros (a structural zero process) and another that generates counts (including zeros) via a Poisson distribution. The hurdle model, on the other hand, separates the modeling of zeros and positive counts into two distinct parts: a binary model that determines whether the count is zero or positive, and a truncated count model (such as Poisson or negative binomial) that governs the distribution of positive counts [3].

This report aims to compare the Zero-Inflated Poisson and hurdle models, both in theory and through practical examples. We will examine their structure, assumptions, and interpretation, and apply them to data scenarios typical of healthcare and hospitality. The goal is to provide guidance on when and how to use each model effectively when analyzing count data with excess zeros.

2 Background

In modeling count data, we often begin with a Poisson regression model, which assumes that the random variable Y follows a Poisson distribution:

$$P(Y = k) = \frac{\lambda^k e^{-\lambda}}{k!}, \quad k = 0, 1, 2, ...$$

Where,

k = number of occurrences

λ = mean number of events within a given time interval

To note, λ also equals the mean/expected value as well as the variance of the distribution such that $\lambda = E(X) = Var(X)$. The Poisson regression model then assumes that the response variable Y has a Poisson distribution, and that the logarithm of its expected value can be modeled by a linear combination of unknown parameters [4]. That is:

$$Y_i \sim Po(\lambda_i)$$

$$log(\lambda_i) = x_i^T \boldsymbol{\beta}$$

Where,

Y_i = count for observation i

λ_i = expected value

x_i = vector of predictor variables

$\boldsymbol{\beta}$ = vector of regression coefficients

Poisson regression is a type of generalized linear model (GLM) which uses a log link function. A GLM is a more general form of an ordinary linear regression model, and the log link function provides a relationship between the linear predictor and the mean of the distribution function, in this case by taking the logarithmic of the mean of the response variable.

While widely used, the Poisson model makes a strong assumption that the mean equals the variance, which is often not the case with real-world data. In many fields, especially healthcare and hospitality, count data often exhibits overdispersion, in which case the variance exceeds the mean. A major cause of this overdispersion is zero inflation, in which the number of observed zeros exceeds what the typical Poisson model would predict. To address this overdispersion, the negative binomial model is often used, which introduces an additional dispersion parameter. This negative binomial model is useful when overdispersion arises from unobserved heterogeneity in the population such as when different subgroups have inherently different rates of event occurrences. However, this model also cannot explicitly account for structural zeros, in which an observation of an event is impossible rather than just being unlikely to occur. As a result, this model may still be a poor fit to the data when overdispersion is driven primarily by excess zeros.

To better handle zero-inflated data, two specialized models are commonly used: the Zero-Inflated Poisson (ZIP) model, and the Hurdle model. Both models explicitly separate the generation of zeros from the generation of positive counts, although they differ slightly in their structure and interpretation.

The ZIP model assumes that the data comes from a mixture of two processes: one that always yield a zero with probably π, and another that follows the standard Poisson distribution with probability $1 - \pi$. That is:

$$P(Y = y_i) = \begin{cases} \pi_i + (1 - \pi_i)e^{-\lambda_i}, & y_i = 0 \\ (1 - \pi_i)\frac{\lambda^{y_i} e^{-\lambda_i}}{y_i!}, & y_i > 0 \end{cases}$$

where the outcome variable y_i represents any non-negative integer value, λ_i is the mean of the standard Poisson distribution, and π_i is the probability of extra zeros [5]. And using z_i as a Bernoulli random variable where

$$z_i = \begin{cases} 1, & y_i \text{ is a structural zero} \\ 0, & y_i \sim Poisson(\lambda_i) \end{cases}$$

From here we can see the mean and variance of this model becomes

$$E(y_i) = E(E(y_i|z_i)) = (1 - \pi_i)\lambda_i$$

$$Var(y_i) = E(Var(y_i|z_i)) + Var(E(y_i|z_i)) = (1 - \pi_i)\lambda_i(1 + \lambda_i\pi_i)$$

In this framework, zeros can be generated from either process, while positive counts arise only from the Poisson part of the model. This allows the model

to account for both structural zeros from the zero-inflation process as well as sampling zeros from the Poisson process.

