Advance Sustainable Science, Engineering and Technology (ASSET)
Vol. 7, No.4, October 2025, pp. 02504024-01 ~ 02504024-015
ISSN: 2715-4211 DOI: https://doi.org/10.26877/asset.v7i4.2282

Evaluating Ordinal Multivariate Models under Multicollinearity via Pairwise Likelihood: A Simulation Perspective

Achmad Fauzan[1,2*], Kusman Sadik[2], Anang Kurnia[2]

[1]Statistics Study Program, Faculty of Mathematics and Natural Science, Universitas Islam Indonesia, Yogyakarta, Indonesia

[2]Study Program of Statistics and Data Science, School of Data Science, Mathematics and Informatics, IPB University, Indonesia

*achmadfauzan@uii.ac.id, fauzanachmad@apps.ipb.ac.id

Abstract. This study examines the effect of multicollinearity on ordinal regression through a two-stage Monte Carlo simulation. A synthetic population of 2,000,000 observations was generated with predictors drawn from a normal distribution, and responses simulated using an ordinal probit model. A Monte Carlo procedure was employed with 10 repetitions, each consisting of 100 random samples of 1,000 observations. Parameter estimation employed Maximum Likelihood Estimation (MLE) for univariate models and Pairwise Likelihood (PL) for multivariate models, with performance assessed using mean squared error (MSE), bias, and computation time. Results show that multicollinearity had negligible impact on estimator bias and MSE, confirming the robustness of both MLE and PL to correlated predictors. However, severe multicollinearity substantially increased computation time, indicating a trade-off between estimator stability and efficiency. These findings highlight PL as a viable approach for analyzing complex ordinal data, particularly in applications such as socio-economic surveys and health metrics where predictor correlation is unavoidable.

Keywords: Latent Variable Modeling, Monte Carlo Simulation, Multivariate Ordinal Regression, Ordinal Probit, Pairwise Likelihood Estimation (PL).

(Received 2025-06-30, Revised 2025-09-03, Accepted 2025-09-18, Available Online by 2025-10-28)

1. Introduction

Multivariate regression represents a statistical methodology designed to assess the simultaneous effects of one or more independent variables on multiple dependent outcomes [1]. In contrast to univariate regression, which focuses on a single response variable, the multivariate framework enables researchers to investigate more intricate patterns of association within systems characterized by interdependent outcomes. This approach is particularly valuable in disciplines such as psychology [2], economics [3], [4], and epidemiology [5], where multiple outcome indicators often exhibit substantial correlations and

therefore require joint modeling for valid inference. One of the principal advantages of multivariate regression is its capacity to capture the interdependence structure among multiple response variables. By modeling all outcome variables simultaneously, the analysis becomes more efficient and produces more accurate parameter estimates, particularly when the response variables are substantively correlated [4]. For categorical response data with more than two levels, two general modeling approaches are commonly used: multinomial regression for unordered (nominal) categories, and ordinal logistic regression for ordered (ordinal) outcomes [6]. The multivariate ordinal regression model, also referred to as multivariate ordinal probit regression, integrates two essential components: the ordinal nature of the response variables and the correlation structure among them [7,8].

As the number of response dimensions increases, full maximum likelihood estimation becomes increasingly prohibitive. Consequently, researchers have explored alternative estimation strategies that are computationally more efficient while maintaining statistical consistency. The pairwise likelihood (PL) has become a well-recognized solution in the statistical literature [7,9]. Instead of evaluating the full joint likelihood of all response variables, PL relies solely on contributions from pairs of response variables, thereby reducing the complexity of high-dimensional integrals to more computationally manageable expressions. This method has demonstrated effectiveness across a wide range of applications, especially in models involving numerous response variables and complex correlation structures [10,11]. More recently, Wieditz [12] and Gambarota & Altoè [13] highlighted the use of simulation-based evaluations to investigate ordinal regression under correlated latent structures, confirming the consistency of PL in practice. Nevertheless, while theoretical advances and software implementations have expanded rapidly, empirical simulation studies that systematically evaluate the robustness of PL, particularly under varying levels of predictor multicollinearity, remain limited.

Although PL has been established as a computationally efficient alternative in multivariate ordinal regression, most existing studies primarily emphasize its theoretical properties or its performance under correlated latent response structures. Recent empirical works have contributed valuable insights into the consistency of PL estimators; however, systematic simulation-based evaluations remain limited in scope. In particular, little attention has been given to how varying levels of predictor multicollinearity influence the accuracy and stability of PL estimates. This study directly addresses the identified research gap by systematically evaluating the impact of predictor multicollinearity on parameter estimation using Pairwise Likelihood Estimation (PLE) in both univariate and multivariate ordinal regression models. While prior works have primarily emphasized latent response correlations, little attention has been given to predictor-side multicollinearity, a pervasive issue in applied research. To fill this gap, we conduct large-scale Monte Carlo simulations that explicitly vary degrees of predictor multicollinearity and assess their effects on estimator bias, variance, mean squared error (MSE), and computation time. In doing so, this research provides new empirical evidence on the robustness of PLE under challenging predictor correlation structures and offers methodological guidance for applied contexts such as socio-economic surveys, psychometric assessments, and epidemiological studies.

2. **Methods**

2.1. Research Methods

This simulation study evaluates the consistency of parameter estimates in univariate and multivariate ordinal regression models under varying levels of predictor multicollinearity. The primary goal is to assess the stability of estimates and the validity of consistency assumptions when predictors are highly correlated. The overall research flow is summarized as follwos. First, ordinal responses are generated using a probit model from latent normally distributed variables, categorized through thresholding. Parameter estimates are repeatedly compared to known true values to assess bias, variance, and systematic deviation. Second, to assess the impact of multicollinearity, predictor variables were designed to reflect three levels of intercorrelation based on Variance Inflation Factor (VIF): categorized as low when $VIF < 5$, moderate when $5 \leq VIF < 10$, and high when $VIF \geq 10$, with additional validation using the Pearson correlation matrix. This design was applied to both univariate and multivariate ordinal

probit regression models, where ordinal outcomes are treated as thresholded realizations of latent variables following a (multivariate) normal distribution. In the multivariate case, inter-response correlations are modeled through the latent covariance structure. An overview of the simulation design is illustrated in Figure 1.

Third, to reflect real-world settings where the true data-generating process is unknown, a synthetic population of 2,000,000 observations was generated. Predictors were drawn from a standard normal distribution, and ordinal responses were created using a probit regression framework. Fourth, a Monte Carlo design with 10 repetitions was employed, where each repetition involved drawing 100 random samples of 1,000 observations each. This setup mimics model evaluation based on holdout samples. Fifth, for every simulation, parameter estimates were obtained across all predictors and resampling iterations, resulting in 100 estimates per predictor per repetition. Mean Squared Error (MSE) was then computed by comparing these estimates with the true parameter values [14].

