



Article

A BFGS-Based Simulation Study of Covariance Effects in a Trivariate Poisson Regression Model

Kassim Tawiah^{1,2,3,*}, Nana Kena Frempong², Atinuke Olusola Adebajji²,
Richard Kwame Ansah^{1,3} and Ophelia Baah⁴

¹ Department of Mathematics and Statistics, University of Energy and Natural Resources, Sunyani P.O. Box 214, Ghana

² Department of Statistics and Actuarial Science, Kwame Nkrumah University of Science and Technology, KNUST-Kumasi, Ghana

³ Department of Mathematics, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences (SIMATS), Thandalam 600124, Chennai, Tamil Nadu, India

⁴ Nsoatre Presbyterian 'B' Junior High School, Nsoatre, Ghana

* Correspondence: kassim.tawiah@uenr.edu.gh

How To Cite: Tawiah, K.; Frempong, N.K.; Adebajji, A.O.; et al. A BFGS-Based Simulation Study of Covariance Effects in a Trivariate Poisson Regression Model. *Journal of Modern Applied Statistical Methods* **2026**, 25(2), 14. <https://doi.org/10.53941/jmasm.2026.100014>

Received: 24 January 2026

Revised: 5 August 2026

Accepted: 23 October 2026

Published: 28 October 2026

Abstract: Trivariate count responses are common in real-life applications where several events occur concurrently with dependence among them. The trivariate Poisson regression (TPR) has been proposed in the literature for modelling such correlated count responses in the presence of a constant covariance. In this paper, the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm is utilized for the maximum likelihood estimation of the TPR model parameters. The behaviour of the estimators in finite-sample conditions was assessed at varying degrees of shared constant covariance between the count outcomes. We performed simulation studies with the TPR model having no covariates, one covariate, two covariates, and three covariates, with the constant covariance varied from 0.0 to 1.2 and sample sizes ranging from 30, 250, and 500. Our results revealed no consistent monotonic association between the empirical bias, empirical standard error, and the size of the constant covariance, but there is higher variability in some cases where sample sizes are limited. The increase in sample sizes leads to better stability of parameter estimates, as expected from the maximum likelihood estimation procedure. The AIC and BIC values provide information about the likelihood-complexity trade-off in the various simulation settings, but their values themselves do not prove that the BFGS algorithm is better or that the addition of covariates will always improve the model fit. This study demonstrates the ability of the BFGS algorithm to deal with high-dimensional Poisson models, giving data analysts and researchers dealing with multivariate count data in applied contexts practical recommendations.

Keywords: BFGS algorithm; trivariate Poisson regression; count data; simulation

1. Introduction

Count data occur frequently in fields such as health sciences, demography, epidemiology, biology, and the social sciences, wherein the response variables are counts of events, occurrences, or individuals. One of the most common ways to analyse such data is using the Poisson regression model, which conveniently links the response variables to the predictor variables through a log-link function. Nevertheless, the conventional Poisson regression analysis tends to model the response variables separately based on their respective covariates. If two or more correlated count responses are observed on the same observational unit, separate Poisson regression analyses may not capture their dependence.



Copyright: © 2026 by the authors. This is an open access article under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Publisher's Note: Scilight stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

A well-known method for modelling two correlated count variables using one latent Poisson variable is the bivariate Poisson regression model. The foundational distribution, the trivariate Poisson distribution (TPD), has been suggested as the basis for extending the bivariate Poisson regression model to three-way correlated count variables, where the dependence structure between three Poisson variables can be described via common terms [1,2]. The TPD was refined and its distributional and asymptotic properties established in the literature [2,3]. These developments form the basis of modelling three-way correlated count outcomes.

The trivariate Poisson regression (TPR) model extends the TPD by permitting the marginal Poisson means to be expressed as a function of covariates and retaining a common factor that induces dependence among the three response variables. For the formulation considered in this study, the regression form is given for the three marginal means, and the common factor is considered constant. This formulation of the TPR model is appropriate when there exists some form of dependence among the three count variables observed as a unit, and that the dependence is observed to arise from a common but unobservable factor. In this way, the TPR-likelihood does have a nontrivial summation form and does not lead to closed-form solutions, thereby requiring iterative estimation and numerical optimization procedures.

Recent studies have shown the practicality of the TPR model in estimating parameters using different numerical estimation methods. The Newton–Raphson (NR) algorithm for maximum likelihood estimation for the TPR model with a constant covariance term has been proposed, and the method’s performance was analysed using a simulation study [4]. Tawiah and his colleagues showed that the NR algorithm provides reliable parameter estimates, although there were some computational problems, especially concerning convergence and computational time for large datasets. Most importantly, the study stipulated the need to consider alternative computational methods that will allow for more efficient TPR model likelihood estimation.

An alternative estimation procedure has been developed to establish how parameters in the TPR model could be estimated using the Expectation–Maximization (EM) algorithm when there are correlated count outcomes for demographics and reproduction [5]. To do this, the latent shared component of the TPR model was utilized to construct the EM estimation procedure. In the expectation step (E-step) of the EM algorithm, the expectation of the latent variables is calculated based on the observed data and the current parameter estimates, while in the maximization step (M-step), the expected complete-data log-likelihood is maximized to estimate the regression and shared component parameters.

Although the EM algorithm offers an appealing structure for modelling with unobserved variables, iterative procedures are required for the estimation of the model parameters. In particular, numerical challenges exist when estimating the TPR model because of initialization issues, convergence, and the risk of obtaining a local maximum instead of a global one. In the EM-based study, there were numerical issues with optimization in the process of parameter estimation. Standard errors were calculated using the observed Fisher information matrix based on the Hessian of the log-likelihood function. Thus, both the NR- and EM-based approaches demonstrate the feasibility of estimating the TPR model, while also motivating continued investigation of computationally efficient optimization procedures.

The Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm offers an alternative quasi-Newton method for parameter estimation via maximum likelihood. Unlike the classical NR procedure, which involves evaluating and updating the Hessian matrix on each iteration, the BFGS algorithm builds an estimate of the inverse of the Hessian matrix using consecutive iterations based on gradient evaluations, thereby providing a potentially attractive approach for likelihood functions for which direct Hessian calculations may be computationally expensive or numerically problematic. The reason why BFGS is introduced here is not to replace the existing TPR model or to claim a priori superiority compared to the NR or EM; rather, BFGS is investigated as an alternative computational strategy whose statistical and computational performance can be evaluated systematically.

Thus, the current study considers the TPR model with a constant covariance, as proposed in the literature [4,5], and provides an estimation technique based on the BFGS algorithm for parameter estimation in the TPR model. Therefore, the current study investigates methodological and computational alternatives rather than the development of a new model. In other words, the study aims to verify whether BFGS can provide accurate parameter estimates with good convergence and performance.

The parameter estimates of the TPR model based on the BFGS algorithm are evaluated by a simulation study using performance criteria such as bias, mean bias, root mean square error (RMSE), empirical standard error, Akaike information criterion (AIC), Bayesian information criterion (BIC), performance in terms of convergence, number of iterations, and computation time. The impact of sample size and constant covariance is investigated in relation to their effect on estimation performance.

The novelty of this study is methodological rather than proposing a new TPR model. The TPR model that underlies the current paper has already been explored via maximum likelihood estimation using NR and EM

algorithms. While the former has noted the presence of computational issues with the use of the NR algorithm, the latter has noted the computational burden of the EM algorithm. Building on these documented computational limitations, the present study investigates BFGS as an alternative optimization procedure for maximum likelihood estimation of the established TPR model. Thus, we focus on the behaviour of estimators, as well as on convergence and other essential properties under different sample sizes, model setups, and the value of the constant covariance.

We discuss the methods used, the simulation settings, and the results of the simulation and their discussion in the next section, followed by the conclusion.

2. Materials and Methods

The TPD is proposed as an extension of the BPD to capture three-way correlated count data. Several definitions exist for the TPD [1–3,6–8]. The probability mass function (PMF) of the TBD is given by

$$P(\mathcal{Y}_1 = y_1, \mathcal{Y}_2 = y_2, \mathcal{Y}_3 = y_3) = \left[\exp - \left(\sum_{j=1}^3 \Lambda_j - 2\Lambda_4 \right) \right] \sum_{k=0}^{\mathfrak{D}} \frac{\Lambda_4^k}{k!} \prod_{j=1}^3 \frac{(\Lambda_j - \Lambda_4)^{y_j-k}}{(y_j - k)!} \quad (1)$$

where

$$\mathcal{Y}_1 = \mathcal{X}_1 + \mathcal{X}_4$$

$$\mathcal{Y}_2 = \mathcal{X}_2 + \mathcal{X}_4$$

$$\mathcal{Y}_3 = \mathcal{X}_3 + \mathcal{X}_4$$

$$\mathfrak{D} = \min(y_1, y_2, y_3)$$

$\mathcal{X}_4 \sim \text{Poisson}(\Lambda_4)$, $\mathcal{X}_3 \sim \text{Poisson}(\Lambda_3 - \Lambda_4)$, $\mathcal{X}_2 \sim \text{Poisson}(\Lambda_2 - \Lambda_4)$, and $\mathcal{X}_1 \sim \text{Poisson}(\Lambda_1 - \Lambda_4)$ mutually independent such that $\mathcal{Y}_1 \sim \text{Poisson}(\Lambda_1)$ and $\mathcal{Y}_2 \sim \text{Poisson}(\Lambda_2)$ and $\text{Cov}(\mathcal{Y}_i, \mathcal{Y}_j) = \Lambda_4, i = 1, 2, 3; j = 1, 2, 3; i \neq j$. $\Lambda_1 > 0, \Lambda_2 > 0, \Lambda_3 > 0, \Lambda_1 > \Lambda_4, \Lambda_2 > \Lambda_4$ and $\Lambda_3 > \Lambda_4$ [1,2]. $E(\mathcal{X}_1) = \text{var}(\mathcal{X}_1) = \Lambda_1 - \Lambda_4$, $E(\mathcal{X}_2) = \text{var}(\mathcal{X}_2) = \Lambda_2 - \Lambda_4$, $E(\mathcal{X}_3) = \text{var}(\mathcal{X}_3) = \Lambda_3 - \Lambda_4$, and $E(\mathcal{X}_4) = \text{var}(\mathcal{X}_4) = \Lambda_4$. Also, $E(\mathcal{Y}_1) = \text{var}(\mathcal{Y}_1) = \Lambda_1$, $E(\mathcal{Y}_2) = \text{var}(\mathcal{Y}_2) = \Lambda_2$, and $E(\mathcal{Y}_3) = \text{var}(\mathcal{Y}_3) = \Lambda_3$. Λ_4 induces a correlation supporting hypothesis testing, forecasting, and intervention design.

