A new version of almost unbiased estimator for the Poisson regression model

Sahar Farook Ibrahim
Department of Mathematics, College of Education for Pure Sciences, University Of Anbar, Iraq, sah21u2003@uoanbar.edu.iq

Mustafa I. Alheety
Department of Mathematics, College of Education for Pure Sciences, University Of Anbar, Iraq

Abstract

When model of Poisson regression's explanatory variables consist of highly correlated, the widely employed maximum-likelihood estimator is unstable with high variance. As a result, the biased estimators are used to decrease the maximum-likelihood instability and it also produces a reduction of the mean squared error matrix. For this reason, in this paper, a new biased estimator is suggested and the statistical properties are found. The performance of suggested estimator is studied and compared with other exist estimators using the mean squared error as a criterion for goodness of fit. In order to support the preference of the new estimator in the theoretical aspect, a real example was given represented by fatality resulting from traffic accidents in Sweden.

Keywords: *Maximum likelihood estimator, Poisson regression, Traffic accidents, Mean squared error matrix, PKL estimator.*

1. INTRODUCTION

The Poisson regression model (PRM) is employed when the outcome variable comes in the form of a count variable. It is applicable in different fields. Examples include the number of patents, takeover bids, bank failures, accident insurance and criminal careers, number of deaths, number of defects and others [5,7].

Most commonly, to estimate the regression coefficient for PRM , the Maximum Likelihood estimator (EML) method is used. In addition, when explanatory variables are correlated, the EML suffers from instability. Effects of Multicollinearity consist of a high correlation coefficient and its variance, neglecting t, in addition to a sizeable R-squared. To solve this problem, possible alternatives like biased estimators are suggested like Manson and others [8] are presented a Poisson regression estimator PRM. The ridge estimator in PRM was introduced by Månsson and Shukur [9]. The modified Liu estimator for PRM was presented by Türkan and Özel [11]. Two new parameters for PRM were developed by Asar and Genç [3]. Recently, the Poisson KL estimator was developed by Lukman et al.[6].

According to that, in this paper, a modified Kibria-Lukman estimator is proposed to solve the problem of multicollinearity in PRM by using the concept of almost unbiasdness. In sec. 2 the methodology for statistics has been given by introducing the concept of Poisson

regression model and the proposed estimator with its properties . Also, the efficiency of proposed the estimator is investigated by comparing it with some biased estimators. In Sec. 3 a real data set as an application for applying Poisson regression model as well as to observe the ability of the proposed estimator under some conditions is given.

2. Methodology for statistics

2-1 Modeling Poisson regression using the maximum likelihood estimator

The model of Poisson regression is used whenever dependent variable y_i is distributed as numerical information and has the form P(μ_i), where μ is a parameter of the Poisson distribution. The Poisson model's mean response function is $\mu_i = exp(x_i\beta)$, where x_i is the i^{th} row of X, which is a $n \times (p + 1)$ data matrix with p explanatory variables and is a $(p + 1) \times 1$ vector of coefficients. It is estimated using the conventional ML . According to this model's log probability,

$$L(\beta; y) = \left(\sum_{i=1}^{n} e^{x_i\beta}\right) + \sum_{i=1}^{n} y_i \log\left(e^{x_i\beta}\right) + log\ \left(\prod_{i=1}^{n} y_i\ !\right) \qquad (1)$$

As a result of solving $L(\beta;\ y)$ with regard to:

$$\frac{\partial L}{\partial \beta} = \sum_{i=1}^{n}(y_i - e^{x_i\beta}))x_i = 0$$