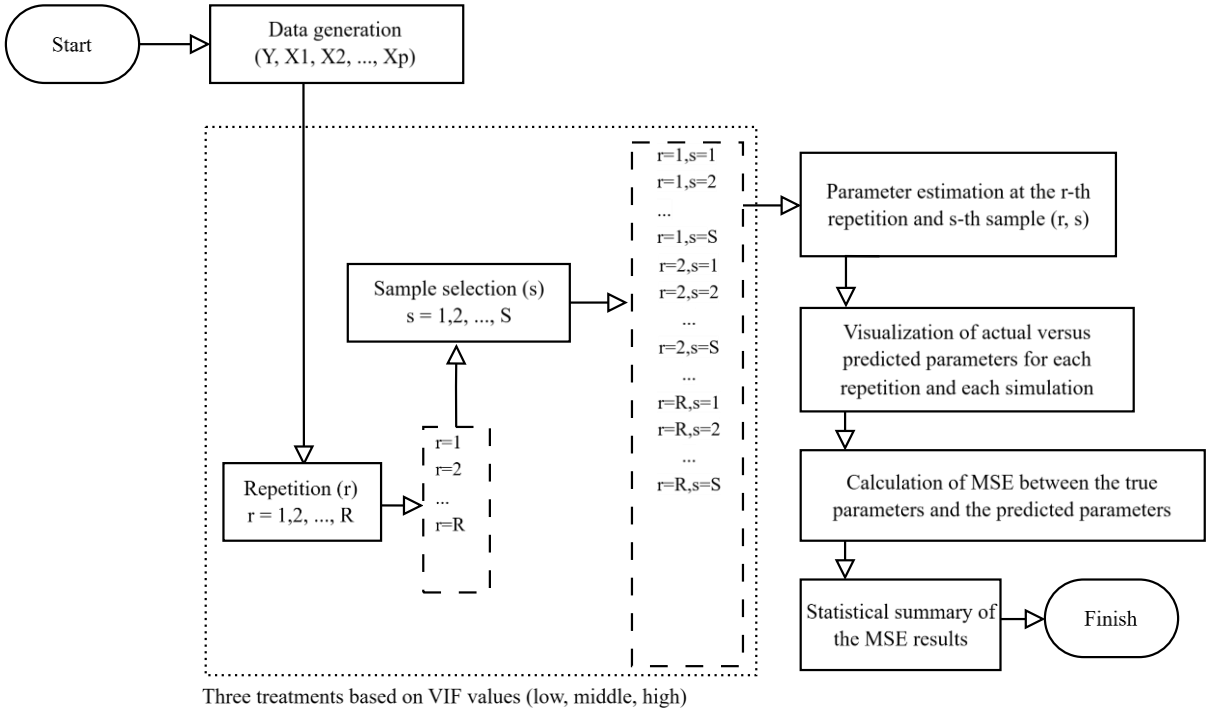


Figure 1. Simulation design.

Sixth, based on the MSE computed in each repetition, summary results are reported for low, moderate, and high multicollinearity scenarios. The simulation results are further illustrated using dot plots (showing parameter distribution across repetitions) and box plots (highlighting estimator variability and deviation from true values). These visualizations enable a comprehensive assessment of estimation accuracy under increasing multicollinearity in both univariate and multivariate settings. All computations were performed in `R` (version 4.4.2) on a system with 16 GB RAM, ensuring scalability for large-scale simulations [15]. The study contributes methodological insights into how ordinal probit regression performs under multicollinearity while emphasizing the computational efficiency of the pairwise likelihood (PL) approach in handling multivariate ordinal data. Numerous `R` packages are available to support the modeling of ordinal data [15]. In the present simulation study, univariate ordinal regression was performed using the `polr()` function in the `MASS` pacakage and for the multivariate ordinal setting, we adopted the `MMO2` specification implemented in the `mvord` package, which accommodates multiple correlated ordinal responses within a unified modeling framework.

2.2. Univariate Ordinal Regression

As a core statistical tool, the Linear Model (LM) offers a basis for analyzing the linear relationship between a response variable and its predictors [16]. The approach was later broadened through the Generalized Linear Model (GLM), which allows for flexible response distributions beyond normality and the use of alternative link functions [17]. Logistic regression represents a specific case of GLM, applicable when the outcome variable is categorical—either binary or polytomous, and nominal or ordinal in nature. Ordinal logistic regression is particularly suited for modeling ordered categorical outcomes with three or more levels. Unlike multinomial logistic regression, which is used for nominal outcomes without inherent ordering, ordinal models leverage the natural rank of the response variable. Depending on the link function used, ordinal models can be estimated via the logit or probit link. The ordinal logit model is widely used due to its interpretability through odds ratios, making it especially relevant for applied research involving relative risk assessment. Meanwhile, the ordinal probit model, also known as the threshold model, assumes normally distributed latent errors and is preferred in contexts such as psychometrics or latent trait modeling, where normality assumptions are theoretically grounded [18,19].

Let Y_i denote an ordinal response variable with c ordered categories, and let $\boldsymbol{x_i}$ represent a vector of p predictor variables for the i^{th}observation, such that $\boldsymbol{x_i} = [\boldsymbol{x_{i1}} \quad \boldsymbol{x_{i2}} \quad \cdots \quad \boldsymbol{x_{ip}}]^T$ with $i = 1,2,\cdots,n$, with n indicating the total count of observations. The ordinal probit model, the estimation process begins with the specification of the structural form, as defined in Equation 1 [12].

$$Y_i^* = \boldsymbol{x_i^T \beta} + \boldsymbol{\varepsilon_i} \tag{1}$$

The latent variable y^* is a continuous, unobserved variable assumed to follow a normal distribution $\boldsymbol{Y^*} \sim N(\boldsymbol{\beta^T x}, \sigma^2)$, $\boldsymbol{x}$ is the vector of predictor variables, $\boldsymbol{\beta}$ denotes the corresponding parameter coefficients, and $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2)$ represents the normally distributed error term. The term "latent" refers to an unobservable construct that cannot be directly measured but is inferred from observed data [19]. As with the ordinal logit model, suppose the response variable consists of c ordered categories. In the ordinal probit model, the observed categorical outcome is derived from the latent variable Y^* through a series of thresholds γ, such that the classification occurs in an ordered manner as Equation 2 [13,20]

$$\begin{aligned} Y_i &= 1 \text{ if } Y_i^* \leq \gamma_1 \\ Y_i &= 2 \text{ if } \gamma_1 < Y_i^* \leq \gamma_2 \\ &\vdots \\ Y_i &= c \text{ if } Y_i^* > \gamma_{c-1} \end{aligned} \tag{2}$$

The thresholds on the latent scale are defined as $-\infty \equiv \gamma_0 < \gamma_1 < \gamma_2 < \cdots < \gamma_{c-1} < \gamma \equiv \infty$ where γ epresents the cutoff points that partition the continuous latent variable into ordered response categories [21]. An illustration of the relationship between the latent variable values and the observed ordinal outcomes in the ordered probit model is presented in Figure 2.

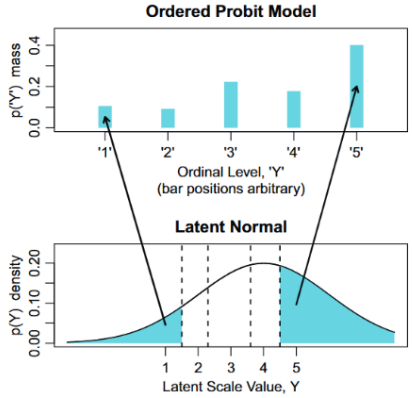


Figure 2. The upper panel displays ordinal outcomes derived from thresholding a cumulative normal distribution, as illustrated in the lower panel [12,22]

The general form of the ordinal probit model can be expressed as Equation 3 [9].

$$\Pr(Y_i \le j) = \Pr\left(\boldsymbol{x_i^T\beta} + \varepsilon_i \le \gamma_j\right) = \Phi\left(\gamma_1 - \boldsymbol{x_i^T\beta}\right) = \pi_1(\boldsymbol{x_i}) + \pi_2(\boldsymbol{x_i}) + \cdots + \pi_j(\boldsymbol{x_i}) \quad (3)$$

where $\Phi(\cdot)$ denotes the cumulative probability function of the standard normal distribution, mapping each value to the probability that a standard normal variable is less than or equal to it. The term $\pi_j(\boldsymbol{x_i})$ represents the probability that the i^{th} observation falls into category j, given the predictor values $\boldsymbol{x_i}$. Parameter estimation in the ordinal probit model is conducted using the MLE method as Equation 4 [23].

$$\begin{aligned} \mathcal{L}(\boldsymbol{\theta}) &= \prod_{i=1}^{n}\prod_{j=1}^{c}\left(\pi_j(\boldsymbol{x_i})\right)^{Y_{ij}} \\ \ell(\boldsymbol{\theta}) &= \ln\left(\mathcal{L}(\theta)\right) = \sum_{i=1}^{n}\sum_{j=1}^{\boldsymbol{c}} Y_{ij}\ln(\pi_j(\boldsymbol{x_i}) \end{aligned} \quad (4)$$

$\boldsymbol{\theta} = [\boldsymbol{\gamma} \quad \boldsymbol{\beta}]^T = [\gamma_1 \quad \gamma_2 \quad \cdots \quad \gamma_{c-1} \quad \beta_1 \quad \beta_2 \quad \cdots \quad \beta_p]^T$. As with the ordinal logit regression model, the first-order partial derivatives of the log-likelihood function in the ordinal probit model are not available in closed form. Consequently, numerical optimization techniques are required for parameter estimation.

