A comparative analysis of count data models

Abid Hussain

Contents

What is count data?

Count data represents the number of occurrences of an event within a specific time period, area, or context. These values are non-negative integers (0, 1, 2, $\cdots$).
Characteristics:

1. Discrete: Can only take on whole number values.
2. Non-negative: Counts cannot be negative.
3. Often skewed: Many observations with low counts, fewer with high counts.

Applications of count data

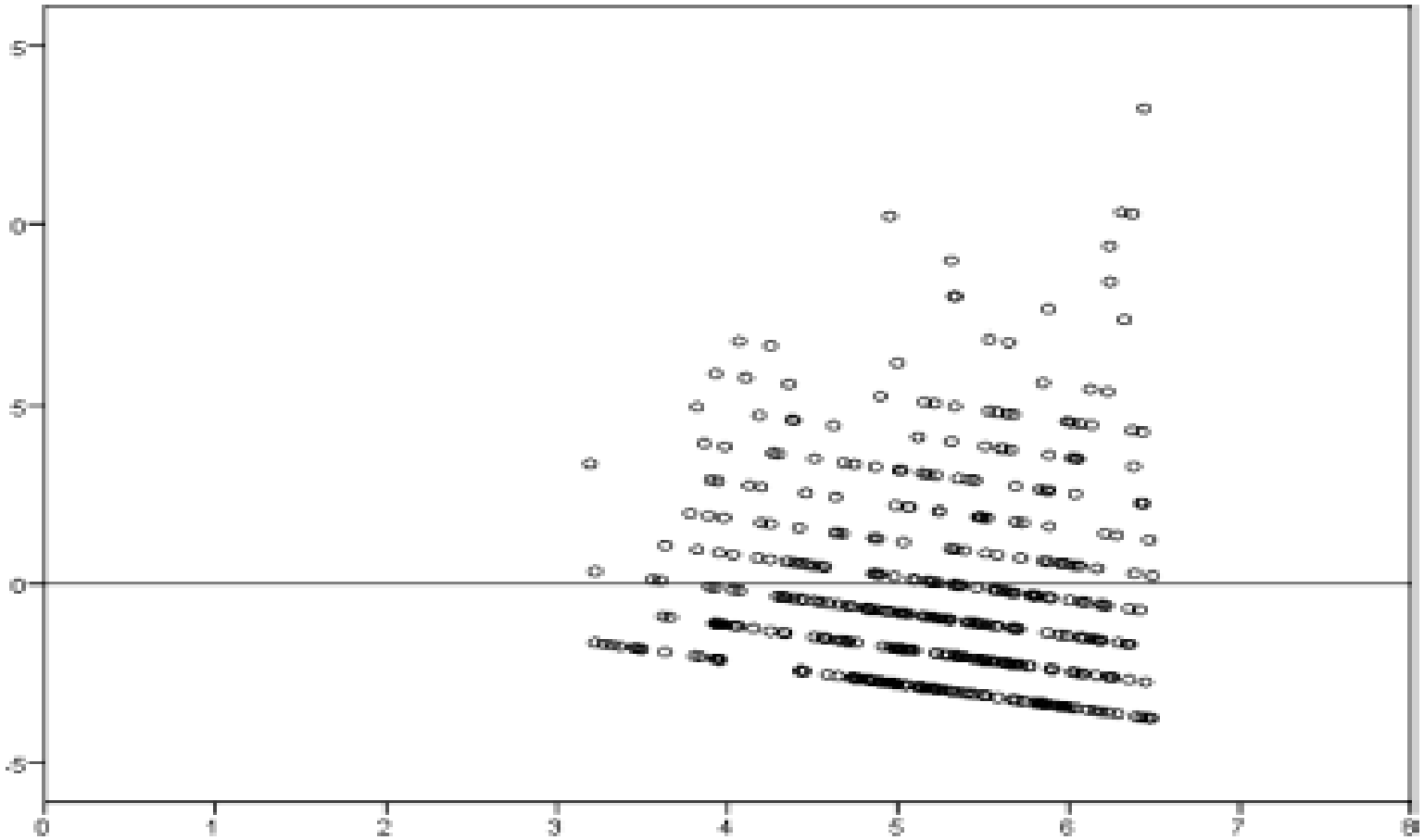
- Number of website visits per day.
- Number of defects in a manufactured product.
- Number of insurance claims filed per year.
- Number of species observed in a survey.
- Number of doctor visits per month.

Why not ordinary least squares (OLS)?

