Modeling 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

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. 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-inflated Poisson (ZIP) model and the negative binomial model. In this paper, the models, Poisson, Negative Binomial, ZIP, and ZINB were been used to analyze rainfall data.

Introduction

Count data reflects the number of occurrence of certain characteristic in a fixed period of time, that is, Count data are non-negative integers {0,1,2,3,...}. Count data becomes popular in a wide areas of interesting sciences; such as finance, marketing, health care, weather, and others. Count data with excessive zeros are prevalent in a wide variety of disciplines, in many of these areas of sciences, 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). 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-inflated Poisson" (ZIP) 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-inflated negative binomial" (ZINB) model discussed by Mwalili (2008).

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[15]. In this paper, I focus on the models, Poisson, Negative Binomial, ZIP, and ZINB 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

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

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

$$p\left(y_i;\frac{1}{\alpha},\mu_i\right)=\frac{\Gamma\left(y_i+\frac{1}{\alpha}\right)}{\Gamma\left(\frac{1}{\alpha}\right)\Gamma(y_i+1)}\left(\frac{1}{1+\alpha\mu_i}\right)^{\frac{1}{\alpha}}\left(\frac{\alpha\mu_i}{1+\alpha\mu_i}\right)^{y_i}$$

$$;y=0,1,2,\ldots \quad (3)$$

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 $X'\beta=\alpha+\beta_1X_{i1}+\cdots+\beta_qX_{iq}$ 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

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

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

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

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

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 (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 (5)

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

Zero-Inflated Models (ZI)

Excess zeros in certain population is lead to Zero-Inflation which is made up two types of data subgroups (data generation), 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) that distributed according to Poisson distribution (Lambert 1992, Van den Broek 1995).

Zero-Inflated Poisson Regression Model (ZIPR)

The "zero-inflated Poisson" regression is used for modelling count data that show over-dispersion and zero counts (excess zeros). This model consider there are two types of data sources, the first source is zero type and the second is comes from data follows Poisson distribution.

According to Lambert (1992), the response variable Y_i is independent with $Y_i \sim 0$ with probability (θ_i) and $Y_i \sim$ Poisson μ_i with probability $(1-\theta_i)$

Where θ_i is the probability that observation (i) is in the always zeros subgroup.

Therefore,

$$\Pr(Y_i = 0) = \theta_i + (1-\theta_i) \times \Pr(Count\ process\ at\ (i) gives\ a\ zero) \quad (6)$$

By assumption the Y_i follows a Poisson distribution with mean μ_i

$$p(y_i; \mu_i | y_i \geq 0) = \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!}$$

Subsequently

The term $\Pr(Count\ process\ at\ (i) gives\ a\ zero)$ is given by

$$p(y_{i=0}; \mu_i | y_i \geq 0) = \frac{e^{-\mu_i}\mu^0}{0!} = e^{-\mu_i}$$

Hence, Equation (6) can now be rewritten as

$$\Pr(Y_i = 0) = \theta_i + (1-\theta_i)e^{-\mu_i} \quad (7)$$

With probability that Y_i is a non-zero count, we have

$$\Pr(Y_i = y_i) = +(1-\theta_i) \times \Pr(Count\ process) \quad (8)$$

Hence, Equation (8) can be rewritten as follow

$$Pr(Y_i = y_i | y_i > 0) = (1-\theta_i)\frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!} \quad (9)$$

Furthermore, The probability density function for a ZIP model is given by

$$p(Y_i = y_i) = \begin{cases} \theta_i + (1-\theta_i)e^{-\mu_i} & if\, y_i = 0 \\ (1-\theta_i)\frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!} & if\, y_i > 0 \end{cases}$$

(10)

By GLM[15], $\mu_i = e^{X'_i\beta_i}$, 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

$$\log(L) = \sum_i^n \begin{bmatrix} I(y_i = 0)\log(\theta_i + (1-\theta_i)e^{-\mu_i}) + \\ I(y_i > 0)(\log(1-\theta_i) - \mu_i + \\ y_i\log(\mu_i) - \log(y_i!)) \end{bmatrix} \quad (11)$$

Subsequently

$$E(y_i|x_i) = \mu_i(1-\theta_i)$$

$$Var(y_i|x_i) = (1-\theta_i)(\mu_i + \theta_i\mu_i^2)$$

<u>Zero-Inflated Negative Binomial Regression Model (ZINB)</u>

In the same way "zero-inflated Negative binomial" regression is used for modelling count data that show over-dispersion and zero counts (excess zeros). This model consider there are two types of data sources, the first source is zero type and the second is comes from data follows Negative binomial distribution.