2.3. Multivariate Ordinal Regression using Pairwise Likelihood

Ordinal multivariate regression models multiple correlated ordinal responses as thresholded latent variables, extending the univariate probit framework to account for inter-response correlations and predictor effects [8,11]. As an illustration, consider $i = 1,2,\cdots,n$ independent observations, each associated with a q-dimensional ordinal response vector $\boldsymbol{Y_i} = \left(Y_{i1}, Y_{i2}, \cdots, Y_{iq}\right)^T$ where $Y_{ij} \in \{1,2,\cdots,K\}$ for $j = 1,2,\cdots,q$. Each $\boldsymbol{Y_i}$ is treated as a realization from a joint distribution $g(\boldsymbol{Y_i})$ that depends on an unknown parameter vector $\boldsymbol{\theta}$. According to the model, each individual i is assumed to possess an underlying continuous latent vector that cannot be directly observed $\boldsymbol{Z}_i = \left(Z_{i1}, Z_{i2}, Z_{i3}, \cdots, Z_{iq}\right)^T$, which follows a multivariate normal distribution $\mathcal{N}(0, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma}$ is a symmetric and positive definite $q \times q$ correlation matrix.

The general structure of $\boldsymbol{\Sigma}$ includes unit values on the diagonal and off-diagonal elements ρ_{rs}, representing the correlation between the r^{th} and s^{th} dimensions of the latent variables. The observed ordinal variable y_{ij} is obtained by discretizing the latent variable z_{ij} through a set of threshold values$\{a_k\}$, such as $y_{ik} = k \Leftrightarrow z_{ij} \in (\alpha_{k-1}, a_k)$. These thresholds are ordered as $-\infty \equiv a_0 < a_1 < a_2 < \cdots < a_K \equiv \infty$ orming a partition of the real line into K mutually exclusive intervals [8]. Based on this structure, the likelihood function for a single observational unit can be expressed as a multivariate normal integral, where the limits of integration are determined by the threshold values associated with each ordinal dimension. Specifically, the joint probability of observing the ordinal response vecto $\boldsymbol{Y_i} = (y_{i1}, y_{i2}, \cdots, y_{iq})$ is given by Equation 5 [8].

$$\Pr\left(Y_{\mathrm{i1}} = \mathrm{y_{i1}}, Y_{\mathrm{i2}} = \mathrm{y_{i2}}, \cdots, Y_{iq} = y_{iq}\right) = \int_{a_{y_{i1}-1}}^{a_{y_{i1}}} \int_{a_{y_{i2}-1}}^{a_{y_{i1}}} \cdots \int_{a_{y_{iq}-1}}^{a_{y_{i1}}} \phi\boldsymbol{\Sigma}\left(z_{i1}, z_{i2}, \cdots, z_{iq}\right) dz_{i1} dz_{i2} \cdots dz_{iq} \quad (5)$$

where $\phi\,\boldsymbol{\Sigma}(\cdot)$ corresponds to the multivariate normal distribution's density, specified by a zero mean vector and covariance structure $\boldsymbol{\Sigma}$. Accordingly, the full log-likelihood function for all n observational units can be expressed as Equation 6.

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log\left\{ \int_{a_{y_{i1}-1}}^{a_{y_{i1}}} \int_{a_{y_{i2}-1}}^{a_{y_{i1}}} \cdots \int_{a_{y_{iq}-1}}^{a_{y_{i1}}} \phi\boldsymbol{\Sigma}\left(z_{i1}, z_{i2}, \cdots, z_{iq}\right) dz_{i1} dz_{i2} \cdots dz_{iq} \right\} \quad (6)$$

From the Equation 6, several parameters are involved in the model: (1) the set of threshold parameters used for categorization $(a_1, a_2, \cdots, a_{K-1})$, (2) the latent correlation parameters ρ_{rs} for all pairs $r, s = 1,2,\cdots,q$ with $r < s$, and (3) the complete parameter vector $\boldsymbol{\theta} = \left(a_1, a_2, \cdots, a_{K-1}, \rho_{12}, \rho_{13}, \cdots, \rho_{q-1,q}\right)$

where $\widehat{\boldsymbol{\theta}}$. However, the likelihood function in the previous equation involves a q-dimensional integral for each individual observation. This implies that evaluating the full likelihood requires computing n multivariate Gaussian integrals of high dimension. Consequently, direct estimation becomes computationally intensive and impractical, particularly when the number of response variables q is large or the sample size n is substantial. To address this limitation, pairwise likelihood (PL) methods employed as a more computationally efficient alternative.

The PL approach is a simplified alternative to the full likelihood function, which only considers bivariate (pairwise) combinations of the response variables. Specifically, the pairwise log-likelihood is constructed by summing the joint probabilities of all possible pairs of ordinal response variables, rather than evaluating the full multivariate distribution. This reduction in complexity makes the estimation process more computationally feasible, particularly in high-dimensional settings. The general form of the pairwise log-likelihood is given by the following Equation 7 [24].

$$\ell^{p}(\boldsymbol{\theta}) = \sum_{i=1}^{n}\sum_{r=1}^{q-1}\sum_{s=r+1}^{q} \log\{\Pr(Y_{ir} = y_{ir}, Y_{is} = y_{is})\} \tag{7}$$

where each term corresponds to the log of the joint probability of observing the pair (Y_{ir}, Y_{is}) for subject i, and the summation runs over all possible unique pairs of response variables. Equation 7 can be expressed in terms of a bivariate integral, as Equation 8.

$$\ell^{p}(\boldsymbol{\theta}) = \sum_{i=1}^{n}\sum_{r=1}^{q-1}\sum_{s=r+1}^{q} \log\left\{\int_{a_{y_{ir}-1}}^{a_{y_{ir}}}\int_{a_{y_{is}-1}}^{a_{y_{is}}} \phi\boldsymbol{\Sigma}(\rho_{r,s})(z_{ir}, z_{is})\, dz_{ir}dz_{is}\right\} \tag{8}$$

The matrix $\boldsymbol{\Sigma}(\rho_{r,s})$ efers to a 2×2 correlation matrix with ones on the diagonal, representing unit variances, and the correlation coefficien ρ_{rs} as the off-diagonal elements, which quantify the association between the two latent variables. The PL approach is particularly advantageous because it replaces the computationally intractable high-dimensional integrals in the full likelihood with a series of more manageable bivariate integrals, which can be efficiently evaluated using standard statistical software. In general, PL operate in a manner analogous to traditional likelihood-based procedures. For instance, the pairwise score vector $U^{P}(\boldsymbol{\theta})$, defined as the derivative of the pairwise log-likelihood with respect to the parameter vector $\boldsymbol{\theta}$, retains the property of unbiasedness. This is because it is constructed as the sum of contributions from all pairs of response variables. The formal expression for the score function is presented in the following Equation 9.

