Research Article

Poisson Regression-Based Mean Estimator

Usman Shahzad,1 Shabnam Shahzadi,2 Noureen Afshan,1 Nadia H. Al-Noor,3 David Anekeya Alilah,4 Muhammad Hanif,1 and Malik Muhammad Anas1

1Department of Mathematics and Statistics, PMAS-Arid Agriculture University, Rawalpindi, Pakistan
2Department of Mathematics and Big Data, Anhui University of Science and Technology, Huainan, China
3Department of Mathematics, College of Science, Mustansiriyah University, Baghdad, Iraq
4Department of Mathematics, Masinde Muliro University of Science and Technology, Kakamega, Kenya

Correspondence should be addressed to David Anekeya Alilah; aliladavid2010@gmail.com

Received 15 July 2021; Accepted 7 October 2021; Published 18 October 2021

Copyright © 2021 Usman Shahzad et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The most frequent method for modeling count responses in numerous investigations is the Poisson regression model. Under simple random sampling, this paper offers utilizing Poisson regression-based mean estimator and discovers its associated formula of the mean square error (MSE). The MSE of the proposed estimator is compared to the MSE of traditional ratio estimators in theory. As a result of these evaluations, the proposed estimator has been proven to be more efficient than traditional estimators. Furthermore, the practical results corroborated the theoretical findings.

1. Introduction

The emphasis on using supplemental/auxiliary data to improve estimate precision might be the criterion that distinguishes sample survey theory from other statistical theories. Auxiliary data are used in almost every important phase of a sample survey, including stratification, selection probability creation, and the formula for the population parameter estimator of interest. The assessment of the total or mean of a population is a common goal of most surveys. Utilizing auxiliary data to improve the estimation of the mean population has been considered by numerous authors, such as Upadhyaya et al. [1], Upadhyaya and Singh [2], Koyuncu [3], Shahzad [4], Abid et al. [5], Shahzad et al. [6–8], Zaman and Bulut [9, 10], Ali et al. [11], and Zaman [12].

When a positive correlation between values of an auxiliary variable with the study variable is available, the ratio-type and regression estimators are good choices for mean estimation. In sampling theory, auxiliary variable’s population information, like the coefficient of variation or kurtosis, is frequently employed to improve the efficiency of the estimate for a population mean for ratio estimators. On the other hand, the presence of outliers or extreme values in the data degrades the efficiency of traditional estimation methods. For minimizing the impact of outliers in ratio-type mean estimators, Kadilar et al. [13] presented a robust regression strategy based on the Huber-M approach. For the mean estimation with a simple random sampling (SRS) technique, Oral and Kadilar [14, 15] used two approaches that are modified maximum likelihood and its integrated method. Through combining the ratio estimators given in Zaman and Bulut [9], Zaman [12] created a new class of robust ratio-type estimators. Using robust regression estimates as well as robust covariance matrices under stratified random sampling, Zaman and Bulut [10] proposed novel regression-type estimators. More recently, with the case of sensitive research under SRS, Ali et al. [11] proposed a class of robust regression-type estimators. Furthermore, with the count data, when the mean is large enough, it is inconvenient to apply a linear regression model to such data, even though the Poisson distribution converges to the normal distribution. Negative prediction
values are possible because the linear model links the predicted value with auxiliary (or explanatory or independent) variables. Also, in linear regression, the validity of hypothesis tests is contingent on the assumption of constant variance of the study (or response or dependent) variable. For count data, these assumptions are invalid. As a result, the Poisson regression model is the most commonly employed approach for modeling count data in the applied sciences.

In Poisson regression, the study variable \( Y_i \) (\( y_i = 0, 1, 2, \ldots \)) is the number of events that occur at a particular period, with a Poisson distribution given by

\[
P(y_i; \lambda_i) = \frac{e^{-\lambda_i} \lambda_i^{y_i}}{y_i!}, \quad \lambda > 0, \tag{1}
\]

and its mean and variance are both the same, \( E(Y_i) = \text{var}(Y_i) = \lambda_i \).