By contrast, the hurdle model treats the occurrence of zero as a separate binary process from the count outcome. It consists of two parts: a binary model, typically as logistic regression, that determines whether the count is zero or positive, and a truncated count model like a Poisson or Negative Binomial model, that models the distribution of positive counts only. The probability mass function (PMF) of the Hurdle model is such that:

$$P(Y_i = y_i) = \begin{cases} p_i, & y_i = 0 \\ (1 - p_i) \cdot \frac{f_p(y_i)}{1 - f_p(0)}, & y_i > 0 \end{cases}$$

Where p_i is the probability that the outcome is zero (coming from the binary process such as a logistic regression model - which satisfies the conditions $0 < p_i < 1$ and $p_i(z_i) = logit(p_i) = log(\frac{p_i}{1-p_i})$), and $f_p(y_i)$ is the PMF of a count distribution. In this case we will stick to the Poisson distribution, and so

$$f_p(y_i) = \frac{\lambda^{y_i} e^{-\lambda}}{y_i!}$$

And the denominator

$$1 - f_p(0) = 1 - e^{-\lambda}$$

Ensures that the positive counts are from the zero-truncated Poisson distribution (that is, it is impossible for the count to be zero since we are rescaling the Poisson PMF to sum to 1 over the positive counts only.) Putting it together, we have the formulation for the Hurdle Poisson model:

$$P(Y_i = y_i) = \begin{cases} p_i, & y_i = 0 \\ (1 - p_i) \cdot \frac{\lambda^{y_i} e^{-\lambda} / y_i!}{1 - e^{-\lambda}}, & y_i > 0 \end{cases}$$

This model assumes that all zeros come from the binary process, and all positive counts are governed by a separate process, which creates a conceptual "hurdle" that observations must cross to register as non-zero [6]. Similar to the ZIP model, we can find the mean and variance of the Poisson Hurdle model as

$$E(Y_i) = \frac{(1 - p_i)\lambda_i}{1 - e^{\lambda_i}}$$

$$(1 - p_i) \cdot \frac{\lambda_i}{1 - e^{-\lambda_i}} \cdot (1 - \frac{\lambda_i e^{-\lambda_i}}{1 - e^{-\lambda_i}}) + (\frac{\lambda_i}{1 - e^{-\lambda_i}})^2 \cdot p_i(1 - p_i)$$

Both the ZIP and Hurdle models provide flexible tools for modeling zero-inflated data. The choice between them depends on the theoretical understanding of the data-generating process and empirical model performance. The next sections will explore these models in greater detail and evaluate their performance through an application of practical examples.

3 Methods and Model Selection

In this section, we describe the methodological framework used to fit and compare the ZIP and Hurdle Poisson models for zero-inflated count data. We rely on a combination of model fitting techniques and diagnostic criteria to evaluate how well each model explains the observed data. Both the ZIP and Hurdle models are estimated using Maximum Likelihood Estimation (MLE).

3.1 Model Selection

To assess model fit selection, we first consider the log-likelihood value, which reflects how probable the observed data are under the model. We aim to maximize this value, and so a higher log-likelihood value indicates a better fit. However, the log-likelihood value on its own does not penalize for model complexity, which is where we turn to other criteria. The Akaike Information Criterion (AIC) is used for comparing model fits and is defined as

$$AIC = -2logL + 2k$$

Where L represents the maximized value of the log-likelihood function and k is the number of estimated parameters. The model with the minimum AIC value is generally preferred. It is easy to see that AIC penalizes an increasing number of parameters, which helps discourage overfitting the models, which is desired since increasing the number or parameters typically improves goodness of fit.

Similarly, the Bayesian Information Criterion (BIC) provides another measure that will penalize model complexity more heavily, particularly in larger datasets [7]. It is defined as

$$BIC = -2logL + k \cdot log(n)$$

Where n is the sample size. Like AIC, BIC favors models with higher log-likelihoods and fewer parameters, but since its penalty term grows with sample size, BIC is more conservative. Again, we generally prefer the model with the lowest BIC value.

Since both the ZIP and Hurdle models are non-nested, we can also apply the Vuong closeness test to compare the ZIP and Hurdle models. This test compares the log-likelihood functions between the two models, specifically testing the null hypothesis that both models fit the data equally well. If we define

$$M_1 : \text{ZIP Model}$$

$$M_2 : \text{Hurdle Model}$$

And let l_{1i} and l_{2i} be the log-likelihood contributions from observation i under each model. We can define a pointwise log-likelihood ratio:

$$d_i = l_{1i} - l_{2i} = logf_1(y_i) - logf_2(y_i)$$

From here we can compute

$$\text{The average:} \quad \overline{d} = \frac{1}{n}\sum_{i=1}^{n} d_i$$

$$\text{The variance:} \quad s_n^2 = \frac{1}{n}\sum_{i=1}^{n}(d_i - \overline{d})^2$$