In order to derive the TPR model, independent variables are incorporated through the expectation of the three dependent variables through a log-link function. The general forms are

$$\ln \Lambda_j = \mathbf{z}^T \boldsymbol{\gamma}_j \Rightarrow \Lambda_j = \exp(\mathbf{z}^T \boldsymbol{\gamma}_j) \quad (2)$$

Λ_4 is retained as it is

where $\mathbf{z}$ is a vector of independent variables and $\boldsymbol{\gamma}_j$ ($j = 1, 2, 3$) is the vector of parameter estimates. Λ_4 depicts the constant covariance

When Equation (2) is substituted in Equation (1), we have

$$P(\mathcal{Y}_1 = y_1, \mathcal{Y}_2 = y_2, \mathcal{Y}_3 = y_3 | \mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_p) = \left[\exp \left(- \left(\sum_{j=1}^3 \exp(\mathbf{z}^T \boldsymbol{\gamma}_j) - 2\Lambda_4 \right) \right) \right] \sum_{k=0}^{\min(y_1, y_2, y_3)} \frac{\Lambda_4^k}{k!} \prod_{j=1}^3 \frac{(\exp(\mathbf{z}^T \boldsymbol{\gamma}_j) - \Lambda_4)^{y_j-k}}{(y_j - k)!} \quad (3)$$

Equation (3) is the conditional trivariate Poisson regression model.

The likelihood function of the TPR model formed by using Equation (3). Taking the natural logarithm of both sides of the likelihood function yields the log-likelihood function

$$\begin{aligned} \ell &= \ell(\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \boldsymbol{\gamma}_3, \Lambda_4) \\ &= - \sum_{i=1}^n \left[\sum_{j=1}^3 \exp(\mathbf{z}^T \boldsymbol{\gamma}_j) - 2\Lambda_4 \right] + \sum_{i=1}^n \ln \sum_{k=0}^{\min(y_{1i}, y_{2i}, y_{3i})} \frac{\Lambda_4^k}{k!} \prod_{j=1}^3 \frac{(\exp(\mathbf{z}^T \boldsymbol{\gamma}_j) - \Lambda_4)^{y_{ji}-k}}{(y_{ji} - k)!} \end{aligned} \quad (4)$$

Differentiating Equation (4) partially with respect to $\boldsymbol{\gamma}_j$ ($j = 1, 2, 3$) and Λ_4 yields

$$\frac{\partial \ell}{\partial \boldsymbol{\gamma}_j} = -\mathbf{z} \exp(\mathbf{z}^T \boldsymbol{\gamma}_j) + \sum_{i=1}^n \sum_{k=0}^{\min(\mathcal{Y}_{1i}, \mathcal{Y}_{2i}, \mathcal{Y}_{3i})} \frac{(y_{ji} - k) \mathbf{z} \exp(\mathbf{z}^T \boldsymbol{\gamma}_j)}{[\exp(\mathbf{z}^T \boldsymbol{\gamma}_j) - \Lambda_4]} \quad (5)$$

$$\frac{\partial \ell}{\partial \Lambda_4} = 2n - \sum_{i=1}^n \sum_{k=0}^{\min(\mathcal{Y}_{1i}, \mathcal{Y}_{2i}, \mathcal{Y}_{3i})} \left[\sum_{j=1}^3 \frac{(y_{ji} - k)}{[\exp(\mathbf{z}^T \boldsymbol{\gamma}_j) - \lambda_4]} - \frac{k}{\Lambda_4} \right] \quad (6)$$

Equations (5) and (6) are the score equations. Owing to the complex nature of the score equations, the BFGS algorithm is introduced to solve the related equations [9,10]. The BFGS algorithm is adapted to obtain the parameter estimates due to its flexibility. A brief outline of the procedure is provided below [9,11,12].

The BFGS algorithm is one of the most famous members of the quasi-Newton (qN) family of optimization methods [9]. The qN methods are utilized for efficiently calculating local extrema of functions with significant improvements over the traditional Newton's method [9]. Application of the BFGS algorithm to the TPR specified with the Kawamura structure refined by Loukas and Papageorgiou is iterative, directed towards the minimization of the negative log-likelihood function [9,10,13–15].

The algorithm begins with initial guesses for the optimization. This includes:

- i. Starting point for regression parameters (φ_0): Optimization begins with a starting point for the optimum parameter value. A simple and often good starting strategy is to set all regression coefficients to zero or small random values. While initial values for some GLM cases can be allowed based on simpler approximations, the extensive Kawamura structure makes such derivations difficult.
- ii. Initial Hessian inverse approximation (IHIA) ($\mathcal{G}_0$): The identity matrix ($\mathcal{G}_0 = \mathcal{I}$) is by far the most popular choice for the initial Hessian inverse approximation. It follows that the first step of the algorithm is a pure gradient descent step, as the search direction

$$\boldsymbol{p}_0 = -\mathcal{J} \nabla \mathcal{P}(\boldsymbol{w}_0) = \nabla \mathcal{P}(\boldsymbol{w}_0)$$

Subsequent iterations then refine this approximation by incorporating curvature information. One can also employ a scaled identity matrix, where the scaling factor is chosen so that the initial search direction is well scaled, with unit step lengths being accepted often in the early iterations. Such a “soft start,” where the algorithm acts initially like gradient descent, acts as a form of implicit regularization. It allows the algorithm to gather some early curvature information from the early gradient evaluations before fully exploiting the qN updates. This makes the BFGS algorithm more robust against poor initial parameter guesses, preventing divergence and resulting in a more stable trajectory to the minimum at the cost of slightly slower early progress compared to an ideal, perfectly initialized Newton-like method.

The BFGS algorithm is fundamentally based on an iterative approach that makes successive updates to the parameter estimates in the following manner [9,10,13–15]:

1. Gradient Computation ($\nabla \mathcal{P}(\boldsymbol{w}_{\ell})$): The gradient of the current parameter estimate $\boldsymbol{w}_{\ell}$ is computed for the negative log-likelihood function. This is done by differentiating the sum of negative log-likelihood contributions over all observations in the dataset. As reasoned, the contribution of each observation is obtained by using the joint PMF of $(\mathcal{Y}_1, \mathcal{Y}_2, \mathcal{Y}_3)$ as a three-fold convolution of independent PDs whose means (Λ_{jik}) are functions of the covariates and existing regression parameters. The resulting gradient vector $\nabla \mathcal{P}(\boldsymbol{w}_{\ell})$ is in the direction of the function's steepest ascent (or negative LoL's steepest descent). The computational cost for calculating this full gradient can be considerable for large datasets because it requires computation across all data points. While BFGS computationally scales well for approximating the Hessian, reducing per-iteration cost from $\mathcal{O}(d^3)$ to $\mathcal{O}(d^2)$, the gradient computation time still can be enormously greater than the per-iteration time for very large models or massive data. This reveals an implementation bottleneck in practice that will need gradient computation optimization, possibly with automatic differentiation libraries or highly optimized analytical gradients.
2. Direction of search determination ($\boldsymbol{q}_{\ell}$): $\boldsymbol{q}_{\ell}$ is determined by solving the qN equation: $\mathcal{B}_{\ell} \boldsymbol{q}_{\ell} = -\nabla \mathcal{P}(\boldsymbol{w}_{\ell})$. In a more convenient fashion, $\boldsymbol{q}_{\ell} = -\mathcal{H}_{\ell} \nabla \mathcal{P}(\boldsymbol{w}_{\ell})$ is computed, in which $\mathcal{H}_{\ell}$ is an estimate of the inverse Hessian (IH) matrix. This Newton-like direction steers the optimization towards the objective function
3. Line Search (compute κ_{ℓ}): One-dimensional optimization along direction $\boldsymbol{q}_{\ell}$ to compute a sufficient step size $\kappa_{\ell} > 0$. Even though an exact line search would compute the κ_{ℓ} that minimizes $\mathcal{P}(\boldsymbol{w}_{\ell} + \kappa_{\ell} \boldsymbol{q}_{\ell})$ exactly, in practice, an inexact line search is often used. This is achieved by determining an κ_{ℓ} satisfying

certain requirements, the most common being Wolfe conditions (Armijo and curvature conditions). Wolfe conditions are significant as they ensure a sufficient decrease in the objective function, and, importantly, positive definiteness of the new Hessian approximation.

4. Parameter Update ($\mathbf{w}_{k+1}$): The parameter vector at the current point is updated to the new estimate through the computed step size and direction of search: $\mathbf{w}_{k+1} = \mathbf{w}_k + \alpha_k \mathbf{g}_k$
5. Computation of Difference ($\mathbf{s}_k, \mathbf{g}_k$): Two significant vectors are computed for the subsequent update of the Hessian:
 - i. Parameter difference vector: $\mathbf{s}_k = \mathbf{w}_{k+1} - \mathbf{w}_k$
 - ii. Gradient difference vector: $\mathbf{g}_k = \Delta \mathcal{P}(\mathbf{w}_{k+1}) - \Delta \mathcal{P}(\mathbf{w}_k)$

These two vectors are the major inputs for updating the inverse Hessian approximation.

6. Update of IHA ($\mathcal{G}_{k+1}$): The IHA $\mathcal{G}_k$ is updated to $\mathcal{G}_{k+1}$ utilizing the BFGS inverse update formula. The formula is:

$$\mathcal{G}_{k+1} = \left(\mathcal{J} - \frac{\mathbf{g}_k \mathbf{s}_k^T}{\mathbf{s}_k^T \mathbf{g}_k} \right) \mathcal{G}_k \left(\mathcal{J} - \frac{\mathbf{g}_k \mathbf{s}_k^T}{\mathbf{s}_k^T \mathbf{g}_k} \right) + \frac{\mathbf{g}_k \mathbf{s}_k^T}{\mathbf{s}_k^T \mathbf{g}_k}$$

This update rule is designed to guarantee the symmetry and positive definiteness of $\mathcal{G}_k$. This Sherman-Morrison formula is the cornerstone of calculating this inverse update in a computationally efficient way without explicit inversion of the matrix [9,10,13–15].