The natural log-likelihood function is defined as follows:

\[
\mathcal{L}(\lambda; y) = \sum_{i=1}^{n} (y_i \ln(\lambda_i) - \lambda_i - \ln(y_i!)). \tag{2}
\]

Let \( X \) be the explanatory variable matrix of order \( n \times (k + 1) \). Then, the relationship between \( Y_i \) and \( i^{th} \) row of matrix \( X, x_i \) associating with \( g(\lambda_i) \), is

\[
\ln(\lambda_i) = \xi_i = x_i^T A, \tag{3}
\]

where \( A = (A_{0i}, A_1, \ldots, A_k) \) are the coefficients of regression parameters. Such a model is well known as the Poisson regression model. \( \hat{A} \) represents the maximum likelihood estimator of \( A \), and it may be obtained by differentiating (3) with respect to \( A \).

\[
\sum_{i=1}^{n} (y_i - \exp(x_i^T A))x_i = 0. \tag{4}
\]

Iterative approaches such as algorithms of Fisher scoring and Newton–Raphson are employed to solve these \( k \) equations (see Cameron and Trivedi [16], Montgomery et al. [17], and Koç [18]).

This article focused on utilizing Poisson regression to estimate the parameter mean of the study variable under SRS. The rest of the article is constructed as follows. First, we introduce the notations and review some existing ratio-type mean estimators with their MSE. Second, we define a novel mean estimator based on Poisson regression and its MSE. Furthermore, the proposed estimator’s condition efficiency is inspected theoretically. Then, using numerical illustrations based on two real datasets, we examine the relative efficiency of the proposed estimator over the adapted estimators. Finally, some concluding remarks are introduced.

2. Some Existing Ratio-Type Estimators of Mean with Their MSE

This section outlines some of the existing population mean estimators that employed SRS and relied on known information on the auxiliary variable’s conventional parameters to improve the mean estimators’ efficiency. Before delving into the specifics of the existing population mean estimators, the notations used are as follows.

\( N \): population size; \( n \): sample size.

\( f \): sample ratio, \( f = n/N \).

\( \overline{Y}, \overline{X} \): population means associated with study and auxiliary variables \( (Y, X) \), respectively.

\( \overline{Y}, \overline{X} \): sample means associated with \( Y \) and \( X \), respectively.

\( C_Y, C_X \): population coefficients of variation associated with \( Y \) and \( X \), respectively.

\( \delta(x) \): population coefficient of the kurtosis associated with \( X \).

\( S_Y^2, S_X^2 \): population variance associated with \( Y \) and \( X \), respectively.

\( S_{xy}, s_{xy} \): sample variance associated with \( Y \) and \( X \), respectively.

\( S_{xy}, s_{xy} \): population and sample covariance between \( X \) and \( Y \), respectively.

\( \rho \): correlation coefficient between \( X \) and \( Y \).

\( B_L, B_C \): coefficient of slope attained by the least squares method, \( B_L = S_{xy}/S_X^2 \) and \( B_C = s_{xy}/s_X^2 \).

As one of the important estimators for the population mean \( \overline{Y} \) in SRS, with assumed known, Kadilar and Cinги [19] proposed the following ratio estimators inspired by Sisodia and Dwivedi [20] and Upadhyaya et al. [1]:

\[
\overline{Y}_{KC1} = \frac{\overline{Y} + b_L(\overline{X} - \overline{X})}{\overline{X}}, \tag{5}
\]

\[
\overline{Y}_{KC2} = \frac{\overline{Y} + b_C(\overline{X} - \overline{X})}{\overline{X} + C_X},
\]

\[
\overline{Y}_{KC3} = \frac{\overline{Y} + b_L(\overline{X} - \overline{X})}{\overline{X} + \delta(x)}, \tag{6}
\]

\[
\overline{Y}_{KC4} = \frac{\overline{Y} + b_C(\overline{X} - \overline{X})}{\overline{X} + \delta(x)} [\overline{X} + \delta(x) + C_X],
\]

\[
\overline{Y}_{KC5} = \frac{\overline{Y} + b_L(\overline{X} - \overline{X})}{\overline{X}C_X + \delta(x)} [\overline{XC} + \delta(x)].
\]