Then the Vuong test statistic is:

$$V = \frac{\sqrt{n}\overline{d}}{s_n}$$

Under the null hypothesis, this statistic asymptotically follows a standard normal distribution [8]. That is $V \sim N(0,1)$. Then we can see that if

$V > 1.96,$ prefer model 1 (ZIP)

$V < -1.96,$ prefer mode 2 (Hurdle)

$|V| \leq 1.96,$ no significant difference

Important to note a common pitfall - when trying to use this statistic for model selection when comparing a zero-inflated model to its non-zero-inflated counterpart. This comes from the fact that this test works only for nested, strictly non-nested, or partially non-nested models, yet a zero-inflated Poisson model and its non-zero-inflated counterpart are a pair of non-nested models falling outside of this scope. Thus, Vuong's test is not a valid test for discriminating between them.

3.2 Goodness-of-Fit Criteria

Beyond likelihood-based measures, we also consider the predictive accuracy of the regression models through metrics such as root mean squared error (RMSE) and mean absolute error (MAE) [9].

Finally, we can consider how well each model predicts the number of zero observations by comparing the number of observed zeros to the number of zeroes predicted under each model. Together, these methods provide a multi-faceted picture of model performance, allowing us to determine which model fits best statistically.

4 Application Study

4.1 Data Generation

Although zero-inflated count data is commonplace among certain industries, finding a neat dataset to use in the application study proved difficult. Therefore, synthetic data was created in order to study the differences between the ZIP

and Hurdle Poisson models for count data. The data was created in R, and simulates 1,000 observations of number of times certain individuals visit the doctor annually.

The simulated dataset was generated using a zero-inflated Poisson (ZIP) model with two components: a count model and a zero-inflation model. Each component uses a set of covariates—age, gender, and smoking level—to influence either the expected number of doctor visits or the probability of being a structural zero. The covariates were simulated as:

$$\text{age} = \text{discrete values between 18-90}$$

$$\text{gender} = \text{binary variable with 2 levels as } \begin{cases} 0, & \text{female} \\ 1, & \text{male} \end{cases}$$

$$\text{smoking} = \text{ordinal variable with 3 levels as } \begin{cases} 1, & \text{non-smoker} \\ 2, & \text{occasional smoker} \\ 3, & \text{frequent smoker} \end{cases}$$

For the count model, the `beta_count` vector governs the Poisson count process and defines how covariates influence the expected number of doctor visits per year for individuals who are not structural zeros. The model uses a log-link, so the coefficients act on the logarithm of the expected count, that follows the structure:

$$log(\lambda_i) = \beta_0 + \beta_1 \cdot \text{age}_i + \beta_2 \cdot \text{gender}_i + \beta_3 \cdot \text{smoking}_{\text{casual},i} + \beta_4 \cdot \text{smoking}_{\text{often},i}$$

And then we can see that

$$\lambda_i = exp(\text{predictor}) = \text{expected doctor visits}$$

In our simulation, the intercept was set to -1.2, representing a low baseline visit rate. Age has a small positive effect ($+0.02$), meaning older individuals are expected to visit the doctor slightly more often. Gender has a negative coefficient (-0.1), indicating that males are expected to have slightly fewer visits than females. And smoking level has a modest positive effect ($+0.1$) and ($+0.2$) for casual smokers and heavy smokers, respectively, implying that heavier smokers tend to have slightly more visits. Note that `smoking` with 3 levels is encoded as 2 dummy variables for these regression models.

The `beta_zero` vector controls the structural zero process via a logistic (logit) link, as

$$\text{logit}(p_{zero,\ i}) = \gamma_0 + \gamma_1 \cdot \text{age}_i + \gamma_2 \cdot \text{gender}_i + \gamma_3 \cdot \text{smoking}_{\text{casual},i} + \gamma_4 \cdot \text{smoking}_{\text{often},i}$$

And then

$$p_{zero,\ i} = \frac{1}{1 + exp(-(\text{predictor}))}$$

These coefficients determine the probability that an individual is a structural zero, that is, someone who will never visit the doctor under any circumstance