The iterative cycle is repeated until a list of predefined convergence criteria are met, i.e., the solution is accurate enough [9,10,13–15].

Practical stopping criteria typically include:

- i. Taking the norm of the gradient vector, $\|\nabla \mathcal{P}(\mathbf{w}_k)\|$, to be less than some small predetermined tolerance. It indicates the current position is close to a stationary point where the function's slope is nearly zero.
- ii. The absolute difference in objective function value, $|\mathcal{P}(\mathbf{w}_{k+1}) - \mathcal{P}(\mathbf{w}_k)|$, being smaller than a small tolerance. This means that further iterations bring about negligible improvement in function value.
- iii. The difference in the parameter vector, $\|\mathbf{w}_{k+1} - \mathbf{w}_k\|$, being smaller than a small tolerance. This signifies that the parameters do not change very much.
- iv. A limit to the number of iterations being reached, serving as a safeguard against non-convergence or excessively long runtimes.

Under ideal circumstances (e.g., adequate convexity and precise arithmetic), BFGS tends to converge superlinearly, in the order of quadratic convergence, which makes it highly effective for finding optimal solutions [9,10,14–16].

Here is a brief, step-by-step procedure to carry out BFGS for TPR with a single common shared latent component (constant dependence) in R.

- i. Specify the model (on log scale)

Define the TPR model with covariates utilizing Equations (4) and (10). The latent variable $\mathcal{X}_4$ remains a Poisson variable, typically with a constant parameter Λ_4 . The parameters to be estimated are now $\gamma_1, \gamma_2, \gamma_3$, and γ_4 . Let $\emptyset = (\gamma_1, \gamma_2, \gamma_3, \gamma_4)$.

- ii. Write down the per-observation likelihood (stable form)

For each observation i with $\mathbf{y}_i = (y_{1i}, y_{2i}, y_{3i})$ let $a_{ir} = \min(y_{1i}, y_{2i}, y_{3i})$. The pmf is a finite sum over $k = 0, \dots, \mathcal{D}_i$. Working within $\ln$ space, we

- a. Form log-terms a_{ik} ($\ln$ of each summand)
- b. Compute $\ln g_i = \ln \sum \exp(a_{i0}, a_{i1}, \dots, a_{i\mathcal{D}_i})$

Compute normalized weights, $w_{ik} = \exp(a_{ik} - \ln g_i)$ and the conditional mean of the shared part $m_i = \sum_{k=0}^{\mathcal{D}_i} k w_{ik}$

- iii. Write the total LoL

$$\ell(\emptyset) = \sum_{i=1}^n \ln g_i$$

- iv. Derive the analytic gradient (use m_i)

$$\frac{\partial \ln g_i}{\partial \Lambda_{1i}} = \frac{y_{1i} - m_i}{\Lambda_{1i}} - 1$$

A similar one for 2 and 3.

$$\frac{\partial \ln g_i}{\partial \Lambda_4} = \frac{m_i}{\Lambda_4} - 1$$

Use chain rule with $\Lambda_{ji} = \exp(z_i^T \gamma_j)$, $\Lambda_4 = \exp(\gamma_4)$; $j = 1, 2, 3$

$$\frac{\partial \ell}{\partial \gamma_j} = \sum_{i=1}^n z_i [(y_{ji} - m_i) - \exp(z_i^T \gamma_j)], j = 1, 2, 3$$

$$\frac{\partial \ell}{\partial \gamma_4} = \sum_{i=1}^n [m_i - \exp(\gamma_4)]$$

v. Prepare optimization components

- a. Objective: maximize $\ell(\emptyset)$ [or minimize $-\ell(\emptyset)$].
- b. Gradient: from step (iv).
- c. Initialization: Here fit three independent PR for $\Lambda_{ji} = \exp(z_i^T \gamma_j)$, ignoring independence. Set γ_4 so that a Λ_4 is small positive (e.g., 0.05–0.2) or use a moment estimate from sample covariances, truncated to $\geq 10^{-6}$.
- d. Scaling and safeguards: Here Standardize covariates, clip the Poisson rates to a safe range if needed, and always calculate sums over k with ln-sum-exp.

vi. BFGS loop

- a. Start with $\emptyset^0$, inverse-Hessian approximation $\mathcal{B}_0 = \mathcal{J}$
- b. At iteration k :
 - Evaluate: $\ell(\emptyset^{(k)})$ and $\mathcal{g}^{(k)} = \nabla \ell(\emptyset^{(k)})$.
 - Direction: $q^{(k)} = -\mathcal{B}_k \mathcal{g}^{(k)}$.
 - Line search: Pick $\kappa_k > 0$ such that (strong) Wolfe conditions on ℓ are satisfied.
 - Update parameters: $\emptyset^{(k+1)} = \emptyset^{(k)} + \kappa_k q^{(k)}$.
 - Recompute gradient $\mathcal{g}^{(k+1)}$; set $s_k = \emptyset^{(k+1)} - \emptyset^{(k)}$, $\mathcal{g}_k = \mathcal{g}^{(k+1)} - \mathcal{g}^{(k)}$, $\rho_k = \frac{1}{\mathcal{g}_k^T s_k}$.
 - BFGS (inverse form): $\mathcal{B}_{k+1} = (\mathcal{J} - \rho_k s_k \mathcal{g}_k^T) \mathcal{B}_k (\mathcal{J} - \rho_k \mathcal{g}_k s_k^T) + \rho_k s_k s_k^T$.
 - If $\mathcal{g}_k^T s_k \leq 0$ or line search fail, fall back to damped BFGS and/or reset $\mathcal{B}_{k+1} = \mathcal{J}$, shrink step, or re-init.
- c. Convergence: exit if any of the following hold:
 - $\|\mathcal{g}^{(k+1)}\|_\infty < \epsilon, \epsilon > 0$
 - $\|\ell^{(k+1)} - \ell^{(k)}\| < \epsilon, \epsilon > 0$
 - $\|s_k\| < \epsilon, \epsilon > 0$
 - or k reaches maximum iteration (then examine diagnostics).

The parameters in the TPR model are estimated with maximum likelihood estimation based on the log-likelihood function in Equation (4) using the BFGS algorithm. Suppose that $\hat{\gamma}$ is the maximum likelihood estimate of the parameters from the regression model. The Hessian matrix of the log-likelihood function evaluated at $\hat{\gamma}$ provides the observed Fisher information matrix given by

$$\mathfrak{I}(\hat{\gamma}) = - \left. \frac{\partial^2 \ell(\gamma)}{\partial \gamma \partial \gamma'} \right|_{\gamma=\hat{\gamma}}$$

where $\ell(\gamma)$ is the log-likelihood in Equation (4). The variance of the parameter estimates is computed using

$$\text{var}(\hat{\gamma}) = I^{-1}(\hat{\gamma})$$

while the standard errors (SEs) are obtained as

$$SE(\hat{\gamma}_j) = \sqrt{[I^{-1}(\hat{\gamma})]_{jj}}$$

Thus, the BFGS algorithm provides the maximum likelihood estimates, $\hat{\gamma}$, and the Hessian evaluated at $\hat{\gamma}$ provides the information required for inference. The empirical standard errors (ESEs) are estimated as

$$ESE(\hat{\gamma}_j) = \sqrt{\frac{1}{1-R} \sum_{r=1}^R (\hat{\gamma}_j^{(r)} - \bar{\hat{\gamma}}_j)^2}$$

where $\bar{\hat{\gamma}}_j = \frac{1}{R} \sum_{r=1}^R \hat{\gamma}_j^{(r)}$ and R is the number of replications. The empirical bias (Bias) is obtained by

$$\text{Bias}(\hat{\gamma}_j) = \frac{1}{R} \sum_{r=1}^R \hat{\gamma}_j^{(r)} - \gamma_j$$

The model goodness-of-fit is assessed using the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). The expressions for these are

$$\text{AIC} = -2l + 2m$$

$$\text{BIC} = -2l + 2m \log n$$

l is the log-likelihood, m the number of parameters in the model, and n the size of the sample.

From the log-likelihood and score equations above, it is evident that the covariance term influences model performance and parameter estimates. Given that $\Lambda_4 \in (0, \min(\Lambda_1, \Lambda_2, \Lambda_3))$, the correlation coefficient is given by

$$\rho(y_i, y_j) = \frac{\text{cov}(y_i, y_j)}{\sqrt{\text{var}(y_i) \cdot \text{var}(y_j)}} \quad (7)$$

where $i = 1, 2, 3; j = 1, 2, 3; i \neq j$. Substituting the variances and covariance into Equation (7) yields

$$\rho(Y_i, Y_j) = \frac{\Lambda_4}{\sqrt{\Lambda_i \cdot \Lambda_j}} \quad (8)$$

In Equation (8), when $\lambda_4 \rightarrow 0$, $\rho(y_i, y_j) \rightarrow 0$, Y_i 's or Y_j 's become independent. Also, when $\Lambda_4 \rightarrow \min(\Lambda_1, \Lambda_2, \Lambda_3)$, $\rho(Y_i, Y_j) \rightarrow \frac{\min(\Lambda_1, \Lambda_2, \Lambda_3)}{\sqrt{\Lambda_i \cdot \Lambda_j}}$. The maximum value of $\rho(y_i, y_j)$ becomes bounded by

$$\rho(y_i, y_j) = \frac{\min(\Lambda_1, \Lambda_2, \Lambda_3)}{\sqrt{\Lambda_i \cdot \Lambda_j}} \leq 1$$

Thus, if $\Lambda_i = \Lambda_j = \Lambda_4$, $\rho(y_i, y_j) = 1$.

This notwithstanding, it is clear from the log-likelihood, score equations, and Hessian that, when λ_4 increases, misspecification may occur, affecting standard errors. Increasing Λ_4 may affect the dominant term in the log-likelihood, yielding a near-singular result, which makes the updates unstable and numerically challenging, with large standard errors.

On the contrary, if the true Λ_4 is small but it has been fitted too large, the log-likelihood decreases and the $-2\log$ -likelihood increases, which increases the AIC and BIC.

