Linear Correction Model for Statistical Inference Analysis

Jing Zhao[1], Zhijiang Zhang[2]*
Kaifeng Vocational College of Culture and Arts, Kaifeng 475000, China[1]
Henan University, Kaifeng 475000, China[2]

***Abstract*—A linear correction model based on joint independent information is proposed to optimize the statistical inference performance in high-dimensional data and small sample scenarios by integrating Fiducial inference and Bayesian posterior prediction methods. The model utilizes multi-source data features to construct a joint independent information framework, combined with an information domain dynamic correction mechanism, significantly improving parameter estimation efficiency and confidence interval coverage. Numerical simulation shows that when the sample size is 30, the posterior prediction method has a coverage rate of 0.927, approaching 95% of the theoretical value, and the coverage probability approaches the ideal level with increasing sample size. Compared with traditional methods, the model exhibits stronger adaptability and stability in high-dimensional noise covariance and dynamic data streams, providing an efficient and robust theoretical tool for statistical inference in complex data environments.**

Keywords—Linear correction model; statistical analysis; fiducial inference; numerical simulation

I. Introduction

Accurate analysis and interpretation of data are crucial in numerous scientific research fields and practical production and life scenarios. Statistical inference, as one of the core components of data analysis, aims to infer the characteristics and patterns of the population through the study of sample data, thereby providing reliable basis for decision-making [1, 2]. However, the actual collected data is often affected by various factors, resulting in a certain degree of error and bias, which poses a challenge to the accuracy of statistical inference. The linear correction model, as an important data processing tool, plays a crucial role in solving data errors and improving the reliability of statistical inference. It is based on the assumption of linear relationships, and by appropriately transforming and adjusting the data, it can effectively correct systematic biases in the data, making statistical inference results more realistic [3]. From a theoretical development perspective, the research on linear correction models has undergone multiple stages of evolution. Early linear correction models were relatively simple, but with the continuous advancement of mathematical theory and computational techniques, the complexity and adaptability of the models have been significantly improved, enabling them to handle more complex data structures and error patterns [4, 5]. Nowadays, linear correction models have been widely applied in many fields such as medicine, economics, environmental science, engineering technology, etc. In the medical field, linear correction of Magnetic Resonance Imaging signals can eliminate device measurement errors and improve the accuracy of clinical diagnostic data. In the field of economics, market forecasting algorithms correct model biases and enhance the reliability of macroeconomic trend analysis. In the field of environmental science, sensor data from air quality monitoring networks is dynamically linearly corrected to reduce the interference of systematic errors on pollution trend analysis. In the field of engineering technology, linear models significantly reduce the computational burden caused by high-frequency updates in real-time satellite clock calibration. In the field of artificial intelligence, high-dimensional sensor data from Internet of Things devices is suppressed by dynamic calibration models to ensure real-time data reliability. With the explosive growth of data volume and the increasing complexity of data types, the stability and efficiency of linear correction models in high-dimensional data environments have decreased. Therefore, to construct a more efficient and stable linear correction model and improve the statistical inference ability of the model in small sample situations, the theoretical system of the linear correction model is systematically reviewed, and a linear correction model based on joint independent information is constructed to statistically infer the independent variables in the data prediction model.

When dealing with high-dimensional data and small sample scenarios, traditional methods assume a single data source or fixed error covariance, making it difficult to effectively integrate heterogeneous data from multiple sources, resulting in low parameter estimation efficiency and significant confidence interval coverage bias. In response to the above issues, this study aims to construct a joint independent information-driven linear correction model, which integrates Fiducial inference and Bayesian posterior prediction methods to achieve the following goals: improve statistical inference efficiency under high-dimensional noise covariance, and solve the problem of insufficient stability of traditional methods in complex data structures. The research results can provide efficient and robust statistical inference tools for fields such as medical image correction, environmental monitoring, and industrial Internet of Things, promoting the theoretical deepening and application expansion of linear correction models in data science.

The study innovatively proposes a linear correction model based on jointly independent information, with its primary advantage lying in the integration of Fiducial inference and Bayesian methods to resolve the issues of low efficiency and significant coverage bias in confidence interval estimation for traditional models under high-dimensional data and small-sample scenarios. The core contributions include constructing a joint independent information framework to integrate multi-source data features, significantly enhancing parameter estimation efficiency. Technical Integration: Inverse parameter distribution analysis is combined with Bayesian posterior predictive correction to optimize confidence interval coverage through Fiducial inference. Adaptability Enhancement: A dynamic information domain correction mechanism is introduced to improve the model's adaptability to complex data structures, thus providing an efficient and stable statistical inference tool for high-dimensional environments.

The main structure of the study is divided into five sections: Section II is a review of the current research status of linear correction models. Section III is the estimation method for the parameter interval of the linear correction model. Section IV is the numerical simulation analysis of model parameter interval estimation. Section V details the discussion. Section VI is a summary of the research content.

