The measure strength of predictive in the poisson regression model with regression correlation coefficient case study of maternal mortality rate in Central Java Province in 2015

K Nisa, S Nurrohmah and M Novita
Department of Mathematics, Faculty of Mathematics and Natural Sciences (FMIPA), Universitas Indonesia, Depok 16424, Indonesia
Corresponding author’s email: snurrohmah@sci.ui.ac.id

Abstract. RCC which stands for regression correlation coefficient is an alternative measure strength of predictive that can be applied to the GLM model in which the distribution of response variable is not only normal. The RCC is constructed based on the definition of correlation coefficient by using generalized linear model (GLM). So, the RCC can be defined as a value that states the strength of the relationship between response variable $Y$ and its conditional expectation given predictor variables $E[Y|X]$. The RCC is one measure of predictable power that can satisfies the property like applicability, interpretability, consistency, and affinity. In general, the explicit form of RCC on GLM is difficult to find. However, when RCC is applied to the Poisson regression model and the predictor variables are assumed to be a normal multivariate distribution, an explicit form is found. This explicit form still contains the unknown parameters derived from the Poisson regression model. Therefore, we need to find an estimate of these parameters to obtain an estimator from the RCC. The Poisson regression model which still contains the unknown parameters are estimated using maximum likelihood method. Application of regression correlation coefficient is done in case of maternal mortality rate in Central Java Province in 2015.

Keywords: Measure strength of predictive, regression correlation coefficient, poisson regression model, maternal mortality rate

1. Introduction
Regression analysis is a statistic method that is used for investigating and modeling relationship between variables [1]. One of the uses of regression analysis is to find out how well the regression model predicts the response variable based on the value of predictor variables by involving a measure. That measure is called the measure strength of predictive. The measure strength of predictive often used in linear regression models is the coefficient of determination $R^2$ [2]. The coefficient of determination $R^2$ that is used in the linear regression model has the assumption that the response variable is a continuous data following the normal distribution.

One of alternative measure on the model whose response variable distribution is not normal that is regression correlation coefficient (RCC) and the model can be used that is Generalized Linear Model (GLM) [3]. RCC is defined as a value that states the strength of the relationship between response variable $Y$ and conditional expectation $E[Y|X]$. Besides applicable in various data types on the response
variable (applicability), RCC also satisfies some other properties i.e. interpretability, consistency, and affinity [4]. Interpretability means that a measure strength of predictive must have convincing criteria or it can be interpreted in a regression model. Consistency is a measure strength of predictive able to compare between data sets different. Finally, affinity means that the measure strength of predictive does not conflict with other measure strength of predictive.

In general, the explicit form of RCC on GLM is difficult to find. However, when the response variable is assumed to be Poisson distributed and the predictor variables are assumed multivariate normal distributed, an explicit form will be found. This explicit form still contains the unknown parameters of Poisson regression model. Therefore, we need to find an estimate of these parameters to obtain an estimator from the RCC. The Poisson regression model which still contains the unknown parameters will be estimated using maximum likelihood method.

2. Materials and method

2.1. Generalized Linear Model (GLM) for count data

Generalized linear models (GLM) are generalization of the classic linear regression models [5]. The main difference from the classical linear regression model is that the distribution of the response variable does not have to be normal, as long as it is still an exponential family distribution. In addition, there is a link function as a connector between the mean response variable with the linear form of the predictor variable component. If the response variable is discrete distribution, the exponential family distribution family that can be used are Bernoulli, Poisson, Binomial, and Negative Binomial. Whereas, if the response variable is continuous distribution, it can use Normal, Inverse-Gaussian, or Gamma.

