Comparing Poisson-Inverse Gaussian Model and Negative Binomial Model on case study: horseshoe crabs data

G N Putri, S Nurrohmah and I Fithriani
Department of Mathematics, Faculty of Mathematics and Natural Sciences (FMIPA), Universitas Indonesia, Kampus UI Depok, Depok 16424, Indonesia

Corresponding author’s e-mail: snurrohmah@sci.ui.ac.id

Abstract. Poisson Regression analysis is commonly used for dependent variable that has non-negative value, called count data. Poisson Regression has an assumption that the mean equal to its variance. On over dispersion case where the variance is greater than mean, poisson regression is inconvenient to used because it may underestimate the standard error of regression parameters and consequently giving misleading inference. Poisson-Inverse Gaussian and Negative Binomial regression model can be used on over dispersion data. This paper will discuss about Poisson-Inverse Gaussian regression model, Negative Binomial regression model and comparing them in terms of Goodness-of-fit (GOF) statistics on case study of horseshoe crabs data. According to the result, pseudo R-squared value of P-IG regression model is greater than the Negative Binomial regression model. It shows that P-IG regression model is better than Negative Binomial regression model.

Keyword: P-IG regression model, negative binomial regression model, over dispersion, goodness-of-fit statistics.

1. Introduction
Let Y is a response variable that state the number of events of interest. The number of events always have a non-negative value, called count data. Count data is commonly modelled using a Poisson Distribution. Poisson distribution has an assumption that the mean equal to its variance. Often on count data exhibit over dispersion, hence Poisson distribution is not good to used. The Negative Binomial distribution and the Poisson-Inverse Gaussian (P-IG) distribution were proposed as alternative to Poisson distribution on over dispersion case. The Negative Binomial is a mixture of Poisson distribution and Gamma distribution while the P-IG distribution is a mixture of Poisson distribution and Inverse Gaussian distribution.

In a few studies such as insurance and medicine, the P-IG distribution has slightly longer tails than Negative Binomial distribution [1]. Hence, P-IG distribution was proposed as a good alternative for modeling an over dispersion and long-tail data than Negative Binomial distribution.

According to that, P-IG regression model is adequate to analyze a relationship between a count response data with associated covariates even on over dispersion case. The objective of this paper is to discuss about P-IG regression model Negative Binomial regression model and comparing them with case study of horseshoe crabs data.

2. Poisson-Inverse Gaussian (P-IG) regression model
The P-IG regression model can be written as a following equation

$$\mu_i = \exp(x'_i \beta) \; ; \; i = 1, 2, ..., m$$

(1)
where
\[ x : \text{the vector of covariates} \]
\[ \beta : \text{regression parameters associated with corresponding covariates} \]

The P-IG regression model has assumptions that the response variable follows P-IG distribution and covariates are fixed.

P-IG distribution is a mixture of Poisson distribution and Inverse Gaussian distribution. Let \( Y|V \) follows Poisson distribution with mean \( \mu V \), where \( V \) follows an Inverse Gaussian distribution with mean equal to 1 and dispersion parameter \( \frac{1}{\tau} \sim IG (1, \frac{1}{\tau}) \). The marginal probability density function (PDF) for \( Y \) is

\[
f(y) = \frac{\mu^y}{y!} \left( \frac{2}{\mu \tau} \right)^{0.5} \exp \left( \frac{1}{\tau} \right) (1 + 2\mu\tau)^{-y/2} K_s(\alpha), \quad y = 0, 1, 2, \ldots
\]

where \( \alpha = \frac{1}{\sqrt{1 + 2\mu\tau}} \) and \( K_s(\alpha) \) is modified Bessel function of second kind [1]. The parameters of Poisson-Inverse Gaussian distribution \( \mu \) and \( \tau \) are always non-negative value.

Log-likelihood function is often used to obtain the maximum likelihood estimators of parameter easily. Denote \( l(\beta, \tau) \) as log-likelihood function is given by

\[
l(\beta, \tau) = \sum_{i=1}^{m} \left[ y_i l(x_i|\beta) - y_i \log(y_i!) - \frac{1}{2} \log(\tau) + \frac{1}{\tau} \left( \frac{2y_i - 1}{4} \right) \log(1 + 2\tau \exp(x_i|\beta)) \right] + \sum_{i=1}^{m} \log K_{y_i-1/2} \left( \frac{1 + 2\tau \exp(x_i|\beta)}{\tau} \right)
\]

