Selecting the best model to fit the Rainfall Count data Using Some Zero Type models with application

Luay Habeeb Hashim, Ahmad Naeem Flaih

Department of Statistics**,** College of Administration and Economics, University of Al-Qadisiyah.

Abstract

Counts data models cope with the response variable counts, where the number of times that a certain event occurs in a fixed point is called count data, its observations consists of non-negative integers values {0,1,2,…}. Because of the nature of count data, the response variables are usually considered doing not follow normal distribution. Therefore, linear regression is not an appropriate method to analysis count data due to the skewed distribution. Hence, using linear regression model to analysis count data is likely to bias the results, under these limitations, Poisson regression model and "Negative binomial regression" are likely the appropriate models to analysis count data. Sometimes researchers may Counts more zeros than the expected. Count data with many Zeros leads to a concept called "Zero-inflation". Data with abundant zeros are especially popular in health, marketing, finance, econometric, ecology, statistics quality control, geographical, and environmental fields when counting the occurrence of certain behavioral and natural events, such as frequency of alcohol use, take drugs, number of cigarettes smoked, the occurrence of earthquakes, rainfall, and etc. Some models have been used to analyzing count data such as the "zero- altered Poisson" (ZAP) model and the "negative binomial" model. In this paper, the models, Poisson, Negative Binomial, ZAP, and ZANB were been used to analyze rainfall data.

Introduction

Count data, including zero counts arise in a wide variety of application, hence models for counts have become widely popular in many fields. In the statistics field, one may define the count data as that type of observation which takes only the non-negative integers value, Sometimes researchers may Counts more zeros than the expected. Excess zero can be defined as Zero-Inflation. Excess zero sometimes may be the reason of occurs Over-dispersion (variance a lot larger than mean). Over-dispersion concept is commonly used in the analysis of discrete data. Therefore, linear regression is not applicable procedure to estimate the parameters of predictors due to the asymmetric distribution of the response variable. Under these limitations, Poisson regression and Negative binomial regression are used to model the Count data.

Lambert (1992) discussed this matter and suggested "zero-inflated Poisson" model with an application in manufacturing quality also suggested by Greene (1994) and "the zero-altered Poisson" model (Another common method to model the excess zeros in count data is to employ hurdle models (also called a zero-altered model) which it developed by Cragg (1971)), that have been suggested to cope with an overabundance of zeros. Models for Zero-Inflation have become of interesting so in this work I focus on the excess zero case.

In some commonly used discrete distributions the mean of the distribution related to the variance, the reason of exhibit Over-dispersion. That is, Over-dispersion appear in the data in which there is evidence that variance of the dependent variable is greater than the mean.

Data with abundant zeros are especially popular in health, marketing, finance, econometric, ecology, statistics quality control, geographical, and environmental fields when counting the occurrence of certain behavioral and natural events, such as frequency of alcohol use, take drugs, number of cigarettes smoked, the occurrence of earthquakes, rainfall, and etc.

Famoye and Consul (1992) proposed "generalized Poisson" distribution which can take consideration of "over-dispersion" of Poisson distribution. The extension of generalized Poisson distribution is "zero-inflated generalized Poisson" (ZIGP) suggested by Famoye and Singh (2006).

Some other models have been used to analyzing count data such as the "zero-altered Poisson" (ZAP) model. In existence of "over-dispersion" in the data "negative binomial" model can be preferred when Poisson mean has a gamma distribution. A normal stretch of "negative binomial" model to accommodate increase zeros in the data is "zero-altered negative binomial" (ZANB) model discussed by Heilbron (1994).

The difference between negative binomial and Poisson models is that negative binomial models can be used when "over-dispersion" exists even in the nonzero part of the distribution. In this paper, I focus on the models, Poisson, Negative Binomial, ZAP, and ZANB to analyze rainfall data.

Poisson Regression Model (PRM)

Poisson regression model is a non-linear (log-linear) regression models and it is convenient for the analysis of count or rate data. Poisson regression is similar to the multiple regression excepting that the response (y) variable is an observed count that follows the "Poisson distribution". Therefore, the possible values of (y) are "non-negative integers". Suppose we have a random sample $y_1,\ldots,y_n$ drawn from Poisson distribution, then the p.m.f of y_i, As follow

$$p(y_i,\mu_i)=\frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!}\quad ;y_i=0,1,2,\ldots. \qquad (1)$$

By assumptions of GLM, We have

$Y_i \sim P(\mu_i)\,; E(Y_i)=\mu_i\,, Var(Y_i)=\mu_i$, and $\mu_i=e^{\eta(X_{i1},\ldots,X_{iq})}=e^{X'\beta}$

Where $X'\beta=\alpha+\beta_1X_{i1}+\cdots+\beta_qX_{iq}$ and $X_{i1},\ldots,X_{iq}$ are the independent variables.