II. Related Works

The linear correction model is mainly used to eliminate systematic errors or biases, thereby improving the accuracy and reliability of data. Zhang R et al. proposed a new mathematical model for correcting the contact and separation conditions to address the inaccuracy of classical piecewise linear models in describing the dynamic behavior of mechanical systems with gaps. The results showed that the new model explained the premature separation and contact hysteresis phenomena of the primary system, and its contact point, separation point, and amplitude frequency response were significantly different from those of the classical model [6]. Sun T et al. proposed a method for calculating asymptotic bias and developing bias correction to address the issue of measurement error impact on matrix data in generalized linear models. The results showed that this method effectively addressed the impact of measurement errors, and its statistical properties have been validated through synthesis and analysis of real datasets [7]. Li H et al. proposed an improved real-time service method using extrapolation algorithms and linear models to address the computational burden and timeliness challenges caused by frequent updates in real-time satellite clock calibration. The results showed that a satellite clock correction sequence with a one-hour arc length was most suitable for fitting the Lauch-Dong-Streebel linear model [8]. ElHorbaty Y S et al. proposed a permutation test method using analysis of variance to test for zero variance components in generalized linear models, which only requires fitting the zero model. The results showed that, through Monte Carlo simulation verification, the new test had a correct Class I error rate and was superior to existing bootstrap score tests [9]. Maksaei N et al. proposed a local influence method based on correction score function and Ridge estimation to evaluate the impact of small data disturbances in linear mixed measurement error models. The results showed that simulation studies and real data applications demonstrated that this method could effectively identify influential observations and demonstrate good diagnostic performance [10].

Wang P et al. proposed a magnetic structure coupling correction model based on classical axial vibration model and image method to investigate the influence of magnetic structure coupling on winding short circuit during transformer axial vibration process. The results showed that compared with the classical model, the vibration amplitude increment of the magnetic structure coupling correction model was smaller [11]. Gibiansky et al. proposed a bivalent binding model considering a 2:1 stoichiometric ratio to address the issue of neglecting double binding sites in monoclonal antibody pharmacokinetic models, and studied its effects through simulation. The results showed that the unit price model could not accurately describe the data of the divalent model, and a model with correct stoichiometric assumptions need to be used [12]. Emami H et al. proposed diagnostic measures based on case deletion, mean shift outlier model, and corrected likelihood to address the issue of identifying influential observations in some linear models. The results showed that both manual examples and real data examples validated the performance of these methods, demonstrating their effectiveness in identifying potential outliers [13]. Chang H et al. proposed an accurate closed form bias correction method to address the bias issue of linear regression estimators in randomized controlled trials pointed out by Freedman. The results showed that the estimator after bias correction had the same limit distribution as the uncorrected estimator [14]. Li L et al. proposed a direct standardization algorithm for transfer component analysis that combines nonlinear and linear correction to address the issue of insufficient prediction accuracy caused by consistency between near-infrared spectrometers. The experimental results showed that the direct standardization algorithm of transfer component analysis was superior to traditional methods on public datasets, significantly improving the model transmission performance [15].

The linear correction model, as an important statistical tool, has demonstrated strong application potential in multiple fields. In recent years, with the development of methods such as dynamic correction, high-dimensional data correction, and robust correction, significant progress has been made in the research of linear correction models. However, linear correction models still have low processing efficiency and noise covariance issues in high-dimensional data processing. Therefore, the study proposes to construct a univariate linear correction model and adjust its confidence interval to improve statistical inference performance.

III. Methods and Materials

A. Linear Correction Model Based on Fiducial Inference

The core of a linear correction model is to establish a linear relationship between input variables and output variables. When using a single variable, its general expression is shown in Eq. (1):

$$y_i = x_i b + \varepsilon \tag{1}$$

In Eq. (1), y_i represents the observation value to be corrected; β represents the input variable; ε represents the vector of correction coefficients to be estimated; ε stands for random error term. The random error term of offline calibration models is usually assumed to follow a normal distribution with a mean of 0. According to different application scenarios, the univariate linear correction model can be adjusted to multivariate joint correction, dynamic linear correction, and linear correction with errors [16]. If the calibration data for the target to be calibrated comes from different experimental data, it is necessary to consider the impact of joint independent information on the linear calibration model. The research assumes that there is a set of data from different experiments with different variances, which satisfies the calibration model as shown in Eq. (2).

$$y_{ij} = a_i + b_i x_j + \varepsilon_{ij}, i = 1,2,\cdots,k; j = 1,1,2,\cdots,n \tag{2}$$

In Eq. (2), k represents the number of data sources; n represents the number of independent response variables; x_j represents the actual value of the j th unit; y_{ij} represents the measurement values related to x_j obtained from the i th experimental plan; a_i and b_i represent unknown parameters. Corresponding to the uncorrected observations from different sources to an explanatory variable, the model also conforms to the predictive model shown in Eq. (3) [17].

$$y_{0i} = a_i + b_i \theta + \varepsilon_{i0}, i = 1,2,\cdots,k \tag{3}$$

In Eq. (3), θ represents the explanatory variable of the unknown value. The random term errors in both the calibration model and the prediction model follow a normal distribution with a mean of 0, and the random error terms of the two models are independent of each other. The univariate linear correction model based on joint independent information can estimate θ based on the response variables in Eq. (2) and Eq. (3). The general steps of the linear correction model are shown in Fig. 1 [18, 19].

In the linear correction model based on joint independent information in Fig. 1, the estimation efficiency of the correction coefficient to be estimated is relatively low. Fiducial inference is a statistical inference method that attempts to provide a different approach to dealing with uncertainty than the frequency school and Bayesian school. It can effectively improve the estimation efficiency and prediction accuracy of linear correction models for estimating parameters. Therefore, the study proposes using Fiducial inference to improve the joint independent information linear correction model. The core of Fiducial inference lies in determining the distribution of parameters through reverse data analysis, rather than directly estimating parameters based on sample data. The application steps of Fiducial inference are shown in Fig. 2 [20, 21].