The ML is now calculated using the iteratively reweighted least squares ($IRLS$) technique and

$$\hat{\beta} = (X'\hat{L}\ X)^{-1}(X'\hat{L}\alpha) = (\ S)^{-1}\acute{X}\hat{L}\alpha \quad , \qquad (2)$$

where $S = \acute{X}\hat{L}X,\ \hat{L} = diag[\hat{\mu}_i]$ and α being the column vector, and

$$\alpha = log\ \hat{\mu}_i + \frac{y_i - \hat{\mu}_i}{\hat{\mu}_i}\ .$$

The $\hat{\beta}$ is asymptotically unbiased estimator . When there is a significant correlation between the explanatory variables, the matrix S is ill-conditioned, leading to ML estimator to be instability and has high variance. For this reason, Mansson and Shukur (2011) presented the Poisson ridge estimator (PR) to address this problem.

$$\hat{\beta}_{PR}(k) = (S + kI_P)^{-1}S\hat{\beta}, k > 0 \qquad (3)$$

The Poisson K-L estimator (PKL) was presented in the following manner by Lukman et al. (2021) as an extension of this concept:

$$\hat{\beta}_{PKL}(k) = (S + kI_P)^{-1}(S - kI_P)\hat{\beta} \qquad . \qquad (4)$$

We can get the PKL estimator as follows:

$$\text{Let } \hat{\beta}_{\text{PKL}}(\text{k}) = \hat{\beta}_{PR}\ (k) + Bias\ \left(\hat{\beta}_{PR}(k)\right)$$

$$= (S + kI_P)^{-1}S\hat{\beta} - k(s + kI_P)^{-1}\beta$$

Since β is unknown, we estimated by $\hat{\beta}$ and therefore,

$$\hat{\beta}_{\text{PKL}}(\text{k}) = (S + kI_P)^{-1}(S - kI_P)\hat{\beta}$$

Another fashion for PKL estimator can be given as:

$$\hat{\beta}_{PKL}(k) = S_k^{-1}(S - kI_P)\hat{\beta}$$

$$= S_k^{-1}(S + kI_P - kI_P + kI_P)\hat{\beta}$$

$$= [I_P - 2kS_k^{-1}]\hat{\beta}$$

$$= T_k\ \hat{\beta}\ , k > 0\ ,$$

where $T_k = [\ I_P - 2kS_k^{-1}]$.

2-2 The proposed estimator

It is known to statisticians that, the biased estimators like PKL estimator have significant amount of bias. For this reason, there is a good chance to reduce the bias in order to increase the efficiency of the biased estimators. Therefore, researchers solve this

problem , by making a correction for bias that significantly has lower amount of bias comparing to biased estimator without using this procedure.

Now , we can say that the idea for the proposed estimator is to suggest a new estimator depending on the PKL estimator by making a correction for bias of PKL estimator. We called the Almost Unbiased PKL estimator (AUPKLE) and has the following form:

$\hat{\beta}_{AUPKLE} = F_k \hat{\beta}$, k > 0 (5)

where $F_k = [I_P - 4k^2(S + kI_P)^{-2}]$, k > 0 .

We can rewrite Eq. (5) as:

$Bias(\hat{\beta}_{PKL}(k)) = -2k(S + kI_P)^{-1}\beta$

Hence, by Kadiyala [1], corrected biased of $\hat{\beta}_{PKL}$ can be defined as follows: $\tilde{\beta}_{AUPKLE} = \hat{\beta}_{PKL}(k) + 2k(S + kI_P)^{-1}\beta$

So, in accordance with Ohtani [1], we substitute the unknown parameter β by $\hat{\beta}_{PKL}$ to obtain the proposed estimator :

$$\tilde{\beta}_{AUPKLE} = [I_P + 2k(S + kI_P)^{-1}]\hat{\beta}_{PKL}$$
$$= [I_P - 4k^2(S + kI_P)^{-2}]\hat{\beta}$$

The following are the AUPKLE's properties :

$E(\hat{\beta}_{AUPKLE}) = F_k\beta,$

where $F_d = I_P - 4k^2(S + kI_P)^{-2}.$

The bias of the (AUPKLE) :

$$Bias(\hat{\beta}_{AUPKLE}) = (F_k - I_P)\beta$$
$$= [(I_P - 4k^2(S + kI_P)^{-2}) - I_P]\beta$$
$$= -4k^2(S + kI_P)^{-2}\beta$$
$$= B_1^* \quad (6)$$