Given the p.m.f in (1) and using the method of maximum likelihood and assuming independence of the observations, We can estimate regression parameters as follow

$$L=\prod_i^n\frac{\mu^{y_i}e^{-\mu_i}}{y_i!}$$

Taking the log of both sides,

$$\log(L)=\sum_i^n(\log(\mu^{y_i}e^{-\mu_i})-\log(y_i))$$
$$=\sum_i^n(\log(\mu^{y_i})+\log(e^{-\mu_i})-\log(y_i!))$$
$$=\sum_i^n(y_i\log(\mu_i)-\mu_i-\log(y_i!))$$
$$=\sum_i^n(y_iX'\beta-e^{X'\beta}-\log(y_i!))$$

By taking partial derivatives of the parameters and equalizing the likelihood equation to zero

$$\frac{\partial \log(L)}{\partial \beta} = \frac{\partial}{\partial \beta}\sum_{i=1}^{n}(y_i X'\beta - e^{X'\beta} - \log(y_i!)) = \sum_{i=1}^{n}(y_i X - X e^{X'\beta}) = 0 \qquad (2)$$

Applying numerical methods such as "Newton Raphson" to solve equation (2).

"Poisson regression model" is suitable for modeling "count data" but in practice, Usually, the variance of count data overrides its mean, resulting Over-dispersion. Count data underlying Over-dispersion and Poisson regression model leads to bias results, and under estimation of the parameters which effects on the standard errors and P-value. This Over-dispersion may be due to a random unobserved variation component in the function of X'.

Negative Binomial Regression Model (NBRM)

Negative binomial regression is one of types of generalized linear models in which the "dependent variable" Y is a count of the number of times an event occurs. Negative binomial regression is similar to the multiple regression excepting that the response variable (y) is an observed count that follows the " negative binomial distribution". Therefore, the possible values of (y) are "nonnegative integers".

To address the problem of "over-dispersion" in "a Poisson regression", "Negative Binomial regression" model has been used, by allowing for the random variation component in Poisson conditional a mean (μ) through the parameter (α). Negative binomial regression is a popularization of Poisson regression which relax the restrictive assumption that the variance is equal to the mean made by the Poisson model. Suppose that $y_1,\ldots,y_n$ are a random sample from the Negative binomial distribution, then the p.m.f of y_1 is expressed as

$$p\left(y_i;\frac{1}{\alpha},\mu_i\right) = \frac{\Gamma\left(y_i+\frac{1}{\alpha}\right)}{\Gamma\left(\frac{1}{\alpha}\right)\Gamma(y_i+1)}\left(\frac{1}{1+\alpha\mu_i}\right)^{\frac{1}{\alpha}}\left(\frac{\alpha\mu_i}{1+\alpha\mu_i}\right)^{y_i} ; y = 0,1,2,\ldots \qquad (3)$$

By assumptions of GLM, We have

$Y_i \sim NB\left(\mu_i,\frac{1}{\alpha}\right)$; $E(Y_i) = \mu_i$,

$Var(Y_i) = \mu_i + \alpha\mu_i^2$

and $\mu_i = e^{\eta(X_{i1},\ldots,X_{iq})} = e^{X'\beta}$

Where $X'\beta = \alpha + \beta_1 X_{i1} + \cdots + \beta_q X_{iq}$ and $X_{i1},\ldots,X_{iq}$ are the independent variables.

Given the p.m.f in (3) and using the method of maximum likelihood and assuming independence of the observations, We can estimate regression parameters as follow

$$L = \prod_i p(y_i;\mu_i)$$

$$L = \prod_i^n \left[\frac{\Gamma(y_i+\frac{1}{\alpha})}{\Gamma(\frac{1}{\alpha})\Gamma(y_i+1)}\left(\frac{1}{1+\alpha\mu_i}\right)^{\frac{1}{\alpha}}\left(\frac{\alpha\mu_i}{1+\alpha\mu_i}\right)^{y_i}\right]$$

$$\log(L) = \sum_{i=1}^{n}\begin{bmatrix} y_i log\left(\frac{\alpha\mu_i}{1+\alpha\mu_i}\right) - \frac{1}{\alpha}\log(1+\alpha\mu_i) \\ +log\Gamma\left(y_i+\frac{1}{\alpha}\right) \\ -log\Gamma(y_i+1) - log\Gamma\left(\frac{1}{\alpha}\right) \end{bmatrix}$$

$$\log(L) = \sum_{i=1}^{n}\begin{bmatrix} y_i log\left(\frac{\alpha e^{X'\beta}}{1+\alpha e^{X'\beta}}\right) - \frac{1}{\alpha}\log\left(1+\alpha e^{X'\beta}\right) \\ +log\Gamma\left(y_i+\frac{1}{\alpha}\right) \\ -log\Gamma(y_i+1) - log\Gamma\left(\frac{1}{\alpha}\right) \end{bmatrix}$$