[10]. The intercept of -1.5 implies a generally low baseline probability of being a structural zero. Age has a slight negative effect (-0.02), meaning older individuals are less likely to be structural zeros. Gender has a positive effect ($+0.6$), suggesting that males are more likely to avoid visiting the doctor entirely. And again smoking has a negative effect (-0.2) and (-0.3) for casual and heavy smokers, respectively, indicating that heavier smokers are less likely to be structural zeros, likely due to a higher underlying need for medical attention. Together, these two sets of coefficients define both the intensity and zero-inflation structure of the simulated doctor visit counts, reflecting plausible real-world behavioral and health patterns. The summary statistics of the regression model was generated below.

	age	gender	smoking	count
Min. :	18.00	0.000	nonsmoker:321	0.000
1st Qu.:	35.75	0.000	sometimes:326	0.000
Median :	53.00	0.000	often :353	1.000
Mean :	53.68	0.493		0.946
3rd Qu.:	71.00	1.000		1.000
Max. :	90.00	1.000		8.000

The observed distribution of doctor visit counts per individual is displayed in Figure 1 below as both a frequency table and histogram. The frequency table provides exact counts for each value, reinforcing the fact that a large proportion of individuals had zero or one visit, while higher visit counts are increasingly rare. The histogram to the right visually highlights the zero-inflated nature of the data, with a pronounced peak at zero and a steep decline as counts increase.

Count	**Frequency**
0	433
1	327
2	148
3	56
4	29
5	5
6	1
8	1

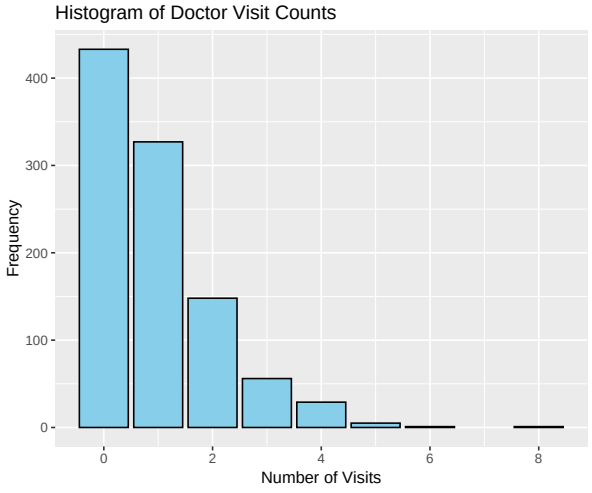

Figure 1: Observed distribution of doctor visit counts.

This distribution motivates the use of models like the Zero-Inflated Poisson (ZIP) and Hurdle models, which explicitly account for excess zeros in count data.

4.2 Data Modeling and Results

Now we can use this data to look at how the ZIP and Hurdle models perform. We used this data to run a regression model using regular Poisson regression, a ZIP model, and a Hurdle Poisson model. Again R was used to generate these 3 models, noting that the ZIP and Hurdle models have a count model and zero model. The summary of coefficients for the intercept and each variable is shown below.

Model	**Intercept**	**Age**	**Gender (Male)**	**Smoking (Casual)**	**Smoking (Often)**
Poisson Count	-1.375	0.023	-0.164	0.063	0.066
ZIP Count	-1.248	0.021	-0.186	0.070	0.111
ZIP Zero	0.997	-0.409	-4.160	6.736	10.909
Hurdle Count	-1.374	0.021	-0.119	0.273	0.228
Hurdle Zero	-1.402	0.036	-0.260	-0.135	-0.128

Table 1: Regression coefficients for each model. Poisson includes only a count component. ZIP and Hurdle models include both count and zero components.

Note that we see much larger coefficient values for the ZIP Zero model component. This logically makes sense, since the logistic model is bounded between 0 and 1 and may need to produce strong logits to push the predicted probability near these values.

In the Poisson model, we observe that age has a small positive effect, suggesting that each additional year of age increases the expected count by approximately 2.3%. Male gender is associated with a decrease in counts ($\hat{\beta}_2 = -0.164$), and both smoking levels show mild positive effects on count frequency. However, this model assumes a single process for both zero and non-zero outcomes, which may be too restrictive for data with zero inflation.

The ZIP model distinguishes between two processes: a logistic model for structural zeros (ZIP Zero) and a Poisson model for counts (ZIP Count). In the ZIP Zero process, age has a strong negative association ($\hat{\gamma}_1 = -0.409$), indicating that older individuals are less likely to be in the always-zero group. Furthermore, male gender has an extremely large negative effect ($\hat{\gamma}_2 = -4.160$), dramatically reducing the odds of structural zeros. Conversely, smoking at either the casual or often level strongly increases the likelihood of being in the structural zero group ($\hat{\gamma}_3 = 6.736, \quad \hat{\gamma}_4 = 10.909$), respectively, suggesting that heavy smokers may be systematically absent from positive counts, possibly due to selection effects, behavioral avoidance, or other model assumptions as in the heterogeneity in patient smoking behavior.

