Modified Regression Correlation Coefficient for Poisson Regression Model

Nattacha Kaengthong, and Uthumporn Domthong

Department of Statistics, Faculty of Science, Khon Kaen University, Khon Kaen 40002, Thailand

E-mail: k_nattacha@kkumail.com, uthudo@kku.ac.th

Abstract. This study gives attention to indicators in predictive power of the Generalized Linear Model (GLM) which are widely used; however, often having some restrictions. We are interested in regression correlation coefficient for a Poisson regression model. This is a measure of predictive power, and defined by the relationship between the dependent variable \( Y \) and the expected value of the dependent variable given the independent variables \( E(Y | X) \) for the Poisson regression model. The dependent variable is distributed as Poisson. The purpose of this research was modifying regression correlation coefficient for Poisson regression model. We also compare the proposed modified regression correlation coefficient with the traditional regression correlation coefficient in the case of two or more independent variables, and having multicollinearity in independent variables. The result shows that the proposed regression correlation coefficient is better than the traditional regression correlation coefficient based on Bias and the Root Mean Square Error (RMSE).

1. Introduction

Statistical model is constructed to analyze the relation of the dependent variables and the independent variables when these variables were related. In order to obtain the appropriate statistical model for analysis and forecast, there must be an indicator of the predictive power of the model. In this paper, we focus on the Generalized Linear Model (GLM) which is one of indicators in predictive power. There are widely used the predictive power of the statistical model; for example, Multiple Correlation Coefficient, Coefficient of Multiple Determination, Adjustment Coefficient of Multiple Determination, Regression Correlation Coefficient, AIC, etc. However, these statistical power constraints often have some limitations for statistical models. The predictive power indicators for generic linear models have been continually developed to suit certain limitations of the statistical model. Through the works was proposed by Zheng and Agresti (2000), the so-called regression correlation coefficient (RCC). In addition, measures of predictive power such as the Adjustments for \( R^2 \)-measures for Poisson regression model were proposed by Mittlbock and Waldhor (2000). Entropy correlation coefficient (ECC) and entropy coefficient of determination (ECD), which rely on Kullback-Leibler information, were proposed by Eshina and Tabata (2007) and Eshina and Tabata (2010). Detrended Cross Correlation Analysis (DCCA) were proposed by Kristoufek (2014). Regression correlation coefficient (RCC) for Poisson regression model were proposed by Takahashi and Kurosawa (2016). These predictive power indicators work as well as the original prediction power indicators. In this work, we focus on the basic measure of power forecasting. RCC is the population value determined by the relationship between the dependent variable \( Y \) and the conditional expected value of \( Y \) given \( X \) \( E(Y | X) \) for GLM.
We have developed a regression correlation coefficient for Poisson regression model to work well in the model with multicollinearity independent variables. In section 3, we then compared our proposed regression correlation coefficients with the traditional regression correlation coefficients in terms of Bias and Root Mean Square Error (RMSE) by using the true value. The simulation results were shown in section 4. The conclusions were summarized in section 5.

2. Adjusted $R^2$

The adjusted multiplier coefficients was developed from the coefficients of multiple determinants. Adding more independent variables into the regression model, it always appears that the Coefficient of Multiple Determination ($R^2$) is higher. However, this does not mean that this regression models are perform well. On the other hand, it may lead us to have model with high Mean Square Error (MSE). Therefore, to reduce the Bias value, the $R^2_{adj}$, ranging between 0 and 1, is proposed as the following:

$$R^2_{adj} = 1 - \frac{(n - 1) \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}{(n - p - 1) \sum_{i=1}^{n} (Y_i - \bar{Y})^2}$$ (1)

When $Y_i$ is the true value of the variable $i$; $i=1,2,3,...,n$, $\hat{Y}_i$ is the estimator value of the variable $i$; $i=1,2,3,...,n$, $\bar{Y}$ is the mean of the variables, $n$ is sample size , and $p$ is the number of independent variables.