As shown in Fig. 2, when applying Fiducial inference, it is necessary to first set up the model, and a linear correction model can be directly used here. After determining the model, parameter inversion is performed and a Fiducial distribution is constructed. Subsequently, the Fiducial distribution is used to calculate the confidence interval of the estimated parameter, adjust the predicted value, and perform prediction correction [22]. Fiducial inference assumes that there is a known random variable distributed in the random error space, and the function of parameters from $\Omega \times \varepsilon$ to χ satisfies the relationship shown in Eq. (4).

$$X^d = h(\eta, E) \tag{4}$$

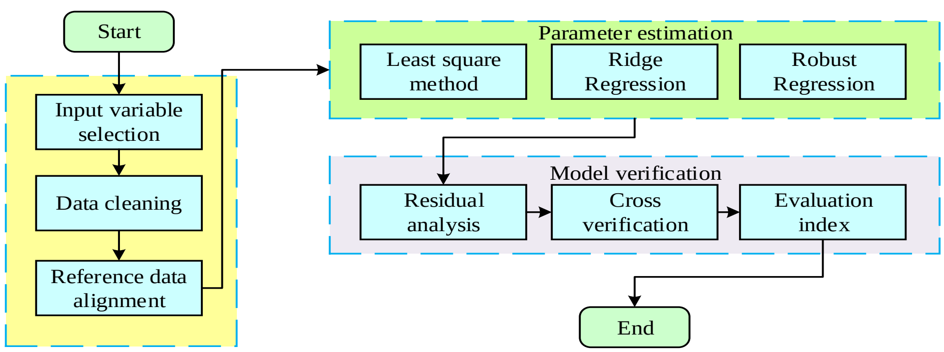

Fig. 1. Building the steps of the linear correction model.

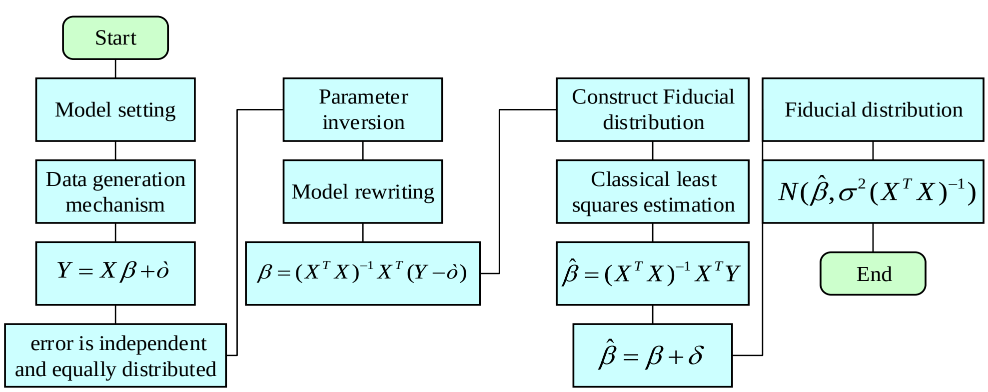

Fig. 2. The Application steps of Fiducial inference.

In Eq. (4), η represents the parameter; E represents known random variables. If all parameters are included in Ω and any $x \in \chi$, $e \in \varepsilon$, $x = h(\eta, e)$ has a unique solution $\eta_x(e)$ in the solution space, the distribution of $\eta_x(E)$ can be called η distribution, and the distribution of $\theta(\eta_x(E))$ can be called $\theta = \theta(\eta)$ distribution. According to the above derivation, the Fiducial distribution of the parameters of the linear correction model is shown in Eq. (5):

$$F_x(\theta) = \Pr\left(\theta\left(\eta_x(E)\right) \le \theta\right) \tag{5}$$

In Eq. (5), $F_x(\theta)$ is in the interval [0, 1], and $F_x(\theta)$ is non-decreasing with respect to θ. At this point, the Fiducial interval of the parameter can be calculated. For any number α in the [0, 1] interval, the 1-α Fiducial lower boundary of θ is shown in Eq. (6):

$$\theta_\alpha x = \sup_{\theta \in \Theta}\left\{\theta : F_x(\theta) < \alpha\right\} \tag{6}$$

In a linear correction model, a function $R(X; x, \xi)$ is defined. X represents a random variable, ξ represents a parameter that is related to the interest and dislike parameters, and x represents the observed value of the random variable x. The distribution of the random variable X is related to the parameter ξ. Function $R(X; x, \xi)$ satisfies two basic conditions. Firstly, the distribution of the function is independent of the parameter $\xi = (\theta, \eta)$. Secondly, the observed values of this function have a low correlation with the aversion parameter. The function $R(X; x, \xi)$ is the generalized pivot quantity and the subset of the sample space of the function satisfies Eq. (7):

$$P\left(R(X; x, \xi) \in C_r\right) = 1 - \alpha, 0 < \alpha < 1 \tag{7}$$

In Eq. (7), C_r represents a subset of the sample space of the function. According to Eq. (7), the subset of parameter space satisfies Eq. (8):

$$\Theta_c(r) = \left\{\theta \in \Theta \middle| R(x; x, \xi) \in C_r\right\} \tag{8}$$

Eq. (8) is the generalized confidence interval of the interest parameter. The confidence interval will have an impact on the linear correction results, and the linear correction problem can be transformed into a G-H problem. In the G-H problem, the confidence coefficient of the bounded confidence interval is 0, and the correction of the confidence interval is based on the models (y_i, x_i) and y_0. Firstly, it is necessary to calculate and $\bar{y}$, as shown in Eq. (9):

$$\begin{cases} \bar{x} = \dfrac{1}{n}\sum_{i=1}^{n} x_i \\ \bar{y} = \dfrac{1}{n}\sum_{i=1}^{n} y_i \end{cases} \tag{9}$$

The least squares estimation of b_0 and b_1 are represented by $\hat{b}_0$ and $\hat{b}_1$, and the calculation of $\hat{b}_0$ and $\hat{b}_1$ is shown in Eq. (10):

$$\begin{cases} \hat{b}_0 = \bar{y} - \hat{b}_1\bar{x} \\ \hat{b}_1 = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \end{cases} \tag{10}$$

Research makes the hypothesis shown in Eq. (11):

$$|\theta - \bar{x}| \le 3\sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} \tag{11}$$

According to Eq. (11), it is assumed that as n approaches infinity, $\dfrac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$ also tends to a constant. At this point, the information domain is introduced, as shown in Eq. (12):

$$\Omega_{1B_n} = \left\{\theta : |\theta - x| \le B_n\right\} \tag{12}$$

In Eq. (12), $B_n = o\left(\sqrt{n}\right)$.