Violation of assumptions: Applying OLS to count data often violates key assumptions:

- Normality of residuals: Count data distributions are typically not normal.
- Homoscedasticity (constant variance): Variance often depends on the mean in count data.
- Linearity: The relationship between predictors and the count variable might not be linear.
- Problematic predictions: OLS can yield negative or non-integer predictions, which are nonsensical for counts.

OLS predicted line for count data



The Poisson distribution: A foundation

Concept: The Poisson distribution models the probability of a given number of events occurring in a fixed interval of time or space if these events occur with a known constant mean rate and independently of the time since the last event.
Probability mass function (PMF):

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

We won't need to use this formula for Poisson regression, but some of its properties.

Cont...

The Poisson distribution is unimodal and skewed to the right. It has a single parameter $\lambda > 0$, which is both its mean and its variance. That is

$$E(Y) = var(Y) = \lambda.$$

Therefore, when the counts are larger, on the average, they also tend to be more variable. If Y = number of conferences attended in the past year has a Poisson distribution, then we observe greater variability in y from person to person when $\lambda = 10.4$ than when $\lambda = 1.2$. Also, λ increases, the skew decreases and the distribution becomes more bell-shaped

Poisson regression: Modeling counts

Goal: To model the relationship between predictor variables and the expected count.
Link Function: Uses a log-link function to ensure the predicted mean is always positive:

$$ln(\lambda) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k,$$

where

- λ is the expected count.
- β_0 is the intercept.
- β_j are the coefficient for the predictor variable X_j, where $j = 1, 2, \cdots, k$.

Interpretation of coefficients

$$\hat{\lambda} = e^{b_0} e^{b_1 X_1} e^{b_2 X_2} \ldots e^{b_k X_k}.$$

- Here the changes in a predictor result in multiplicative changes in the predicted count.
- This contrasts with OLS regression in which changes in the predictor result in additive changes in the predicted value.
- For a 1-unit increase in X_1, the predicted count ($\hat{\lambda}$) is multiplied by e^{b_1}, holding all other variables constant.

Problems with Poisson regression

Real-life count data often exhibit two (related) characteristics: overdispersion and zero-inflation. Overdispersion refers to an excess of variability in the data (i.e., the variance exceeds the mean), while zero-inflation refers to an excess of zeros.
Overdispersion occurs when the variance of the count variable is significantly greater than its mean (i.e. $Var(Y) > E(Y)$).
Causes:

- Unobserved heterogeneity: Missing important predictor variables that influence the count.
- Clustering of events: Events may not be independent.
- Excess zeros: More zero counts than predicted by the Poisson distribution.

Consequences for Poisson regression

- Underestimation of standard errors.
- Inflated Type I error rates (incorrectly rejecting the null hypothesis).
- In the presence of overdispersion, the Poisson regression model is not reliable and can lead to biased parameter estimates.

The negative binomial distribution: Handling overdispersion

The negative binomial distribution is a generalization of the Poisson distribution that allows for overdispersion. It introduces an additional parameter, often denoted as α (the dispersion parameter), that controls the variance independently of the mean. The PMF is:

$$P[Y = y] = \frac{\Gamma(y + \frac{1}{\alpha})}{\Gamma(\frac{1}{\alpha})y!}(\frac{1}{1 + \lambda\alpha})^{1/\alpha}(\frac{\lambda\alpha}{1 + \lambda\alpha})^{y}.$$

The mean and variance for this probability model are:
$E(Y) = \lambda$ and $V(Y) = \lambda(1 + \alpha\lambda)$.
This clearly indicates $Var(Y) > E(Y)$.

Negative binomial regression

The negative binomial model adjusts for Poisson overdispersion; it cannot be used to model underdispersed Poisson data. The negative binomial allows us to model a far wider range of variability than the Poisson.

- Model: Similar to Poisson regression, but uses the negative binomial distribution for the response variable.
- Link function: Typically uses a log-link function for the mean ($ln(\lambda)$).
- Handling overdispersion: The dispersion parameter α can be estimated from the data.
- Interpretation of coefficients: Similar to Poisson regression.

Zeros: False, structural and random

Zeros have multiple origins in a dataset: false zeros occur due to errors in the experimental design or the observer; structural zeros are related to the system under study; and random zeros are the result of the sampling variability. Identifying the type of zeros and their relation with overdispersion and/or zero inflation is key to select the most appropriate statistical model.
Zero-inflated count models provide a way to both model the excess zeros and the overdispersion.

Zero-Inflated Poisson (ZIP) Model

A mixture model that combines two processes:

1. A Bernoulli process: Determines whether the count is zero or comes from the Poisson process. Let π be the probability of being in the "always zero" state.
2. A Poisson process: Generates the count (including zero) with mean λ for those not in the "always zero" state.