The(AUPKLE) variance covariance matrix is given as:

$Cov(\hat{\beta}_{AUPKLE}) = F_k S^{-1}F_k$ (7)

2-3 The properties of AUPKLE estimator:

The mean squared error matrix (MSE) and the scalar mean square error (SMSE) of AUPKLE estimator is given as follows:

$MSE(\hat{\beta}_{AUPKLE}) = (I_P - 4k^2(S + kI_P)^{-2})\, S^{-1}(I_P - 4k^2(\acute{S} + kI_P)^{-2}) + B_1\acute{B}_1$, (8)

where $B_1 = 4k^2\,(S + kI_P)^{-2}\,\beta$.

The SMSE is given as follows:

$SMSE(\tilde{\beta}) = tr(MSE(\tilde{\beta}))$ (9)

where, *tr* is the sum of the diagonal elements of the square matrix .

So,

$SMSE(\hat{\beta}_{AUPKLE}) = \sum_{i=1}^{P}(\frac{(\mu_i+k)^2-4k^2}{(\mu_i+k)^2})^2.\frac{1}{\mu_i} + 4^2k^4\sum_{i=1}^{P}\frac{\beta_i^2}{(\mu_i+k)^4}$, (10)

where μ_i is the i[th] eigen value of S.

According to Asar and Genc¸ [2] , the MSE and $SMSE$ of PKL:

$MSE(\hat{\beta}_{PKL}(k)) = [I_P - 2k(S + kI_P)^{-1}]S^{-1}[\, I_P - 2k(\, S + kI_P)^{-1}] + B_2\acute{B}_2$,

$SMSE(\hat{\beta}_{PKL}(k)) = \sum_{i=1}^{P}(\frac{\mu_i-k}{\mu_i+k})^2.\frac{1}{\lambda_i} + \sum_{i=1}^{P}(\frac{4k^2}{\mu_i+k)^2})$,

where $B_2 = Bias\left(\hat{\beta}_{PKL}(k)\right) = (S + kI_P)^{-1}2k.$

The MSE and SMS of the MLE are defined accordingly as follows:

$$MSE(\hat{\beta}) = S^{-1}.$$

$$SMSE(\hat{\beta}) = \sum_{i=1}^{P} \frac{1}{\mu_i}$$

2-4 The efficiency of AUPKLE estimator

2-4-1. Comparing the EML with AUMPKLE

The comparison of EML and AUPKLE utilizing the mean squared error (MSE) matrix is shown:

$$MSE(\hat{\beta}_{AUPKLE}) = F_k S^{-1} F_k + B_1 B_1'$$

To demonstrate that, we state the following theorem.

Theorem 2.1. Under MSE criterion, when k < λ_i , the AUPKLE is superior to EML, namely:

MSE($\hat{\beta}$) − MSE($\hat{\beta}_{AUPKLE}$) ≥ 0 if and only if: $B_1'[S^{-1} - F_k S^{-1} F_k]B_1 \leq 1$

Proof :

$$\Delta^*_1 = MSE(\hat{\beta}) - MSE(\hat{\beta}_{AUPKLE})$$

$$= S^{-1} - (F_k S^{-1} F_k + B_1 B_1')$$

$$= D_1^* - B_1 B_1' ,$$

where $D_1^* = S^{-1} - F_k S^{-1} F_k$.

Let $D_1^* = P\gamma P' = Pdiag\{\gamma_1, \dots\dots\dots, \gamma_P\} P'$, where $\gamma_i = \frac{1-(1-4k^2(\mu_i+k)^{-2})^2}{\mu_i}$, $i = 1, \dots\dots, p$ and P is the eigen vectors of S.