The Forming of GLM has three components [6], namely:

1. The random component is the random response variables \(Y_1, Y_2, \ldots, Y_n\), which have independent distribution in an exponential family. In addition, mean of dependent variable denoted \(\mu_i\).
2. The systematic component or the “linear predictor” is built with \(p + 1\) parameters \(\beta = (\beta_0, \beta_1, \ldots, \beta_p)^T\) and with \(p\) explanatory variables \(\eta_i = \beta_0 + \sum_{k=1}^{p} x_{ik}\beta_k, i = 1, 2, ..., n\).
3. The “link function” between the random and systematic components, namely a monotonic and differentiable function \(g\), so that \(\eta_i = g(\mu_i) = \beta_0 + \sum_{k=1}^{p} x_{ik}\beta_k, \ i = 1, 2, ..., n\).

The most used GLM for count data is the Poisson regression model. The Poisson regression model is written as follows:

\[ Y_i = \mu_i + \epsilon_i, \quad i = 1, 2, ..., n, \]

with \(Y_i\) is dependent variable that states the number of events in the form of count data based on Poisson’s distribution, \(\mu_i\) is the mean of the dependent variable, and \(\epsilon_i\) is error for the \(i\)-th observation.

Note that in the Poisson regression model, the conditional variable response of the predictor variable is assumed to be Poisson distributed. Since \(Y | X\) is Poisson distributed, mean \(\mu_i\) must exist at interval \((0, \infty)\), while the value of \(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{pi}\) lies at interval \((-\infty, \infty)\). Therefore, the link function \(g\) is needed to solve the problem, so that the values of the left and right sides of the equation at the same interval, i.e. \((-\infty, \infty)\). The appropriate link function \(g\) is a logarithm, so that it is obtained:

\[ g(\mu_i) = \ln(\mu_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{pi}. \]

\[ \mu_i = \exp(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{pi}). \]

Since \(Y | X \sim \text{Poisson}(\theta)\), then we have

\[ E[Y|X] = Var(Y|X) = \exp(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{pi}) \]

The estimation of the parameters is achieved through the maximum likelihood method.
2.2. Regression Correlation Coefficient (RCC)

2.2.1. RCC. The Regression Correlation Coefficient (RCC) can be defined as a value that states the strength of the relationship between response variable $Y$ and its conditional expectation $E[Y|X]$. RCC is a measure strength of predictive to determine predictive ability of predictor variable to the response variable. The value of this RCC lies in $[0,1]$, where values closer to 1 represent the better model.

The theorem below will prove that RCC is built on the correlation coefficient formula in GLM.

**Theorem 1** Let $Y$ and $X$ be a random variable with $0 < \text{Var}(Y) < \infty$ and $X = (X_1, X_2, ..., X_p)^T$ is a vector of random variables. Then, RCC can be rewritten as

$$RCC(Y, X) = \frac{\text{Corr}(Y, E[Y|X])}{\sqrt{\text{Var}(Y)/\text{Var}(Y)}}$$

**Proof:**

The first, it will be proved that $\text{Cov}(Y, E(Y|X))$ can be accepted as $\text{Var}(E[Y|X])$. In addition, the following definitions and properties will be used:

1. $\text{Cov}(Y, X) = E((Y - E[Y])(X - E[X]))$.
2. $\text{Corr}(Y, X) = \frac{\text{Cov}(Y, X)}{\sqrt{\text{Var}(Y)\text{Var}(X)}}$.
3. The law of iterative expectation $E[Y] = E[E[Y|X]]$.
4. The law of total variance $\text{Var}(E[Y|X]) = \text{Var}(Y) - \text{Var}(E[Y|X])$.