It has the following PMF:

$$P(Y = y) = \begin{cases} \omega + (1 - \omega)e^{-\lambda}; & \text{if } y = 0 \\ (1 - \omega)e^{-\lambda}y^{\lambda}/y!; & \text{if } y > 0, \end{cases}$$

where $0 \leq \omega \leq 1$.

Cont...

The mean and variance of ZIP model are:

$$E(Y) = \lambda(1-\omega) = \mu$$

and

$$Var(Y) = \mu + \frac{\omega}{1-\omega}\mu^2.$$

Since the variance is larger than the mean, the ZIP distribution is overdispersed with respect to the Poisson, and it will be appropriate when overdispersion is due to a large number of zeros.

Zero-Inflated negative binomial (ZINB) Model

When there are other sources of overdispersion different from the excess of zeros, a ZINB model could be more appropriate.
It has the following PMF:

$$P(Y=y)=\begin{cases}\omega+(1-\omega)(\frac{1}{1+\lambda\alpha})^{1/\alpha}; & \text{if } y=0\\ (1-\omega)\frac{\Gamma(y+\frac{1}{\alpha})}{\Gamma(\frac{1}{\alpha})y!}(\frac{1}{1+\lambda\alpha})^{1/\alpha}(\frac{\lambda\alpha}{1+\lambda\alpha})^{y}; & \text{if } y>0,\end{cases}$$

where $0 \leq \omega \leq 1$.

Cont...

The mean and variance for the ZINB are:

$$E(Y) = \lambda(1-\omega) = \mu$$

and

$$Var(Y) = \mu + (\frac{\omega}{1-\omega} + \frac{\alpha}{1-\omega})\mu^2.$$

The overdispersion comes from the ratio $\frac{\omega}{1-\omega}$, related with the proportion of structural zeros, and it also comes from the dispersion parameter of the underlying NB distribution which is related to $\frac{\alpha}{1-\omega}$.

Hurdle Models (Two-Part Models)

Similar to zero-inflated models, hurdle models also handle excess zeros by modeling two distinct processes:

1. A binary process: Determines whether the count is zero or positive (often using a logistic or probit model).
2. A truncated count process: Models the magnitude of the positive counts (e.g., a truncated Poisson or negative binomial distribution, excluding zeros).

Key difference from zero-inflated models

Hurdle models assume that if an individual crosses the "hurdle" (i.e., has a positive count), the zero count is no longer possible from the second process. Zero-inflated models allow for zeros to arise from both processes.