By taking partial derivatives of the parameters and equalizing the likelihood equation to zero

$$\frac{\partial \log(L)}{\partial \beta} = \frac{\partial}{\partial \beta}\left[\sum_{i=1}^{n}\begin{bmatrix} y_i log\left(\frac{\alpha e^{X'\beta}}{1+\alpha e^{X'\beta}}\right) - \frac{1}{\alpha}\log\left(1+\alpha e^{X'\beta}\right) \\ +log\Gamma\left(y_i+\frac{1}{\alpha}\right) \\ -log\Gamma(y_i+1) - log\Gamma\left(\frac{1}{\alpha}\right) \end{bmatrix}\right] = 0 \qquad (4)$$

$$\frac{\partial \log(L)}{\partial \alpha} = \frac{\partial}{\partial \alpha}\left[\sum_{i=1}^{n}\begin{bmatrix} y_i log\left(\frac{\alpha e^{X'\beta}}{1+\alpha e^{X'\beta}}\right) - \frac{1}{\alpha}\log\left(1+\alpha e^{X'\beta}\right) \\ +log\Gamma\left(y_i+\frac{1}{\alpha}\right) \\ -log\Gamma(y_i+1) - log\Gamma\left(\frac{1}{\alpha}\right) \end{bmatrix}\right] = 0 \qquad (5)$$

Applying numerical methods such as "Newton Raphson" to solve equations (4) and (5).

Zero-Altered Models (ZA)

Zero-altered models known as a two-part models, Where the first part is a binary outcome model governs with binomial probability, and the second part is a truncated count model. In zero-inflated models assumed that count data consist of two types of data subgroups, the first subgroup is a set of only zeros count (true zeros and false zeros), and the second subgroup is a set of count variables (with true zeros). While, zero-altered models do not discriminate between the types of zeros; they are simply zeros. 'The basic idea for the zero-altered models is that the outcomes are treated as absence and presence zeros data'. This means that the outcomes are divided into two groups, the first includes all zeros, the second includes non-zero counts.

Where, The binomial distribution is used to model the absence and presence, and a Poisson (or negative binomial) distribution for the counts. To measure a non-zero count should be modified the distribution and exclude the possibility of a zero observation, and this is called a zero-truncated distribution.

Assume that the zeros are follow the probability mass function (p.m.f) $f_1(.)$ with $P(y=0) = f_1(0)$ and $P(y>0) = 1 - f_1(0)$, while the positive outcomes are formed by the probability mass function truncated at zero given by

$f_2(y|y>0) = f_2(y)/[1-f_2(0)]$.

Hence, the Hurdle (Altered) probability mass function as follow

$$P(y) = \begin{cases} f_1(0) & ; y = 0 \\ \frac{1-f_{1(0)}}{1-f_2(0)} f_2(y) & ; y > 0 \end{cases} \qquad (6)$$

Zero-Altered Poisson Model (ZAPM)

Suppose that the probability of measuring zero observation in the first part of Hurdle structure is modelled with a binomial distribution, Where θ_i is the probability that $y_i = 0$.

Suppose that be the response variable for the positive counts' (truncated at zero) with Poisson probability mass function (1).

Furthermore, let the probability of observing $y_i = 0$ in the first part of Hurdle model (zero count) as follow

$$P(y_i = 0) = f_1(0) = \theta_i \qquad (7)$$

Where, the probability of observing ($y_i > 0$) in the second part of Hurdle model (positive counts) as follow

$$P(y_i; \mu_i | y_i > 0) = f_2(y) = \frac{\mu^{y_i} e^{-\mu_i}}{y_i!} \qquad (8)$$

Therefore, substituting (1), (7), and (8) in Zero-Altered (6), we have

$$P(Y_i = y_i) = \begin{cases} \theta_i & ; y_i = 0 \\ (1-\theta_i)\frac{\mu^{y_i} e^{-\mu_i}}{(1-e^{-\mu_i})y_i!} & ; y_i > 0 \end{cases} \qquad (9)$$

By GLM, $\mu_i = e^{X'_i \beta_i}$, where X'$_i$ are knows independent variables, Lambert (1992) suggested the functional form for modelling the parameter θ_i as logistic function, which is given by

$$Log\left(\frac{\theta_i}{1-\theta_i}\right) = z'_i \gamma_i$$

and therefore,

$$\theta_i = \frac{e^{z'_i \gamma_i}}{1 + e^{z'_i \gamma_i}} > 0$$

Where; Z : the covariates and γ : are regression coefficients.

