Japanese Journal of Statistics and Data Science
https://doi.org/10.1007/s42081-025-00303-3

ORIGINAL PAPER

A new algorithm for variable selection in general linear models

Fayyaz Bahari[1]

Received: 22 January 2025 / Revised: 17 March 2025 / Accepted: 11 April 2025

Abstract
In regression studies, especially in general linear models, variable selection is key. It is vital for statistical inferences. Recently, Chatterjee's correlation has been introduced as a novel measure of dependence. This criterion can measure the nonlinear or functional relationship between two variables. On the other hand, the general linear model does not necessarily require a linear relationship between the response variable and covariates. This paper proposes a new algorithm that selects the appropriate variables by sequential tests for the general linear model based on Chatterjee's correlation. In classical algorithms, changing the functional form of the general linear model may lead to the selection of different covariates. But, in the new algorithm, selected variables aren't affected by changing the functional form of the general linear model. Also, a theorem discusses the algorithm's properties, and simulations show the excellent performances of the new proposed algorithm. Finally, we applied our method to real data.

Keywords General Linear Model · Variable Selection · Chatterjee's correlation · Sequential tests · New Algorithm

1 Introduction

In most regression studies, it is not appropriate to consider multiple linear relationships between the response variable and covariates. For this reason, many researchers have considered the following general linear model (see Bahari et al. (2021), Ben-Dor and Banin (1995), Chatterjee (2021), Chatterjee (2024), Creemers et al. (2011)):

$$y = g^T(\mathbf{x})\beta + \epsilon, \tag{1}$$

✉ Fayyaz Bahari
fayyaz.bahari@uma.ac.ir

[1] Department of Statistics and Applications, Faculty of Mathematical Sciences, University of Mohaghegh Ardabili, Ardabil, Iran

Published online: 16 June 2025

Where, y is a response variable, $\mathbf{x}$ is a vector of covariates with dimension k, $g(\cdot)$ is a vector function of dimension p, β is an unknown vector parameter of dimension p, and ϵ is the measurement error with mean 0 and constant variance, σ^2. Moreover, it is assumed that $E(\epsilon|x) = 0$ and $E(\epsilon^2|x) < \infty$. If one applies the general linear model, two essential problems will arise. The first problem is related to choosing the functional form of the function $g(\cdot)$, and the second one is related to the variable selection, the covariates should be included in the model.

In the special case for the multiple regression, the functional form of $g(\cdot)$ is known, and it is only enough to select the suitable covariates in the model. There are many methods in the literature to overcome this problem. For example, Efroymson (1960) used the stepwise methods, especially the forward algorithm, for variable selection in the model. Such algorithms for variable selection aim to find covariates that are highly correlated with the response variable. For more details on the use of monotonic associations, such as Pearson correlation, in variable selection, see also Guyon and Elisseeff (2003), and Kutner et al. (2005). Indeed, the covariates remain in the model with the highest partial correlation in comparison to others. In general, the functional form of the general linear model is not necessarily linear. Therefore, using the classical stepwise algorithms for variable selection in the general linear model may lead to an inappropriate model. When the functional form of the general linear model is known, using a stepwise algorithm can be helpful. But, in applications, it rarely happens. By the way, it is reasonable to assume that the functional form of $g(\cdot)$ is unknown.

Some researchers have proposed statistical tests to check the goodness of fit of the general linear model (see Hardle et al. (1998), Holm (1979), Kutner et al. (2005), Lin and Han (2023), Montgomery et al. (2021), Pei et al. (2024), Shi et al. (2022), Thottolil et al. (2023), Zhu and Chu (2005)). But, they didn't study the variable selection problem. The main reason that such researchers didn't study the variable selection problem is that Pearson correlation and other criteria, derived from it, measure the strength of the linear relationship between the response variable and covariates. In the general linear model, the response variable relationship with covariates can be defined by numerous $g(\cdot)$ functions. This means that variable selection by classical algorithms strongly depends on the $g(\cdot)$ function. Therefore, using Pearson criteria is not suitable for checking dependency of them. This problem motivated us to find some way to select effective covariates in the general linear model.

In recent years, Chatterjee (2021) has introduced a criterion that measures the functional relationship of two variables. If we have two variables such as Y and X, the functional relationship of X to Y can be measured. Consider the paired case of a sample by size n ($n > 2$). If there exist no ties in the observations, consider the sorted paired of two samples based on the x_is as $(x_{(1)}, y_{(1)}), (x_{(2)}, y_{(2)}), \ldots, (x_{(n)}, y_{(n)})$, where $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$. Since there exist no ties in data, we can do it uniquely. Also, we assume that r_i be the rank of $y_{(i)}$, where it is the number j such that $y_{(j)} \leq y_{(i)}$. Chatterjee (2021) proposed a new correlation criterion, now known as Chatterjee's correlation.

$$\zeta_n(X, Y) = 1 - \frac{3\sum_{i=1}^{n-1} |r_{i+1} - r_i|}{n^2 - 1}. \tag{2}$$