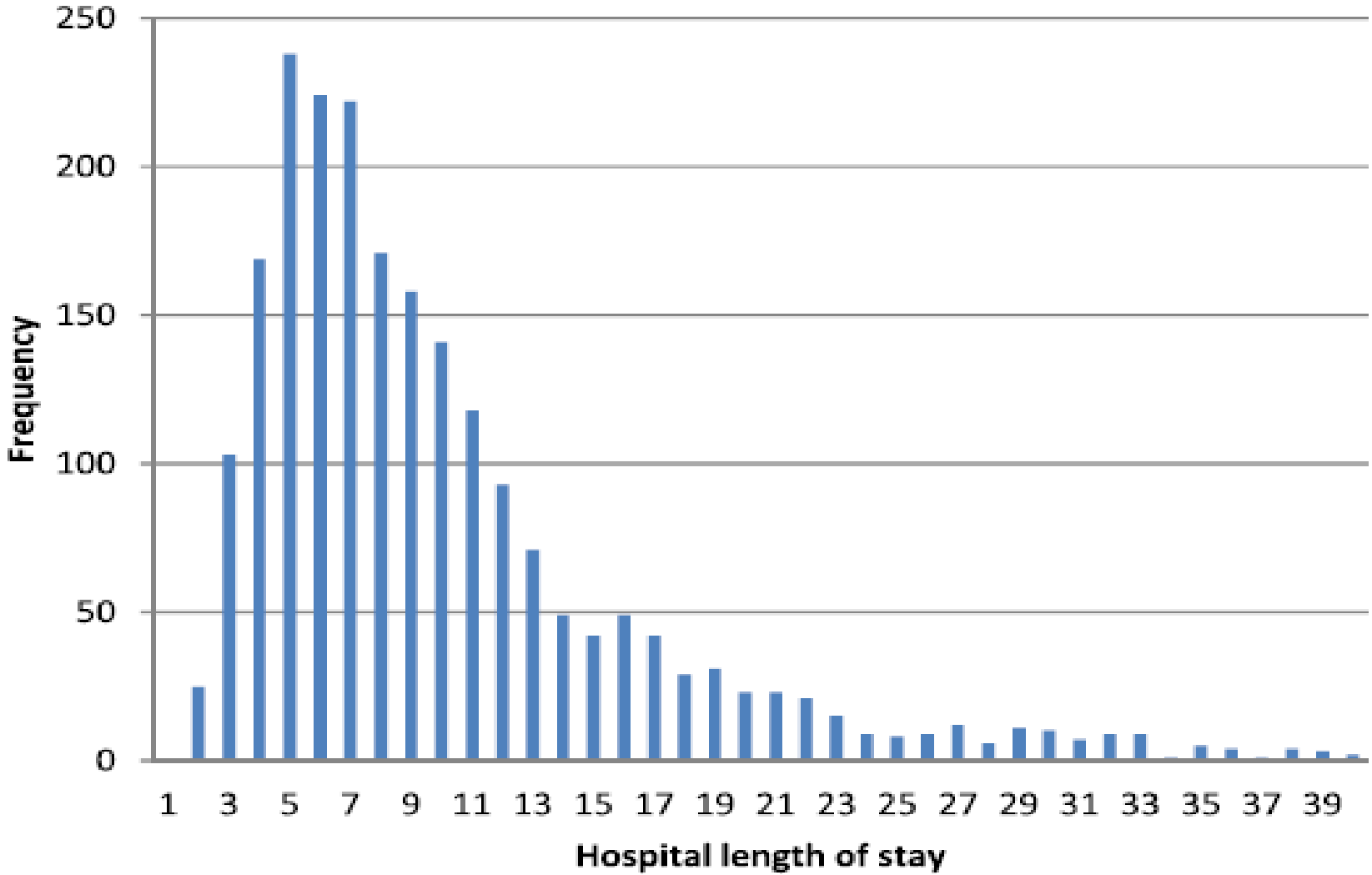
An empirical study

The empirical study population consisted of 2,167 patients admitted in hospitals with a diagnosis of Asthma selected from MIMIC dataset using ICD-9 code 49,390. This dataset is used in the study of Fernandez and Vatcheva (2022). We present the main demographic characteristics of the study population.
The distribution of the variable hospital length of stay (LOS) was positively skewed, with values ranging from 0 to 40 days. The mean LOS, 8.0 days, was much lower than the variance of 43.10. The larger sample variance compared to the sample mean suggested a deviation from the Poisson regression model's assumption for equal variance and mean.

- Fernandez, G.A., and Vatcheva, K.P. (2022). A comparison of statistical methods for modeling count data with an application to hospital length of stay. BMC Medical Research Methodology, 22(1), 211.

Demographic and clinical characteristics of the study

Characteristic		
Age (years), mean (SD)		62.3 (40.66)
Sex, n (%)		
	Male	864 (39.87)
	Female	1303 (60.13)
Hospital admission type, n (%)		
	Elective	378 (17.44)
	Emergency	1748 (80.66)
	Urgent	41 (1.89)
Health insurance type, n (%)		
	Government	96 (4.43)
	Medicaid	304 (14.03)
	Medicare	961 (44.35)
	Private	789 (36.41)
	Self-Pay	17 (0.78)
LOS, mean (SD)		8.0 (6.56)



Histogram of hospital length of stay for patients with asthma diagnosis, $n = 2{,}167$

Statistical analysis

Poisson, NB, ZIP, and ZINB regression models were fitted for LOS on the predictor variables age, sex, health insurance, and admission type. The Pearson dispersion statistic, calculated by dividing the model's Pearson Chi-square statistic by the corresponding degrees of freedom, was used as a criterion for assessing model's misspecification or an overdispersed response variable. When the resultant value is greater than one, the model is considered to be overdispersed. AIC and BIC were used to compare the models. Furthermore, the models estimated coefficients, standard errors and their significance where examined, giving special attention to the difference in findings and conclusions across the models.

Findings from fitted Poisson, NB, ZIP, and ZINB regression models for hospital LOS, n = 2,167