B. Confidence Interval Correction and Parameter Interval Estimation of Linear Correction Model

After establishing the linear correction model, it is necessary to revise its confidence interval and estimate the parameter interval. When performing confidence interval correction on a linear calibration model, it is necessary to first provide the generalized confidence interval of the linear calibration model. For the data in the linear correction model, it is necessary to first calculate statistics such as $\bar{y}_i$, $\bar{x}$, S_{xx}, and S_i^2, and use the least squares method to estimate $\hat{a}_i$ and $\hat{b}_i$. Based on the properties of the model, Eq. (13) can be obtained [23]:

$$\begin{cases} y_{io} - \bar{y}_i : N\left(b_i(\theta - \bar{\bar{x}}), (1 + \frac{1}{n})\sigma_i^2\right) \\ \hat{b} : N_k\left(b, \frac{\Sigma}{S_{xx}}\right), \frac{S_i^2}{\sigma_i^2} : x^2(n-2) \end{cases} \tag{13}$$

By using variable transformation, and $b_1 = b\sqrt{S_{xx}}$, $\theta_1 = \dfrac{(\theta - \bar{x})}{\sqrt{\left(1 + \frac{1}{n}\right)S_{xx}}}$ can obtain Eq. (14),

$$\begin{cases} \hat{b}\sqrt{S_{xx}}^{\,d} = \gamma + BE_1 \\ \dfrac{Y_0-\bar{Y}}{\sqrt{1+\dfrac{1}{n}}}^{\,d} = \gamma\theta_1 + BE_2 \\ L = BE_3 \end{cases} \tag{14}$$

In Eq. (14), γ represents the disliked parameter in generalized *p*-value calculation, and the elements in $E_1, E_2 \colon N_k(0, I_k)$ and E_3 satisfy the conditions shown in Eq. (15):

$$e_{ii}^2 : x^2(n-2) \tag{15}$$

All elements in E_3 are independent of each other, and based on this, a generalized pivot quantity $R(Y; y, \theta_1, b_1, \Sigma)$ can be constructed. The construction steps of the generalized pivot quantity are shown in Fig. 3 [24].

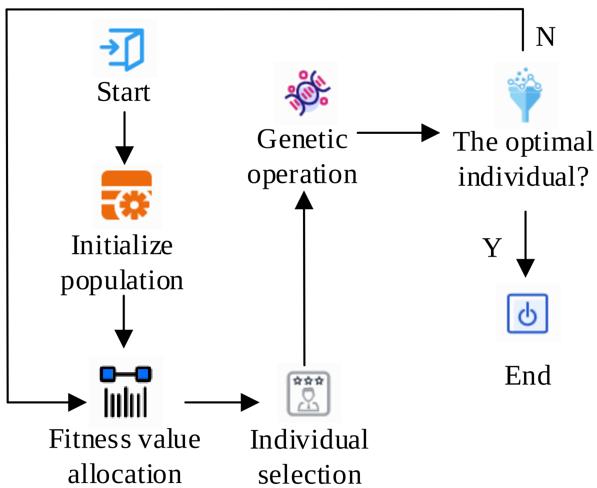

Fig. 3. Construction steps of generalized pivot quantity.

After constructing the generalized pivot quantity, the generalized confidence interval of the model can be determined. It assumes that $\hat{R}_L$ and $\hat{R}_U$ are the $100\frac{a}{2}\%$ quantile and $100\left(1-\frac{a}{2}\right)\%$ quantile of the generalized pivot distribution, respectively, then Eq. (16) can be obtained:

$$P(\hat{R}_L < R(Y; y, \theta_1, b_1, \textstyle\sum) < \hat{R}_U) = 1-a \tag{16}$$

According to Eq. (16), $G(Y) = \left(\hat{R}_L, \hat{R}_U\right)$ is the $100(1-a)$ generalized confidence interval of the interest parameter θ. Monte Carlo simulation is a computational method based on probability and statistics theory, also known as statistical simulation method. Its core is to use a large number of random numbers for simulation experiments, and the generalized confidence interval of the interest parameter θ can be obtained using this method. $\bar{x}_1$, $\bar{y}_1$, and I_s are calculated for a given set of data $(x_1, x_2, \cdots, x_n)$ and $y = (y_1, y_2, \cdots, y_n, y_0)$. For $i = 1, 2, \cdots, k$, Eq. (17) can be generated:

$$E_1 \sim N_k(0, I_k), E_{2j} \sim N_k(0, I_k), e_{jj}^2 \sim \chi^2(n-2) \tag{17}$$

By calculating the generalized pivot quantity, the generalized confidence interval of θ_1 can be obtained. The modified generalized confidence interval is shown in Eq. (18):

$$CG(Y) = \left(\max\left\{\hat{R}_L, \bar{x}-B_n\right\} \min\left\{\hat{R}_U, \bar{x}-B_n\right\}\right) \tag{18}$$