Kadilar and Cinги [19] also provided the following formula for the MSE associated with their estimators:

\[
\text{MSE}(\overline{Y}_{KC}) = \frac{1-f}{n} \left[ \frac{R_{KC}^2 S_Y^2 + 2B_L R_{KC} S_Y^2}{S_X^2} + B_C^2 (S_Y^2 - 2R_{KC} s_{xy} - 2B_L s_{xy} + S_Y^2) \right],
\]

\[
i = 1, 2, \ldots, 5,
\]

where the population ratios \( R_{KC}, l = 1, 2, \ldots, 5 \) can be gained as follows:
\[ \mathcal{R}_{KC1} = \mathcal{R} = \frac{Y}{X}, \]
\[ \mathcal{R}_{KC2} = \frac{Y}{X + C_x}, \]
\[ \mathcal{R}_{KC3} = \frac{Y}{X + \delta(x)}, \]
\[ \mathcal{R}_{KC4} = \frac{Y \delta(x)}{X \delta(x) + C_x}, \]
\[ \mathcal{R}_{KC5} = \frac{Y C_x}{X C_x + \delta(x)}. \]  
(7)

Recently, Koç [18] offered his idea to improve the aforementioned estimators, introducing new ratio estimators based on Poisson regression as follows:
\[ \mathcal{R}_{H1} = \frac{Y + b_p (X - \bar{x})}{X}, \]
\[ \mathcal{R}_{H2} = \frac{Y + b_p (X - \bar{x})}{X + C_x}, \]
\[ \mathcal{R}_{H3} = \frac{Y + b_p (X - \bar{x})}{X + \delta(x)}, \]
\[ \mathcal{R}_{H4} = \frac{Y + b_p (X - \bar{x})}{X \delta(x) + C_x}, \]
\[ \mathcal{R}_{H5} = \frac{Y + b_p (X - \bar{x})}{X C_x + \delta(x)}, \]  
(8)

where \( b_p \) represents the coefficient of slopes attained by Poisson regression.

Koç [18] also provided the following formula for the MSE associated with his suggested estimators that is the same as the MSE equation in (6), but it is evident that \( B_\ell \) should be changed by \( B_p \), whose value is acquired from the Poisson regression model.

\[
\text{MSE}(\bar{Y}_H) = \frac{1 - f}{n} \left[ \frac{\mathcal{R}_{KCI}^2 S_x^2 + 2 B_p \mathcal{R}_{KCI} S_x^2}{\mathcal{R}_{KCI}^2 S_x^2 + 2 B_p \mathcal{R}_{KCI} S_x^2 + B_p^2 \mathcal{R}_{KCI} S_{xy} - 2 B_p S_{xy} + S_y^2} \right],
\]
(9)

In addition, Koç [18] demonstrated that his estimators are more efficient than Kadilar and Cingi [19] estimators if any of the following conditions are satisfied:
\[
B_p - B_\ell < -2 \mathcal{R}_{KCI},
\]
\[
B_p - B_\ell > -2 \mathcal{R}_{KCI}.
\]  
(10)

2.1. Proposed Estimator Based on Poisson Regression and Its MSE. The newly constructed estimator of the mean population can be arranged in the frameworks of Zaman and Bulut [10] and Ali et al. [11]. But we are implementing their frameworks based on Poisson regression as
\[
\mathcal{R}_U = \mathcal{R} = \frac{Y}{X}.
\]  
(11)

Furthermore, taking advantage of established results, with some basic algebra, and eliminating tedious or futile calculations, we mention the MSE expression of the proposed mean estimator as
\[
\text{MSE}(\bar{Y}_U) = \frac{1 - f}{n} \left[ Y^2 C_y^2 - 2 B_p X Y \rho_{xy} C_x C_y + B_p^2 \mathcal{R}_{KCI} S_{xy}^2 \right].
\]  
(12)

2.2. Efficiency Comparisons. The efficiency condition for the proposed mean estimator may be found by comparing the MSE of the proposed estimator in (12) to the MSE of Haydar’s estimators in (9):
\[
\text{MSE}(\bar{Y}_U) < \text{MSE}(\bar{Y}_H), \quad i = 1, 2, \ldots, 5,
\]
\[
Y^2 C_y^2 - 2 B_p X Y \rho_{xy} C_x C_y + B_p^2 \mathcal{R}_{KCI} S_{xy}^2 < \mathcal{R}_{KCI} S_x^2 + 2 B_p \mathcal{R}_{KCI} S_x^2 + B_p^2 \mathcal{R}_{KCI} S_{xy}^2 - 2 B_p S_{xy} + S_y^2.
\]  
(13)