According to Lambert (1992), response variable Y_i is independent with

$Y_i \sim 0$ with probability (θ_i) and $Y_i \sim Negative\ binomial\ (\mu_i, \frac{1}{\alpha})$ with probability $(1+\theta_i)$

Therefore,

$$\Pr(Y_i = 0) = \theta_i + (1-\theta_i) \times \Pr(Count\ process\ at\ (i) gives\ a\ zero) \quad (12)$$

by assuming the Y$_i$ follows a Negative binomial distribution with mean μ_i

$$p\left(y_i; \frac{1}{\alpha}, \mu_i | y_i \geq 0\right) = \frac{\Gamma(y_i + \frac{1}{\alpha})}{\Gamma(\frac{1}{\alpha})\Gamma(y_i+1)}\left(\frac{1}{1+\alpha\mu_i}\right)^{\frac{1}{\alpha}}\left(\frac{\alpha\mu_i}{1+\alpha\mu_i}\right)^{y_i}$$

Subsequently

The term $\Pr(Count\ process\ at\ (i) gives\ a\ zero)$ is given by

$$p\left(y_i = 0; \frac{1}{\alpha}, \mu_i \middle| y_i \geq 0\right) = \left(\frac{1}{1+\alpha\mu_i}\right)^{\frac{1}{\alpha}}$$

Hence, Equation (12) can now be written as

$$\Pr(Y_i = 0) = \theta_i + (1-\theta_i)\left(\frac{1}{1+\alpha\mu_i}\right)^{\frac{1}{\alpha}} \quad (13)$$

for the probability that Y$_i$ is a non-zero count;

$$\Pr(Y_i = y_i) = (1-\theta_i) \times \Pr(Count\ process) \qquad (14)$$

Hence, Equation (14) can be rewritten as follows

$$p(Y_i = y_i | y_i > 0) = (1-\theta_i)\frac{\Gamma\left(y_i + \frac{1}{\alpha}\right)}{\Gamma\left(\frac{1}{\alpha}\right)\Gamma(y_i+1)} \left(\frac{1}{1+\alpha\mu_i}\right)^{\frac{1}{\alpha}} \left(\frac{\alpha\mu_i}{1+\alpha\mu_i}\right)^{y_i} \qquad (15)$$

Therefore, the probability density function for a ZINB model is given by

$$P(Y_i = y_i) = \begin{cases} \theta_i + (1-\theta_i)\left(\frac{1}{1+\alpha\mu_i}\right)^{\frac{1}{\alpha}} & if\, y_i = 0 \\ (1-\theta_i)\frac{\Gamma\left(y_i+\frac{1}{\alpha}\right)}{\Gamma\left(\frac{1}{\alpha}\right)\Gamma(y_i+1)}\left(\frac{1}{1+\alpha\mu_i}\right)^{\frac{1}{\alpha}} \left(\frac{\alpha\mu_i}{1+\alpha\mu_i}\right)^{y_i} & if\, y_i > 0 \end{cases} \qquad (16)$$

By GLM[15], $\mu_i = e^{X'_i\beta_i}$, 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 of (16) is given as follow

$$\log(L) = \sum_i^n \begin{bmatrix} I(y_i=0)log\left(\theta_i + (1-\theta_i)\left(\frac{1}{1+\alpha\mu_i}\right)^{\frac{1}{\alpha}}\right) + \\ I(y_i>0)(\log((1-\theta_i) + \\ \log\left(\frac{\Gamma\left(y_i+\frac{1}{\alpha}\right)}{\Gamma\left(\frac{1}{\alpha}\right)\Gamma(y_i+1)}\right) - \\ \left(y_i+\frac{1}{\alpha}\right)\log(1+\alpha\mu_i) + \\ y_i\log(\alpha\mu_i)) \end{bmatrix} \qquad (17)$$