We can get, $(1 - 4k^2(\mu_i + k)^{-2}) < 1$; when k < μ_i , then $1 - (1 - 4k^2(\mu_i + k)^{-2})^2 > 0$

$\therefore \gamma_i > 0$, $\forall i$.

This indicates that D_1^* is positive definite. The following Lemma can be used to complete the proof.

Lemma 2.1 (See Farebrother, [4]). Let M be a positive definite matrix and α be a vector, then $M - \alpha\alpha' \geq 0$ if and only if $\alpha' M^{-1}\alpha \leq 1$.

As a result, the proof is finished by using Lemma 2.1.

2-4-2 Comparison between the PKL and AUPKLE estimators

The following methods are used to determine PKL's properties:

$$Bias(\hat{\beta}_{PKL}(k)) = -2k(S + kI_P)^{-1}\beta$$

$$= B_2$$

And

$$Cov(\hat{\beta}_{PKL}(k)) = T_k S^{-1} T_k$$

The MSE for the PKL is as follows:

$$MSE(\hat{\beta}_{PKL}(k)) = T_k S^{-1} T_k + B_2 \acute{B}_2 \ . \quad (11)$$

The comparison between PKL and $AUPKLE$ is illustrated by the subsequent theorem.

Theorem 2.2. For > 0 , under Poisson regression model, the $AUPKLE$ is better than PKL in terms of MSE if and only if :

$B_1^T D_2^{-1} B_1 \leq 1$

Proof: Let $\Delta_2 = MSE(\hat{\beta}_{PKL}(k)) - MSE(\hat{\beta}_{AUPKLE})$

$$= PD_2\acute{P} + B_2\acute{B}_2 - B_1\acute{B}_1$$

$$= P\ diag\left\{\frac{(\mu_i-k)^2}{\mu_i(\mu_i+k)^2} - \frac{(1-4k^2(\mu_i+k)^{-2})^2}{\mu_i}\right\}_{i=1}^{P} \acute{P} + B_2\acute{B}_2 - B_1\acute{B}_1,$$

where $D_2 = [\, I_P - 2k(\Lambda + kI_P)^{-1}]\Lambda^{-1}[I_P - 2k(\Lambda + kI_P)^{-1}] - [1 - 4k^2(\Lambda + kI_P)^{-2}]\Lambda^{-1}[1 - 4k^2\,(\Lambda + kI_P)^{-2}]$

$B_2\acute{B}_2$ is nonnegative definite, thus we concentrate on the quantity.

$$\frac{(\mu_i - k)^2}{\mu_i(\mu_i + k)^2} - \frac{[1 - 4k^2(\mu_i + k)^{-2}]^2}{\mu_i}$$

Therefore, D_2 is positive definite if

$$\frac{(\mu_i - k)^2}{\mu_i(\mu_i + k)^2} \geq \frac{[1 - 4k^2(\mu_i + k)^{-2}]^2}{\mu_i}$$

After the above expression has been simplified, we get

$$(\mu_i - k)^2 + (\mu_i + k)^{-2}\,\frac{k}{\mu_i} \geq 0$$

Since $k > 0$ and $\mu_i > 0$, The proof is completed after using Lemma 2.1.

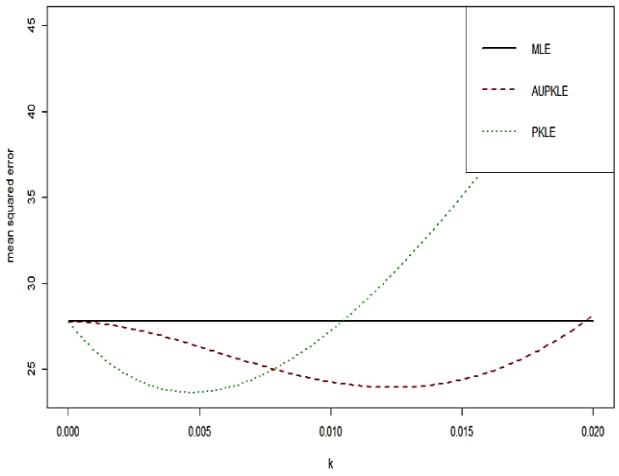
3. Application

In this section, the performance of the proposed estimator for this paper is illustrated by using the Swedish traffic. The fatality statistics for 2019 are analyzed. We classify all into three different age (15-24 years, 25-64 years and bigger than 64 years) [10]. The finding results of the model are given in Table I