The Hurdle model differs in that all zero counts arise from the binary component, and all positive counts are modeled using a truncated-at-zero Poisson. The count component (Hurdle Count) shows a notably higher sensitivity to smoking: casual ($\hat{\beta}_3 = 0.273$) and often smoking ($\hat{\beta}_4 = 0.228$). Both have much stronger effects than in the Poisson or ZIP models, reflecting that, conditional on having a nonzero count, smokers tend to have considerably higher expected counts. In the Hurdle Zero component, age has a positive coefficient ($\hat{\gamma}_1 = 0.036$), indicating a slight increase in the probability of a zero count with age. Smoking has a negative effect on the odds of a zero count ($\hat{\gamma}_4 = -0.128$ for often smokers), implying that smokers are more likely to cross the hurdle and produce a positive count, contrasting with the ZIP model where smoking predicts structural zeros.

Taken together, the models reveal differing assumptions and interpretations of the zero-generation mechanism. The ZIP model attributes excess zeros to a separate latent class, whereas the Hurdle model treats zeros as behavioral, arising from a distinct participation decision. The ZIP zero model suggests heavy smokers are disproportionately in the always-zero group, while the Hurdle model suggests they are more likely to produce positive counts. These diverging implications underscore the importance of understanding the data-generating process and selecting a model that reflects the underlying scientific or behavioral mechanisms.

The following table summarizes the model metrics for Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), log-likelihood, Mean Absolute Error (MAE) and Root Square Mean Error (RMSE).

Model	**AIC**	**BIC**	**Log-likelihood**	**MAE**	**RMSE**
Poisson	2462.5	2487.1	-1226.3	0.781	1.001
ZIP	2454.9	2504.0	-1217.4	0.771	0.998
Hurdle	2467.6	2516.6	-1223.8	0.781	1.001

Table 2: Model comparison metrics. ZIP and Hurdle models more accurately reflect zero inflation in the data.

Lower values for the AIC and BIC score, MAE, and RMSE, and higher values for log-likelihood indicate better fits of the model to the observed data. Among the three models, the Zero-Inflated Poisson (ZIP) model demonstrates the best overall performance, achieving the lowest AIC score (2454.9) and the highest log-likelihood (-1217.4), as well as the lowest mean absolute error (0.771). These results suggest that the ZIP model provides the best balance between goodness-of-fit and model complexity.

While the Hurdle model also accounts for zero inflation, it yielded a higher AIC (2467.6) than even the regular Poisson model, and a lower log-likelihood than the ZIP model (-1223.8), but slightly higher than the regular Poisson model, indicating a slightly inferior fit for this dataset. This suggests that the ZIP model yielded modest but consistent improvements in prediction over the standard Poisson model.

In addition, we also the BIC score, which applies a stronger penalty for model complexity by incorporating sample size. While the ZIP model achieved the lowest AIC and highest log-likelihood, its BIC score (2504.0) was higher than that of the standard Poisson model (2487.1), reflecting its greater complexity. These results highlight an important trade-off: although the ZIP model fits the data best, the BIC indicates that the simpler Poisson model may be more favorable under stricter complexity penalties, especially in smaller datasets.

The Vuong test comparing the ZIP to Hurdle model produced a test statistic of $z = 2.85$ with a p-value of 0.057. This indicates that the ZIP model provides a slightly better fit to the data than the Hurdle model, which is consistent with the previous results.

The following table details the predict percentage of zeros in the data, noting that the number of observed zeros in the original dataset is $\sim 43.3\%$.

Model	**Predicted % Zeros**
Poisson	42.6%
ZIP	42.9%
Hurdle	3.9%

Table 3: Comparison of predicted zero counts from each model.

Once again the ZIP model has the closest predicted zero percentage to the observed zero rate. The standard Poisson model slightly underfit the data, and the Hurdle model severely underpredicted the number of zeros, suggesting that its assumption, in which all zeros arise from the separate binary process, does not accurately reflect the data's underlying structure in this case.

Finally, Figure 2 also visualizes the alignment between observed count levels and the corresponding average predicted values across the three models. Each bar represents the mean predicted count for individuals within each observed count group.