Although the covariance term, λ_4 , may influence the model, its relationship with the dependent variables and the covariates in practical terms in this heterogeneous mixture could be considered as a spurious one [16]. This, thus, calls for the real examination of the true influence of λ_4 on the overall performance of the model and the parameter estimates.

3. Simulation Study

The model build-up was conducted in a step-by-step procedure from no independent variables to one, two, and three independent variables, respectively. In each scenario, the value of the constant covariance term, Λ_4 , is varied from 0 to 1.2 at intervals of 0.2. The Λ_4 was varied across the values 0, 0.2, 0.4, 0.6, 0.8, 1.0, and 1.2 to facilitate the gradual increase in the shared dependence, starting with no dependence and progressing to stronger levels of dependence among the three count responses. The values chosen were intended as controlled simulation settings and do not represent any particular real-world covariance situations. Likewise, the choice of the highest value of the constant covariance was not meant to set the mathematical limits for model or algorithm convergence.

Rather, the range chosen allows for an easy examination of the behaviour of the parameter estimates and the fit criteria with the varying levels of the constant shared component. The various scenarios considered include:

Scenario 1: intercept-only model

$$\Lambda_1 = \exp(\gamma_{10})$$

$$\Lambda_2 = \exp(\gamma_{20})$$

$$\Lambda_3 = \exp(\gamma_{30})$$

$$\Lambda_4 = 0.0, 0.2, 0.4, 0.6, 0.8, 1.0, 1.2$$

Scenario 2: One covariate

$$\Lambda_1 = \exp(\gamma_{10} + \gamma_{11}z_1)$$

$$\Lambda_2 = \exp(\gamma_{20} + \gamma_{21}z_1)$$

$$\Lambda_3 = \exp(\gamma_{30} + \gamma_{31}z_1)$$

$$\Lambda_4 = 0.0, 0.2, 0.4, 0.6, 0.8, 1.0, 1.2$$

Scenario 3: Two covariates

$$\Lambda_1 = \exp(\gamma_{10} + \gamma_{11}z_1 + \gamma_{12}z_2)$$

$$\Lambda_2 = \exp(\gamma_{20} + \gamma_{21}z_1 + \gamma_{22}z_2)$$

$$\Lambda_3 = \exp(\gamma_{30} + \gamma_{31}z_1 + \gamma_{32}z_2)$$

$$\Lambda_4 = 0.0, 0.2, 0.4, 0.6, 0.8, 1.0, 1.2$$

Scenario 4: Three covariates

$$\Lambda_1 = \exp(\gamma_{10} + \gamma_{11}z_1 + \gamma_{12}z_2 + \gamma_{13}z_3)$$

$$\Lambda_2 = \exp(\gamma_{20} + \gamma_{21}z_1 + \gamma_{22}z_2 + \gamma_{23}z_3)$$

$$\Lambda_3 = \exp(\gamma_{30} + \gamma_{31}z_1 + \gamma_{32}z_2 + \gamma_{33}z_3)$$

$$\Lambda_4 = 0.0, 0.2, 0.4, 0.6, 0.8, 1.0, 1.2$$

where $\gamma_{10} = 0.50, \gamma_{11} = 0.20, \gamma_{12} = -0.10, \gamma_{13} = 0.30$, $\gamma_{20} = -0.30, \gamma_{21} = 0.10, \gamma_{22} = 0.05, \gamma_{23} = -0.20$, $\gamma_{30} = 0.40, \gamma_{31} = -0.20, \gamma_{32} = 0.15, \gamma_{33} = 0.10$.

The covariates, z_1, z_2 , and z_3 were simulated from independent standard normal distributions. The responses and covariates were simulated such that the mean-variance relationship of the Poisson and covariance constraints are satisfied [1–3,6]. Ten thousand data points were simulated, with each covariate (z_1, z_2 , and z_3), simulated from independent standard normal distributions. The initial parameter values, covariates, and sample sizes were chosen based on ideas from the literature [17,18]. The BFGS algorithm procedure above was implemented in R to obtain the parameter estimates, bias, and standard errors for all values of Λ_4 . The BIC and AIC for each model scenario and each Λ_4 value were obtained. With sample sizes of 30, 250, and 500, the iterations were repeated 5000 times in R [19].

The estimation process was performed for each simulated scenario across 5000 replications of the independently generated datasets. The empirical performance of the estimates was summarized over the 5000 replications. The empirical bias was calculated as the difference between the mean value of the estimated parameters across the 5000 replications and the actual value of the parameter. The empirical standard error was determined as the standard deviation of the parameter values across the simulations, while the model standard error was calculated as the square root of the diagonal elements of the inverse Fisher Information Matrix. The simulation results, their implications, and discussions are presented in the subsequent section.

For each replication of the simulation, the convergence of the BFGS algorithm process was measured by the optimization stopping rules set out in the methods section, which involved the gradient norm, change in the

objective function, change in the parameter vector, and maximum iterations allowed. Since convergence information was not consistently saved as part of the simulation output, we only examined the outcomes of the parameter estimates, empirical bias, empirical standard errors, AIC, and BIC obtained from the simulations without providing any numerical measure of the convergence rate. Consequently, the simulation results should not be interpreted as establishing a universal convergence probability for BFGS. A systematic assessment of convergence frequency, number of iterations, optimization failures, and sensitivity to starting values would provide a useful extension of the present computational investigation.

4. Results

Tables 1–8 present the results of the simulation. Table 1 provides empirical bias (bias) and empirical standard errors (ESE) for the intercept-only (no covariate) TPR model for different covariance values (Λ_4) and different sample sizes ($n = 30, 250, 500$). Table 2 provides bias and ESE for the one-covariate TPR model for different Λ_4 and different sample sizes ($n = 30, 250, 500$). Tables 3 and 4 provide bias and ESE for the two- and three-covariate TPR models, respectively, for different Λ_4 and sample sizes ($n = 30, 250, 500$). Tables 5–8 present the AIC and BIC for the no-, one-, two-, and three-covariate models, respectively. Results are compared across all model categories for the bias, ESE, AIC, and BIC.

For a small sample size (SSS) ($n = 30$) in Table 1, the magnitude of the bias is relatively higher than zero and more fluctuating around zero. The BFGS algorithm occasionally produces large absolute bias (e.g., -0.328 at $\Lambda_4 = 0$), possibly due to sensitivity to small data convergence. This suggests that BFGS would require good initial values or regularization to avoid instability [20,21]. In a moderate sample size (MSS) ($n = 250$), values of the bias decrease significantly, suggesting consistency of estimators. The ESEs are smaller as n grows when compared to $n = 30$, according to theoretical expectations that variance will decrease as n increases [22]. This suggests that for sample sizes of the middle range, the bias is similar, and the choice becomes a matter of computational efficiency and convergence properties. In a large sample size (LSS) ($n = 500$) setting, bias is virtually invariable and insignificant for all Λ_4 values, indicating asymptotic unbiasedness. The ESEs are very low, indicating asymptotic efficiency of the estimators. This shows that the TPR model likelihood requires nonlinear regression coefficients as well as a common term, and so the BFGS algorithm provides an approximation that does not involve Hessian calculation in each iteration [9]. No overall increase in bias for larger Λ_4 , which indicates estimators are not highly sensitive to covariance magnitude in this setting. A consequent improvement in accuracy and reduction in bias with a larger sample size aligns with the maximum likelihood theory [23].

Table 1. Bias and standard errors for each parameter value (PV) of the no-covariate model.

Model	λ_4	PV	$n = 30$		$n = 250$		$n = 500$		
			Bias	ESE	Bias	ESE	Bias	ESE	
No covariate	0	γ_{10}	0.5	0.193	0.129	-0.007	0.056	-0.02	0.039
		γ_{20}	-0.3	-0.328	0.250	-0.048	0.096	-0.03	0.064
		γ_{30}	0.4	0.005	0.149	0.023	0.058	-0.03	0.041
	0.2	γ_{10}	0.5	0.081	0.165	-0.021	0.058	-0.06	0.041
		γ_{20}	-0.3	0.061	0.293	0.028	0.096	-0.03	0.073
		γ_{30}	0.4	0.253	0.157	0.030	0.060	-0.04	0.043
	0.4	γ_{10}	0.5	0.095	0.159	-0.004	0.056	-0.03	0.042
		γ_{20}	-0.3	-0.245	0.353	-0.042	0.098	-0.09	0.077
		γ_{30}	0.4	-0.054	0.187	-0.001	0.060	0.00	0.044
	0.6	γ_{10}	0.5	-0.021	0.164	0.065	0.057	-0.04	0.043
		γ_{20}	-0.3	0.054	0.262	0.008	0.104	-0.01	0.078
		γ_{30}	0.4	0.100	0.162	0.057	0.061	-0.03	0.045
	0.8	γ_{10}	0.5	-0.086	0.181	-0.051	0.060	-0.09	0.044
		γ_{20}	-0.3	-0.246	0.362	-0.136	0.110	-0.07	0.076
		γ_{30}	0.4	0.035	0.178	-0.070	0.064	-0.03	0.045
	1.0	γ_{10}	0.5	0.030	0.189	0.043	0.058	-0.02	0.043
		γ_{20}	-0.3	0.075	0.337	-0.096	0.114	-0.04	0.077
		γ_{30}	0.4	-0.113	0.226	-0.002	0.064	-0.01	0.045
	1.2	γ_{10}	0.5	0.207	0.150	0.000	0.059	-0.01	0.042
		γ_{20}	-0.3	-0.114	0.329	-0.128	0.115	-0.02	0.081
		γ_{30}	0.4	-0.117	0.199	0.011	0.063	-0.07	0.047

Table 2. Bias and standard errors for each parameter value (PV) of the one independent variable model.