Table I. SMSE of the EML, PKLE and AUPKLE

k	EML	PKLE	AUPKLE
0.0079	27.75217	25.05549	24.96747
0.0091	27.75217	26.20149	24.51436
0.0112	27.75217	28.82665	24.00333
0.0119	27.75217	29.85099	23.93426
0.0127	27.75217	31.10052	23.92721
0.0134	27.75217	32.25691	23.98831
0.0159	27.75217	36.79437	24.76774
0.0176	27.75217	40.17494	25.83732
0.0196	27.75217	44.38280	27.68610

Figure I. Empirical estimated SMSE of the EML, PKLE and AUPKLE.



The eigenvalues of X′X matrix are 2010, 223.2, 186.54, 14.157, 0.581, 0.224 and 0.047. The condition index, $\mathrm{CI} = \sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}} = 207.77$, which means that severe issues of multicollinearity exist. For this reason, we use the biased estimators to reduce the effect of this problem on estimation. For that, we can see form Table 1, when we add a small value k , the SMS for PKLE and AUPKLE are starting for decreasing comparing to MLE. Also, the performance of the proposed estimator AUPKLE is better compared to PKLE estimator for some small values of k and we can observe that by looking for Figure1.

Reference

[1]-Alheety, Mustafa I, Qasim, M., Mansson, K., Kibria, , B. M. G., (2021). Modifed almost unbiased two-parameter estimator for the Poisson regression model with an application to accident data. SORT, 45, (2), 121-142.

[2]-Asar, Y. and Genc¸, A. (2018). A new two-parameter estimator for the Poisson regression model. Iranian Journal of Science and Technology, Transactions A: Science, 42, 793-803

[3]- Asar, Y., Eris¸oglu, M. and Arashi, M. (2017). Developing a restricted two-parameter Liu- type estimator: A comparison of restricted estimators in the binary logistic regression model. Communications in Statistics-Theory and Methods, 46, 6864-6873.

[4]- Farebrother, R. W. (1976). Further results on the mean square error of ridge regression. Journal of the Royal Statistical Society. Series B (Methodological), 38, 248-250.

[5]- K. Månsson, and G. Shukur, A Poisson ridge regression estimator. Econ. Model 28 (2011), pp. 1475–1481.

[6]-Lukman AF, Adewuyi MK, Kibria BMG: A new estimator for the multicollinear Poisson regression model: simulation and application. Sci Rep. 2021; 11:3732.

[7]-M. Amin, M.N. Akram, and M. Amanullah, On the James-Stein estimator for the poisson regression model. Commun. Stat. -Simul. Comput.(2020).

[8]- Mansson, K. (2012). On ridge estimators for the negative binomial regression model. Economic Modelling, 29, 178-184.

[9]- Mansson, K., and Shukur, G. (2011). A Poisson ridge regression estimator. Economic Modelling, 28, 1475-1481.

[10]-Stipdonk, H., Bijleveld, F., Van Norden, Y. and Commandeur, J. (2013). Analysing the development of road safety using demographic data. Accident Analysis and Prevention, 60, 435-444.

[11]-Türkan, S. & Özel, G. (2016). A new modified Jackknifed estimator for the Poisson regression model. J. Appl. Stat. 43, 1892–1905.

[12]-Wiklund, M., Simonsson, L. and Forsman, A. (2012). Traffic safety and economic fluctuation: long-term and short-term analyses and a literature survey.