Subsequently

$$E(Y_i) = \mu_i(1-\theta_i)$$
$$Var(Y_i) = (1-\theta_i)(\mu_i + \alpha\mu_i^2) + \mu_i^2(\theta_i^2 + \theta_i)$$

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[19], 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

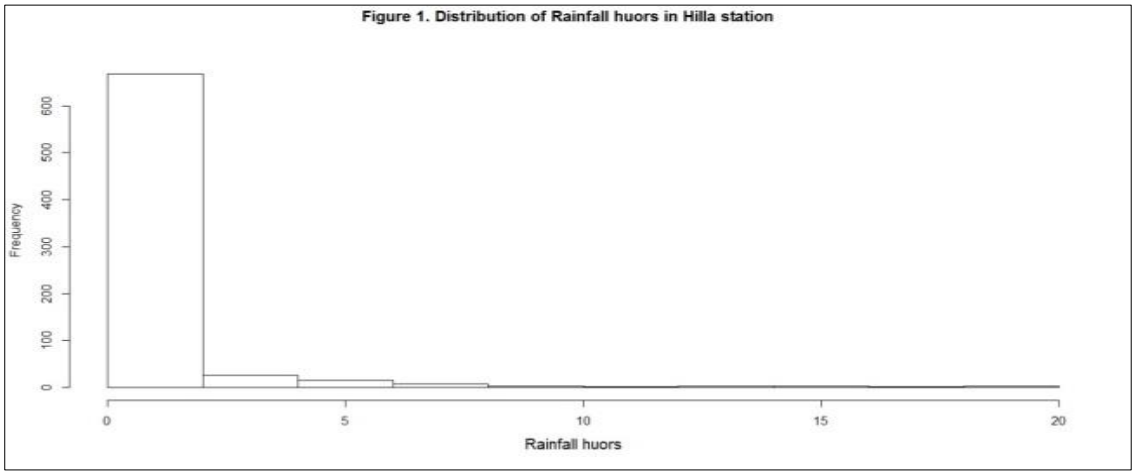
Data were collected from database of the Meteorology and Seismology Organization in Iraq for Hilla weather station. The weather station are located in central Iraq, specifically in the city of Hilla (about 116 kilometers south of Baghdad).

The count response variable of interest to be modeled "Rainfall hours" measured at Hilla 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 variable used in our count data regression models in Hilla weather station.

variables	Minimum value	First quarter	Median	Mean	Third quarter	Maximum value
Wind speed (m/s)	**0**	**0.6**	**1.4**	**1.619**	**2.3**	**9.3**
Temperature (°C)	**3**	**15.8**	**25**	**23.97**	**32.85**	**40.5**
Station pressure (1bar/1000)	**0.9908**	**1.0007**	**1.0068**	**1.0074**	**1.0131**	**1.3804**
Sea pressure (1bar/1000)	**0.9947**	**1.0046**	**1.0108**	**1.0109**	**1.0171**	**1.0287**
Humidity (%)	**17**	**31.8**	**40.6**	**44.54**	**56**	**94**
Rainfall (hours)	**0**	**0**	**0**	**0.6553**	**0**	**20**

The distribution of the number of non-rainfall hours in Hilla 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	Hilla weather station
-2Log Likelihood	1466.649
AIC	1478.649
BIC	11506.216

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

Parameter	Estimate	Standard Error	z Value	Pr > |z|
Intercept	43.305973	14.941807	2.898	0.00375
Wind speed	0.2477	0.0253	9.79	<2e-16
Temperature	0.031051	0.017090	1.817	0.06922
Station pressure	-4.573125	20.646534	-0.221	0.82471
Sea pressure	-45.002287	25.235835	-1.783	0.07454
Humidity	0.096406	0.004648	20.742	<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 Hilla weather station, Chi-Square test statistic= 579.6014 ,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	Hilla weather station
-2Log Likelihood	892.7366
AIC	906.7386
BIC	938.8995