Now
\[ S_y^2 - 2 B_p S_{xy} + B_p^2 S_x^2 < R_{KG1}^2 S_x^2 + 2 B_p R_{KG1} S_x^2 + B_p^2 S_x^2 - 2 B_p S_{xy} + S_y^2, \]
\[ 2 R_{KG1} S_{xy} - 2 R_{KD1} S_x^2 < 2 B_p R_{KG1} S_x^2, \]
\[ (14) \]
\[ 2 R_{KG1} S_{xy} - 2 R_{KD1} S_x^2 < B_p S_{xy} + S_y^2, \]
\[ \frac{S_{xy}}{S_x^2} < 2 R_{KG1} < B_p. \]

The proposed Poisson mean estimator \( \hat{y} = \hat{y} + b_p (\bar{X} - \bar{Y}) \) is more efficient than Haydar’s estimators, if condition \( B_p > (S_{xy}/S_x^2) - (1/2) R_{KG1} \) is satisfied.

2.3. Numerical Illustrations. Here, we use two real datasets to evaluate the performance of proposed and existing estimators.

Population I (Pop-I). We consider the dataset collected between 2006 and 2010 from the Afyon Respiratory Disease Hospital and the Afyon Environmental Department Air Pollution Unit, which was used in [18]. The number of patients admitted to the hospital on a weekly basis was taken as the dependent variable \( Y \), and PM10 was taken as the explanatory variable \( X \).

Population II (Pop-II). We consider the dataset obtained from TUIK of 81 provinces in 2019 used in [18]. The number of people who died due to traffic accidents was taken as the dependent variable \( Y \), and the number of motor vehicles was taken as the explanatory variable \( X \).

For these datasets, using SRS, we consider \( n = 40 \) and \( n = 20 \). The characteristics of the two populations, as well as the values of population ratios, are given in Tables 1 and 2.

Based on the MSE values of the proposed estimator and reviewed estimators, we calculate the relative efficiency (RE) values of the proposed estimators with respect to reviewed estimators, say \( (\hat{Y}_i = \hat{Y}_{KCI}, \hat{Y}_{HI}, i = 1, 2, \ldots, 5) \), as follows:

\[ R(\hat{Y}_i) = \frac{\text{MSE}(\hat{Y}_i)}{\text{MSE}(\hat{Y})}, \quad i = 1, 2, \ldots, 5. \]  

Tables 3 and 4 give the outcomes for the relative efficiencies.

After assessing the relative efficiency values, in Tables 3 and 4, it is observed that all values exceed 100. As a result, the proposed Poisson regression estimator is more efficient than the reviewed estimators. On the other hand, it is observed that all relative efficiency values of proposed estimator with respect to Kadilar and Cingi [19] estimators are greater than corresponding values with respect to Haydar estimators, and thus

\[ R(\hat{Y}_i) > R(\hat{Y}_{HI}) > R(\hat{Y}_{KCI}), \quad i = 1, 2, \ldots, 5. \]  

Table 1: Characteristics of two populations (Pop-I and Pop-II).

| Notation | Pop-I | Pop-II |
|----------|-------|--------|
| \( N \)  | 213   | 81     |
| \( \eta \) | 40    | 20     |
| \( \bar{Y} \) | 4.676056338 | 82.40740741 |
| \( \bar{X} \) | 116.1915 | 274308 |
| \( C_p \) | 0.633555159 | 1.02550449 |
| \( C_s \) | 0.439846 | 1.90331 |
| \( \delta(\bar{x}) \) | -0.6908 | 35.77509 |
| \( S_y \) | 2.962539616 | 84.50913823 |
| \( S_x \) | 51.10636 | 522093 |
| \( S_{xy} \) | 53.51706 | 38124520.6 |
| \( \rho_{xy} \) | 0.353 | 0.864077723 |
| \( B_p \) | 0.02049002 | 0.02049002 |
| \( B_p \) | 0.004 | 6.131E - 07 |

Table 2: Values of population ratios.