In Eq. (18), B_n represents information domain related parameters that satisfy $|\theta-\bar{x}| \le B_n$'s adjustable correction interval for values. In summary, when obtaining the generalized confidence interval of a linear correction model and correcting it, the relevant statistics and parameter estimates are first calculated based on the model data, and the generalized pivot quantity is constructed through variable transformation. Due to the difficulty in obtaining its distribution, Monte Carlo simulation is used to determine the generalized confidence interval by generating random numbers for a specific distribution multiple times. To optimize the performance of the confidence interval, the information domain is introduced for correction. The modified generalized confidence interval is obtained by taking the maximum value of the range related to the original generalized confidence interval and the information domain. After obtaining and correcting the generalized confidence interval, it is necessary to modify the post confidence interval of the linear correction model. The determination of prior distribution requires the use of Fisher information matrix to obtain the Jeffreys priors of a, b, and $\overset{-1}{\Sigma}$ in the univariate linear correction model with joint independent information, as shown in Eq. (19):

$$\pi(\alpha, \beta, \overset{-1}{\textstyle\sum}) \propto |\overset{-1}{\textstyle\sum}|^{\frac{1}{2}} \tag{19}$$

Based on the linear correction model, an expression for the prior distribution is derived, which in turn leads to the posterior distribution of a given b and σ_i, and the posterior distribution of b 's and σ_i^{-1} given σ_i^{-1}. On this basis, the posterior predictive distribution of Y_i^{rep} $(i = 1, 2, \cdots, n+1)$ is obtained. When b and Σ are known, the least squares estimation of the interest parameter θ can be obtained. When θ and Σ are unknown, the corresponding estimation values are introduced to $\hat{\theta}$, and a generalized pivot quantity is constructed. Due to the difficulty in obtaining the distribution of pivot quantities, numerical Monte Carlo simulations are used to obtain the confidence interval for $\hat{\theta}$. Its confidence interval is adjusted based on the information domain. When constructing a posterior prediction distribution, it is necessary to use the Fisher information matrix to determine the prior of the relevant parameters. Parameter posterior prediction distribution is derived based on prior knowledge. Then, based on b and whether it is known or not, an estimated value of interest parameter b is obtained and a pivot quantity is constructed. The confidence interval of interest parameter θ is obtained through Monte Carlo simulation. Finally, to optimize the confidence interval, the information domain is used for correction, resulting in a more reasonable posterior prediction interval and improving the accuracy and reliability of estimating the interest parameter θ. The study uses numerical simulation methods to validate the interval estimation method of the linear correction model, as shown in Fig. 4.

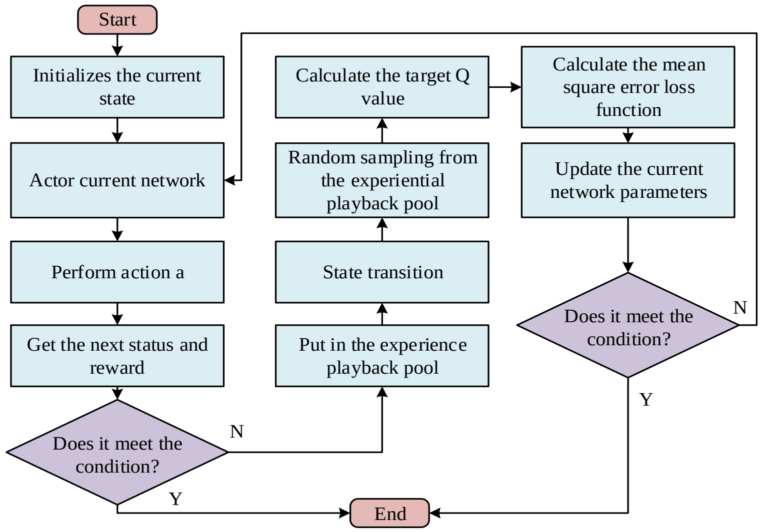

Fig. 4. Numerical simulation validation steps for interval estimation.

In Fig. 4, the numerical simulation operation steps can be divided into three parts. Firstly, the simulation setting is carried out using Monte Carlo simulation method, with 2500 × 5000 cycles. The explanatory variables are set to have a mean of 0.5 and a variance of 1. From this distribution, specific values are selected for different sample sizes and other model parameters are determined. Next is to determine the value of θ_1. Finally, there is simulation calculation, which calculates the coverage probability and interval length of each confidence interval to evaluate its goodness.

IV. Results

When verifying the feasibility of the interval estimation method for the linear correction model designed in the study, the Monte Carlo simulation method was used. This method is a numerical calculation method based on probability and statistics theory. Its core idea is to simulate various uncertain factors using random numbers through a large number of random experiments, and then solve problems in fields such as mathematics, physics, and engineering. This method requires determining a probability model or stochastic process related to the problem, so that the problem to be solved can be represented by certain statistical features of this model. Next, computer-generated random numbers conforming to a particular distribution were used to simulate multiple repeated trials of that probabilistic model or stochastic process. Finally, an approximate solution to the problem was obtained by statistically analyzing the results of these simulations. The values of the explanatory variables for the numerical simulation setup of the study are shown in Table I.

In Table I, the study set up three different sets of data, with the first set consisting of 10 data points, the second set consisting of 20 data points, and the third set consisting of 30 data points. The highest x value for the first set of data was 2.3, and the lowest was -0.6. The second set of data had a maximum x value of 1.5 and a minimum x value of -2.3. The highest x value for the third set of data was 2.6, and the lowest was -1.4. In the numerical simulation process, let k be 3, intercept a be (1, 1, 1), and Σ be diagonal matrices with diagonals of 1, 2, and 3, respectively. The numerical simulation results of the first set of data are shown in Table II.