The corresponding Log-Likelihood function is given as follow

$$\log(L) = \sum_i^n \begin{bmatrix} I(y_i = 0)\log(\theta_i) + \\ I(y_i > 0)(\log(1-\theta_i) \\ -\mu_i + y_i \log(\mu_i) \\ -\log(1-e^{-\mu_i}) - \log(y_i!)) \end{bmatrix} \qquad (10)$$

The mean and variance for ZAP are

$$E(Y_i) = \frac{1-\theta_i}{1-e^{-\mu_i}}\mu_i$$

$$Var(Y_i) = \frac{1-\theta_i}{1-e^{-\mu_i}}(\mu_i + \mu_i^2) - (\frac{1-\theta_i}{1-e^{-\mu_i}}\mu_i)^2$$

Zero-Altered Negative binomial Model (ZANBM)

The same procedure can be easily generalized to "Zero-Altered Negative Binomial regression" model.

Suppose that the probability of measuring zero observation in the first part of Hurdle structure is modelled with a binomial distribution', Where θ_i is the probability that $y_i = 0$.

Suppose that be the response variable for the positive counts' (truncated at zero) with Negative binomial probability mass function (3).

Furthermore, let the probability of observing $y_i = 0$ in the first part of Hurdle model (zero count) as follow

$$P(y_i = 0) = f_1(0) = \theta_i \qquad (11)$$

Where, the probability of observing ($y_i > 0$) in the second part of Hurdle model (positive counts) as follow

$$p(y_i; \mu_i | y_i > 0) = f_2(y) = \frac{\Gamma(y_i + \frac{1}{\alpha})}{\Gamma(\frac{1}{\alpha})\Gamma(y_i+1)} \left(\frac{1}{1+\alpha\mu_i}\right)^{\frac{1}{\alpha}} \left(\frac{\alpha\mu_i}{1+\alpha\mu_i}\right)^{y_i} \qquad (12)$$

Therefore, substituting (3), (11), and (12) in Zero-Altered (6), we have

$$P(Y_i = y_i) = \begin{cases} \theta_i & ; y_i = 0 \\ \dfrac{(1-\theta_i)\frac{\Gamma(y_i+\frac{1}{\alpha})}{\Gamma(\frac{1}{\alpha})\Gamma(y_i+1)}\left(\frac{1}{1+\alpha\mu_i}\right)^{\frac{1}{\alpha}}\left(\frac{\alpha\mu_i}{1+\alpha\mu_i}\right)^{y_i}}{\left(1-\left(\frac{1}{1+\alpha\mu_i}\right)^{\frac{1}{\alpha}}\right)} & ; y_i > 0 \end{cases} \qquad (13)$$

By GLM, $\mu_i = e^{X'_i\beta_i}$, where X'$_i$ are knows independent variables, Lambert (1992) suggested the functional form for modelling the parameter θ_i as logistic function, which is given by

$$Log\left(\frac{\theta_i}{1-\theta_i}\right) = z'_i\gamma_i$$

and therefore,

$$\theta_i = \frac{e^{z'_i\gamma_i}}{1+e^{z'_i\gamma_i}} > 0$$

Where; Z : the covariates and γ : are regression coefficients.

The corresponding Log-Likelihood function is given as follow

$$\log(L) = \sum_i^n \left[\begin{array}{c} I(y_i = 0)\log(\theta_i) + \\ I(y_i > 0) log \left(\dfrac{(1-\theta_i)\frac{\Gamma(y_i+\frac{1}{\alpha})}{\Gamma(\frac{1}{\alpha})\Gamma(y_i+1)}\left(\frac{1}{1+\alpha\mu_i}\right)^{\frac{1}{\alpha}}\left(\frac{\alpha\mu_i}{1+\alpha\mu_i}\right)^{y_i}}{\left(1-\left(\frac{1}{1+\alpha\mu_i}\right)^{\frac{1}{\alpha}}\right)} \right) \end{array} \right] \qquad (14)$$

The mean and variance for ZANB are

$$E(Y_i) = \frac{1-\theta_i}{1-P_0}\mu_i \qquad \text{where} \qquad P_0 = \left(\frac{1}{1+\alpha\mu_i}\right)^{\frac{1}{\alpha}}$$

$$Var(Y_i) = \frac{1-\theta_i}{1-P_0}(\mu_i^2 + \mu_i + \alpha\mu_i^2) - \left(\frac{1-\theta_i}{1-P_0}\mu_i\right)^2$$