3. Regression Correlation Coefficient (RCC)

The RCC is used to measure the predictive power of the model. The value show the relationship between ($Y$) and conditional expectation of $Y$ given $X$ [\(E(Y|X)\)] where $X$ is an independent variable. The value RCC close to 1 indicates that the model is appropriate; on the other hand, if this value closes to 0, it indicates that the model is improper.

3.1 Properties of the RCC

1.) The RCC, introduced by Zheng and Agresti (2000), is defined as following:

$$RCC(Y, X) = cor(Y, E(Y|X)).$$

where $Y$ is dependent variable, and $X$ is independent variable vector, size $(p \times 1)$.

2.) RCC is ranged between 0 to 1.

3.) RCC can be rewritten as

$$RCC(Y, X) = \left(1 - \frac{E(Var(Y|X))}{Var(Y)}\right)^{1/2}.$$

Estimators of RCC, proposed by Zheng and Agresti (2000), are

- Sample correlation estimator, $\hat{R}_{cor} = Cor(Y, \hat{Y})$.
- Jack-knife estimator, $\hat{R}_{jack} = Cor(Y, \hat{Y})_{jack} = N Cor(Y, \hat{Y}) - (N - 1)Cor(Y, \hat{Y})(.)$,

where $Cor(Y, \hat{Y})(.) = \frac{1}{N} \sum_{i=1}^{N} Cor(Y_i, \hat{Y}_i)$, and $Cor(Y, \hat{Y})^{(-i)}$ is a correlation without the $i^{th}$ element of the sample.
• Adjustment Jack-knife estimator is defined as $\hat{R}_{\text{jack}}^0 = \text{Cor}(Y, \hat{Y})^0_{\text{jack}} = \max \left[ 0, \text{Cor}(Y, \hat{Y})_{\text{jack}} \right].$

• Leave-one-out cross validation estimator is defined as $\hat{R}_{\text{crs}} = \text{Cor}(Y, \hat{Y}_{\text{crs}}).$

Vector $\hat{Y}_{\text{crs}}$ is $\hat{Y}^{(1)}, \hat{Y}^{(2)}, \ldots, \hat{Y}^{(N)}$ where $\hat{Y}^{(i)}$ is estimated value of $Y_i$ using the maximum estimates $\hat{\alpha}$, and $\hat{\beta}$ is obtained by deleting the $i^{th}$ observation.

### 3.2 RCC for poisson regression model

The RCC for poisson regression model, proposed by Takahashi and Kurosawa (2016), was developed from RCC for GLM model [2]. To get the form of the RCC, the expected value of $Y$ ($E(Y)$) and the variance of $Y$ ($\text{Var}(Y)$) are defined in proposition 1.

**Proposition 1** Let $\Sigma$ be a positive definite matrix. Assume that $Y \mid X$ follows a Poisson regression model, and a vector of the predictor variables $X$ has a multivariate normal distribution $N \left( \mu, \Sigma \right)$. Then, the expected value of $Y$ ($E(Y)$) and the variance of $Y$ ($\text{Var}(Y)$) are defined as following [2]:

$$
E(Y) = E(E(Y \mid X)) = \exp(\alpha + \mu^T \beta + \frac{1}{2} \beta^T \Sigma \beta),
$$

$$
\text{Var}(Y) = E(Y) \left\{ 1 + E(Y)(\exp(\beta^T \Sigma \beta) - 1) \right\}.
$$

**Theorem 2** Let $\Sigma$, $Y$, and $X$ be defined as proposition 1. Then, $\text{RCC}(Y, X) = \text{RCC}(Y, X; \alpha, \beta)$ for Poisson regression model is [2]

$$
\text{RCC}(Y, X; \alpha, \beta) = \left( \frac{\exp(\alpha + \mu^T \beta + \frac{1}{2} \beta^T \Sigma \beta)(\exp(\beta^T \Sigma \beta) - 1)}{1 + \exp(\alpha + \mu^T \beta + \frac{1}{2} \beta^T \Sigma \beta)(\exp(\beta^T \Sigma \beta) - 1)} \right)^{1/2}.
$$