Based on the properties 3 above will be obtained as follows:

$$\text{Cov}(Y, E[Y|X]) = E((Y - E[Y])(E[Y|X] - E[E[Y|X]]))$$

$$= E((Y - E[Y])(E[Y|X] - E[Y]))$$

$$= E(Y, E[Y|X]) - Y, E[Y] - E[Y]. E[Y|X] + (E[Y])^2$$

$$= E(Y, E[Y|X]) - E(Y, E[Y]) - E(E[Y].E[Y|X]) + E(E[Y]^2)$$

$$= E(Y, E[Y|X]) - E(Y^2) - E[Y]E(E[Y|X]) + (E[Y])^2$$

$$= E(Y, E[Y|X]) - E[Y]E(Y|X)$$

$$= E((E[Y|X]^2) - (E(E[Y|X]))^2 = \text{Var}(E[Y|X]).$$

Based on the definition that $E[Y|X] = \int_{-\infty}^{\infty} y f(x, y)/f_1(x) dy$ which is a function of $X$, it can be proven that $E(Y, E[Y|X]) = E((E[Y|X])^2)$.

$$E(Y, E[Y|X]) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y, E[Y|X]f(x, y)dy dx$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y \left( \int_{-\infty}^{\infty} \frac{f(x, y)}{f_1(x)} dy \right) f(x, y) dy dx$$

(2)
\[ \text{Next, with use properties 2, properties 4 and equation 1, then will be obtained} \]

\[ \text{Corr}(Y, E[Y|X]) = \frac{\text{Var}(E[Y|X])}{\sqrt{\text{Var}(Y)\text{Var}(E[Y|X])}} = \frac{\sqrt{\text{Var}^2(Y)\text{Var}(E[Y|X])}}{\text{Var}(Y)} = \sqrt{1 - \frac{E[\text{Var}(Y|X)]}{\text{Var}(Y)}} \]

2.2.2. The Properties of RCC

- From Theorem 1 it can be seen that the RCC is the ratio of the expectation \( E[\text{Var}(Y|X)] \) and \( \text{Var}(Y) \). \( E[\text{Var}(Y|X)] \) is a small value that makes the RCC value close to 1 implies the regression model estimates the response variable well. This interpretation allows the RCC to be included in the nature of interpretability.

- Since RCC is constructed from the correlation coefficient, where RCC is applied to the GLM which assumes exponential family distributed response variables, this measure can be applied to various response variable type and the value of RCC has the meaning. Therefore, the RCC satisfies property of applicability.

- Based on Kullback-Leibler Information [7], AIC (Akaike Information Criteria) is a popular measure strength of predictive. The model with the smallest AIC value will be selected as the model. But, it is no guarantee that the model is the best. Based on these limitations, it says that AIC is a relative measure. This is different from the measure of the RCC, which RCC value can be compared for different data sets. The ability to compare between different sets of data is included in properties of the consistency.

- Based on the previous section in Theorem 1 that The RCC is an alternative measure strength of predictive built from the correlation coefficient on the GLM. So this can be interpreted that the measure of the RCC can satisfy the property of affinity which does not conflict with other measure strength of predictive.

2.3. Apply the RCC to the Poisson regression model

In this section, RCC will be applied to the Poisson regression model. In general, the explicit form of RCC on GLM is difficult to find. But, when the response variable is assumed to be Poisson distributed and the predictor variables \( X \) are assumed multivariate normal distributed, an explicit form of RCC will be obtained. First, we will look for explicit forms of \( E(Y) \) and \( \text{var}(Y) \) in order to obtain an explicit form of RCC.

**Proposition 1** Let \( \Sigma \) be a positive definite matrix. Assume that \( Y|X \) follows a Poisson regression model and a vector of the predictor variables \( X = (X_1, X_2, ..., X_p)^T \) has a multivariate normal distribution \( N_p(\mu, \Sigma) \). Then

\[ E(Y) = E(Y|X) = \exp(\alpha + \beta^T \mu + \frac{1}{2}\beta^T \Sigma \beta) \]
\[ \text{Var}(Y) = E(Y\{1 - E(Y)(\exp(\beta^T \Sigma \beta) - 1) \}. \quad (4) \]

**Sketch of Proof:**

The expectation value \( E[Y] \) is written as

\[ E(Y) = E(E(Y|X)) = E(\exp(\alpha + \beta^T X)) = \exp(\alpha) E(\exp(\beta^T X)). \quad (5) \]

Note that

\[ E[\exp(\beta^T X)] = M_X(\beta) = \exp \left( \frac{1}{2} (2\beta^T \mu + \beta^T \Sigma \beta) \right), \quad (6) \]

where \( M_X(\cdot) \) is moment generating function of \( X \). By substituting equation 5 and equation 6, we obtain the first statement.