$$U^{P}(\boldsymbol{\theta}) = \frac{\partial \ell^{p}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \sum_{i=1}^{n}\sum_{r=1}^{q-1}\sum_{s=r+1}^{q} \frac{1}{P_{irs}(\theta)}\frac{\partial}{\partial\theta}\left\{\int_{a_{y_{ir}-1}}^{a_{y_{ir}}}\int_{a_{y_{is}-1}}^{a_{y_{is}}} \phi\boldsymbol{\Sigma}(\rho_{r,s})(z_{ir}, z_{is})\, dz_{ir}dz_{is}\right\} \tag{9}$$

Parameter estimation in the PL framework, denoted as $\widehat{\boldsymbol{\theta}}^{P}$, can be obtained either by maximizing the pairwise log-likelihood function$\ell^{p}(\boldsymbol{\theta})$, or equivalently, by solving the score equation $U^{P}(\boldsymbol{\theta}) = 0$, similar to the standard maximum likelihood estimation approach. The PL estimator is consistent and asymptotically normally distributed as the sample size n becomes large $n \to \infty$ [25]. More formally, the asymptotic distribution of the estimator is $\widehat{\boldsymbol{\theta}}^{P} \sim \mathcal{N}(\theta, G(\theta)^{-1})$, $G(\boldsymbol{\theta}) = W(\boldsymbol{\theta})J(\boldsymbol{\theta})^{-1}W(\boldsymbol{\theta})$ referred to as the Godambe information matrix, which serves as an analogue to the Fisher information matrix in full likelihood settings and is used to approximate the asymptotic variance of the estimator. Godambe information matrix is constructed from the combination of two key components: the sensitivity matrix $\boldsymbol{W}(\boldsymbol{\theta}) = \mathbb{E}_{\theta}\left\{\frac{-\partial U^{P}(\theta)}{\partial}\right\}$ and the variability matrix $\boldsymbol{J}(\boldsymbol{\theta}) = \boldsymbol{Var}_{\boldsymbol{\theta}}\{U^{p}(\boldsymbol{\theta})\}$ [24]. In the multivariate setting, parameter estimation is performed using the PL method. This approach approximates the full likelihood by considering only the bivariate marginal distributions of each pair of ordinal responsesBy reducing the high-dimensional integral in the full likelihood to multiple two-dimensional integrals, the PL method significantly alleviates computational complexity and is well-suited for high-dimensional multivariate settings. As an illustration, we have included an appendix demonstrating parameter estimation using both the univariate and the multivariate ordinal regression. The appendix provides a comparative

example of how parameters are estimated under each link function, along with a simulated dataset and the corresponding ordinal logistic regression models.

Beyond its theoretical foundations, PL has also been applied in diverse empirical domains, demonstrating its practical utility. In psychometrics, Robitzsch (2024) showed that PL can yield nearly unbiased estimates in the two-parameter logistic (2PL) item response theory model, even when the local independence assumption is violated due to stimulus-based dependencies across items. In computational biology, Mazo et al. (2024) introduced a randomized pairwise likelihood (RPL) approach to efficiently analyze transcriptomic count data, reducing computational burden while maintaining estimator consistency [27]. In health metrics, Tong et al. (2022) developed a privacy-preserving distributed conditional logistic regression (dCLR) algorithm that leverages PL to integrate electronic health records from 230 hospitals during the COVID-19 pandemic, achieving greater robustness compared to meta-analysis methods, especially under rare-event settings [28]. These applied studies underscore the versatility of PL as not only a theoretically sound estimation method but also a robust and scalable solution for addressing correlation structures, data heterogeneity, and computational challenges in real-world applications.

3. **Result**

3. 1. Univariate Ordinal regression Simulation

As shown in the flow diagram, the simulation begins with a univariate ordinal probit model and adopts a Monte Carlo approach to assess the stability of parameter estimates. Due to computational constraints, the simulation was limited to $r = 10$ repetitions. In each repetition, a large synthetic population ($n = 2{,}000{,}000$) was generated, from which $s = 100$ random samples (each ≈1,000 observations, or 0.05%) were drawn to mimic data sparsity conditions. To ensure valid ordinal model estimation, subsampling was controlled so that all outcome categories were present in each training set. Predictor variables X_1 to X_4 ($p = 4$) were drawn from a standard normal distribution and standardized using z-score transformation. True regression coefficients (β) were randomly varied across iterations by perturbing a baseline vector [-2, 3, -4, and 0.6] with noise from $\mathcal{N}(0,1)$. In the initial simulation scenario, the predictors were designed to exhibit low multicollinearity, as indicated by VIF below 5.

For clarity of exposition, the simulation workflow is summarized as a pseudocode in Algorithm 1. The pseudocode details the application of Pairwise Likelihood Estimation (PLE) in both univariate and multivariate ordinal probit models, evaluated under varying degrees of predictor multicollinearity. By structuring the steps of data generation, resampling, estimation, and performance assessment, the pseudocode provides a transparent and reproducible framework for examining the robustness and efficiency of PLE.

Algorithm 1

Input:

- N: synthetic population size (e.g., 2,000,000)
- n: sample size per estimation (e.g., 1,000)
- R: number of Monte Carlo replications (e.g., 10)
- S: number of resamples per replication (e.g., 100)
- p: number of predictors (e.g., 3–4)
- β^{true}: true regression coefficients (intercept included)
- Threshold probabilities:
  - Univariate: $(0.15, 0.35, 0.70, 0.85)$
  - Multivariate: response-specific (e.g., $1/3, 2/3, 0.25, 0.5, 0.75$)
- Σ: latent residual covariance matrix (for multivariate responses)
- VIF thresholds: low (< 5), moderate ($5 - 10$), high (≥ 10)
- T_{max}: maximum resampling attempts to ensure complete category coverage

Output:

- Estimated parameters $(\widehat{\beta},\ \widehat{\zeta})$for univariate and multivariate models

- Variance Inflation Factors (VIF)
- Bias, variance, and Mean Squared Error (MSE) summaries
- Visual diagnostics (boxplots of estimates, MSE curves)

Procedure

1. **Initialization**
   1.1. Set random seed; load required libraries (`MASS, mvord, dplyr, car, ggplot2, mvtnorm`)
   1.2. Define simulation parameters (N, n, R, S, T_{max}) and true values of β^{true}.
2. **Population Generation**
   2.1. Generate baseline predictors $X_j \sim N(0,1), j = 1,2 \cdots, p$
   2.2. For each multicollinearity level $L \in \{\text{low}, \text{moderate}, \text{high}\}$
   - Construct predictors with target VIF structure:
     a. Low: independent normal predictors.
     b. Moderate/High: introduce linear dependencies among predictors.
   - Standardize predictors and compute VIF; log results.

3. **Response Construction**
   Univariate Case:
   3.1. Compute latent variable $Y^* = X\beta^{true} + \varepsilon, \varepsilon \sim N(0,1)$
   3.2. Derive thresholds $\tau = \Phi^{-1}(probs)$
   3.3. Categorize $Y = cut\ (Y^*, \tau)$ into 5 ordered categories (*for example*).
   Multivariate Case
   3.4. For each response $r = 1,2, \cdots, q$:
   - Compute latent variable $Z_r = X\beta_r^{true} + \varepsilon_r$, with $\varepsilon \sim N_q(0, \Sigma)$.
   - Apply response-specific thresholds $\tau = \Phi^{-1}(probs)$
   - Categorize $Y_r = cut\ (Y^*, \tau_r)$ into C_r ordered categories (*for example*).

4. **Monte Carlo Sampling**
   For each replication $i = 1,2, \cdots, R$:
   4.1. (*Optional*) Perturb β^{true} slightly to mimic parameter variability.
   4.2. For each subsample $s = 1,2, \cdots, S$:
   - Draw random sample of size n from population.
   - If any category is missing, resample until all categories are represented ($\leq T_{max}$)
   - Estimate parameters:
     a. Univariate model: Maximum Likelihood Estimation via `polr(..., method="probit")`.
     b. Multivariate model: Pairwise Likelihood Estimation via `mvord(..., link="mvprobit")`.
   - Store coefficient estimates, thresholds, and standard errors.