Then, the estimator of RCC for Poisson regression model use the explicit form (4) where $\mu$ is the vector of the mean of the independent variables $X$, size $(p \times 1)$. $\Sigma$ is the matrix covariance of independent variables $X$, size $(p \times p)$. $\alpha$ is the intersection of the scale regression model, size $(1 \times 1)$. $\beta$ is the vector of regression coefficients, size $(p \times 1)$ [2]. Thus, the estimator $\hat{R}$ of the RCC for Poisson regression model is

$$
\hat{R} = \left( \frac{\exp(\hat{\alpha} + \hat{\mu}^T \hat{\beta} + \frac{1}{2} \hat{\beta}^T \hat{\Sigma} \hat{\beta})(\exp(\hat{\beta}^T \hat{\Sigma} \hat{\beta}) - 1)}{1 + \exp(\hat{\alpha} + \hat{\mu}^T \hat{\beta} + \frac{1}{2} \hat{\beta}^T \hat{\Sigma} \hat{\beta})(\exp(\hat{\beta}^T \hat{\Sigma} \hat{\beta}) - 1)} \right)^{1/2}.
$$

Additionally, the new estimators for $E(Y)$ and $\text{Var}(Y)$ are as following:

$$
\hat{E}(Y) = \exp(\hat{\alpha} + \hat{\mu}^T \hat{\beta} + \frac{1}{2} \hat{\beta}^T \hat{\Sigma} \hat{\beta}),
$$

$$
\hat{\text{Var}}(Y) = \hat{E}(Y) \left\{ 1 + \hat{E}(Y)(\exp(\hat{\beta}^T \hat{\Sigma} \hat{\beta}) - 1) \right\}.
$$
3.3 Modified RCC for Poisson Regression Model

Adjustment of the approximation of the regression coefficients of the Poisson regression model have two approaches. The first approach is to adjust RCC when two or more variables are added into the model. The second approach is to adjust RCC when independent variables have multicollinearity under Poisson model. The guidelines for modification are as follow.

We have improved the RCC estimator of Takahashi and Kurosawa (2016) by using the term $(n-p-1)/(n-1)$. According to section 2, the proposed modified regression correlation coefficient was created by the idea from the adjusted $R^2$ that used the term $(n-p-1)/(n-1)$ which is less than 1 and multiply to the existing term. $R^*$ will be approximated as

$$R^* = \frac{(n-p-1)}{(n-1)} \left[ \frac{E(Y)(\exp(\beta^T \Sigma \beta) - 1)}{1 + E(Y)(\exp(\beta^T \Sigma \beta) - 1)} \right]^{1/2}$$

where $n$ is sample size, and $p$ is the number of independent variables.

Replacing $E(Y)$ and $\text{Var}(Y)$ for Poisson regression model from (2), (3) [2], we have

$$R^* = \frac{(n-p-1)}{(n-1)} \left[ \frac{\exp(\alpha + \mu^T \beta + \frac{1}{2} \beta^T \Sigma \beta)(\exp(\beta^T \Sigma \beta) - 1)}{1 + \exp(\alpha + \mu^T \beta + \frac{1}{2} \beta^T \Sigma \beta)(\exp(\beta^T \Sigma \beta) - 1)} \right]^{1/2}$$

We propose a new estimator by substituting the maximum likelihood estimators $\hat{\alpha}$ and $\hat{\beta}$ into $\alpha$ and $\beta$ in (9), respectively. Therefore, have the following new estimator $\hat{R}^*$ of the RCC.

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

If $\mu$ and $\Sigma$ in (6), (7), and (10) are unknown, we use the sample average, and sample covariance matrix of $X$ for $\mu$ and $\Sigma$; respectively.