Based on the fact that \( E[Y|X] = \text{Var}(Y|X) \), will be obtained

\[ \text{Var}(Y) = E[Y] + \text{Var}(E[Y|X]) \quad (7) \]

Based on the law of iterative expectations, then

\[ \text{Var}(E[Y|X]) = E[E[Y|X]^2] - (E(Y))^2 \quad (8) \]

with

\[ E[E[Y|X]^2] = E[(\exp(\alpha + \beta^T X))^2] = \exp(2\alpha) M_X(2\beta) \]

\[ (E(Y))^2 = \{\exp(\alpha) E(\exp(\beta^T X))^2\} = \exp(2\alpha) (M_X(\beta))^2. \]

So, by substituting equation 6 and equation 8, we obtain

\[
\begin{align*}
\text{Var}(Y) &= E[Y] + \text{Var}(E[Y|X]) \\
&= E[Y] + \exp(2\alpha) \left( M_X(2\beta) - (M_X(\beta))^2 \right) \\
&= E[Y] + E[Y]^2 (\exp(\beta^T X) - 1) \\
&= E(Y\{1 - E(Y(\exp(\beta^T \Sigma \beta - 1) \}. \]
\]

**Theorem 2** Let \( \Sigma, Y, \) and \( X \) be as in Proposition 1. Then, \( \text{RCC}(Y, X) = \text{RCC}(Y, X; \alpha, \beta) \) for a Poisson regression model is

\[
\text{RCC}(Y, X; \alpha, \beta) = \sqrt{ \frac{\exp \left( \alpha + \beta^T \mu + \frac{1}{2} \beta^T \Sigma \beta \right) (\exp(\beta^T \Sigma \beta) - 1)}{1 + \exp \left( \alpha + \beta^T \mu + \frac{1}{2} \beta^T \Sigma \beta \right) (\exp(\beta^T \Sigma \beta) - 1) \}} \quad (9)
\]

**Proof:**

\[
\text{RCC}(Y, X) = \sqrt{ \frac{\text{Var}(E[Y|X])}{\text{Var}(Y)} } = \sqrt{ \frac{E[Y]^2 (\exp(\beta^T X) - 1)}{E(Y)\{1 - E(Y(\exp(\beta^T \Sigma \beta - 1) \}}} \]

We propose an estimator of the RCC using the explicit form. By substituting the maximum likelihood estimators \( \hat{\alpha} \) and \( \hat{\beta} \) into \( \alpha \) and \( \beta \) in (9), respectively, we get the following estimator of the RCC for the Poisson regression model:
3. Results and discussion

In this study, the data will be applied using maternal mortality data in the Health Service Profile of Central Java Province in 2015 [8]. Maternal mortality is the number of maternal deaths that occur due to the process of pregnancy, childbirth and childbed. This data is in the form of count, so model that can be used is Poisson regression model. In addition, the software used to perform data analysis is R program version 3.4.3 [9].

Here are the variables used in the study:
1. Response variable (Y) is maternal mortality rate in Central Java Province in 2015 at 35 regency / city.
2. Predictable variables (measured in percentage)
   - \(X_1\) is the number of pregnant mother who received health services at least 4 times a visit.
   - \(X_2\) is the number of pregnant mother given Fe3 tablets.
   - \(X_3\) is the number of postpartum mother receiving health care.
   - \(X_4\) is the number of pregnant mother who have neonatal complications, namely infant health problems in the womb.
   - \(X_5\) is the number of PHBS that is people who live clean and healthy behavior.
   - \(X_6\) is the number of active Integrated Healthcare Center (posyandu) managed by the community.