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

Parameter	Estimate	Standard Error	z Value	Pr > |z|
Intercept	**72.08011**	**46.31368**	**1.556**	**0.12**
Wind speed	**0.43955**	**0.09276**	**4.738**	**2.15e-06**
Temperature	**-0.05902**	**0.04534**	**-1.302**	**0.193**
Station pressure	**-5.48177**	**47.6056**	**-0.115**	**0.908**
Sea pressure	**-70.94321**	**65.8155**	**-1.078**	**0.281**
Humidity	**0.0921**	**0.01372**	**6.715**	**1.88e-11**
Alpha	**0.15**	**0..0248**		

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 87.8% of the responses in Hilla weather station , have non-Rainfall days (rainfall hours are zeros)), we develop Zero-inflated regression to handle excessive number of zeros.

Zero-Inflated Regression Models

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

Zero-Inflated Poisson Regression (ZIPR) Model

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

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

criterions	Hilla weather station
-2Log Likelihood	**841**
AIC	**865.0707**
BIC	**880.5665**

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

Parameter	Estimate	Standard Error	z Value	Pr > |z|
Poisson _ Intercept	**-1.784e+00**	**3.533e+01**	**-0.05**	**0.96**
Poisson _ Wind speed	**-2.866e-02**	**2.556e-02**	**-1.122**	**0.262**
Poisson _ Temperature	**4.288e-02**	**6.296e-02**	**0.681**	**0.496**
Poisson _ Station pressure	**1.481e+02**	**5.419e+03**	**0.027**	**0.978**
Poisson _ Sea pressure	**-1.469e+02**	**5.428e+03**	**-0.027**	**0.978**
Poisson _ Humidity	**3.300e-02**	**4.345e-03**	**7.595**	**3.09e-14**
Logit _ Intercept	**-150.15360**	**54.76032**	**-2.742**	**0.00611**
Logit _ Wind speed	**-0.65697**	**0.10337**	**-6.356**	**2.07e-10**
Logit _ Temperature	**0.07375**	**0.05397**	**1.366**	**0.1718**
Logit _ Station pressure	**3.67363**	**78.0384**	**0.047**	**0.96245**
Logit _ Sea pressure	**151.68983**	**85.93739**	**1.765**	**0.07754**
Logit _ Humidity	**-0.10703**	**0.01641**	**-6.521**	**6.96e-11**

Zero-Inflated Negative Binomial Regression (ZINBR) Model

We used the same explanatory variables in both parts of the ZINBR model. The model fit statistics and estimated coefficients of ZINBR model are given in Table 8 and Table 9.

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

criterions	Hilla weather station
-2Log Likelihood	**774.8**
AIC	**800.7555**
BIC	**814.3665**

Table 9. Estimated coefficients of Zero-Inflated Negative Binomial Regression (ZINBR) model, 2016-2017 Rainfall count data in Hilla weather station

Parameter	Estimate	Standard Error	z Value	Pr > |z|
NB _ Intercept	**4.450441**	**28.89126**	**0.154**	**0.877577**
NB _ Wind speed	**-0.01452**	**0.047221**	**-0.307**	**0.758472**
NB _ Temperature	**0.025615**	**0.023359**	**1.097**	**0.272823**
NB _ Station pressure	**18.075792**	**27.86469**	**-0.627**	**0.520412**
NB _ Sea pressure	**-23.29534**	**28.50922**	**0.775**	**0.418057**
NB _ Humidity	**0.032441**	**0.007135**	**4.547**	**5.44e-06**
Logit _ Intercept	**-144.89025**	**51.27398**	**-2826**	**0.00472**
Logit _ Wind speed	**-0.66723**	**0.10557**	**-6.320**	**2.61e-10**
Logit _ Temperature	**0.07038**	**0.04973**	**1.415**	**0.15701**
Logit _ Station pressure	**-0.12663**	**31.50265**	**-0.004**	**0.99679**
Logit _ Sea pressure	**150.17569**	**59.40599**	**2.528**	**0.01147**
Logit _ Humidity	**-0.10567**	**0.01591**	**-6.643**	**3.07e-11**
Log (Alpha)	**0.937272**	**0.280696**	**3.339**	**0.000841**