5. **Performance Evaluation**
   5.1. Compute Bias, Variance, and MSE for each parameter:

$$MSE_k = \frac{1}{R.S} \sum_{i=1}^{R} \sum_{s=1}^{S} \left(\hat{\beta}_i, s, k - \beta_k^{true}\right)^2$$

   5.2. Summarize results by multicollinearity level: mean, SD, min, max of MSE.
   5.3. Visualize
   - Boxplots of estimated coefficients with true values indicated.
   - Line plots of MSE across replications.
6. **Output**
   - Consolidate results into tables (parameter estimates, VIF log, MSE summary).

- Report computational time and goodness-of-fit indices (log-PL, CLAIC, CLBIC for multivariate).
- Provide diagnostic plots for interpretation.

7. **End of Algorithm**

All simulation results were compiled into a panel data structure, enabling both the visualization of estimation dynamics and the computation of performance metrics such as MSE across conditions. Figure 3 provides a visual comparison between the true parameters and their estimates, illustrating the accuracy and stability of the estimation process under varying sampling conditions.

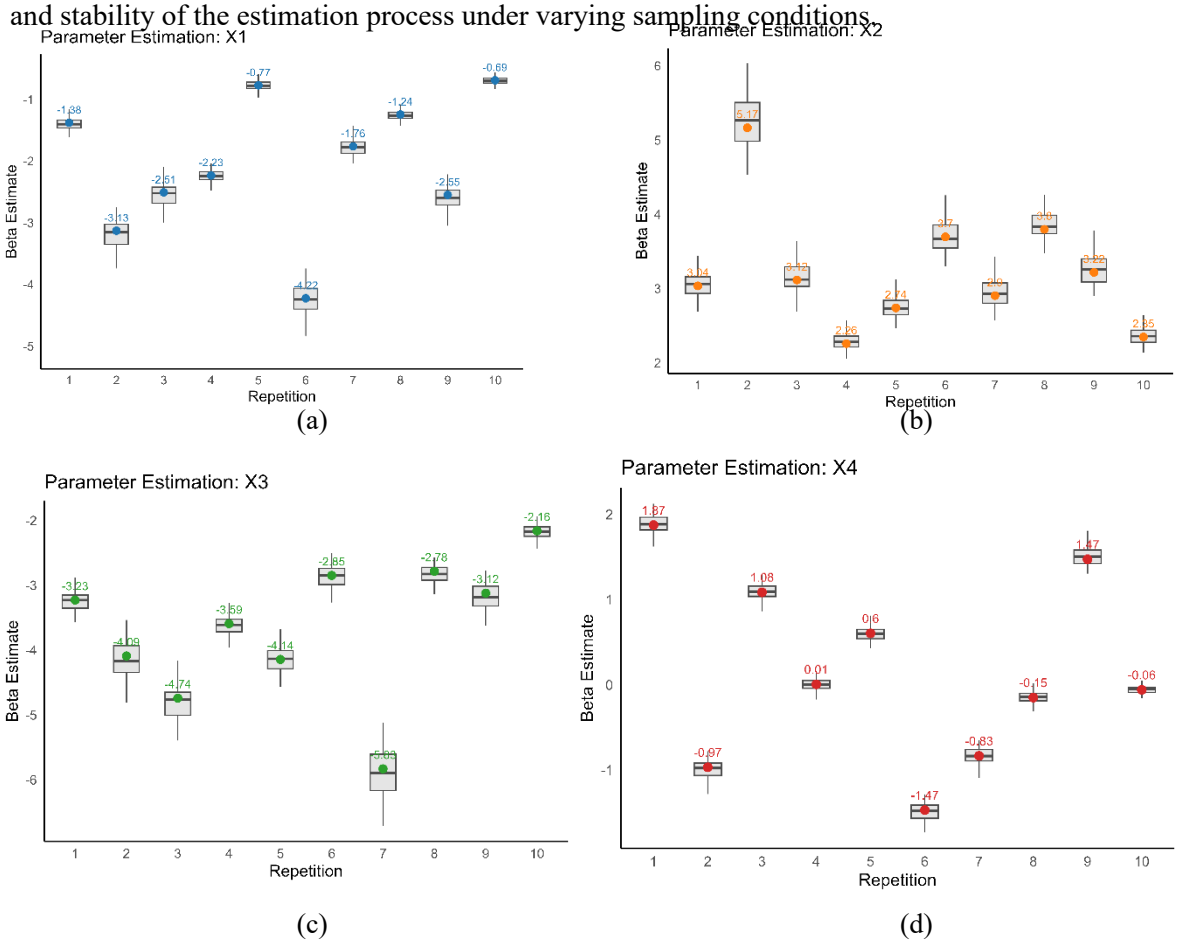


Figure 3. Parameter estimates across Monte Carlo repetitions, (a) X_1, (b) X_2, (c) X_3, and (d) X_4.

Based on Figure 3, the estimated parameter values appear to be closely aligned with the true (simulated) parameter values, suggesting that the estimators are unbiased. At a glance, the associated standard errors also appear relatively small. To further support this observation, two additional simulation scenarios were conducted: one under moderate multicollinearity ($5 < \text{VIF} \leq 10$), and another under high multicollinearity ($\text{VIF} > 10$). Both simulations followed the same design and procedural steps as the first. MSE for each predictor variable, calculated by comparing the estimated regression coefficients with their corresponding true parameter values, a quantitative assessment of estimation accuracy across simulation repetitions ($MSE_{r,p}$) as Equation 10.

$$MSE_{r,p} = \frac{1}{s}\sum_{i=1}^{s}\left(\beta_{r,p}^{true} - \hat{\beta}_{r,s,p}\right)^2 \tag{10}$$

$MSE_{r,p}$ denote the MSE for the p-th variable in the r-th repetition. The term $\beta_{r,p}^{true}$ refers to the true parameter value, while $\hat{\beta}_{r,s,p}$ represents the estimated parameter value obtained from the s-th resamping within the r-th repetition for the p-th predictor variable. Here, $r = 1,2,\cdots,10$, $s = 1,2,\cdots,100$, and $p = 1,2,3,4$. An illustration of this formulation is presented in Equation 11, which summarizes the

computation of MSE arranged by repetition (r) and variable (p).

$$MSE_{1,1} = \frac{1}{100}\sum_{s=1}^{100}\left(\beta_{1,1} - \hat{\beta}_{1,s,1}\right)^2 = \frac{1}{100}\left(\left(\beta_{1,1} - \hat{\beta}_{1,1,1}\right)^2 + \left(\beta_{1,1} - \hat{\beta}_{1,2,1}\right)^2 + \cdots + \left(\beta_{1,1} - \hat{\beta}_{1,100,1}\right)^2\right)$$
$$\vdots \quad (11)$$
$$MSE_{10,4} = \frac{1}{s}\sum_{i=1}^{100}\left(\beta_{10,4} - \hat{\beta}_{10,s,4}\right)^2 = \frac{1}{100}\left(\left(\beta_{10,4} - \hat{\beta}_{10,1,4}\right)^2 + \left(\beta_{10,4} - \hat{\beta}_{10,2,4}\right)^2 + \cdots + \left(\beta_{10,4} - \hat{\beta}_{10,100,4}\right)^2\right)$$

Subsequently, the MSE values for each variable across all repetitions were summarized and presented in Table 1. This summary provides an overview of the estimation performance and variability for each predictor over the course of the simulation study.