The multiple Poisson regression model is

\[
\hat{R} = \frac{\exp \left( \alpha + \mu^T \hat{\beta} + \frac{1}{2} \hat{\beta}^T \Sigma \hat{\beta} \right) (\exp(\hat{\beta}^T \Sigma \hat{\beta}) - 1)}{\sqrt{1 + \exp \left( \alpha + \mu^T \hat{\beta} + \frac{1}{2} \hat{\beta}^T \Sigma \hat{\beta} \right) (\exp(\hat{\beta}^T \Sigma \hat{\beta}) - 1)}}
\]

Using the program R version 3.4.3 obtained the maximum likelihood estimation of the parameters. So the estimation of the Poisson regression model is:

\[
\log E(Y_i | X_i = x_i) = \log (\mu_i(x_i)) = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{5i} + \beta_6 x_{6i}
\]

Furthermore, a multivariate-normal test is performed. Multivariate-normal test is needed because in regression correlation coefficient, must be assumed that the predictor variables are multivariate-normal distributed. The test used is the Mardia test. Using the R program, obtained the results of a multivariate-normal test that the decision of predictor variables \(X_1, X_2, \ldots, X_6\) are normal multivariate distribution. This is shown from the figure 1 that the distribution of plots close to the model (straight line), which means that the data is normal distributed.

Next, the significance of the model is tested to determine whether the model can be used to describe the relationship between maternal mortality and the predictor variables, the number of pregnant mother who received health services and given Fe3 tablets, the number of postpartum mother receiving health services, the number of pregnant mother who have neonatal complications, the number PHBS, and the number of active posyandu. The significance test for all models uses the likelihood ratio test and for each parameter uses the Wald test. Based on the significance test at the significance level \(\alpha = 0.05\) the number of PHBS and the number of active posyandu has contributed to maternal mortality.

Regression correlation coefficient (RCC) is used to measure of the predictive strength on Poinsson regression model. The value of RCC approaching one can be stated that the predictive power of the model is strong. Based on the test of significance, it has been found that the predictor variables that have
contributed to maternal mortality are $X_5$ and $X_6$. To fulfil the purpose of this study, that explains the RCC can satisfies of interpretability, consistency, and affinity. Three possible combination of Poisson regression model will be made and then the value of predictive strength ability will be search. From the R program, obtained the results the parameter estimation of $\beta_5$ and $\beta_6$ for each model. Thus, 3 possible Poisson regression models are formed

1. $\log(\hat{\mu}_1(x_{5i}, x_{6i})) = 2.599153 - 0.005071 x_{5i} - 0.009963 x_{6i}$
2. $\log(\hat{\mu}_1(x_{5i})) = 2.463452 - 0.011266 x_{5i}$
3. $\log(\hat{\mu}_1(x_{6i})) = 2.288100 - 0.011204 x_{6i}$

Based on the result of R program, $R$ value which is estimator of regression correlation coefficient from each model above summarized in table 1.

From table 1 it can be seen that the R value for all models has a low value that is below 0.5. A low R value implies that the predictive power of each predictor variable in the model is low. Thus, it can be concluded that the three models are not exactly describes the relationship between maternal mortality with the number of PHBS and the number of active posyandu.

In addition to the RCC values, AIC value can also be used to measure strength of predictive. The model that is chosen is the model with the smallest AIC value. The AIC values for each model can be seen in the output of the R program summarized in table 2.