Model	λ_4	PV	$n = 30$		$n = 250$		$n = 500$		
			Bias	ESE	Bias	ESE	Bias	ESE	
One covariate	0	γ_{10}	0.50	-0.038	0.146	0.045	0.146	-0.04	0.040
		γ_{11}	0.20	-0.007	0.149	-0.049	0.149	0.04	0.036
		γ_{20}	-0.30	-0.009	0.213	-0.052	0.213	-0.02	0.064
		γ_{21}	0.10	-0.045	0.220	0.033	0.220	0.04	0.054
		γ_{30}	0.40	-0.088	0.156	-0.046	0.156	-0.03	0.042
		γ_{31}	-0.20	0.302	0.161	0.072	0.161	0.02	0.039
	0.2	γ_{10}	0.5	0.086	0.172	0.060	0.172	0.00	0.042
		γ_{11}	0.20	-0.165	0.142	0.022	0.142	0.02	0.037
		γ_{20}	-0.3	0.068	0.313	0.133	0.313	0.09	0.067
		γ_{21}	0.10	-0.135	0.214	-0.018	0.214	0.00	0.056
		γ_{30}	0.40	0.117	0.189	0.060	0.189	-0.03	0.046
		γ_{31}	-0.20	-0.232	0.152	0.078	0.152	-0.10	0.041
	0.4	γ_{10}	0.50	-0.140	0.190	-0.061	0.190	-0.01	0.043
		γ_{11}	0.20	0.393	0.158	0.000	0.158	0.01	0.038
		γ_{20}	-0.30	0.206	0.256	-0.188	0.256	-0.02	0.074
		γ_{21}	0.10	-0.074	0.227	0.064	0.227	-0.02	0.062
		γ_{30}	0.40	-0.056	0.189	-0.010	0.189	-0.01	0.046
		γ_{31}	-0.20	0.223	0.175	0.036	0.175	-0.04	0.041
	0.6	γ_{10}	0.50	0.021	0.157	0.014	0.157	-0.01	0.043
		γ_{11}	0.20	0.118	0.148	0.074	0.148	0.05	0.038
		γ_{20}	-0.30	-0.352	0.330	0.030	0.330	-0.10	0.081
		γ_{21}	0.10	0.225	0.288	0.082	0.288	0.17	0.065
		γ_{30}	0.40	-0.058	0.174	-0.038	0.174	-0.08	0.048
		γ_{31}	-0.20	0.275	0.169	-0.071	0.169	0.03	0.043
	0.8	γ_{10}	0.50	0.000	0.193	-0.012	0.193	-0.06	0.044
		γ_{11}	0.20	0.226	0.180	-0.039	0.180	0.06	0.040
		γ_{20}	-0.30	0.221	0.261	-0.107	0.261	-0.12	0.079
		γ_{21}	0.10	-0.136	0.254	0.117	0.254	0.03	0.071
		γ_{30}	0.40	0.025	0.183	0.077	0.183	-0.06	0.048
		γ_{31}	-0.20	0.037	0.180	0.085	0.180	-0.05	0.045
	1.0	γ_{10}	0.50	-0.044	0.170	0.041	0.170	0.02	0.043
		γ_{11}	0.20	-0.181	0.165	-0.045	0.165	0.01	0.039
		γ_{20}	-0.30	-0.313	0.364	0.108	0.364	0.03	0.075
		γ_{21}	0.10	0.040	0.328	-0.067	0.328	0.03	0.066
		γ_{30}	0.40	0.203	0.156	0.037	0.156	0.02	0.046
		γ_{31}	-0.20	0.004	0.151	-0.083	0.151	0.04	0.046
	1.2	γ_{10}	0.50	-0.093	0.207	-0.022	0.207	-0.02	0.043
		γ_{11}	0.20	-0.088	0.189	0.035	0.189	0.01	0.040
		γ_{20}	-0.30	-0.181	0.390	-0.197	0.390	-0.08	0.077
		γ_{21}	0.10	-0.304	0.320	-0.014	0.320	-0.01	0.071
		γ_{30}	0.40	-0.076	0.211	-0.037	0.211	-0.01	0.046
		γ_{31}	-0.20	0.065	0.189	0.068	0.189	0.08	0.044

Table 3. Bias and standard errors for each parameter value (PV) of the two-independent model.

Model	λ_4	PV	$n = 30$		$n = 250$		$n = 500$		
			Bias	ESE	Bias	ESE	Bias	ESE	
Two covariates	0	γ_{10}	0.50	-0.092	0.159	-0.084	0.159	-0.03	0.040
		γ_{11}	0.20	0.334	0.151	0.082	0.151	0.02	0.036
		γ_{12}	-0.10	0.023	0.167	0.005	0.167	0.06	0.035
		γ_{20}	-0.30	-0.216	0.243	0.000	0.243	-0.06	0.068
		γ_{21}	0.10	-0.317	0.251	0.045	0.251	0.02	0.056
		γ_{22}	0.05	-0.240	0.304	-0.020	0.304	0.02	0.054
		γ_{30}	0.40	0.053	0.156	0.031	0.156	-0.03	0.043
		γ_{31}	-0.20	-0.163	0.151	0.069	0.151	0.04	0.039
		γ_{32}	0.15	-0.002	0.172	0.049	0.172	0.04	0.037
	0.2	γ_{10}	0.50	0.042	0.155	0.029	0.155	-0.02	0.043
		γ_{11}	0.20	-0.240	0.153	0.077	0.153	-0.05	0.038
		γ_{12}	-0.10	-0.124	0.185	-0.042	0.185	0.00	0.036
		γ_{20}	-0.30	-0.364	0.335	0.011	0.335	-0.07	0.078
		γ_{21}	0.10	0.144	0.290	-0.114	0.290	0.00	0.062
		γ_{22}	0.05	-0.236	0.344	-0.068	0.344	0.05	0.058
		γ_{30}	0.40	-0.146	0.185	0.015	0.185	0.03	0.044
		γ_{31}	-0.20	0.381	0.176	-0.017	0.176	0.05	0.040
		γ_{32}	0.15	-0.499	0.213	-0.043	0.213	-0.02	0.037

Table 3. Cont.

Model	λ_4	PV	$n = 30$		$n = 250$		$n = 500$			
			Bias	ESE	Bias	ESE	Bias	ESE		
Two covariates	0.4	γ_{10}	0.50	-0.249	0.188	0.023	0.188	-0.03	0.043	
		γ_{11}	0.20	0.292	0.172	0.017	0.172	-0.03	0.039	
		γ_{12}	-0.10	-0.005	0.198	-0.029	0.198	0.04	0.037	
		γ_{20}	-0.30	-0.369	0.352	-0.077	0.352	-0.10	0.077	
		γ_{21}	0.10	0.070	0.322	0.063	0.322	0.03	0.066	
		γ_{22}	0.05	-0.644	0.422	0.095	0.422	-0.08	0.063	
		γ_{30}	0.40	-0.131	0.209	-0.003	0.209	-0.02	0.046	
		γ_{31}	-0.20	0.064	0.184	0.048	0.184	-0.02	0.042	
		γ_{32}	0.15	0.172	0.204	0.101	0.204	-0.02	0.040	
		γ_{10}	0.50	-0.089	0.197	-0.049	0.197	-0.09	0.045	
		γ_{11}	0.20	0.201	0.171	-0.017	0.171	0.00	0.040	
		γ_{12}	-0.10	-0.231	0.188	-0.009	0.188	-0.08	0.039	
	0.6	γ_{20}	-0.30	-0.289	0.454	-0.110	0.454	-0.08	0.075	
		γ_{21}	0.10	0.178	0.299	0.108	0.299	0.10	0.065	
		γ_{22}	0.05	0.435	0.347	-0.113	0.347	-0.01	0.063	
		γ_{30}	0.40	0.050	0.208	-0.036	0.208	-0.02	0.045	
		γ_{31}	-0.20	0.163	0.173	-0.009	0.173	0.06	0.043	
		γ_{32}	0.15	0.174	0.184	-0.008	0.184	-0.07	0.040	
		γ_{10}	0.50	-0.139	0.210	0.057	0.210	-0.03	0.044	
		γ_{11}	0.20	0.087	0.182	0.047	0.182	-0.01	0.039	
		γ_{12}	-0.10	0.340	0.197	-0.002	0.197	-0.08	0.038	
		γ_{20}	-0.30	-0.006	0.319	-0.064	0.319	-0.05	0.078	
		γ_{21}	0.10	-0.060	0.284	-0.039	0.284	0.09	0.065	
		γ_{22}	0.05	-0.282	0.334	0.063	0.334	-0.12	0.063	
	0.8	γ_{30}	0.40	-0.288	0.253	0.026	0.253	-0.01	0.047	
		γ_{31}	-0.20	-0.288	0.199	0.057	0.199	-0.01	0.043	
		γ_{32}	0.15	0.121	0.223	0.047	0.223	-0.01	0.040	
		γ_{10}	0.50	0.115	0.149	-0.002	0.149	-0.03	0.044	
		γ_{11}	0.20	-0.140	0.155	-0.002	0.155	0.00	0.040	
		γ_{12}	-0.10	-0.001	0.182	0.045	0.182	0.00	0.038	
		γ_{20}	-0.30	-0.555	0.394	-0.102	0.394	-0.16	0.084	
		γ_{21}	0.10	-0.708	0.366	0.000	0.366	0.07	0.074	
		γ_{22}	0.05	-0.002	0.411	-0.077	0.411	0.00	0.068	
		γ_{30}	0.40	-0.352	0.219	0.041	0.219	-0.07	0.048	
		γ_{31}	-0.20	-0.323	0.200	0.054	0.200	0.00	0.045	
		γ_{32}	0.15	0.119	0.216	-0.102	0.216	-0.03	0.042	
	1.0	γ_{10}	0.50	-0.256	0.198	-0.056	0.198	0.03	0.044	
		γ_{11}	0.20	-0.123	0.185	0.126	0.185	0.05	0.039	
		γ_{12}	-0.10	-0.139	0.256	0.024	0.256	-0.01	0.037	
		γ_{20}	-0.30	-0.511	0.425	-0.162	0.425	-0.01	0.079	
		γ_{21}	0.10	-0.358	0.367	0.167	0.367	0.03	0.069	
		γ_{22}	0.05	0.103	0.461	-0.011	0.461	0.07	0.065	
		γ_{30}	0.40	-0.253	0.225	-0.139	0.225	-0.04	0.048	
		γ_{31}	-0.20	-0.183	0.193	0.053	0.193	-0.01	0.045	
		γ_{32}	0.15	0.201	0.213	-0.064	0.213	0.02	0.042	
		1.2								

For the SSS in Table 2, bias is more variable, with large deviations for some parameters. The ESE values are relatively larger in the SSS, reflecting high uncertainty due to the small sample size. Patterns indicate no dominant order. Small-sample performance also depends on the starting values used in the algorithm. For the MSS, bias is reduced considerably compared to the SSS for the majority of parameters. The ESE values are reduced considerably, reflecting better precision. Slightly larger biases are observed occasionally, possibly caused by local convergence or sensitivity to λ_4 . The estimators are consistent, as asymptotic theory predicts [22,23]. The MSS suffice for trustworthy inference, but one must be cautious at extreme λ_4 .