Partial derivatives of the log-likelihood function

\[
\frac{\partial l(\beta, \tau)}{\partial \beta_k} = \sum_{i=1}^{m} x_{ki} \left( y_i - \frac{R_{y_i-1/2}(\alpha)}{(1 + 2\tau \exp(x_i|\beta))^{1/2}} \exp(x_i|\beta) \right) = 0 \quad ; \quad k = 1, \ldots, n \text{ and } x_{0i} = 1
\]

\[
\frac{\partial l(\beta, \tau)}{\partial \tau} = \sum_{i=1}^{m} \left[ -\frac{1}{\tau^2} - \frac{y_i}{\tau} + \frac{R_{y_i-1/2}(\alpha)(1 + \tau \exp(x_i|\beta))}{\tau^2 (1 + 2\tau \exp(x_i|\beta))^{1/2}} \right] = 0
\]

The equations are nonlinear so the maximum log-likelihood equations cannot be solved explicitly. It is often use the Newton-Raphson to estimate the regression parameter. The procedure for Newton-Raphson iteration method are [2]

1. Define an initial estimation of \( (\beta, \tau) \) denote \( (\beta, \tau)^{(0)} \)
2. Define an estimation of \( (\beta, \tau)^{(t)} \) for \( t = 1, 2, \ldots \) by following iterative equation

\[
(\beta, \tau)^{(t+1)} = (\beta, \tau)^{(t-1)} - \left[ H \left( (\beta, \tau)^{(t-1)} \right) \right]^{-1} U \left( (\beta, \tau)^{(t-1)} \right)
\]

where
\( U \left( (\beta, \tau)^{(t-1)} \right) \) is a vector with the elements of first partial derivatives
\( H \left( (\beta, \tau)^{(t-1)} \right) \) is the Hessian matrix with the elements of second partial derivatives

3. Stop the iteration of equation (6) if \( \left\| (\beta, \tau)^{(t)} - (\beta, \tau)^{(t-1)} \right\| < 10^{-5} \)
4. The result of the optimization \( (\beta, \tau)^{(t)} \) is an estimation of \( (\beta, \tau) \)
Table 1. Variables.

| Notation | Variables |
|----------|-----------|
| Y        | A number of satellites |
| X₁       | Female crab’s colour (0 = light, 1 = dark) |
| X₂       | Female crab’s spine condition (0 = good, 1 = broken) |
| X₃       | Female crab’s carapace width (cm) |
| X₄       | Female crab’s weight (kg) |

3. Negative Binomial regression model

The Negative Binomial regression model can be written as a following equation

\[ \mu_i = \exp(x'_i \beta) ; i = 1, 2, ..., m \]  

where

- \( x \) : the vector of covariates
- \( \beta \) : regression parameters associated with corresponding covariates

The Negative Binomial regression model has assumptions that the response variable follows Negative Binomial distribution and covariates are fixed.

Negative Binomial distribution is a mixture of Poisson and Gamma distribution. Let \( Y | \Lambda \) is a random variable that follows Poisson distribution which is conditional on its mean \( \mu \). In overdispersion case, \( \lambda \) is a value of other random variable \( \Lambda \) that follows Gamma distribution. The marginal probability density function (PDF) for \( Y \) can be written as

\[ f(y) = \frac{\Gamma \left( y + \frac{1}{a} \right)}{\Gamma \left( \frac{1}{a} \right) y!} \left( \frac{\mu}{\mu + \frac{1}{a}} \right)^y \left( \frac{\frac{1}{a}}{\mu + \frac{1}{a}} \right)^{\frac{1}{a}} \]  

where \( \mu > 0 \) and \( a > 0 \) are the parameters of Negative Binomial distribution.

Log-likelihood function for Negative Binomial regression model is given by

\[ l(\beta, \tau) = \sum_{i=1}^{m} \sum_{r=0}^{y_i-1} \log(1 + a r) - \log (y_i!) + y_i (x'_i \beta) - \left( y_i + \frac{1}{a} \right) \log (1 + a \exp(x'_i \beta)) \]  

Partial derivatives of the log-likelihood function

\[ \frac{\partial l(\beta, \tau)}{\partial \beta_k} = \sum_{i=1}^{m} x_{ki} \left( y_i - \exp(x'_i \beta) \right) \left( 1 + a \exp(x'_i \beta) \right)^{-1}, \text{ dimana } x_{0i} = 1 \]  

\[ \frac{\partial l(\beta, \tau)}{\partial a} = \sum_{i=1}^{m} \sum_{r=0}^{y_i-1} \left( \frac{r}{1 + a r} + \frac{1}{a^2} \log(1 + a \exp(x'_i \beta)) \right) - \left( y_i + \frac{1}{a} \right) \frac{\exp(x'_i \beta)}{1 + a \exp(x'_i \beta)} \]  

As same as P-IG model, it uses the Newton-Raphson method to estimate the regression parameter.

4. Case study: horseshoe crabs data

4.1. Data description

The data from Agresti [4] presents 173 observations of female horseshoe crab that for each female horseshoe crab had a male crab resident in her nest called satellites. The response variable is the number of satellites, i.e. \( y = 0, 1, 2, ..., 15 \). This paper will analyze factors affecting number of satellites, such as the female crab’s color, spine condition, carapace width, and weight.