![Chi-Square Q-Q Plot]

**Figure 1.** Q-Q plot normal of $X_1, X_2, ..., X_6$

### Table 1. Result of $R$ values for maternal mortality models

|                              | $\log(\hat{\mu}_1(x_{5i}, x_{6i}))$ | $\log(\hat{\mu}_1(x_{5i}))$ | $\log(\hat{\mu}_1(x_{6i}))$ | $R$  |
|------------------------------|--------------------------------------|--------------------------------|--------------------------------|------|
|                              | 0.4098438                            | 0.2866707                      | 0.4359659                      | 0.4359659 |

### Table 2. Result of AIC value for maternal mortality models

|                              | $\log(\hat{\mu}_1(x_{5i}, x_{6i}))$ | $\log(\hat{\mu}_1(x_{5i}))$ | $\log(\hat{\mu}_1(x_{6i}))$ | AIC  |
|------------------------------|--------------------------------------|--------------------------------|--------------------------------|------|
|                              | 220.54                               | 223.44                         | 219.04                         |      |
From table 2 it can be seen that the value for the third model has the smallest value compared to others, so according to the AIC value the third model will be chosen to illustrate the relationship between maternal mortality with the number of PHBS and number of active posyandu.

Note, based on table 1 and table 2 have the same model selection order rate that are the third, first and second models. This means that the RCC as a measure strength of predictive has an affinity property that is not contrary to other measure in choosing a model.

Although the smallest AIC value can choose the third model to describe maternal mortality data, the smallest AIC values and the difference in AIC values between the models do not explain anything to the model. That is, there is no guarantee that the model is best. Unlike the RCC value, from table 1 the RCC value for the third model is 0.43 which means 43% of the predictor variable's ability in the model in predicting the response variable. So, using the value of RCC in choosing a model will give the right decision. This implies that the RCC has interpretability property and consistency.

Note that in the Poisson regression model, there is an assumption that must be satisfied, that is the assumption of equidispersion where the value of variance is equal to the average. However, in the analysis of data counts, these assumptions are often not met, i.e., the value of variance is greater than the average or is called overdispersion. The overdispersion makes the Poisson regression model formed to be less precise to use [10]. The RCC for all three models in table 1 have low value maybe because of overdispersion. However, in this study overdispersion testing in not conducted.

4. Conclusion
Determination of the explicit form of regression correlation coefficient (RCC) on Poisson regression model by first determining the explicit form of $E[Y]$ and $Var(Y)$. In the process of determining the explicit form, it is assumed that the predictor variable to be multivariate normal distribution, so that the explicit form can be obtained.

The regression correlation coefficient as an alternative measure strength of predictive has the advantage of being able to fulfill several properties such as interpretability, applicability, consistency, and affinity.

Application of regression correlation coefficient is done in case of maternal mortality rate in Central Java in 2015 and obtained the conclusion that Poisson regression model that formed less precise to describe data of mother mortality rate because RCC of the model has low value so that the ability of predictive variable in the regression model is low.

References
[1] Montgomery D C, Peck E A dan Vining G G 2001 *Introduction to Linear Regression Analysis* (United States: John Wiley & Sons)
[2] Takahashi A and Kurosawa T 2015 *Comput. Stat. Data An.* **98** 71-8
[3] Nelder J A and Wedderburn R W M 1972 *J. R. Statist. Soc. A* **135** 370-84
[4] Zheng B and Agresti A 2000 *Statist. Med.* **19** 1771-81
[5] Covrig M, Mircea J, Zbaganu G, Coser A and Tindeche A 2015 *Romanian Statistical Review* **63** 33-45
[6] de Jong P and Heller G Z 2008 *Generalized Linear Models For Insurance Data* (United States: Cambridge University Press)
[7] Akaike H 1973 *Information Theory and Extension of The Maximum Likelihood Principle* (New York: Springer)
[8] Dinas Kesehatan Provinsi Jawa Tengah 2015 *Profil Kesehatan Provinsi Jawa Tengah Tahun 2015* available at http://dinkesjatengprov.go.id/v2018/dokumen/profil2015/mobile/index.html
[9] R Core Team 2015 *R: A Language and Environment for Statistical Computing* available at https://www.gbif.org/tool/81287/r-a-language-and-environment-for-statistical-computing
[10] Wang W and Famoye F 1997 *J. Popul. Econ.* **10** 273-83