4. Simulation Study

We compare the estimator of the traditional regression correlation coefficient [2] with our proposed estimator regression correlation coefficient ($R^*$) by using R program version 3.3.3 for simulation. Poisson regression model is defined as

$$\log(E(Y | X)) = \alpha + \beta X; \quad Y | X \sim P(\theta), \quad X_y \sim N(0,1)$$

where $\theta = E(Y | X)$. The number of independent variables ($p$) are 2, 4, and 6. The parameters $\alpha$ is set as 0.1, 0.5, 0.8, 1, and $\beta_i = 0.3, \beta_2 = 0.03, \beta_3 = 0.04, \beta_4 = -0.02, \beta_5 = -0.002, \beta_6 = 0.0004$. The correlation of independent variables ($\rho$) is 0.40, 0.70, 0.90, and 0.99. The sample size ($n$) is 50, 100, 250, and 500.
The criteria, used to compare the performance of the estimated regression coefficient, is Bias and RMSE. The number of Monte Carlo replications is 5000 times.

We first compare the estimators \( \hat{R} \) and \( \hat{R}' \) of the RCC for poisson regression model. Table 1 shows the Bias and RMSE for two estimators with large sample size, \( n = 500 \). We found that if \( \alpha \) is increased, RMSE value decreases. If the number of independent variables increases, RMSE value increases. However, \( \hat{R} \) has larger Bias and RMSE than \( \hat{R}' \) for all situations.

### Table 1 Comparison of the estimators \( \hat{R} \) and \( \hat{R}' \) of the RCC in terms of Bias and RMSE when \( \alpha = 0.1, 0.5, 0.8, 1 \) (\( \rho = 0.99, n = 500 \)).

| \( \alpha \) | 0.1 | 0.5 | 0.8 | 1 |
|-------------|-----|-----|-----|---|
| \( p = 2 \) |     |     |     |   |
| Bias        | \( \hat{R} \) | 0.0589 | 0.0408 | 0.0519 | 0.0278 |
|             | \( \hat{R}' \) | 0.0587 | 0.0407 | 0.0517 | 0.0277 |
| RMSE        | \( \hat{R} \) | 0.0585 | 0.0602 | 0.0594 | 0.0600 |
|             | \( \hat{R}' \) | 0.0583 | 0.0599 | 0.0592 | 0.0598 |
| \( p = 4 \) |     |     |     |   |
| Bias        | \( \hat{R} \) | 0.0684 | 0.0603 | 0.0166 | 0.0778 |
|             | \( \hat{R}' \) | 0.0678 | 0.0598 | 0.0164 | 0.0771 |
| RMSE        | \( \hat{R} \) | 0.0623 | 0.0620 | 0.0621 | 0.0611 |
|             | \( \hat{R}' \) | 0.0618 | 0.0615 | 0.0616 | 0.0606 |
| \( p = 6 \) |     |     |     |   |
| Bias        | \( \hat{R} \) | 0.0732 | 0.0648 | 0.0303 | 0.0433 |
|             | \( \hat{R}' \) | 0.0723 | 0.0641 | 0.0300 | 0.0428 |
| RMSE        | \( \hat{R} \) | 0.0658 | 0.0651 | 0.0648 | 0.0639 |
|             | \( \hat{R}' \) | 0.0650 | 0.0643 | 0.0640 | 0.0631 |

Table 2 shows that given \( \alpha = 1, \rho = 0.70, p = 2 \), the Bias and RMSE of \( \hat{R}' \) is lower than those of \( \hat{R} \) in all cases of sample size. Similarly, in table 3, given \( \alpha = 0.5, \rho = 0.90, p = 4 \), the Bias and RMSE of \( \hat{R}' \) is lower than those of \( \hat{R} \) in all cases of sample size.