Table 1. Summary of MSE between estimated and true parameter values from each repetition

No	VIF	Variable	Mean	Standar Deviation	Minimum	Maximum
1	$VIF \leq 5$	X_1	0.024	0.023	0.004	0.073
2		X_2	0.043	0.036	0.013	0.137
3		X_3	0.053	0.041	0.011	0.139
4		X_4	0.009	0.004	0.002	0.014
5	$5 < VIF \leq 10$	X_1	0.038	0.016	0.019	0.061
6		X_2	0.036	0.020	0.011	0.081
7		X_3	0.043	0.037	0.009	0.138
8		X_4	0.021	0.005	0.011	0.029
9	> 10	X_1	0.100	0.035	0.053	0.162
10		X_2	0.038	0.023	0.012	0.093
11		X_3	0.037	0.027	0.009	0.101
12		X_4	0.073	0.023	0.043	0.117

The Table 1 presents a summary of the MSE values obtained from comparing the estimated parameters with the true values across three levels of multicollinearity, categorized based on VIF thresholds in the univariate ordinal regression simulations. Overall, there is a noticeable trend indicating that higher VIF levels are associated with increased average MSE values—most prominently observed in predictors X_1 and X_4. For instance, the average MSE for X_1 rises substantially from 0.024 (VIF ≤ 5) to 0.100 (VIF > 10). Nevertheless, this pattern does not hold consistently across all predictors. Variables such as X_2 and X_3 exhibit relatively minor variations in MSE across multicollinearity conditions. Furthermore, the range of MSE values (minimum to maximum) within each VIF category shows modest fluctuation, suggesting that while multicollinearity does impact estimation stability, its effect may not be universally severe. These findings imply that, despite the presence of multicollinearity, the parameter estimates may remain approximately unbiased.

3. 2. Multivariate *Ordinal regression Simulation*

In this simulation study, a Monte Carlo simulation is conducted involving two ordinal response variables $q = 2$. The first response variable consists of three ordinal categories, while the second has four, and the response vector for the *i*-th individual is denoted as $\boldsymbol{Y_i} = (Y_{i1}, Y_{i2})^T$, representing a bivariate ordinal outcome. For each repetition $r = 1,2,\cdots,R$, a population of N=2,000,000 observations is generated. Each observation is associated with a predictor vector $\boldsymbol{X} = (\boldsymbol{X_1}, \boldsymbol{X_2}, \boldsymbol{X_3}, \boldsymbol{X_4}) \in \mathbb{R}^4$, where the predictors are sampled from a standard normal distribution. The regression parameters for each response variable, denoted as $\boldsymbol{\beta_1} = (\beta_{10}, \beta_{11}, \beta_{12}, \beta_{13}, \beta_{14})$ and $\boldsymbol{\beta_2} = (\beta_{20}, \beta_{21}\beta_{22}, \beta_{23}, \beta_{24})$, are randomly generated for each repetition based on the specification provided in Equation 12.

$$\beta_k = (\beta_{k0}, \beta_{k1}, \cdots, \beta_{k4}) = (1, \mu_k + \varepsilon_k), \varepsilon_k \sim \mathcal{N}_4(\mathbf{0}, \sigma^2 \boldsymbol{I}) \quad (12)$$

For $k = 1,2$, with a standard deviation of $\sigma = 0.2$, the true regression coefficients are fixed at randomly selected values, defined as $\mu_1 = (0.3, -0.4, 0.8, -0.5)$ and $\mu_2 = (0.2, -0.2, 0.5, -0.7)$, respectively. Based on these coefficient vectors, the corresponding latent variables are constructed

following the specification outlined in Equation 13. Subsequently, the latent variable Y_k^* was transformed into an ordinal response variable Y_k using a quantile-based thresholding procedure.

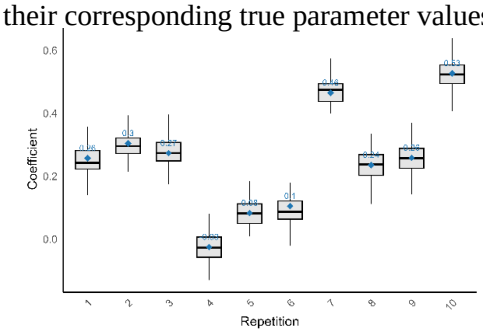
$$\begin{matrix} Y_1^* = \boldsymbol{X\beta_1} + \boldsymbol{\varepsilon_1} \\ Y_2^* = \boldsymbol{X\beta_2} + \boldsymbol{\varepsilon_2} \end{matrix} \quad \text{dengan } \begin{bmatrix}\varepsilon_1 \\ \varepsilon_2\end{bmatrix} \sim \mathcal{N}_2(\mathbf{0}, \boldsymbol{\Sigma}), \boldsymbol{\Sigma} = \begin{bmatrix}1 & \rho \\ \rho & 1\end{bmatrix}, \rho = 0.7 \qquad (13)$$

From the full dataset consisting of two million observations, a resampling procedure was conducted 100 times per iteration, with each resample consisting of a fixed sample size of $n = 1000$ (0.05%). For each resampled subset, parameter estimation was performed using a multivariate ordinal probit model specified. Although the model includes an intercept term, the analysis focuses solely on the regression coefficients $(\beta_{k1}, \beta_{k2}, \beta_{k3}, \beta_{k4})$. For each estimation, both the estimated coefficient $\hat{\beta}_{kj}$ and the corresponding true parameter value β_{kj}^{true} are recorded. The estimation error is defined as $e_{kj} = \hat{\beta}_{kj} - \beta_{kj}^{\text{true}}$. These results are then compiled into a comprehensive evaluation table to assess the bias and stability of the parameter estimates under controlled data structures and random variability. To facilitate interpretation, the estimated regression coefficients are presented separately for each response variable. This partial presentation allows a more focused evaluation of parameter behavior specific to each outcome (low multicollinearity) (see Table 2).

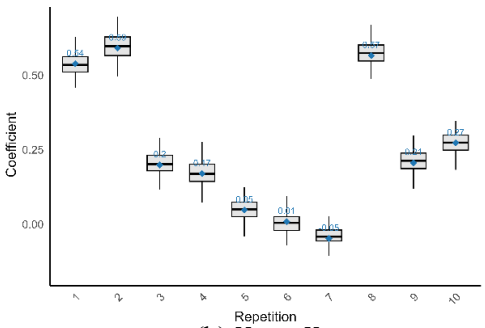
Table 2. Summary of simulation results for estimated regression coefficients in the multivariate ordinal regression model.

No	Response	Predictor	Repetition	Resampling Repetition	$\boldsymbol{\beta_{kj}^{\text{true}}}$	$\boldsymbol{\hat{\beta}_{kj}}$	$\boldsymbol{\hat{\beta}_{kj} - \hat{\beta}}$
1	Y_1	X_1	1	1	0.26	0.28	0.03
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
100	Y_1	X_1	1	100	0.26	0.27	0.01
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
400	Y_1	X_4	1	100	-0.34	-0.38	-0.04
401	Y_2	X_1	1	1	0.54	0.49	-0.05
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
8000	Y_2	X_4	10	100	-0.49	-0.42	0.07

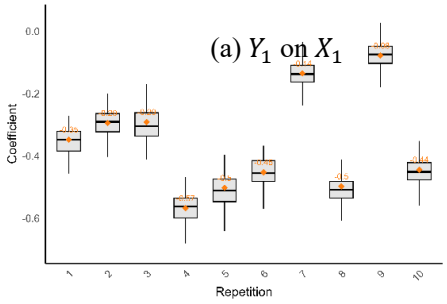
Based on Table 2, to facilitate interpretation, Figure 4 provides a visual representation of the estimated regression coefficients $\widehat{\beta}_{kj}$ across all repetitions and resampling iterations, in comparison to their corresponding true parameter values $(\beta_{kj}^{\text{true}})$.