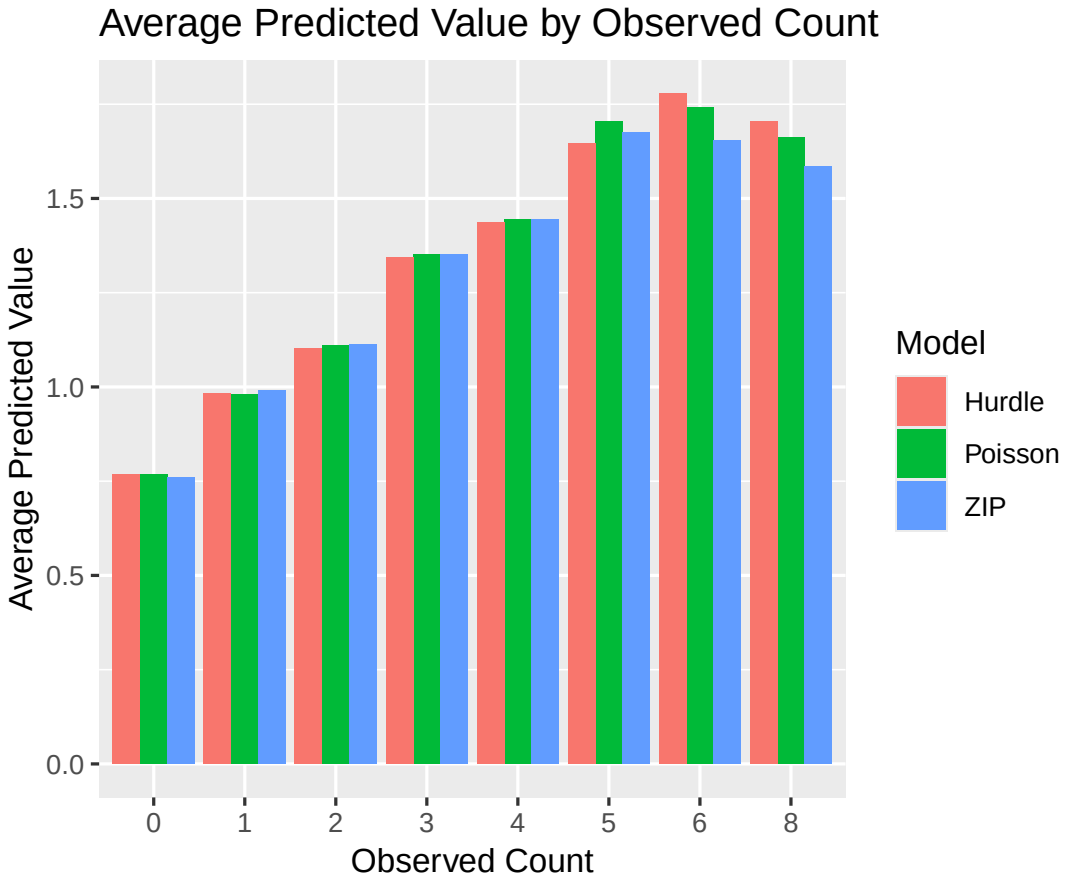

Figure 2: Histogram of simulated zero-inflated count data

The trend confirms that all three models successfully capture the overall monotonic relationship: as the observed count increases, the average predicted value also rises. However, the rate of increase is sublinear, especially for higher counts, where the models tend to systematically underestimate the results. For example, individuals with 4 or more observed visits are often assigned average predictions closer to 1.5, highlighting the conservative nature of Poisson-based models when fitting infrequent extreme values. This underestimation at the tails is a well-known property of GLMs, which shrink extreme predictions toward the center due to their exponential link structure and the goal of minimizing overall error. The limited number of high-count observations also contributes to greater prediction shrinkage, since the model receives less information about those regions during training. Additionally, even for observed zeros, predicted values remained around 0.7-0.8, reflecting the non-zero mean behavior of Poisson-based distributions.

5 Discussion

This analysis demonstrates the comparative performance of three count data models for the standard Poisson regression, Zero-Inflated Poisson (ZIP), and Hurdle models, as applied to a synthetic dataset with known zero inflation. While all models aim to capture the distributional structure of count outcomes, they differ in their assumptions and flexibility regarding excess zeros and count generation mechanisms.