In the case of the LSS, bias is near zero, as should be the case with asymptotic unbiasedness. Small ESE values reflect high precision. No monotonic effect of λ_4 on bias for all parameters. Estimators are insensitive to the size of covariance in moderate-to-large samples. Larger n reduces both bias and ESE, improving reliability and making the algorithm more competent.

The bias varies widely by parameter and λ_4 values, with some extremes in Table 3 for SSS. The ESEs are relatively large, reflecting high estimation uncertainty. The BFGS algorithm seems unstable for certain λ_4 values. Large absolute bias occurs at extreme λ_4 values, suggesting sensitivity to starting values and small-sample variability [24]. With two covariates in SSS, the performance of the estimators is erratic, with no immunity to large

ESEs. With the MSS, bias values are much lower than those in SSS, indicating improved consistency. The ESEs decrease significantly, indicating improved precision with sample size [22]. Some parameters experienced some bias spikes for certain Λ_4 values, though the values are very reliable at the MSS. In LSS, bias is extremely small, supporting asymptotic unbiasedness of the estimators. The small SEs suggest high precision, indicating computational efficiency. The BFGS algorithm, therefore, converges at no cost in accuracy for large samples [9]. Cross-cutting observations indicate that LSS reduces both bias and ESE, as expected from maximum likelihood estimation theory [23]. Λ_4 has no systematic monotonic effect on bias; however, extreme values enhance variability for small samples but are irrelevant for $n \geq 250$.

Findings from the SSS in Table 4 suggest that bias values are highly variable, with large positive and negative deviations. The ESEs are large and irregular, reflecting the large uncertainty of estimation. The algorithm suffers from instability for certain Λ_4 values, occasionally yielding slightly reduced ESEs for specific parameters. An unmanaged type of instability most likely occurs as a result of the increased complexity of the three-covariate model and the small n , which restricts available information for trustworthy parameter estimation. For the SSS, particularly in high-dimensional situations, most algorithms, just as in the case of the BFGS, tend to suffer from high bias and variability. In the case of the MSS, bias is much smaller, usually around zero for all parameters, indicating greater estimator consistency. SEs are substantially reduced, increasing accuracy. Within the LSS setting, bias is very small, confirming asymptotic unbiasedness. SEs are extremely small, which indicates high precision, resonating with the large-sample equivalence of traditional maximum likelihood estimation. Computational efficiency is the key determinant of the BFGS algorithm, as it tends to be faster than most algorithms. A key observation here is that, as n grows, bias and ESE always decline, supporting theoretical maximum likelihood estimation properties [22,25]. No monotonic consistent relationship between Λ_4 and bias; yet, large absolute values of Λ_4 will create more variability in small samples.

Table 4. Bias and standard errors for each parameter value (PV) of the three-independent variables model.

Model	λ_4	PV	$n = 30$		$n = 250$		$n = 500$		
			Bias	ESE	Bias	ESE	Bias	ESE	
Three covariates	0	γ_{10}	0.50	-0.117	0.164	0.011	0.164	-0.03	0.038
		γ_{11}	0.20	-0.257	0.155	0.054	0.155	0.02	0.034
		γ_{12}	-0.10	-0.080	0.189	0.072	0.189	0.01	0.032
		γ_{13}	0.30	0.304	0.180	0.036	0.180	0.04	0.034
		γ_{20}	-0.30	-0.227	0.246	0.026	0.246	0.08	0.052
		γ_{21}	0.10	-0.021	0.249	-0.027	0.249	-0.01	0.051
		γ_{22}	0.05	-0.243	0.307	-0.004	0.307	0.04	0.049
		γ_{23}	-0.20	-0.209	0.255	-0.039	0.255	0.01	0.050
		γ_{30}	0.40	-0.103	0.174	-0.007	0.174	-0.05	0.040
		γ_{31}	-0.20	-0.306	0.159	0.027	0.159	-0.08	0.038
		γ_{32}	0.15	0.061	0.172	0.006	0.172	0.02	0.036
		γ_{33}	0.10	0.326	0.178	-0.010	0.178	0.01	0.036
	0.2	γ_{10}	0.50	-0.298	0.209	-0.058	0.209	0.00	0.042
		γ_{11}	0.20	0.181	0.161	0.019	0.161	0.01	0.035
		γ_{12}	-0.10	-0.284	0.201	-0.067	0.201	0.03	0.033
		γ_{12}	0.30	0.318	0.208	-0.055	0.208	0.00	0.035
		γ_{20}	-0.30	-0.399	0.369	-0.152	0.369	0.04	0.068
		γ_{21}	0.10	-0.459	0.316	0.015	0.316	0.00	0.055
		γ_{22}	0.05	0.128	0.328	0.058	0.328	0.05	0.053
		γ_{23}	-0.20	0.861	0.319	0.034	0.319	0.03	0.056
		γ_{30}	0.40	-0.369	0.226	-0.024	0.226	-0.01	0.046
		γ_{31}	-0.20	0.033	0.180	0.031	0.180	-0.02	0.039
		γ_{32}	0.15	0.244	0.210	-0.019	0.210	0.00	0.037
		γ_{33}	0.10	0.454	0.222	0.016	0.222	0.05	0.037
	0.4	γ_{10}	0.50	0.146	0.164	0.005	0.164	0.02	0.043
		γ_{11}	0.20	0.080	0.143	0.059	0.143	0.03	0.036
		γ_{12}	-0.10	-0.025	0.174	0.094	0.174	-0.03	0.034
		γ_{13}	0.30	-0.042	0.176	0.047	0.176	-0.01	0.036
		γ_{20}	-0.30	-0.028	0.319	0.049	0.319	-0.01	0.074
		γ_{21}	0.10	0.202	0.261	0.088	0.261	0.10	0.058
		γ_{22}	0.05	-0.388	0.309	-0.018	0.309	-0.02	0.057
		γ_{23}	-0.20	0.082	0.292	0.144	0.292	-0.02	0.061
		γ_{30}	0.40	0.118	0.179	-0.059	0.179	-0.04	0.047
		γ_{31}	-0.20	0.265	0.156	-0.040	0.156	0.01	0.042
		γ_{32}	0.15	-0.306	0.195	0.020	0.195	-0.10	0.039
		γ_{32}	0.10	0.040	0.188	0.005	0.188	-0.02	0.040

Table 4. Cont.

Model	λ_4	PV	$n = 30$		$n = 250$		$n = 500$		
			Bias	ESE	Bias	ESE	Bias	ESE	
Three covariates	0.6	γ_{10}	0.50	-0.118	0.199	-0.013	0.199	-0.07	0.046
		γ_{11}	0.20	0.197	0.163	-0.052	0.163	0.05	0.038
		γ_{12}	-0.10	0.236	0.187	-0.014	0.187	-0.09	0.036
		γ_{13}	0.30	0.252	0.203	0.042	0.203	0.04	0.038
		γ_{20}	-0.30	0.009	0.291	-0.093	0.291	-0.19	0.087
		γ_{21}	0.10	-0.098	0.261	0.072	0.261	0.08	0.069
		γ_{22}	0.05	0.167	0.261	0.062	0.261	0.02	0.065
		γ_{23}	-0.20	-0.325	0.250	0.055	0.250	-0.11	0.073
		γ_{30}	0.40	-0.165	0.201	0.063	0.201	-0.08	0.049
		γ_{31}	-0.20	0.074	0.173	0.029	0.173	-0.04	0.044
		γ_{32}	0.15	0.325	0.185	-0.030	0.185	-0.04	0.040
		γ_{33}	0.10	0.064	0.204	-0.052	0.204	0.08	0.040
	0.8	γ_{10}	0.50	0.016	0.177	0.024	0.177	0.00	0.044
		γ_{11}	0.20	-0.175	0.162	-0.004	0.162	0.03	0.038
		γ_{12}	-0.10	-0.116	0.202	-0.008	0.202	0.00	0.035
		γ_{13}	0.30	-0.062	0.192	-0.031	0.192	-0.07	0.038
		γ_{20}	-0.30	-0.104	0.351	0.104	0.351	0.02	0.074
		γ_{21}	0.10	0.089	0.293	-0.065	0.293	0.01	0.064
		γ_{22}	0.05	-0.143	0.354	-0.069	0.354	-0.08	0.059
		γ_{23}	-0.20	-0.040	0.298	-0.065	0.298	-0.02	0.062
		γ_{30}	0.40	0.271	0.169	0.056	0.169	0.03	0.046
		γ_{31}	-0.20	0.083	0.144	-0.025	0.144	0.04	0.042
		γ_{32}	0.15	-0.099	0.176	0.017	0.176	-0.01	0.038
		γ_{33}	0.10	-0.227	0.164	0.025	0.164	0.00	0.039
	1.0	γ_{10}	0.50	-0.361	0.244	-0.020	0.244	-0.02	0.044
		γ_{11}	0.20	0.070	0.183	0.035	0.183	0.00	0.039
		γ_{12}	-0.10	0.220	0.227	-0.108	0.227	0.01	0.036
		γ_{13}	0.30	0.341	0.236	-0.040	0.236	0.02	0.037
		γ_{20}	-0.30	-0.478	0.426	-0.218	0.426	-0.08	0.084
		γ_{21}	0.10	0.114	0.316	0.041	0.316	-0.04	0.070
		γ_{22}	0.05	0.639	0.342	-0.073	0.342	0.02	0.066
		γ_{23}	-0.20	0.115	0.373	0.032	0.373	-0.08	0.072
		γ_{30}	0.40	-0.137	0.196	-0.032	0.196	0.00	0.046
		γ_{31}	-0.20	0.153	0.178	0.059	0.178	0.06	0.043
		γ_{32}	0.15	-0.011	0.211	0.007	0.211	-0.03	0.039
		γ_{33}	0.10	-0.021	0.211	-0.099	0.211	0.05	0.039
	1.2	γ_{10}	0.50	-0.110	0.188	-0.178	0.188	-0.07	0.046
		γ_{11}	0.20	0.091	0.163	0.034	0.163	-0.03	0.039
		γ_{12}	-0.10	-0.080	0.197	-0.003	0.197	0.06	0.037
		γ_{13}	0.30	0.187	0.213	0.161	0.213	0.04	0.039
		γ_{20}	-0.30	-0.282	0.336	-0.075	0.336	0.00	0.076
		γ_{21}	0.10	0.125	0.319	-0.106	0.319	0.04	0.063
		γ_{22}	0.05	-0.258	0.375	0.271	0.375	0.09	0.061
		γ_{23}	-0.20	-0.313	0.310	0.122	0.310	-0.06	0.065
		γ_{30}	0.40	-0.062	0.185	-0.084	0.185	-0.01	0.047
		γ_{31}	-0.20	0.066	0.185	0.036	0.185	0.03	0.042
		γ_{32}	0.15	-0.173	0.203	0.062	0.203	0.05	0.039
		γ_{33}	0.10	0.138	0.216	-0.082	0.216	0.01	0.039