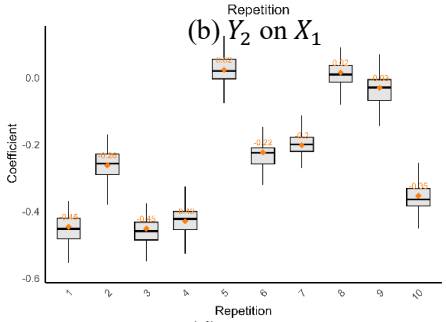
(a) Y_1 on X_1



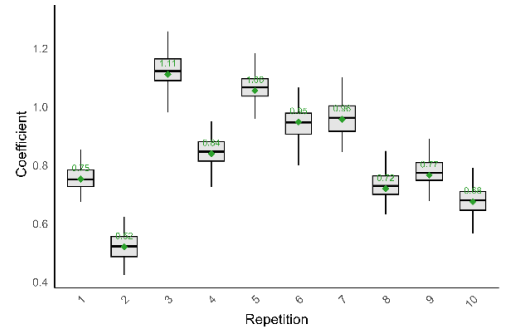
(b) Y_2 on X_1



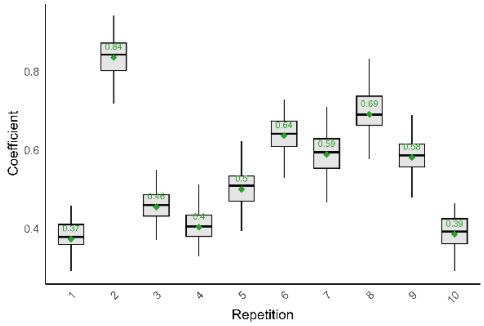
(c) Y_1 on X_2



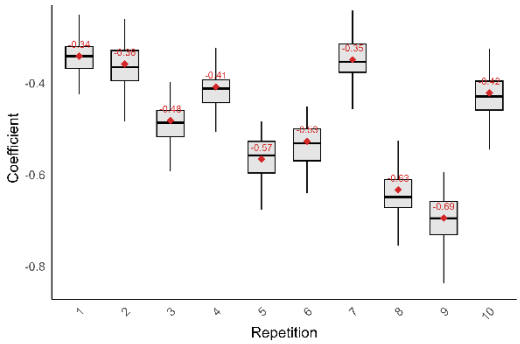
(d) Y_2 on X_2



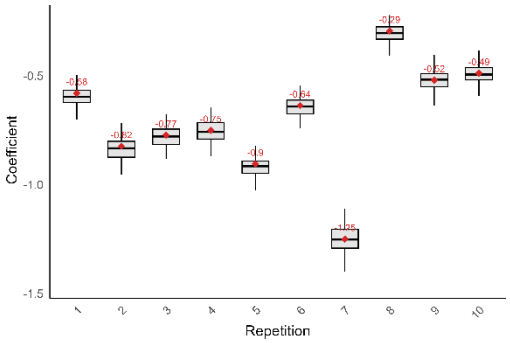
(e) Y_1 on X_3



(f) Y_2 on X_3



(g) Y_1 on X_4



(h) Y_2 on X_4

Figure 4. Deviation of the estimated regression coefficients $\widehat{\beta}_{kj}$ from the true parameter β_{kj}^{true} in low multicollinearity (Left: Effect of predictors on Y_1; Right: Effect of predictors on Y_2).

The distribution of parameter estimates in datasets with high multicollinearity does not differ substantially from those with low multicollinearity. A more in-depth evaluation is provided through the Mean Squared Error (MSE) results, which are summarized in Table 3 for each response variable across the three levels of multicollinearity. Overall, the results presented in the table indicate that multicollinearity does not exert a substantial impact on the accuracy of parameter estimates in multivariate ordinal regression. The Mean Squared Error (MSE) values across the three levels of multicollinearity remain relatively stable, particularly for predictors X_3 and X_4, which demonstrate consistent alignment with the true parameter values.

Table 3. Summary of MSE between estimated and true parameter values from each repetition.

No	VIF	Y	X	Mean	Standar Deviation	Minimum	Maximum
1.	$VIF < 5$	Y_1	X_1	0.0019	0.0002	0.0015	0.0022
2.		Y_1	X_2	0.0022	0.0004	0.0017	0.0032
3.		Y_1	X_3	0.0028	0.0006	0.0019	0.0038
4.		Y_1	X_4	0.0022	0.0004	0.0016	0.0031
5.		Y_2	X_1	0.0017	0.0003	0.0013	0.0021
6.		Y_2	X_2	0.0018	0.0002	0.0016	0.0021
7.		Y_2	X_3	0.0021	0.0005	0.0015	0.0029
8.		Y_2	X_4	0.0024	0.0006	0.0017	0.0038
9.	$5 \leq VIF < 10$	Y_1	X_1	0.0160	0.0025	0.0118	0.0200
10.		Y_1	X_2	0.0214	0.0041	0.0165	0.0317
11.		Y_1	X_3	0.0030	0.0008	0.0016	0.0045

No	VIF	Y	X	Mean	Standar Deviation	Minimum	Maximum
12.		Y_1	X_4	0.0208	0.0026	0.0178	0.0272
13.		Y_2	X_1	0.0141	0.0023	0.0110	0.0177
14.		Y_2	X_2	0.0168	0.0019	0.0130	0.0193
15.		Y_2	X_3	0.0021	0.0005	0.0015	0.0029
16.		Y_2	X_4	0.0195	0.0024	0.0160	0.0231
17.	$VIF \geq 10$	Y_1	X_1	0.0369	0.0067	0.0278	0.0472
18.		Y_1	X_2	0.0219	0.0046	0.0157	0.0321
19.		Y_1	X_3	0.0032	0.0009	0.0018	0.0048
20.		Y_1	X_4	0.0220	0.0034	0.0168	0.0274
21.		Y_2	X_1	0.0321	0.0042	0.0253	0.0380
22.		Y_2	X_2	0.0179	0.0024	0.0135	0.0208
23.		Y_2	X_3	0.0021	0.0004	0.0016	0.0029
24.		Y_2	X_4	0.0204	0.0024	0.0174	0.0232

Although the MSE values tend to be slightly higher under moderate and high multicollinearity compared to the low-multicollinearity setting, the spread of parameter estimates under severe multicollinearity still centers around the true values. This is evidenced by the narrow range between minimum and maximum values and the moderate standard deviations observed across repetitions. These findings reinforce the notion that, within the context of this simulation, multicollinearity does not inherently introduce systematic bias into parameter estimation, suggesting that the estimators remain essentially unbiased. Nonetheless, while the accuracy of the estimates is largely unaffected, high multicollinearity does lead to increased computational complexity, as reflected in longer estimation times and higher resource demands. These results are consistent with prior findings that multicollinearity can degrade algorithmic performance and complicate parameter estimation, especially in high-dimensional contexts [29–31]. In scenarios with VIF ≥ 10, estimation procedures required noticeably more processing time compared to low-multicollinearity conditions. This increase is likely due to strong intercorrelations among predictors, which can slow convergence in iterative optimization algorithms, particularly in multivariate models with correlated response structures. Therefore, even though the quality of the parameter estimates remains robust, the presence of multicollinearity should still be considered a critical factor from the perspective of computational efficiency and resource allocation.

Conclusion

As outlined in the introduction and further elaborated in the results and discussion sections, this study investigated parameter estimation in both univariate and multivariate ordinal regression models under three levels of multicollinearity among predictor variables: low, moderate, and high. While an increase in MSE was observed under high multicollinearity, the deviation between true and estimated values remained relatively small, indicating that multicollinearity does not substantially distort parameter estimation. Both Maximum Likelihood Estimation (MLE) for univariate ordinal regression and Pairwise Likelihood (PL) for multivariate ordinal regression yielded parameter estimates that were essentially unbiased, although computation time increased considerably in high-correlation settings. The broader implication of these findings is that MLE and PL remain robust estimation strategies even when predictor variables exhibit strong correlations, a condition frequently encountered in applied contexts such as large-scale surveys, psychometric assessments, and socio-economic studies. Nonetheless, the generalizability of these results is limited by the simulation design, which assumed normally distributed latent traits and balanced data structures. Real-world data often involve heteroskedasticity, non-normal latent distributions, or incomplete responses, which may affect estimation performance. Future research should therefore extend this work by incorporating such complexities, particularly examining estimator

behavior under heteroskedastic errors, skewed or heavy-tailed latent distributions, and missing data mechanisms. Exploring these scenarios will enhance methodological guidance for practitioners analyzing complex ordinal data and ensure that estimation techniques remain robust across diverse empirical applications.