TABLE I. THE VALUES OF THE EXPLANATORY VARIABLES FOR THE NUMERICAL SIMULATION

n	x									
10	-0.4	0.1	0.6	1.4	-0.8	0.4	2.3	-0.6	-0.4	1.7
20	0.3	0.4	1.6	0.6	0.4	-0.6	-1.3	-2.1	-2.1	-1.6
2.6	1.3	1.2	-2.0	1.5	1.7	0.5	-1.6	-2.2	0.4	2.3
	-0.4	0.3	0.5	1.6	-0.6	-1.2	1.3	2.4	2.5	-1.3
30	-1.2	1.3	2.0	0.6	-1.3	-0.4	2.1	1.5	1.7	1.8
	-0.6	-1.3	-0.9	1.6	1.3	-0.7	1.6	-1.8	0.5	0.6

TABLE II. RESULTS OF THE NUMERICAL SIMULATIONS OF THE FIRST SET OF DATA

n	θ	b_1 (1,2,3)	b_1 (2,1,5)	b_1 (3,4,10)	Method
10	1	0.892	0.924	0.946	Generalized confidence intervals and corrections
		0.931	0.944	0.942	Confidence interval and correction of posterior prediction method
	0.5	0.952	0.945	0.932	Generalized confidence intervals and corrections
		0.944	0.942	0.956	Confidence interval and correction of posterior prediction method
	0	0.974	0.964	0.944	Generalized confidence intervals and corrections
		0.941	0.948	0.943	Confidence interval and correction of posterior prediction method
	-0.5	0.935	0.942	0.946	Generalized confidence intervals and corrections
		0.933	0.953	0.948	Confidence interval and correction of posterior prediction method
	-1	0.883	0.921	0.934	Generalized confidence intervals and corrections
		0.924	0.946	0.942	Confidence interval and correction of posterior prediction method

As shown in Table II, after correction, its confidence interval was more in line with the requirements. When b_1 was (1, 2, 3) and b_1 was -1, the confidence interval coverage probability of the posterior prediction method was 0.924, while the generalized confidence interval was only 0.883. The confidence intervals of the posterior prediction method were closer, and as b_1 increased, the confidence intervals of both methods gradually approached 0.95. When b_1 was (1, 2, 3), taking different values resulted in significant fluctuations in the coverage probability of the generalized confidence interval; When b_1 was (3, 4, 10), the coverage probability was relatively closer to 95%. Overall, in terms of sample size, confidence intervals based on posterior prediction distributions outperformed generalized confidence intervals in terms of coverage probability. The numerical simulation results of the second set of data are shown in Table III.

Table III shows the coverage probabilities of different confidence intervals in the linear correction model. This table compares the coverage probabilities of generalized confidence intervals and corrections, based on posterior prediction methods at different values. When b_1 was (1, 2, 3) and θ was 1, the confidence interval coverage probability of the posterior prediction method was 0.916, while the generalized confidence interval was 0.896. Meanwhile, the coverage probability of the generalized confidence interval might differ significantly from 95% in some cases, but as it increased, its coverage probability gradually approached 95%. For example, when b_1 was (3, 4, 10), the coverage probability of the generalized confidence interval was closer to 95% for each value. In addition, as the sample size increased from 10 to 20, the coverage probabilities of both confidence intervals became closer to 95%. This indicated that when the sample size was 20, the confidence interval of the posterior prediction distribution performed better in terms of coverage probability, and an increase in sample size helped to improve the accuracy of the confidence interval coverage probability. The numerical simulation results of the third set of data are shown in Table IV.

TABLE III. RESULTS OF THE NUMERICAL SIMULATIONS OF THE SECOND SET OF DATA

n	θ	b_1 (1,2,3)	b_1 (2,1,5)	b_1 (3,4,10)	Method
20	1	0.896	0.912	0.945	Generalized confidence intervals and corrections
		0.916	0.943	0.946	Confidence interval and correction of posterior prediction method
	0.5	0.952	0.945	0.933	Generalized confidence intervals and corrections
		0.945	0.952	0.956	Confidence interval and correction of posterior prediction method
	0	0.977	0.965	0.947	Generalized confidence intervals and corrections
		0.942	0.946	0.943	Confidence interval and correction of posterior prediction method
	-0.5	0.937	0.944	0952	Generalized confidence intervals and corrections
		0.936	0.955	0.943	Confidence interval and correction of posterior prediction method
	-1	0.888	0.928	0.936	Generalized confidence intervals and corrections
		0.923	0.948	0.946	Confidence interval and correction of posterior prediction method

TABLE IV. RESULTS OF THE NUMERICAL SIMULATIONS OF THE THIRD SET OF DATA

n	θ	b_1			Method
		(1,2,3)	(2,1,5)	(3,4,10)	
30	1	0.892	0.911	0.945	Generalized confidence intervals and corrections
		0.912	0.943	0.949	Confidence interval and correction of posterior prediction method
	0.5	0.920	0.932	0.936	Generalized confidence intervals and corrections
		0.946	0.947	0.959	Confidence interval and correction of posterior prediction method
	0	0.976	0.965	0.944	Generalized confidence intervals and corrections
		0.942	0.941	0.947	Confidence interval and correction of posterior prediction method
	-0.5	0.936	0.943	0955	Generalized confidence intervals and corrections
		0.938	0.955	0.946	Confidence interval and correction of posterior prediction method
	-1	0.886	0.926	0.933	Generalized confidence intervals and corrections
		0.924	0.943	0.941	Confidence interval and correction of posterior prediction method