### Table 2 Comparison of the estimators of the RCC in terms of Bias and RMSE according to sample size (\( \alpha = 1, \rho = 0.70, p = 2 \)).

| \( N \) | \( \hat{R} \) | \( \hat{R}' \) | \( \hat{R} \) | \( \hat{R}' \) |
|---------|--------------|--------------|--------------|--------------|
| 50      | 0.1569       | 0.1505       | 0.1710       | 0.1640       |
| 100     | 0.1661       | 0.1628       | 0.1592       | 0.1559       |
| 250     | 0.1672       | 0.1658       | 0.1526       | 0.1514       |
| 500     | 0.1468       | 0.1462       | 0.1505       | 0.1499       |

### Table 3 Comparison of the estimators of the RCC in terms of Bias and RMSE according to sample size (\( \alpha = 0.5, \rho = 0.90, p = 4 \)).

| \( N \) | \( \hat{R} \) | \( \hat{R}' \) | \( \hat{R} \) | \( \hat{R}' \) |
|---------|--------------|--------------|--------------|--------------|
| 50      | 0.1909       | 0.1753       | 0.1794       | 0.1648       |
| 100     | 0.1508       | 0.1447       | 0.1508       | 0.1447       |
| 250     | 0.0961       | 0.0945       | 0.1300       | 0.1279       |
| 500     | 0.1149       | 0.1140       | 0.1227       | 0.1218       |

Table 4 shows that given \( \alpha = 1, n = 500, p = 6 \), both RMSE and Bias of \( \hat{R}' \) is less than those of \( \hat{R} \) in all cases of \( \rho \).

Table 4
Table 4 Comparison of the estimators of the RCC in terms of Bias and RMSE according to $\rho$ ($\alpha = 1, n = 500, p = 6$).

| $\rho$ | Bias $\hat{R}$ | Bias $\hat{R}'$ | RMSE $\hat{R}$ | RMSE $\hat{R}'$ |
|-------|----------------|-----------------|----------------|----------------|
| 0.4   | 0.1148         | 0.1134          | 0.1233         | 0.1218         |
| 0.7   | 0.1433         | 0.1415          | 0.1542         | 0.1524         |
| 0.9   | 0.1120         | 0.1107          | 0.1299         | 0.1284         |
| 0.99  | 0.0433         | 0.0428          | 0.0639         | 0.0631         |

5. Conclusions
In this paper, we proposed a new regression correlation coefficient for Poisson regression model, and compared performance with the traditional regression correlation coefficient by using the simulation study [2]. Based on the results of the simulation study, it was found that when the number of independent variables is 2, 4, and 6, the sample size of 50, 100, 250, and 500, the correlation of the independent variables equals 0.4, 0.7, 0.9, and 0.99, and $\alpha$ equals 0.1, 0.5, 0.8, and 1, then new RCC outperforms the traditional RCC. The Bias of our proposed RCC is lower than the traditional RCC in all cases of simulation study. When independent variables and the correlation of independent variables increase, the RMSE of $\hat{R}'$ decreases, and smaller than RMSE of $\hat{R}$. In all scenario that we show in this study, it is confirm that the proposed $\hat{R}'$ perfectly outperform the $\hat{R}$.

Acknowledgments
The author is grateful to thesis advisor for her helpful comments, and thanks to scholarships from the Graduate School of Khon Kaen University.

References
[1] Zheng, B. and Agresti, A. 2000 Summarizing the predictive power of a generalized linear model Stat. Med. 19 p 1771-1781
[2] Takahashi, A. And Kurosawa, T. 2016 Regression correlation coefficient for a Poisson regression model Comput. Statist. Data Anal. 98 p 71-78
[3] Mittlbock, M. and Waldhor, T. 2000 Adjustments for $R^2$ -measures for Poisson regression model Comput. Statist. Data Anal. 34 p 461-472
[4] Eshima, N. and Tabata, M. 2007 Entropy correlation coefficient for measuring predictive power of generalized linear model Statist. Probab. Let. 77 p 588-593
[5] Eshima, N. And Tabata, M. 2010 Entropy coefficient determination for generalized linear model Comput. Statist. Data Anal. 54 p 1381-1389
[6] Ladislav Kristoufek 2014 Measuring correlations between non-stationary series with DCCA coefficient Physica. A. 402 p 291-298