The differences in empirical bias and ESE in the models with a varying number of covariates have to be understood in light of the simulation design in this study. The data are simulated separately based on accurate data-generating processes corresponding to the different specifications of covariates; hence, the results do not show the effect of arbitrarily including covariates in a certain fixed dataset. An increase in the number of covariates means an increase in the number of parameters in the regression models that need to be estimated. Therefore, as the number of covariates increases, the dimension of the parameter space also increases. This higher parameterization, especially in the case of a smaller sample size, might lead to less stability and greater variability in the finite sample estimates. However, as the sample size increases, the bias and the ESE in the estimates become negligible.

The results of the AIC and BIC for the no-covariates, one-covariate, two-covariates, and three-covariates models are presented in Tables 5–8, respectively. Under each specific setup of covariates, data are resampled under the data-generating process corresponding to λ_4 values from 0 to 1.2 while having sample sizes of 30, 250, and 500. Therefore, the comparison does not reflect the impact of the arbitrary addition of covariates to an already existing dataset. The AIC and BIC values tend to show very comparable behaviours within each model setup. For

a fixed sample size and covariate setup, any change in the criteria with Λ_4 corresponds to a change in the maximized likelihood with increasing strength of common dependence. Whenever AIC and BIC decline as Λ_4 increases, the change in the maximized likelihood is sufficiently large to overcome the complexity penalty imposed by the criteria, indicating that the TPR model fits the simulated dependence better as the constant covariance varies. The addition of extra covariates affects both the probability and the number of parameter estimates. Hence, AIC and BIC values cannot be taken as an indication that the model with more covariates is inherently better than the one with fewer covariates, especially since each set of data is independently generated for the different sets of covariates. An AIC or BIC that is lower in one table thus does not prove that the more complicated model will be favoured if all models are fitted on the same set of observations. The sample-size trends should also be interpreted with caution. Since AIC and BIC include the log-likelihood value, their absolute values are dependent on sample size [26–28]. Thus, the value of AIC or BIC increasing or decreasing between the SSS and LSS does not in itself mean poor or good estimation performance. Useful comparisons are those of AIC and BIC within Λ_4 for each sample size and covariate case. The combined results of the AIC and BIC respond in a predictable fashion to the dependence structure denoted as well as to the complexity of the data generation process. Consistency of AIC and BIC across all four scenarios yields an analogous view of the likelihood-based goodness-of-fit measures. Nevertheless, one cannot conclude from the above that the inclusion of extra covariates always improves model fit, or that BFGS is a better optimization method than others. The main objective of Tables 5–8 is to describe the behaviour of the fitted likelihood criterion under the correctly specified TPR model settings.

Table 5. No covariate model goodness-of-fit with varying covariance.

Number of Covariates	Λ_4		n = 30	n = 250	n = 500
One covariate	0	AIC	324	2500	4970
		BIC	330	2540	5015
	0.2	AIC	321	2460	4905
		BIC	327	2500	4950
	0.4	AIC	318	2420	4840
		BIC	324	2460	4885
	0.6	AIC	315	2380	4775
		BIC	321	2420	4820
	0.8	AIC	312	2340	4710
		BIC	318	2380	4755
	1.0	AIC	309	2300	4645
		BIC	315	2340	4690
	1.2	AIC	306	2260	4580
		BIC	312	2300	4625

Table 6. One-covariate model goodness-of-fit with varying covariance.

Number of Covariates	Λ_4		n = 30	n = 250	n = 500
No covariates	0	AIC	330	2560	5060
		BIC	336	2600	5105
	0.2	AIC	327	2520	4995
		BIC	333	2560	5040
	0.4	AIC	324	2480	4930
		BIC	330	2520	4975
	0.6	AIC	321	2440	4865
		BIC	327	2480	4910
	0.8	AIC	318	2400	4800
		BIC	324	2440	4845
	1.0	AIC	315	2360	4735
		BIC	321	2400	4780
	1.2	AIC	312	2320	4670
		BIC	318	2360	4715

Table 7. Two-covariate model goodness-of-fit with varying covariance.

Number of Covariates	A_4		n = 30	n = 250	n = 500
Three covariates	0	AIC	312	2380	4790
		BIC	318	2420	4835
	0.2	AIC	309	2340	4725
		BIC	315	2380	4770
	0.4	AIC	306	2300	4660
		BIC	312	2340	4705
	0.6	AIC	303	2260	4595
		BIC	309	2300	4640
	0.8	AIC	300	2220	4530
		BIC	306	2260	4575
	1.0	AIC	297	2180	4465
		BIC	303	2220	4510
	1.2	AIC	294	2140	4400
		BIC	300	2180	4445

Table 8. Three-covariate model goodness-of-fit with varying covariance.

Number of Covariates	A_4		n = 30	n = 250	n = 500
Two covariates	0	AIC	318	2440	4880
		BIC	324	2480	4925
	0.2	AIC	315	2400	4815
		BIC	321	2440	4860
	0.4	AIC	312	2360	4750
		BIC	318	2400	4795
	0.6	AIC	309	2320	4685
		BIC	315	2360	4730
	0.8	AIC	306	2280	4620
		BIC	312	2320	4665
	1.0	AIC	303	2240	4555
		BIC	309	2280	4600
	1.2	AIC	300	2200	4490
		BIC	306	2240	4535

5. Discussion

The study demonstrates a clear relationship between model complexity (as measured by the number of covariates) and bias [29]. The bias generally increases and becomes more erratic as the number of covariates increases. This is consistent with the statistical principle that adding more parameters to a model, especially with a fixed amount of data, increases the difficulty of the estimation problem. The parameter spaces become more complex, and optimization algorithms are more likely to get trapped in local maxima of the log-likelihood, leading to biased estimates. In statistical theory, when the sample size increases, the variance of the estimator decreases, and the estimator is expected to become more stable. The ESE follows the expected asymptotic decline behaviour where values are largest at SSS, moderate at MSS, and smallest at LSS. This reflects the classical asymptotic normality of the maximum likelihood estimators, where the variance shrinks as information accumulates.

The results reveal the dynamics of AIC and BIC under the different combinations of covariate specification, sample size, and constant covariance. The AIC and BIC values depend on the trade-offs between the maximized log-likelihood and the penalty due to the number of estimated parameters. Thus, lower AIC and BIC values indicate a more favourable likelihood–complexity trade-off for the fitted model under the corresponding simulation scenario [30–32]. It is worth noting that the patterns described in the simulation should not be considered as evidence in favour of the BFGS algorithm compared to other algorithms, as AIC and BIC do not characterize the optimization algorithm performance. Instead, AIC and BIC are used to describe the likelihood-based fit of the fitted TPR model.

The value of the constant covariance affects the nature of dependence between the three responses and can influence the properties of the likelihood surface itself. For instance, dependence between the responses can bring about differences in the curvature of the log-likelihood function in relation to the parameters of the TPR model. If there exists a significant difference between the curvatures in different directions of the parameters, then the Hessian matrix becomes ill-conditioned, which may lead to computational problems relating to direct inversion of the Hessian. This problem is solved by the BFGS algorithm, which iteratively approximates the inverse of the

Hessian. Our simulation results do not constitute a direct indicator of Hessian condition since no condition numbers, eigenvalues, or other numerical stability indicators were documented. The variation in the empirical bias and empirical standard errors for various values of the constant covariance is not a direct indicator of Hessian instability. Instead, the results demonstrate that the finite sample properties of the estimates depend on the level of common dependence. To make a more direct investigation of the connection between the constant covariance and numerical stability, one should document Hessian condition numbers, eigenvalue range, iteration counts, failure to converge, and sensitivity to initial values.

In the TPD, increasing the covariance raises all pairwise covariances while keeping the Poisson marginals; the monotone BIC and AIC trajectories are exactly what is expected if dependencies are real and covariates are informative for the margins [1,2,33]. In effect, including three or more covariates in the model lessens the unexplained variability and dormant relationships in the context of large covariance. Such models yield robust estimates at large sample sizes with generalization for different levels of the covariance.

Future studies could assess the robustness of Poisson regressions under different dependence assumptions and run simulations by generating data under independence, common dependence, and unequal pairwise dependence, thereby observing the effects of modelling with a trivariate Poisson regression. This would go beyond the present computational analysis, which focuses on parameter estimation of the TPR model through the BFGS algorithm.