| Ratio     | Pop-I          | Pop-II          |
|-----------|----------------|-----------------|
| \( R_{KCI} \) | 0.040244       | 0.000300419 |
| \( R_{KCI} \) | 0.040093       | 0.000300417 |
| \( R_{KCI} \) | 0.040485       | 0.000300380 |
| \( R_{KCI} \) | 0.040466       | 0.000300419 |
| \( R_{KCI} \) | 0.040796       | 0.000300399 |

Table 3: Relative efficiencies with respect to Kadilar and Cingi estimators.

| Estimator | Pop-I          | Pop-II          |
|-----------|----------------|-----------------|
| \( \hat{\theta}_1 \) | 141.7171       | 1624.092       |
| \( \hat{\theta}_2 \) | 141.3403       | 1624.081       |
| \( \hat{\theta}_3 \) | 142.3213       | 1624.086       |
| \( \hat{\theta}_4 \) | 142.2736       | 1624.091       |
| \( \hat{\theta}_5 \) | 143.1064       | 1624.083       |

Table 4: Relative efficiencies with respect to Haydar estimators.

| Estimator | Pop-I          | Pop-II          |
|-----------|----------------|-----------------|
| \( \hat{\theta}_1 \) | 109.0927       | 125.2937       |
| \( \hat{\theta}_2 \) | 108.8701       | 125.2912       |
| \( \hat{\theta}_3 \) | 109.4509       | 125.2454       |
| \( \hat{\theta}_4 \) | 109.4225       | 125.2921       |
| \( \hat{\theta}_5 \) | 109.9185       | 125.2689       |

The proposed estimator and Haydar’s estimators outperform Kadilar and Cingi [19] estimators, which is an expected result due to the employment of the count data in this analysis, and the proposed estimator has the best performance. The consequences of relative efficiency values are also provided graphically in Figures 1–3.

Furthermore, we investigated the efficiency condition of the proposed estimator as follows.

For Pop-I:
Therefore, the condition $B_p = 0.004 > (S_{xy}/S_x^2) - (1/2)R_{KC_i}, \ i = 1, 2, \ldots, 5$ is fulfilled.

For Pop-II:

\[
\frac{S_{xy}}{S_x^2} - \frac{1}{2}R_{KC1} = -0.0000103495,
\]

\[
\frac{S_{xy}}{S_x^2} - \frac{1}{2}R_{KC2} = -0.0000103485,
\]

\[
\frac{S_{xy}}{S_x^2} - \frac{1}{2}R_{KC3} = -0.0000103395,
\]

\[
\frac{S_{xy}}{S_x^2} - \frac{1}{2}R_{KC4} = -0.0000103495,
\]

\[
\frac{S_{xy}}{S_x^2} - \frac{1}{2}R_{KC5} = -0.0000103395.
\]

Therefore, the condition $B_p = 0.0000006131 > (S_{xy}/S_x^2) - (1/2)R_{KC_i}, \ i = 1, 2, \ldots, 5$ is fulfilled.

### 3. Concluding Remarks

In this article, the mean estimator based on the Poisson regression model has been proposed under simple random sampling. It is found that the proposed estimator produced more efficient results than the ratio estimators proposed in Koç [18] and Kadilar and Cingi [19]. For the numerical investigation, the performances of these estimators are compared based on two real populations, and it is seen that the new proposed estimator in terms of relative efficiency is more proficient than the reviewed estimators (see Figures 1–3), as all values surpass 100. As a result, we recommend emphatically using the proposed mean estimator over the other estimators considered in this study for such count data analysis. Furthermore, the estimator developed here can be utilized to calculate new estimates in other count models. The proposed estimator can also be derived in future studies using the concepts of Zaman and Bulut and Ali et al. [10, 11].

### Data Availability

The data used to support the findings of this study are included within the article.

### Conflicts of Interest

The authors declare that there are no conflicts of interest.

### References

[1] L. N. Upadhyaya, H. P. Singh, and J. W. E. Vos, “On the estimation of population means and ratios using supplementary information,” *Statistica Neerlandica*, vol. 39, no. 3, pp. 309–318, 1985.