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 Hilla weather station

Model	Vuong Statistic	Preferred model
ZIP vs P	**6.969103**	**ZIP**
ZIP vs NB	**1.4564090**	**NONE**
ZIP vs ZINB	**-2.579764**	**ZINB**
ZINB vs P	**7.092766**	**ZINB**
ZINB vs NB	**5.943327**	**ZINB**

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 Hilla weather station

Model	Likelihood Ratio Test (p-value)	Preferred model
P vs NB	**0.99**	**NB**

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 Hilla weather station

models	criterions		
	-2Log Likelihood	AIC	BIC
Poisson regression	**1466.649**	**1478.649**	**11506.216**
NB regression	**892.7366**	**906.7386**	**938.8995**
ZIPR	**841**	**865.0707**	**880.5665**
ZINBR	**774.8***	**800.7555***	**814.3665***

*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-Inflated Negative Binomial (ZINB) regression model was the best count data model for our data, Although it is hard to distinguish Negative Binomial, and Zero-Inflated Poisson (ZIP) 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- Broek, V., (1995), "A Score Test for Zero Inflation in a Poisson Distribution", Biometrics, Vol. 51, No. 2, PP. 738-743.

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

– Theory and Methods, Vol. 21, PP.89-109.

4- Dudley, L. P., (2003), "Using Zero-inflated Count Regression Models To Estimate The Fertility Of U. S. Women", Journal of Modern Applied Statistical Methods, Vol. 2, No. 2, PP. 371-379.

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

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

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

11- Mwalili, S.M., (2008), "The zero-inflated negative binomial regression model with correction for misclassification: An example in caries research", Statistical Methods in Medical Research, Vol. 17, PP. 123-139.

12- Naya, H. & Others, (2008), "A comparison between Poisson and zero-inflated Poisson regression models with an application to number of black spots in Corriedale sheep", Genetics Selection Evolution, Vol. 40, PP. 379-394.

13- Oliveira, M. & Others, (2016), "Zero-inflated regression models for radiation-induced chromosome aberration data: A comparative study", Biometrical Journal, Vol. 58, No. 2, PP. 259–279.

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

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

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

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

18- 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.

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

نمذجة بيانات العد لتساقط الأمطار 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

بيانات العد، بما في ذلك التعدادات الصفرية تنشأ في مجموعة متنوعة واسعة من التطبيقات، وبالتالي أصبحت نماذج العد شائعة على نطاق واسع في العديد من المجالات. وفي مجال الإحصائيات ، يمكن تعريف بيانات العد بأنها ذلك النوع من المشاهدة الذي لا يأخذ سوى قيمة الأعداد الصحيحة غير السلبية. في بعض الأحيان قد يقوم الباحثون بحساب أصفار أكثر من المتوقع. ويمكن تعريف الصفر الزائد (زيادة الاصفار) على أنه تضخم صفري. البيانات ذات الأصفار الوفيرة (الكثيرة) تحظى بشعبية خاصة في مجالات الصحة والتسويق والتمويل والاقتصاد القياسي وعلم البيئة واحصاءات مراقبة الجودة والمجالات الجغرافية والبيئية عند حساب حدوث بعض الأحداث السلوكية والطبيعية ، مثل تكرار تعاطي الكحول وتناول الأدوية وعدد السجائر المدخنة وحدوث الزلازل و هطول الأمطار ، إلخ. وقد استخدمت بعض النماذج لتحليل بيانات العد مثل نموذج بوايسون للتضخم الصفري والنموذج ثنائي الحدين السالب. في هذه الورقة ، تم استخدام نماذج بوايسون و ثنائي الحدين السالب و بوايسون للتضخم الصفري ، و ثنائي الحدين السالب للتضخم الصفري لتحليل بيانات هطول الأمطار.