4.2. Variables

Table 1 shows some variables that used in this paper
### Table 2. Result

| Parameter | Estimation Result of P-IG | Estimation Result of NB |
|-----------|---------------------------|-------------------------|
| $\beta_0$ | -2.0311                   | -3.8868                 |
| $\beta_1$ | -0.3375                   | -0.2636                 |
| $\beta_2$ | 0.0543                    | -0.0339                 |
| $\beta_3$ | 0.0640                    | 0.1986                  |
| $\beta_4$ | 0.5679                    | -0.0886                 |
| $\tau$   | 0.8889                    | 0.9396                  |

![Figure 1](image_url)

**Figure 1.** The data, Negative Binomial distribution (blue line), and P-IG distribution (red line).

#### 4.3. Analysis of the data

The sample mean and variance for the counts of number of satellites equals 2.92 and 9.912 respectively. The variance is larger than the mean that indicates over dispersion. Here plot of a number of satellites against P-IG distribution and Negative Binomial distribution to fit an appropriate distribution to the data. The fit of Negative Binomial and P-IG distribution are shown in figure 1.

Figure 1 shows that the P-IG distribution is more fit to the data than the Negative Binomial distribution, especially on long tail. To analyze which factors affect the response variable we use P-IG regression model. Since the P-IG regression model has an equation as same as Negative Binomial regression model, so the equation that used to modeling data is

$$
\ln(\mu_i) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i}; \quad i = 1, 2, ..., 173
$$

To estimate the parameter $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4$ and $\tau$ it used the Newton-Raphson algorithm on MATLAB R2015a software. The result of the optimization, i.e., log-likelihood maximization, were table 2 shows result of the experiment, the estimation of each regression model’s parameter can be substituted to the equation (4). So, the estimated equation of P-IG regression model is

$$
\ln(\mu_i) = -2.0311 - 0.3375X_{1i} + 0.0543X_{2i} + 0.0640X_{3i} + 0.5679X_{4i}, \text{ for } i = 1, 2, ..., 173
$$

and the estimated equation of Negative Binomial regression model is

$$
\ln(\mu_i) = -3.8868 - 0.2636X_{1i} - 0.0339X_{2i} + 0.1986X_{3i} - 0.0886X_{4i}, \text{ for } i = 1, 2, ..., 173
$$
4.4. Goodness-of-fit test of P-IG regression model

4.4.1. Pseudo R-Squared statistic. The value of pseudo R-squared statistics for P-IG model equals 0.91538650, that means the model is good enough to the data.

4.4.2. Likelihood ratio test and Wald test. The value of likelihood ratio statistics for P-IG model equals 8206.44 is greater than the value of $\chi^2_{0.05.4}$. So, we reject null hypothesis and it means that there’s at least one of explanatory variables which significant to the response variable with significance level of $\alpha$. After that, Wald test is conducted to evaluate the significance of individual explanatory variables, with significance level of $\alpha$. There is only one explanatory variable, which significant to the response variable, which is female crab’s colour.

4.4.3. Significance dispersion parameter test. The value of D statistics equals 190.6164 is greater than the value of $\chi^2_{0.05.1}$. So, we reject null hypothesis and it means that equidispersion is not fulfilled and the estimation of $\tau$ is greater than 0. It is summed up as over dispersion data.

4.4.4. Comparison of Goodness-of-fit. The P-IG regression model have a greater pseudo R-squared value than the Negative Binomial regression model (0.81211088) [2]. So that the P-IG regression model is better fit than the Negative Binomial regression model to analyse the horseshoe crabs data.

5. Conclusions
This paper conclude that the Negative Binomial regression model and the P-IG regression model are reliable alternative for over dispersion data. Maximum likelihood method is used to estimate regression parameter in both models. Due to the nonlinear equation of log-likelihood function, Newton Raphson method is used to estimate regression parameter. According to the result of horseshoe crabs data, P-IG regression model is better than Negative Binomial regression model in case of its long-tail data. This conclusion is backed up by the result of comparison of Goodness-of-fit that shows the greater pseudo R-squared value than the Negative Binomial regression model.

References
[1] Zha L, Lord D and Zou Y 2016 J. Transp. Saf. Secur. 8 18–35
[2] Shafira S 2011 Estimating Parameter of Negative Binomial Distribution on Overdispersion Case (Depok: Universitas Indonesia) Undergraduate Thesis
[3] Hilbe J M 2011 Negative Binomial Regression 2nd ed (New York: Cambridge University Press)
[4] Agresti A 2007 An Introduction to Categorical Data Analysis 2nd edition (Hoboken: John Wiley & Sons)