Parameter	Poisson		NB		ZIP		ZINB	
	Estimate (SE)	*P*-Value	Estimate (SE)	*P*-Value	Estimate (SE)	*P*-Value	Estimate (SE)	*P*-Value
Age	-0.001 (0.0002)	< .0001	-0.001 (0.0004)	0.0546	-0.001 (0.0002)	0.0002	-0.001 (0.0004)	0.0561
Sex								
Female	0.03 (0.02)	0.1086	0.03 (0.03)	0.4651	0.02 (0.02)	0.1247	0.03 (0.03)	0.4414
Male	reference		reference		reference		reference	
Health Insurance								
Government	0.32 (0.11)	0.0045	0.32 (0.21)	0.1228	0.26 (0.11)	0.0215	0.26 (0.21)	0.2129
Medicaid	0.49 (0.11)	< .0001	0.49 (0.20)	0.0131	0.44 (0.11)	< .0001	0.43 (0.20)	0.0319
Medicare	0.46 (0.11)	< .0001	0.46 (0.19)	0.0189	0.41 (0.11)	0.0001	0.40 (0.20)	0.0445
Private	0.41 (0.11)	0.0001	0.40 (0.19)	0.0372	0.35 (0.11)	0.0009	0.35 (0.20)	0.0796
Self-pay	reference		reference		reference		reference	
Admission type								
Elective	-0.30 (0.05)	< .0001	-0.29 (0.12)	0.0159	-0.29 (0.05)	< .0001	-0.29 (0.12)	0.0160
Emergency	-0.17 (0.05)	0.0011	-0.16 (0.12)	0.1649	-0.16 (0.05)	0.0025	-0.16 (0.12)	0.1667
Urgent	reference		reference		reference		reference	
Zero Model	Estimate (SE)	*P*-Value	Estimate (SE)	*P*-Value	Estimate (SE)	*P*-Value	Estimate (SE)	*P*-Value
Age					0.003 (0.004)	0.5015	0.07 (0.13)	0.5722
Sex								
Female					-0.09 (0.43)	0.8437	18.10 (4082.56)	0.9965
Male					reference		reference	
Health Insurance								
Government					-14.35 (528.33)	0.9787	-22.23 (16,904.72)	0.9990
Medicaid					-1.55 (1.23)	0.2057	-34.36 (4533.48)	0.9940
Medicare					-1.56 (1.17)	0.1800	-37.75 (4039.53)	0.9925
Private					-2.22 (1.19)	0.0662	-22.36 (5537.88)	0.9968
Self-pay					reference		reference	
Admission type								
Elective					12.99 (810.38)	0.9871	1.81 (20,295.23)	0.9999
Emergency					12.72 (810.38)	0.9873	17.23 (19,158.48)	0.9993
Urgent					reference		reference	
Pearson Chi-Square (value/degrees of freedom)	5.3016		1.1815		4.9001		1.1868	
AIC	17,675.4989		13,134.7955		17,560.4274		13,150.2305	
BIC	17,726.6288		13,191.6065		17,662.6872		13,258.1713	

Comparison of fitted Poisson, NB, ZIP, and ZINB regression models

The previous table presents the results of fitted Poisson, NB, ZIP, and ZINB regression models for the outcome variable LOS on the patient level predictor variables age, sex, type of hospital admission, and health insurance status. In the zero-inflated models the same predictors were used to fit both the count model and the logistic (zero) model.
Based on the results:

- The NB regression model provided the best fit to the data since it resulted the smallest AIC and BIC values.
- The second best model was ZINB, followed by the ZIP model.
- The Poisson regression model resulted with the worst fit to the data according to the AIC and BIC values.

Cont...

- The Pearson dispersion statistic in Poisson regression model was 5.3016, greater than 1, suggesting overdispersion.
- The fitted NB regression model had the smallest dispersion statistic of 1.1815.
- The regression coefficient estimates and their respective standard errors differed across the models.
- It is quite noticeable in table the tendency for the Poisson, and ZIP regression models to produced smaller standard errors of the regression coefficient estimates than NB and ZINB regression models.

Cont...

- Overdispersion may cause standard errors of the regression coefficient estimates to be underestimated and therefore contributing to discrepancies in significant regression coefficients findings between the models. For instance, at a 5% significance level, only based on the fitted Poisson and ZIP regression models there were significant association between age and log LOS, controlling for the effect of sex, health insurance type, and admission type variables included in the models.
- In relation to the logistic part (zero-model), none of the variables in both ZIP and ZINB regression models had significant contribution to the structural zero-generating process of LOS.

Choosing the right model

1. Start with Poisson regression: If the equidispersion assumption holds (i.e., mean $\approx$ variance), Poisson might be sufficient.
2. Test for overdispersion: If variance is significantly greater than the mean, consider negative binomial regression.
3. Assess excess zeros: If there are more zeros than predicted by Poisson or negative binomial, explore zero-inflated or hurdle models.
4. Conceptual understanding: Does the process generating zeros seem fundamentally different from the process generating positive counts? If so, zero-inflated or hurdle models might be appropriate.
5. Model comparison: Use statistical tests (e.g., Vuong test) and information criteria (AIC, BIC) to compare model fit.