TABLE V. COMPARATIVE ANALYSIS OF LINEAR CORRECTION MODELS

Method	Applicable scenarios	Computing efficiency	Confidence interval coverage (sample size=30)
Traditional Fiducial inference	Low dimensional data, large sample size	Medium	0.898
Bayesian calibration model	Small sample, stable data structure	Low	0.912
Generalized linear model	Medium-dimensional data	High	0.885
Dynamic linear correction model	Real time data stream	High	0.902
This method	High dimensional data, small sample size	High	0.927

According to Table IV, after correction using the posterior prediction method, the confidence interval was closer to 0.95 and the coverage level was better. The confidence interval coverage probability of the posterior prediction method was 0.927, while the generalized confidence interval was 0.898. The coverage probability of the generalized confidence interval might deviate significantly from 95%, but as it increased, its coverage probability approached 95%. For example, when b_1 was (1, 2, 3), the generalized confidence interval coverage probability fluctuated significantly. b_1 had a relatively stable coverage probability of (3, 4, 10) and was closer to 95%. As the sample size increased from 10 and 20 to 30, the coverage probability of both confidence intervals approached 95%, indicating that increasing the sample size can improve the performance of coverage probability. Overall, when the sample size was 30, the confidence interval based on posterior prediction distribution performed better in terms of coverage probability. The study further compared the performance of the research-designed method with other current methods, and the results are shown in Table V.

The research-designed model achieved a coverage rate of 0.927 at a sample size of 30, significantly better than traditional Fiducial methods and generalized linear models, highlighting its superiority in small sample scenarios. Although Bayesian methods performed moderately, they relied on prior distributions and had limited flexibility. Integrating Fiducial inference (reverse parameter analysis) with Bayesian posterior prediction reduced redundant iterations and achieved efficiency comparable to GLM while maintaining high accuracy.

V. DISCUSSION

A. The Significance of Research Results

The joint independent information linear correction model proposed in the study significantly improved the accuracy and efficiency of statistical inference in high-dimensional data and small sample scenarios by integrating Fiducial inference and Bayesian methods. Simulation experiments showed that when the sample size was 30, the confidence interval coverage of the posterior prediction method reached 0.927, approaching the theoretical 95% confidence level, which has significant advantages over traditional Fiducial methods and generalized linear models. This result validated the effectiveness of the model in reducing coverage bias and enhancing the robustness of parameter estimation. Especially in high-dimensional noise covariance and dynamic data stream scenarios, the model achieved adaptive adjustment of complex data structures through information domain dynamic correction mechanism.

B. Comparative Advantages with Existing Methods

Compared with traditional methods, the innovation of research-designed model mainly lies in their high-dimensional adaptability: although existing methods support real-time data streams, their ability to handle high-dimensional noise covariance is limited. This study significantly improved computational stability in high-dimensional environments by integrating multi-source data features through a joint independent information framework. Dynamic correction capability: The information domain dynamic correction mechanism integrates background knowledge and data features to solve the limitations of traditional Bayesian methods that rely on fixed prior distributions, demonstrating stronger scene adaptability in industrial manufacturing multi-sensor systems. Balancing efficiency and accuracy: Fiducial's reverse parameter distribution analysis reduces the iterative redundancy of Bayesian posterior prediction, allowing the model to maintain high coverage while maintaining computational efficiency comparable to generalized linear models, making it suitable for resource constrained real-time systems.

C. Actual Application Potential

The research-designed method has broad application prospects in medical image analysis, environmental monitoring, and industrial Internet of Things. In the medical field, B1 non-uniformity correction of MRI signals can be combined with dynamic correction in the information domain to improve image signal-to-noise ratio. In the field of environmental monitoring, in the fusion of multi-source precipitation data, the model can dynamically integrate meteorological station and satellite data to reduce systematic bias. In the field of industrial Internet of Things, dynamic calibration of real-time sensor data streams can optimize manufacturing process monitoring and reduce equipment anomaly false alarm rates.

VI. Conclusion

The study successfully improved the accuracy and reliability of statistical inference by constructing a linear correction model based on joint independent information and combining it with Fiducial inference method. The research results indicated that the confidence interval correction method designed in the study had better performance, especially in small sample situations, where the confidence interval coverage probability of the posterior prediction method was closer to the 95% confidence level. The specific data showed that when the sample size was 10, the confidence interval coverage probability of the posterior prediction method was 0.923, while the generalized confidence interval was 0.889. When the sample size increased to 30, the confidence interval coverage probability of the posterior prediction method was 0.927, and the generalized confidence interval was 0.898. In addition, as the sample size increased, the coverage probabilities of both confidence intervals were closer to 95%, indicating that an increase in sample size helps to improve the accuracy of the confidence intervals. The study also found that in some cases, the generalized confidence interval had a large deviation between the coverage probability and 95%, but as the sample size increased, its coverage probability gradually approached 95%. For example, when the sample size was 30, the fluctuation of the coverage probability of the generalized confidence interval was significantly reduced, and the coverage probability was relatively stable and closer to 95%. This indicated that confidence intervals based on posterior prediction distributions performed better in larger sample sizes. The study validated the effectiveness and stability of a linear correction model based on joint independent information in processing high-dimensional data through numerical simulations. The research results provide new theoretical support and methodological improvements for linear correction models in high-dimensional data environments, and have important theoretical and practical value. However, the statistical inference correction model designed for research cannot meet the requirements for model inference correction when dealing with multivariate problems. Future research will focus on highly coupled multivariate data and propose to introduce tensor decomposition techniques and graph model structures to address the problem of insufficient representation of complex correlation structures in current models by characterizing the nonlinear relationships and topological dependencies between variables. At the same time, for large-scale data, it plans to combine distributed computing architecture and hardware acceleration technology to design layered iterative algorithms to reduce memory usage. Low-level approximations and sparse representations of model parameters are also explored to compress computational complexity while ensuring statistical performance.