Model Selection

It is important that we have one or more a criterion to consider the best results and choose the appropriate model for data representation. There are several methods that provide a measure for selecting the appropriate model, The following four methods will be used: AIC is an evaluating model fit for a given data among different types of non-nested models , and its formula is given as $AIC = -2logL + 2k$, BIC is another estimator for evaluating model fit for a given data among different types of non-nested models, and its formula is given as $BIC = -2logL + klogn$, Likelihood ratio test (LR) is a statistical test used to compare two nested models, its formula is given as $LR = -2\log(L_1/L_2)$, and

Vuong test (V) is a statistical test used to compare non-nested models, It is defined as :

$$V = (\sqrt{n}(\frac{1}{n}\sum_{i}^{n} m_i)/(\sqrt{\frac{1}{n}\sum_{i}^{n}(m_i - \bar{m})^2})$$

Where $m_i = \log(P_1(Y_i|X_i)) - \log(P_2(Y_i|X_i))$. If V>1.96 , then the first model is preferred. If V<-1.96, then the second one is preferred. If | V | <1.96, none of the models are preferred.

Data Analysis

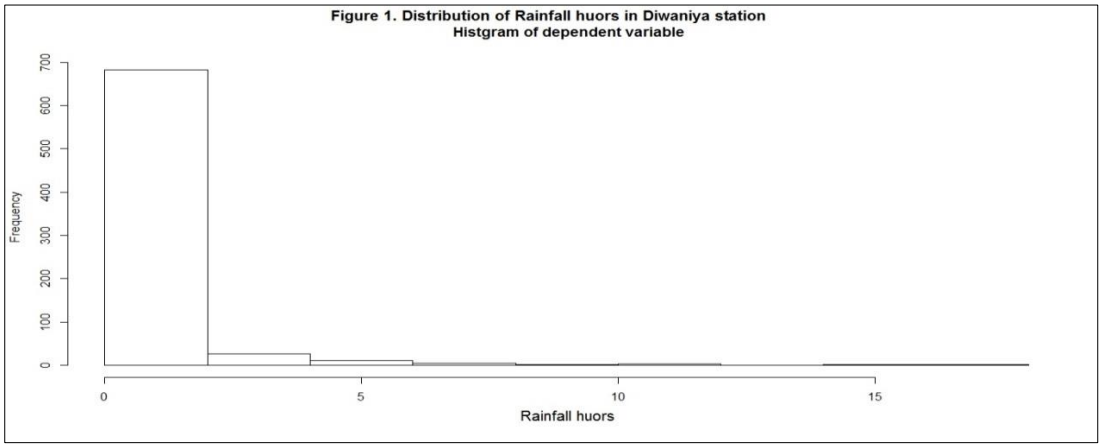
Data were collected from database of the Meteorology and Seismology Organization in Iraq for Diwaniya weather station. The weather station are located in central Iraq, specifically in the city of Diwaniya (about 116 kilometers south of Baghdad).

The count response variable of interest to be modeled "Rainfall hours" measured at Diwaniya weather station. The predictor variables consists of six climate variables derived from Iraqi Meteorological Organization and Seismology database, which include measurements of rainfall, sea pressure, station pressure, wind speed, temperature, and humidity, as shown in Table (1). Data contain observations of (731) for two years (2016 and 2017).

Table 1. Summary statistics of explanatory variables and response used in our count data regression models in Diwaniya weather station.

variables	Minimum value	First quarter	Median	Mean	Third quarter	Maximum value
Wind speed (m/s)	**0**	**1.5**	**2**	**2.098**	**2.5**	**7.7**
Temperature (°C)	**4.2**	**17**	**27.3**	**25.9**	**35.4**	**42.6**
Station pressure (1bar/1000)	**0.9933**	**1.0033**	**1.0095**	**1.0095**	**1.0157**	**1.0274**
Sea pressure (1bar/1000)	**0.9959**	**1.0057**	**1.0120**	**1.0120**	**1.0185**	**1.0301**
Humidity (%)	**16.5**	**29.55**	**38**	**43.74**	**55.9**	**95.4**
Rainfall (hours)	**0**	**0**	**0**	**0.4186**	**0**	**17**

The distribution of the number of non-rainfall hours in Diwaniya weather stations for the two years is shown in figure 1



Poisson Regression

The model fit statistics and estimated coefficients of Poisson regression model are given in Table 2 and Table 3.