[2] L. N. Upadhyaya and H. P. Singh, “Use of transformed auxiliary variable in estimating the finite population mean,” *Biometrical Journal*, vol. 41, no. 5, pp. 627–636, 1999.
[3] N. Koyuncu, “Efficient estimators of population mean using auxiliary attributes,” *Applied Mathematics and Computation*, vol. 218, no. 22, pp. 10900–10905, 2012.

[4] U. Shahzad, “On the estimation of population mean under systematic sampling using auxiliary attributes,” *Oriental Journal Physical Sciences*, vol. 1, no. 1–2, pp. 21–25, 2016.

[5] M. Abid, N. Abbas, H. Z. Nazir, and Z. Lin, “Enhancing the mean ratio estimators for estimating population mean using non-conventional location parameters,” *Revista Colombiana de Estadística*, vol. 39, no. 1, pp. 63–79, 2016.

[6] U. Shahzad, M. Hanif, N. Koyuncu, and A. V. Garcia Luengo, “A new family of estimators for mean estimation alongside the sensitivity issue,” *Journal of Reliability and Statistical Studies*, vol. 10, pp. 43–63, 2017.

[7] U. Shahzad, M. Hanif, N. Koyuncu, and A. V. G. Luengo, “A regression type estimator for mean estimation under ranked set sampling alongside the sensitivity issue,” *Communications Faculty of Science University of Ankara Series A1 Mathematics and Statistics, Series A1: Mathematics and Statistics*, vol. 65, no. 2, pp. 2037–2049, 2019.

[8] U. Shahzad, P. F. Perri, and M. Hanif, “A new class of ratio-type estimators for improving mean estimation of nonsensitive and sensitive variables by using supplementary information,” *Communications in Statistics - Simulation and Computation*, vol. 48, no. 9, pp. 2566–2585, 2019.

[9] T. Zaman and H. Bulut, “Modified ratio estimators using robust regression methods,” *Communications in Statistics-Theory and Methods*, vol. 48, no. 8, pp. 2039–2048, 2019.

[10] T. Zaman and H. Bulut, “Modified regression estimators using robust regression methods and covariance matrices in stratified random sampling,” *Communications in Statistics-Theory and Methods*, vol. 49, no. 14, pp. 3407–3420, 2020.

[11] I. Ali, I. Ahmad, M. Hanif, and U. Shahzad, “Robust-regression-type estimators for improving mean estimation of sensitive variables by using auxiliary information,” *Communications in Statistics-Theory and Methods*, vol. 50, no. 4, pp. 979–992, 2021.

[12] T. Zaman, “Improvement of modified ratio estimators using robust regression methods,” *Applied Mathematics and Computation*, vol. 348, pp. 627–631, 2019.

[13] C. Kadilar, M. Candan, and H. Cingi, “Ratio estimators using robust regression,” *Hacettepe Journal of Mathematics and Statistics*, vol. 36, no. 2, pp. 181–188, 2007.

[14] E. Oral and C. Kadilar, “Improved ratio estimators via modified maximum likelihood,” *Pakistan Journal of Statistics*, vol. 27, no. 3, pp. 269–282, 2011.

[15] E. Oral and C. Kadilar, “Robust ratio-type estimators in simple random sampling,” *Journal of the Korean Surgical Society*, vol. 40, no. 4, pp. 457–467, 2011.

[16] A. C. Cameron and P. K. Trivedi, *Regression Analysis of Count Data*, Cambridge University Press, Cambridge, 1998.

[17] D. C. Montgomery, E. A. Peck, and G. G. Vining, *Introduction to Linear Regression Analysis*, John Wiley and Sons, Hoboken, 4th edition, 2006.

[18] H. Koç, “Ratio-type estimators for improving mean estimation using Poisson regression method,” *Communications in Statistics - Theory and Methods*, vol. 50, no. 20, pp. 4685–4691, 2021.

[19] C. Kadilar and H. Cingi, “Ratio estimators in simple random sampling,” *Applied Mathematics and Computation*, vol. 151, no. 3, pp. 893–902, 2004.

[20] B. V. S. Sisodia and V. K. Dwivedi, “A modified ratio estimator using coefficient of variation of auxiliary variable,” *Journal of the Indian Society of Agricultural Statistics*, vol. 33, no. 2, pp. 13–18, 1981.