References

[1] Xie W, Yi S, Leng C. Two-Stage Multi-Source Precipitation Data Merging Method Combining Bias Correction and Dynamic Constrained Linear Regression Model. Journal of Geo-Information Science, 2024, 26(11): 2506-2528.

[2] Ouyang Y, Taljaard M, Forbes A B, Li F. Maintaining the validity of inference from linear mixed models in stepped-wedge cluster randomized trials under misspecified random-effects structures. Statistical Methods in Medical Research, 2024, 33(09): 1497-1516.

[3] Zhou J, Claeskens G. Automatic bias correction for testing in high-dimensional linear models. Statistica Neerlandica, 2023, 77(01): 71-98.

[4] Sun P Z. Quasi-steady-state (QUASS) reconstruction enhances T1 normalization in apparent exchange-dependent relaxation (AREX) analysis: A reevaluation of T1 correction in quantitative CEST MRI of rodent brain tumor models. Magnetic Resonance in Medicine, 2024, 92(01): 236-245.

[5] Kaitan R G, Bellec P C. Noise covariance estimation in multi-task high-dimensional linear models. Bernoulli, 2024, 30(03): 1695-1722.

[6] Zhang R, Shen Y, Han D. Correction and dynamical analysis of classical mathematical model for piecewise linear system. Lixue Xuebao/Chinese Journal of Theoretical and Applied Mechanics, 2024, 56(01): 225-235.

[7] Sun T, Li W, Lin L. Matrix-variate generalized linear model with measurement error. Statistical Papers, 2024, 65(06): 3935-3958.

[8] Li H, Luojie D, Ding H. Real-time service performances of BDS-3 and Galileo constellations with a linear satellite clock correction models. Satellite Navigation, 2023, 4(03): 72-81.

[9] ElHorbaty Y S. A Monte Carlo permutation procedure for testing variance components in generalized linear regression models. Computational Statistics, 2024, 39(05): 2605-2621.

[10] Maksaei N, Rasekh A, Babadi B. Local influence in linear mixed measurement error models with ridge estimation. Communications in Statistics - Simulation and Computation, 2024, 53(12): 5899-5912.

[11] Wang P, Teng F, Geng J, Liu Y. Axial vibration characteristics analysis of transformer windings based on magnetic-structural coupling correction model. IET Electric Power Applications, 2024, 18(10): 1408-1420.

[12] Gibiansky L, Gibiansky E. Note on importance of correct stoichiometric assumptions for modeling of monoclonal antibodies. Journal of Pharmacokinetics and Pharmacodynamics, 2024, 51(04): 307-317.

[13] Emami H. Diagnostics for partially linear measurement error models. Communications in Statistics - Theory and Methods, 2024, 53(17): 6224-6239.

[14] Chang H, Middleton J A, Aronow P M. Exact bias correction for linear adjustment of randomized controlled trials. Econometrica, 2024, 92(05): 1503-1519.

[15] Li L, Wang Z, Chen J, Lu F. A model transfer method based on transfer component analysis and direct correction. Spectroscopy and Spectral Analysis, 2024, 44(12): 3399-3405.

[16] Cardot H, Mas A, Sarda P. Correction: CLT in functional linear regression models (Probability Theory and Related Fields, (2007), 138, 3-4, (325-361), 10.1007/s00440-006-0025-2). Probability Theory and Related Fields, 2023, 187(1-2): 519-522.

[17] Wan W, Bai Y, Lu Y, Ding L. A hybrid model combining a gated recurrent unit network based on variational mode decomposition with error correction for stock price prediction. Cybernetics and Systems, 2024, 55(05): 1205-1229.

[18] Hong S, Jiang J, Jiang X, Wang H. Inference for possibly misspecified generalized linear models with nonpolynomial-dimensional nuisance parameters. Biometrika, 2024, 111(04): 1387-1404.

[19] Bariffi J, Bartz H, Liva G, Rosenthal J. Error-correction performance of regular ring-linear LDPC codes over Lee channels. IEEE Transactions on Information Theory, 2024, 70(11): 7820-7839.

[20] Haschka R E. Robustness of copula-correction models in causal analysis: Exploiting between-regressor correlation. IMA Journal of Management Mathematics, 2024, 36(01): 161-180.

[21] Wu Q, Gong P, Liu S, Li Y. B1 inhomogeneity corrected CEST MRI based on direct saturation removed omega plot model at 5T. Magnetic Resonance in Medicine, 2024, 92(02): 532-542.

[22] Saraswat S P, Addad Y. A comprehensive examination of the linear and numerical stability aspects of the bubble collision model in the TRACE-1D two-fluid model applied to vertical disperse flow in a PWR core channel under loss of coolant accident conditions. Nuclear Engineering and Technology, 2024, 56(08): 2974-2989.

[23] Lu Q, Wang B, Huang Z. A dual electrode mixed potential SO2 sensor with humidity self-correction function utilizing multiple linear regression model. IEEE Electron Device Letters, 2024, 45(07): 1293-1296.

[24] Simicic D, Zollner H J, DaviesJenkins C W, Hupfeld K E, Edden R A E, Oeltzschner G. Model-based frequency-and-phase correction of 1H MRS data with 2D linear-combination modeling. Magnetic Resonance in Medicine, 2024, 92(05): 2222-2236.