Table 2. Fit statistics of Poisson regression model, 2016-2017 Rainfall count data

criterions	Diwaniya weather station
-2Log Likelihood	1098.952
AIC	1110.952
BIC	1138.518

Table 3. Estimated coefficients of Poisson regression model, 2016-2017 Rainfall count data in Diwaniya weather station

Parameter	Estimate	Standard Error	z Value	Pr > |z|
Intercept	55.395032	18.716753	2.96	0.00308
Wind speed	0.492547	0.046956	10.49	<2e-16
Temperature	0.003711	0.020349	0.182	0.85528
Station pressure	-18.668579	137.082777	-0.136	0.89168
Sea pressure	-42.194381	140.465486	-0.3	0.76388
Humidity	0.075247	0.005224	14.404	<2e-16

Since the variance of count data usually exceeds the conditional mean, the equality of variance and mean should always be checked after the development of a Poisson regression. We conducted a test of over-dispersion and The results of this test are shown below likelihood ratio test of H_0: Poisson, as restricted NB model, Critical value of test statistic at the alpha= 0.00 level: 2.7055, For Diwaniya weather station, Chi-Square test statistic= 337.7449, p-value = <2.2e-16. The significance of X^2-statistics implies the existence of over-dispersion. Therefore, in the next section, we develop Negative Binomial model to handle the issue of over-dispersion.

Negative Binomial Regression

In order to address the issue of over-dispersion, we used The model fit statistics and estimated coefficients of Negative Binomial regression model are given in Table 4 and Table 5.

Table 4. Fit statistics of Negative Binomial regression model, 2016-2017 Rainfall count data

criterions	Diwaniya weather station
-2Log Likelihood	761.2069
AIC	775.2069
BIC	807.3677

Table 5. Estimated coefficients of Negative Binomial regression model, 2016-2017 Rainfall count data in Diwaniya weather station

Parameter	Estimate	Standard Error	z Value	Pr > |z|
Intercept	**94.75651**	**47.39353**	**1.999**	**0.0456**
Wind speed	**0.70176**	**0.13645**	**5.143**	**2.71e-07**
Temperature	**-0.07932**	**0.04606**	**-1.722**	**0.0851**
Station pressure	**424.33591**	**544.11577**	**0.78**	**0.4355**
Sea pressure	**-520.93171**	**549.80107**	**-0.947**	**0.3434**
Humidity	**0.06081**	**0.01285**	**4.734**	**2.20e-06**
Alpha	**0.1592**	**0.0311**		

Lambert (1992) and Mullahy (1986) indicated that Negative Binomial regression might not be an appropriate model for count data with excess zeros because it increases the probabilities of both zero and non-zero counts. Since the initial data analysis of our data implied excess zeros (more than 89.5% of the responses in Diwaniya weather station , have non- Rainfall days (rainfall hours are zeros)), we develop Zero-inflated regression to handle excessive number of zeros.

Zero-Altered Regression Models (ZARM)

To fixable the excess zeros problem in non-Rainfall days (rainfall hours are zeros), We used Zero-Altered regression models.

Zero- Altered Poisson Regression (ZAPR) Model

We used the same "explanatory variables" in both parts of the ZAPR 'model. The model fit statistics and estimated coefficients of ZAPR model are given in Table 6 and Table 7.

Table 6. Fit statistics of Zero- Altered Poisson Regression (ZAPR) model, 2016-2017 Rainfall count data

criterions	Diwaniya weather station
-2Log Likelihood	**656**
AIC	**680.0924**
BIC	**695.5665**

Table 7. Estimated coefficients of Zero- Altered Poisson Regression (ZIPR) model, 2016-2017 Rainfall count data in Diwaniya weather station

Parameter	Estimate	Standard Error	z Value	Pr > |z|
Poisson _ Intercept	**-44.686641**	**22.170210**	**-2.016**	**0.04384**
Poisson _ Wind speed	**0.037957**	**0.052565**	**0.722**	**0.47023**
Poisson _ Temperature	**0.059552**	**0.022478**	**2.649**	**0.00807**
Poisson _ Station pressure	**-17.80448**	**172.049507**	**-0.103**	**0.91758**
Poisson _ Sea pressure	**60.294751**	**176.933391**	**0.341**	**0.73327**
Poisson _ Humidity	**0.025634**	**0.005281**	**4.854**	**1.21e-06**
Logit _ Intercept	**221.94421**	**52.48495**	**4.229**	**2.35e-05**
Logit _ Wind speed	**0.99552**	**0.15605**	**6.380**	**1.78e-10**
Logit _ Temperature	**-0.14313**	**0.04938**	**-2.898**	**0.00375**
Logit _ Station pressure	**608.01873**	**625.13393**	**0.973**	**0.33074**
Logit _ Sea pressure	**-830.79921**	**634.53527**	**-1.309**	**0.19043**
Logit _ Humidity	**0.08042**	**0.0141**	**5.703**	**1.18e-08**

Zero- Altered Negative Binomial Regression (ZANBR) Model

We used the same explanatory variables in both parts of the ZANBR 'model. The model fitting statistics and parameters estimation of ZANBR model are given in Table 8 and Table 9.