The results confirm that the standard Poisson model is inadequate when the data exhibits substantial zero inflation. In our case, the Poisson model predicted a lower proportion of zeros (42.6%) compared to the observed rate (43.3%), and its relatively high AIC and RMSE values suggest both underfitting in the zero range and potential overfitting of the positive counts. This, then, is a consequence of the model's assumption that all of the outcomes arise from a single Poisson process, which forces the model to compromise between fitting the frequent zero values and the spread of positive counts. In practice, this structural rigidity can lead to biased inferences, particularly when zeros are generated by different latent processes.

The Hurdle model, which assumes all zeros come from a separate binary mechanism, and where all positive counts from a truncated count distribution, showed surprising limitations in this study. It significantly underpredicted the number of zero counts (3.9%), and yielded a higher AIC than the Poisson model. This poor performance suggests that the hurdle model's underlying assumptions were misaligned with the structure of our synthetic data. When zero inflation arises not from a distinct data-generating mechanism, but rather as an excess from a single process with occasional structural zeros, the hurdle model may perform poorly by misallocating probability mass to positive counts.

While it would have been straightforward to construct a synthetic dataset that aligns well with the assumptions of the Hurdle model (such as generating zeros exclusively from a separate binary process), the data was intentionally designed to more closely mimic real-world scenarios, where zeros may arise from both structural and sampling processes. This choice was made to evaluate how well each model generalizes to more complex, mixed-generation mechanisms often encountered in applied settings, rather than optimizing for a specific model's assumptions.

In contrast, the ZIP model provided the best overall performance across all evaluated criteria. Its flexible two-part structure allowed it to capture the excess zeros, while still somewhat accurately modeling the positive count distribution. This is reflected in its lowest AIC (2454.9), lowest RMSE (0.998), and closest alignment to the observed zero proportion (42.9%). The Vuong test further supported the ZIP model as the preferred choice, with a test statistic of $z = 2.85$ suggesting a better fit than the Hurdle model, though the p-value of 0.057 is just above typical significance levels.

While these results are encouraging, they should be interpreted with caution. The use of synthetic data, while valuable for controlled experimentation, may not capture the full complexity and noise of real-world phenomena. For ex-

ample, the structure of the zero-generating process was explicitly defined, and model covariates were chosen with known relationships. In applied contexts, the latent processes driving zero inflation may be more nuanced or confounded with unobserved variables. Therefore, while the ZIP model performed best in this setting, it may not generalize as effectively to real-world data with different forms of sparsity or overdispersion.

Moreover, these models are sensitive to choices made during development. Model fit can be influenced by the selection of covariates in the count versus zero model components, interaction terms, or whether overdispersion is appropriately modeled (e.g., using Negative Binomial variants). In our case, we assumed a Poisson base for both the ZIP and Hurdle models. Future work could explore the impact of overdispersion by extending to zero-inflated or hurdle negative binomial models. Additionally, the metrics used (AIC, BIC, MAE, RMSE) each emphasize different trade-offs between bias, variance, and complexity, and interpretation should account for these distinctions.

Practically, the findings underscore the importance of model choice when working with sparse count data. The ZIP model's ability to closely approximate both the frequency of zeros and the distribution of positive counts suggests it may be a robust default when excess zeros are suspected, but are not clearly attributable to a distinct structural process. However, model diagnostics, predictive accuracy, and domain knowledge should always guide the final model selection.

6 Conclusion

This study examined and compared the performance of three statistical models for count data-the standard Poisson regression, Zero-Inflated Poisson (ZIP), and Hurdle Poisson models, using a simulated dataset that mimics real-world healthcare scenarios with excess zeros. The aim was to evaluate how well each model accommodates zero inflation, interprets covariate effects, and fits observed data in a realistic setting.

The results demonstrate that the ZIP model consistently outperformed both the Poisson and Hurdle models across multiple metrics, including AIC, log-likelihood, MAE, RMSE, and predicted proportion of zeros. The ZIP model's strength lies in its ability to model zeros from two distinct processes: one that generates structural zeros and another that governs the count data. This flexibility allows the model to better represent heterogeneous patterns in the data, particularly when zeros arise from a mix of behavioral, structural, and stochastic mechanisms. Even though the BIC slightly favored the simpler Poisson model due to its parsimony, the ZIP model achieved superior predictive accuracy and goodness-of-fit, suggesting it is more appropriate for overdispersed data with structural zeros.