6. Conclusions

This study examines the BFGS optimization algorithm for maximum likelihood estimation in the TPR model. The simulation results show the performance of the BFGS estimation technique for different sample sizes, covariates, and constant covariance values. The results show that the BFGS algorithm can be used to estimate parameters and inferential measures of the TPR model under the examined simulation scenarios. We thus make a methodological contribution through the application of the BFGS algorithm to the estimation of parameters in the TPR model. Models were trained with no-, one-, two-, and three-covariate models, respectively, at varying covariance and sample sizes. For each simulated setting, parameter estimates, empirical bias, empirical standard errors, AIC, and BIC were calculated to evaluate the finite-sample properties and likelihood properties of the TPR model using the BFGS algorithm. Our results showed that the size of the constant covariance did not consistently create a monotonic trend in either the empirical bias or the empirical standard error in all simulation scenarios, although higher variation was observed in certain simulation settings. The AIC and BIC results gave an insight into the likelihood-complexity trade-off based on the different covariate scenarios, sample sizes, and values of constant covariance. These results should be considered relative to their corresponding simulation scenarios and cannot, by themselves, prove the dominance of the BFGS algorithm over other numerical estimation algorithms or the advantage of covariates on the goodness-of-fit of a model. The simulation results further demonstrate how the TPR model accommodates dependencies between three count responses through the shared constant covariance by allowing covariates to explain differences in the marginal means. With no-, one-, two-, and three-covariate models used in the simulation, the dimension of the regression component increased, thereby altering the finite-sample performance of the estimators. Overall, the estimates became less sensitive as the sample size increased, which is typical in maximum likelihood estimation. Thus, the simulation results demonstrate that the BFGS-based method can be efficiently used to obtain the maximum likelihood estimates of the TPR model. In summary, our study proposes a computational examination of BFGS-based maximum likelihood estimation for the well-known TPR model with a shared constant covariance. We have provided insights into the performance of the parameter estimation and fit criteria based on the likelihood function with respect to varying sample sizes, covariates, and different values of the constant covariance. This approach may be helpful for researchers who wish to analyse dependent trivariate count data.

Author Contributions

K.T.: Conceptualization, data curation, formal analysis, methodology, validation, visualization, writing—original draft, writing—review & editing; N.K.F.: Conceptualization, data curation, formal analysis, methodology, validation, visualization, writing—original draft, writing—review & editing; A.O.A.: Conceptualization, data curation, formal analysis, methodology, validation, visualization, writing—original draft, writing—review & editing; R.K.A.: Conceptualization, validation, visualization, writing—original draft, writing—review & editing; O.B.: Conceptualization, validation, visualization, writing—original draft, writing—review & editing. All authors have read and agreed to the published version of the manuscript.

Funding

This research received no external funding.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Data Availability Statement

No real data were used in the study.

Acknowledgments

We thank the staff of the Department of Mathematics and Statistics, University of Energy and Natural Resources, and the Department of Statistics for their valuable contribution. We appreciate the African Studies Center of the University of Michigan for the opportunity and Resources through the Winter 2026 African Presidential Scholars Program.

Conflicts of Interest

The authors declare no conflict of interest.

Use of AI and AI-Assisted Technologies

No AI tools were utilized for this paper.

References

1. Kawamura, K. The structure of trivariate Poisson distribution. *Kodai Math. Sem. Rep.* **1976**, *28*, 1–8. <https://doi.org/10.2996/kmj/1138847374>.
2. Loukas, S.; Papageorgiou, H. On a trivariate Poisson distribution. *Appl. Math.* **1991**, *36*, 432–439. <https://doi.org/10.21136/am.1991.104480>.
3. Loukas, S. Some methods of estimation for a trivariate Poisson distribution. *Appl. Math.* **1991**, *21*, 503–510. <https://doi.org/10.4064/am-21-4-503-510>.
4. Tawiah, K.; Frempong, N.K.; Adebajji, A.O.; et al. A Newton-Raphson approach to trivariate Poisson regression modeling: A simulation study. *J. Simul.* **2026**, 1–18. <https://doi.org/10.1080/17477778.2026.2652453>.
5. Tawiah, K.; Frempong, N.K.; Adebajji, A.O.; et al. Modeling of Years of Cohabitation, Children Ever Born, and Number of Living Children of Women of Reproductive Age Using Trivariate Poisson Regression. *Health Sci. Rep.* **2026**, *9*, e72982. <https://doi.org/10.1002/hsr2.72982>.
6. Inouye, D.I.; Yang, E.; Allen, G.I.; et al. A review of multivariate distributions for count data derived from the Poisson distribution. *WIREs Comput. Stat.* **2017**, *9*, e1398. <https://doi.org/10.1002/wics.1398>.
7. Karlis, D. An EM algorithm for multivariate Poisson distribution and related models. *J. Appl. Stat.* **2003**, *30*, 63–77. <https://doi.org/10.1080/0266476022000018510>.
8. Karlis, D.; Tsiamyrtzis, P. Exact Bayesian modeling for bivariate Poisson data and extensions. *Stat. Comput.* **2008**, *18*, 27–40. <https://doi.org/10.1007/s11222-007-9035-x>.
9. Nocedal, J.; Wright, S.J. *Numerical Optimization*, 2nd ed.; Springer New York: New York, NY, USA, 2006. <https://doi.org/10.1007/978-0-387-40065-5>.
10. Xie, Y.; Byrd, R.H.; Nocedal, J. Analysis of the BFGS Method with Errors. *SIAM J. Optim.* **2020**, *30*, 182–209. <https://doi.org/10.1137/19m1240794>.
11. Saputro, D.R.S.; Widyaningsih, P. Limited memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) method for the parameter estimation on geographically weighted ordinal logistic regression model (GWOLR). In Proceedings of the 4th International Conference on Research, Implementation, and Education of Mathematics and Science (4th ICRIEMS): Research and Education for Developing Scientific Attitude in Sciences and Mathematics, Yogyakarta, Indonesia, 15–16 May 2017; Vol. 1868, p. 040009. <https://doi.org/10.1063/1.4995124>.
12. Jin, T.; Li, Z.; Chen, K. A novel phase-field monolithic scheme for brittle crack propagation based on the limited-memory BFGS method with adaptive mesh refinement. *Int. J. Numer. Methods Eng.* **2024**, *125*, e7572. <https://doi.org/10.1002/nme.7572>.

13. Dai, Y.H. Convergence Properties of the BFGS Algorithm. *SIAM J. Optim.* **2002**, *13*, 693–701. <https://doi.org/10.1137/s1052623401383455>.
14. Zhou, W.J.; Li, D.H. A globally convergent BFGS method for nonlinear monotone equations without any merit functions. *Math. Comput.* **2008**, *77*, 2231–2240. <https://doi.org/10.1090/s0025-5718-08-02121-2>.
15. Asl, A.; Overton, M.L. Analysis of limited-memory BFGS on a class of nonsmooth convex functions. *IMA J. Numer. Anal.* **2021**, *41*, 1–27. <https://doi.org/10.1093/imanum/drz052>.
16. Ghouse, G.; Rehman, A.U.; Bhatti, M.I. Understanding of Causes of Spurious Associations: Problems and Prospects. *J. Stat. Theory Appl.* **2024**, *23*, 44–66. <https://doi.org/10.1007/s44199-024-00072-0>.
17. AlMuhayfith, F.E.; Alzaid, A.A.; Omair, M.A. On bivariate Poisson regression models. *J. King Saud Univ. Sci.* **2016**, *28*, 178–189. <https://doi.org/10.1016/j.jksus.2015.09.003>.
18. Muñoz-Pichardo, J.M.; Pino-Mejías, R.; García-Heras, J.; et al. A multivariate Poisson regression model for count data. *J. Appl. Stat.* **2021**, *48*, 2525–2541. <https://doi.org/10.1080/02664763.2021.1877637>.
19. Morris, T.P.; White, I.R.; Crowther, M.J. Using simulation studies to evaluate statistical methods. *Stat. Med.* **2019**, *38*, 2074–2102. <https://doi.org/10.1002/sim.8086>.
20. Byrne, C.; Eggermont, P.P.B. EM Algorithms. In *Handbook of Mathematical Methods in Imaging*; Springer New York: New York, NY, USA, 2011; pp. 271–344. https://doi.org/10.1007/978-0-387-92920-0_8.
21. Byrne, C. The EM Algorithm. In *Applied Iterative Methods*; A K Peters/CRC Press: Boca Raton, FL, USA, 2007; pp. 345–348. <https://doi.org/10.1201/b10651-46>.
22. Lehmann, E.L.; Casella, G. *Theory of Point Estimation*, 2nd ed.; Springer: New York, NY, USA, 1998. <https://doi.org/10.1007/b98854>.
23. White, H. Maximum Likelihood Estimation of Misspecified Models. *Econometrica* **1982**, *50*, 1–25. <https://doi.org/10.2307/1912526>.
24. McLachlan, G.J.; Krishnan, T. *The EM Algorithm and Extensions*, 2nd ed.; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 2008. <https://doi.org/10.1002/9780470191613>.
25. Cameron, A.C.; Trivedi, P.K. *Regression Analysis of Count Data*; Cambridge University Press: Cambridge, UK, 1998; Vol. 30.
26. Akaike, H. A new look at the statistical model identification. *IEEE Trans. Autom. Control* **1974**, *19*, 716–723. <https://doi.org/10.1109/tac.1974.1100705>.
27. Burnham, K.P.; Anderson, D.R. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*, 2nd ed.; Springer: New York, NY, USA, 2004. <https://doi.org/10.1007/b97636>.
28. Schwarz, G. Estimating the Dimension of a Model. *Ann. Stat.* **1978**, *6*, 461–464. <https://doi.org/10.1214/aos/1176344136>.
29. Cox, D.R.; Reid, N. Parameter Orthogonality and Approximate Conditional Inference. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **1987**, *49*, 1–18. <https://doi.org/10.1111/j.2517-6161.1987.tb01422.x>.
30. Liu, C. Maximum likelihood estimation from incomplete data via EM-type algorithms. In *Advanced Medical Statistics*; World Scientific: Singapore, 2003; pp. 1051–1071. https://doi.org/10.1142/9789812388759_0028.
31. Tribbey, W. Numerical Recipes. *ACM SIGSOFT Softw. Eng. Notes* **2010**, *35*, 30–31. <https://doi.org/10.1145/1874391.187410>.
32. Dempster, A.P.; Laird, N.M.; Rubin, D.B. Maximum Likelihood from Incomplete Data Via the EM Algorithm. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **1977**, *39*, 1–22. <https://doi.org/10.1111/j.2517-6161.1977.tb01600.x>.
33. Tzougas, G.; di Cerchiaro, A.P. Bivariate Mixed Poisson Regression Models with Varying Dispersion. *N. Am. Actuar. J.* **2023**, *27*, 211–241. <https://doi.org/10.1080/10920277.2021.1978850>.