Table 8. Fit statistics of Zero- Altered Negative Binomial Regression (ZANBR) model, 2016-2017 Rainfall count data

criterions	Diwaniya weather station
-2Log Likelihood	**633.4**
AIC	**659.4968**
BIC	**672.9665**

Table 9. Estimated coefficients of Zero- Altered Negative Binomial Regression (ZANBR) model, 2016-2017 Rainfall count data in Diwaniya weather station

Parameter	Estimate	Standard Error	z Value	Pr > |z|
NB _ Intercept	**-47.442209**	**31.52839**	**-1.505**	**0.13239**
NB _ Wind speed	**0.043539**	**0.074526**	**0.584**	**0.55908**
NB _ Temperature	**0.064675**	**0.031909**	**2.027**	**0.04267**
NB _ Station pressure	**6.625742**	**286.919839**	**0.023**	**0.98158**
NB _ Sea pressure	**38.253228**	**293.890203**	**0.13**	**0.89644**
NB _ Humidity	**0.028987**	**0.007821**	**3.706**	**0.00021**
Logit _ Intercept	**221.94421**	**52.48495**	**4.229**	**2.35e-05**
Logit _ Wind speed	**0.99552**	**0.15605**	**6.38**	**1.78e-10**
Logit _ Temperature	**-0.14313**	**0.04938**	**-2.898**	**0.00375**
Logit _ Station pressure	**608.01873**	**625.13393**	**0.973**	**0.33074**
Logit _ Sea pressure	**-830.79921**	**634.53527**	**-1.309**	**0.19043**
Logit _ Humidity	**0.08042**	**0.0141**	**5.703**	**1.18e-08**
Log (Alpha)	**1.291631**	**0.412397**	**3.132**	**0.00174**

Model Comparison

We used Vuong test to compare non-nested models and Likelihood ratio test to compare nested models, The results of all the Vuong tests are summarized in Table 10 and the results of all Likelihood ratio tests are summarized in Table 11. Furthermore, the results of all information criterions (fit statistics) for all models were summarized in Table 12.

Table 10. Model comparison by Vuong test for non-nested models for Diwaniya weather station

Model	Vuong Statistic	Best model
ZAP vs NB	**3.821809**	**ZAP**
ZANB vs ZAP	**1.514966**	**NONE**
ZANB vs P	**7.495628**	**ZANB**

Note: "If V > 1.96, the first model is preferred. If V < -1.96, then the second one is preferred. If |V |<1.96, none of the models are preferred ".

Table 11. Model comparison by likelihood ratio test for nested models for Diwaniya weather station

Model	Likelihood Ratio Test (p-value)	Best model
P vs NB	**0.7**	**NB**
P vs ZAP	**1.03**	**ZAP**
NB vs ZANB	**0.37**	**ZANB**

Note:

H_o : the simpler model is preferred.

H_1 : the more complex model is preferred.

If p-value < 0.05, we reject H_o , H_1 is preferred.

Table 12. Fit statistics of all models, 2016-2017 Rainfall count data Diwaniya weather station

models	criterions		
	-2Log Likelihood	AIC	BIC
Poisson regression	**1098.952**	**1110.952**	**1138.518**
NB regression	**761.2069**	**775.2069**	**807.3677**
ZAPR	**656**	**680.0924**	**695.5665**
ZANBR	**633.4***	**659.4968***	**672.9665***

*The best model.

Application results

After estimating the regression parameters for all models using real counting data. The test criteria values for all models were obtained for the purpose of comparing these models and selecting the best ones to represent our data. The results in Table 12 indicated that Zero-Altered Negative Binomial (ZANBR) regression model was the best count data model for our data, Although it is hard to distinguish Negative Binomial, and Zero-Altered Poisson (ZAPR) regression models, they are better than Poisson regression model.

References

1- Bozdogan, H., (2000), "Akaike's information criterion and developments in information complexity" Journal of Mathematical Psychology, Vol. 44, PP. 62-91.

2- Christopher, J. W., (1996), "Evaluating zero-inflated and hurdle Poisson specifications", Midwest Political Science Association.

3- Consul, P.C. & Famoye, F., (1992), "Generalized Poisson regression

model", Communications in Statistics – Theory and Methods, Vol. 21, PP.89-109.

4- Dalrymple, M. L. & Hudson, I. L. & Ford, R. P. K., (2003), "Finite Mixture, Zero-inflated Poisson and Hurdle models with application to SIDS", Computational statistics & Data Analysis, Vol. 41, PP. 491-504.