By contrast, the Hurdle model, while theoretically appealing, underperformed in this application. It significantly underestimated the number of zero outcomes and produced a worse AIC than the Poisson model. This indicates

a poor match between the model's assumptions and the data-generating mechanism. Specifically, the hurdle model assumes that all zeros are generated by a separate binary process, which may not reflect more nuanced zero-generation behavior observed in the data. These findings suggest that hurdle models may be ill-suited in contexts where zeros do not arise solely from participation-type processes but may instead result from latent heterogeneity or probabilistic behavior within the population.

The Poisson model, though simple, somewhat lacked the flexibility to account for overdispersion and structural zeros, leading to underestimation of zero counts and poorer fit, when compared to the other models. While the Poisson model performed reasonably well and even produced a close approximation of the observed zero rate, its assumption of a single underlying process limited its flexibility. Compared to the ZIP model, which explicitly accounts for structural zeros, the Poisson model showed slightly inferior fit and predictive accuracy. Nonetheless, its simplicity and competitive BIC suggest it may still be a viable baseline in settings where model parsimony is prioritized.

Overall, this analysis highlights the importance of matching model structure to the data-generating process. In cases where excess zeros are anticipated, especially when driven by unobserved heterogeneity or behavioral avoidance, the ZIP model provides a valuable framework. However, model choice should ultimately be guided by domain knowledge, diagnostic metrics, and the specific nature of the zeros in the dataset. Future work may expand on these findings by incorporating real-world data and exploring extensions such as zero-inflated negative binomial models, hierarchical frameworks, or non-parametric approaches to improve robustness and generalization across diverse applications.

References

[1] D. Lambert *"Zero-inflated Poisson Regression, with an Application to Defects in Manufacturing*, *Technometrics*, vol. 34, no. 1, pp. 1-14, Feb. 1992. Available: `https://doi.org/10.2307/1269547`

[2] J. Mullahy *"Specification and Testing of Some Modified Count Data Models,"*, *Journal of Econometrics*, vol. 33, issue 3, pp. 341-365, Dec. 1986. Available: `https://doi.org/10.1016/0304-4076(86)90002-3`

[3] D.S. Young, E.S. Roemmele, and P. Yeh *"Zero-inflated modeling part I: Traditional zero-inflated count regression models, their applications, and computational tools,"*, *WIREs Computational Statistics*, vol. 14, no. 1, Art. no. e1541, Jan. 2022. Available: `https://doi.org/10.1002/wics.1541`

[4] Y. Liu, W. Li, X. Zhang *"A marginalized zero-truncated Poisson regression model and its model averaging prediction,"*, *Communications in Mathematics and Statistics*, vol. 13, pp. 527-570, 2025. Available: `https://doi.org/10.1007/s40304-022-00312-8`

[5] C. X. Feng *"A comparison of zero-inflated and hurdle models for modeling zero-inflated count data,"*, *Journal of Statistical Distributions and Applications*, vol. 8, Art. no. 8, Jun. 2021. Available: https://doi.org/10.1186/s40488-021-00121-4

[6] M. Asghar, S. Ali, I. Shah *"Poisson hurdle model for monitoring the inflation of zeros,"*, *Quality and Reliability Engineering International*, vol. 39, issue 6, pp. 2152-2161, Oct 2023. Available: https://doi.org/10.1002/qre.3310

[7] S. Vrieze *"Model Selection and Psychological Theory: A Discussion of the Differences Between the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC),"* *Psychological Methods*, vol. 17, issue 2, pp. 228-243, Jun 2012. Available: https://doi.org/10.1037/a0027127

[8] B. Desmarais, J. Harden *"Testing for zero inflation in count models: Bias correction for the Vuong test,"*, *The Stata Journal*, vol 13, issue 4, pp. 810-835, Dec 2013. Available: https://doi.org/10.1177/1536867X1301300408

[9] H. Agrawal, P. Jain, A. Joshi *"Machine learning models for non-invasive glucose measurement: towards diabetes management in smart healthcare,"*, *Health and Technology*, vol. 12, pp. 955-970, Aug 2022. Available: https://doi.org/10.1007/s12553-022-00690-7

[10] Y. Wang, Y. Han *"Score Tests for Overdispersion in Marginalized Zero-Inflated Poisson Regression Based on Marginalized Zero-Inflated Generalized Poisson Model,"*, *Statistical Analysis and Data Mining: The ASA Data Science Journal*, vol. 18, issue 2, Apr 2025 Available: https://doi-org.proxy1.library.jhu.edu/10.1002/sam.70019