Acknowledgements

The author wishes to express his sincere gratitude. First, to the Statistics Study Program, Faculty of Mathematics and Natural Sciences, Universitas Islam Indonesia, for their academic support. Second, to the Program on Statistics and Data Science, School of Data Science, Mathematics, and Informatics, IPB University, for their valuable guidance and constructive feedback. Third, to the *Beasiswa Pendidikan Indonesia* (The Indonesian Education Scholarship); the *Pusat Pelayanan Pembiayaan dan Asesmen Pendidikan Tinggi* (Center for Higher Education Funding and Assessment), Ministry of Higher Education, Science, and Technology of the Republic of Indonesia; and the *Lembaga Pengelola Dana Pendidikan* (Endowment Fund for Education Agency), Ministry of Finance of the Republic of Indonesia, for their generous support.

References

[1] Ganesh S. Multivariate Linear Regression. International Encyclopedia of Education, Elsevier; 2010, p. 324–31. https://doi.org/10.1016/B978-0-08-044894-7.01350-6.

[2] Bonnini S, Borghesi M. Relationship between Mental Health and Socio-Economic, Demographic and Environmental Factors in the COVID-19 Lockdown Period—A Multivariate Regression Analysis. Mathematics 2022;10:3237. https://doi.org/10.3390/math10183237.

[3] Kunkler M. Multilateral exchange rates: A multivariate regression framework. J Econ Bus 2023;125–126:106132. https://doi.org/10.1016/j.jeconbus.2023.106132.

[4] Cui J, Yi GY. Variable selection in multivariate regression models with measurement error in covariates. J Multivar Anal 2024;202:105299. https://doi.org/10.1016/j.jmva.2024.105299.

[5] Hernáez Á, Rogne T, Skåra KH, Håberg SE, Page CM, Fraser A, et al. Body mass index and subfertility: multivariable regression and Mendelian randomization analyses in the Norwegian Mother, Father and Child Cohort Study. Human Reproduction 2021;36:3141–51. https://doi.org/10.1093/humrep/deab224.

[6] Liang J, Bi G, Zhan C. Multinomial and ordinal Logistic regression analyses with multi-categorical variables using R. Ann Transl Med 2020;8:982–982. https://doi.org/10.21037/atm-2020-57.

[7] Pagui ECK, Canale A, Genz A, Azzalini A. ***PLordprob***: Title Multivariate Ordered Probit Model via Pairwise Likelihood 2025.

[8] Kenne Pagui EC, Canale A. Pairwise likelihood inference for multivariate ordinal responses with applications to customer satisfaction. Appl Stoch Models Bus Ind 2016;32:273–82. https://doi.org/10.1002/asmb.2147.

[9] Hirk R, Hornik K, Vana L. mvord: An R package for fitting multivariate ordinal regression models. J Stat Softw 2020;93. https://doi.org/10.18637/jss.v093.i04.

[10] Lindsay BG. Composite likelihood methods. Comtemporary Mathematics, vol. 80, American Mathematical Society; 1988, p. 221–39. https://doi.org/10.1090/conm/080/999014.

[11] Varin C, Reid N, Firth D. AnOverview of Composite Likelihood Methods. Stat Sin 2011;21:5–42.

[12] Wieditz J, Miller C, Scholand J, Nemeth M. A Brief Introduction on Latent Variable Based Ordinal Regression Models With an Application to Survey Data. Stat Med 2024;43:5618–34. https://doi.org/10.1002/sim.10208.

[13] Gambarota F, Altoè G. Ordinal regression models made easy: A tutorial on parameter interpretation, data simulation and power analysis. International Journal of Psychology 2024;59:1263–92. https://doi.org/10.1002/ijop.13243.

[14] Mualifah LNA, Soleh AM, Notodiputro KA. Comparison of GARCH, LSTM, and Hybrid GARCH-LSTM Models for Analyzing Data Volatility. International Journal of Advances in Soft Computing and Its Applications 2024;16:150–65. https://doi.org/10.15849/IJASCA.240730.10.
[15] R Core Team. R: A Language and Environment for Statistical Computing 2024.
[16] Labambe M. Predicting Waste Production Trends in Palu City Using Linear Regression Analysis. Advance Sustainable Science Engineering and Technology 2024;6:0240306. https://doi.org/10.26877/asset.v6i3.523.
[17] Rusyana A, Kurnia A, Sadik K, Wigena AH, Sumertajaya IM, Sartono B. Comparison of GLM, GLMM and HGLM in Identifying Factors that Influence the District or City Poverty Level in Aceh Province. J Phys Conf Ser 2021;1863:012023. https://doi.org/10.1088/1742-6596/1863/1/012023.
[18] Agresti A. Foundations of Linear and Generalized Linear Models. Wiley & Sons; 2015.
[19] Skrondal A, Rabe-Hesketh S. Generalized Latent Variable Modeling : Multilevel, Longitudinal, and Structural Equation Models. CRC Press; 2004.
[20] Greene WH. Econometric analysis. Prentice Hall; 2003.
[21] Tutz Gerhard. Regression for categorical data. Cambridge University Press; 2012.
[22] Liddell TM, Kruschke JK. Analyzing ordinal data with metric models: What could possibly go wrong? J Exp Soc Psychol 2018;79:328–48. https://doi.org/10.1016/j.jesp.2018.08.009.
[23] Croux C, Haesbroeck G, Ruwet C. Robust estimation for ordinal regression. J Stat Plan Inference 2013;143:1486–99. https://doi.org/10.1016/j.jspi.2013.04.008.
[24] Bravo M, Canale A. Pairwise likelihood inference for the multivariate ordered probit model 2019.
[25] Molenberghs G, Verbeke G. Models for Discrete Longitudinal Data. New York: Springer-Verlag; 2005. https://doi.org/10.1007/0-387-28980-1.
[26] Robitzsch A. Pairwise Likelihood Estimation of the 2PL Model with Locally Dependent Item Responses. Applied Sciences 2024;14:2652. https://doi.org/10.3390/app14062652.
[27] Mazo G, Karlis D, Rau A. A Randomized Pairwise Likelihood Method for Complex Statistical Inferences. J Am Stat Assoc 2024;119:2317–27. https://doi.org/10.1080/01621459.2023.2257367.
[28] Tong J, Luo C, Islam MN, Sheils NE, Buresh J, Edmondson M, et al. Distributed learning for heterogeneous clinical data with application to integrating COVID-19 data across 230 sites. NPJ Digit Med 2022;5:76. https://doi.org/10.1038/s41746-022-00615-8.
[29] Chan JY-L, Leow SMH, Bea KT, Cheng WK, Phoong SW, Hong Z-W, et al. Mitigating the Multicollinearity Problem and Its Machine Learning Approach: A Review. Mathematics 2022;10:1283. https://doi.org/10.3390/math10081283.
[30] Dertli HI, Hayes DB, Zorn TG. Effects of multicollinearity and data granularity on regression models of stream temperature. J Hydrol (Amst) 2024;639:131572. https://doi.org/10.1016/j.jhydrol.2024.131572.
[31] Sundus KI, Hammo BH, Al-Zoubi MB, Al-Omari A. Solving the multicollinearity problem to improve the stability of machine learning algorithms applied to a fully annotated breast cancer dataset. Inform Med Unlocked 2022;33:101088. https://doi.org/10.1016/j.imu.2022.101088.