5- Famoye, F. & singh, K.P., (2006), "Zero-Inflated Generalized Poisson Regression Model with an Application to Domestic Violence Data", Journal of Data Science, Vol. 4, PP. 117-130.

6- Greene, W.H.,(1994), "Accounting for Excess Zeros and Sample Selection in Poisson and Negative Binomial Regression Models", Leonard N. Stern School of Business, New York University.

7- Guisan, A. & Hastie, T. ,(2002), "Generalized linear and generalized additive models in studies of species distributions: setting the scene", Ecological Modelling, Vol. 157, PP. 89-100.

8- Heilbron, D. C., (1994), "Zero-Altered and Other Regression Models for Count Data with Added Zeros", Biometrical Journal, Vol. 36, PP. 531-547.

9- Hilbe, J.M., (2011), "Negative Binomial Regression", 2nd Edition, Cambridge University Press, New York.

10- Lambert, D., (1992), "Zero-Inflated Poisson Regression, with an Application to Defects in Manufacturing", Technometrics, Vol. 34, No. 1, PP. 1-14.

11- Lee, L. F., (1986), "Specification Test for Poisson Regression models", International Economic Review, Vol. 27, No. 3, PP. 689-706.

12- Ping, J., (2013), " Count Data Models for Injury Data from the National Health Interview Survey (NHIS)", Thesis, the Graduate School, The Ohio State University.

13- Schwarz, G., (1978), "Estimating the Dimension of a Model", The Annals of Statistics, Vol. 6, No. 2, PP. 461-464.

14- Vincent, P. J. & Haworth, J. M., (1983), "Poisson Regression Models of Species Abundance", Journal of Biogeography, Vol. 10, No. 2, PP. 153-160.

15- Vuong, Q.H., (1989), " Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses", Journal of Econometrics, Vol. 57, No. 2, PP. 307-333.

16- Yang, S. & Others, (2017), " A Comparison of Different Methods of Zero-Inflated Data Analysis and an Application in Health Surveys", Journal of Modern Applied Statistical Methods, Vol. 16, No. 1, PP. 518-543.

17- Zuur, A.F. & others, (2009), "Mixed effects models and extensions in ecology with R", Springer, NY: New York.

اختيار افضل نموذج لملائمة بيانات العد لهطول الامطار باستخدام بعض النماذج الصفرية مع تطبيق عملي

لؤي حبيب هاشم، أ.م. د. احمد نعيم فليح

قسم الاحصاء، كلية الادارة والاقتصاد، جامعة القادسية

نبذة مختصرة

تتعامل نماذج بيانات العد مع أعداد متغير الاستجابة ، حيث يُطلق على عدد المرات التي يحدث فيها حدث معين في نقطة ثابتة بيانات العدد ، وتتكون مشاهداته من قيم صحيحة غير سالبة {0،1،2،...}. نظرًا لطبيعة بيانات العد ، فإن متغيرات الاستجابة تُعتبر عادةً لا تتبع التوزيع الطبيعي . لذلك ، لا يعد الانحدار الخطي طريقة مناسبة لتحليل بيانات العد بسبب التوزيع المنحرف . وبالتالي ، من المرجح أن يؤدي استخدام نموذج الانحدار الخطي لتحليل بيانات العد إلى التحيز في النتائج ، في ظل هذه القيود ، من المرجح أن يكون نموذج الانحدار بوايسون و الانحدار ثنائي الحدين السالب هو النموذج المناسب لتحليل بيانات العد . قد يقوم الباحثون أحيانًا بحساب أصفار أكثر من المتوقع . إن بيانات العد مع العديد من الأصفار يؤدي إلى مفهوم يسمى "التضخم الصفري" . البيانات ذات الأصفار الوفيرة تظهر بصورة واسعة خاصة في مجالات الصحة والتسويق والتمويل والاقتصاد القياسي والإيكولوجيا ومراقبة جودة الإحصاءات والمجالات الجغرافية والبيئية وكذلك عند حساب بعض الأحداث السلوكية والطبيعية ، مثل تكرار تعاطي الكحول وتناول الأدوية وعدد من السجائر المدخنة و وقوع الزلازل وهطول الأمطار وغيرها . وقد استخدمت بعض النماذج لتحليل بيانات العد مثل نموذج (Zero-Altered Poisson) (ZAP) ونموذج "ذو الحدين السالب". في هذه الورقة ، تم استخدام نموذج بوايسون ونموذج ثنائي الحدين السالب ونموذج Zero-Altered Poisson ونموذج ZAP ، ونموذج (Zero-Altered Negative binomial) ZANB لتحليل بيانات هطول الأمطار.