In general case, if there exist ties in the observations, choose an increasing rearrangement as above by breaking ties uniformly at random. r_is defined as before and define the l_is to be the number j such that $y_{(j)} \geq y_{(i)}$. Finally, the Chatterjee's correlation can generally be defined as follows:

$$\zeta_n(X, Y) = 1 - \frac{n \sum_{i=1}^{n-1} |r_{i+1} - r_i|}{2 \sum_{i=1}^{n} l_i (n - l_i)}. \tag{3}$$

Note that, unlike the Pearson correlation, in the Chatterjee's correlation, $\zeta_n(X, Y) \neq \zeta_n(Y, X)$. Indeed, $\zeta_n(Y, X)$ measures the dependency rate of Y to X. Moreover, $\zeta_n(X, Y)$ belongs to $[0, 1]$. It is equal to 0 if and only if X is independent of Y, and it is equal to 1 if and only if there exists a real-valued and measurable function f such that $Y = f(X)$, almost surely. Some authors such as Azadkia and Chatterjee (2021), Shi et al. (2022) and Lin and Han (2023) have improved some properties of the functional correlation. Some other researchers, such as Fuchs and Wang (2024) have used the properties of functional correlation for hierarchical variable clustering, and Thottolil et al. (2023) have also used functional correlation for establishing the relationship between HSIs and network density and to build a prediction model. Recently, Pei et al. (2024) employed Chatterjee's correlation in place of Pearson's correlation, proposing a new test for time series studies, while some of the properties and generalizations of Chatterjee's correlation have been reviewed by Chatterjee (2024).

In the next section, new algorithm has been introduced for variable selection in the general linear model. In Sect. 3, the performances of the proposed algorithm have been studied by simulation of some models. In Sect. 4, we will see applications of our proposed algorithm for real data. In the final section, we have discussed some applications and theoretical properties of our proposed algorithm.

2 Theorical approches

As noted in the last section, we face two key problems in fitting the general linear model. To select the suitable covariates for the general linear model, we will use the Chatterjee's correlation. It measures the rate of dependency between covariates and the response variable. Also, to select the appropriate covariates for the model, we will introduce a new algorithm that selects the variables by the sequential tests. In the new algorithm the sequentially rejective Bonferroni scheme which is introduced by Holm (1979) is used. Using this scheme in similar algorithms is very common, and is powerful than the simple Bonferroni scheme. By applying a similar sequential test scheme in the new algorithm, we guarantee that the level of whole tests in the algorithm is at most α. Moreover, this scheme is more powerful than the Bonferroni scheme. On the other hand, after selecting appropriate covariates to overcome the functional form of $g(\cdot)$, we can use different strategies to fit the final general linear model. In this paper, we will use the Akaike Information Criterion (AIC) and the Bayesian Information criterion (BIC) to determine the final general linear model in a real data study. In statistical studies, it is common to consider a class of models and select a model which is optimal among them. Because the class of functional form of

models, especially the form of $g(\cdot)$ function in general linear model is very wide. The following algorithm is proposed to select appropriate variables for the general linear model.

New algorithm for variable selection in general linear model:

Step 1. Calculate the Chatterjee's correlation between the response variable and the covariates, i.e. $\zeta_n(X_1, Y)$, $\zeta_n(X_2, Y)$, $\cdots$, $\zeta_n(X_k, Y)$ based on the observations.

Step 2. Consider the following hypotheses based on the biggest to the smallest Chatterjee's correlation:

$$
\begin{aligned}
&H_0^{(1)}: \zeta_n^{(k)} = 0 \quad vs \quad H_1^{(1)}: \zeta_n^{(k)} \neq 0,\\
&H_0^{(2)}: \zeta_n^{(k-1)} = 0 \quad vs \quad H_1^{(2)}: \zeta_n^{(k-1)} \neq 0,\\
&\cdot\\
&\cdot\\
&\cdot\\
&H_0^{(k)}: \zeta_n^{(1)} = 0 \quad vs \quad H_1^{(k)}: \zeta_n^{(1)} \neq 0,
\end{aligned}
$$

where, $\zeta_n^{(1)} \leq \zeta_n^{(2)} \leq ... \leq \zeta_n^{(k)}$, which are the ordered values of $\zeta_n(X_1, Y)$, $\zeta_n(X_2, Y), \cdots, \zeta_n(X_k, Y)$.

Step 3. Calculate the significant level of the hypotheses $H_0^{(j)}$, $j = 1, 2, ..., k$ as follows:

$$
p.value^{(j)} = 2\left(1 - P\left(Z \leq |\frac{\zeta_n^{(k-j+1)}}{\sqrt{\frac{2}{5n}}}|\right)\right), \quad j = 1, 2, \cdots, k, \qquad (4)
$$

where, Z follows N(0,1).

Step 4. Select the suitable covariates in the model as follows:

4.1: If $p.value^{(1)} > \frac{\alpha}{k}$, then no covariates enter the general linear model. If not, enter the covariate into the general linear model which its Chatterjee's correlation matches to the $\zeta_n^{(k)}$. Then, go to the next step.

4.2: If $p.value^{(2)} > \frac{\alpha}{k-1}$, then stop. If not, enter the covariate into the general linear model which its Chatterjee's correlation matches to the $\zeta_n^{(k-1)}$. Then, go to the next step.

.
.
.

4.j: If $p.value^{(j)} > \frac{\alpha}{k-j+1}$, then stop. If not, enter the covariate into the general linear model which its Chatterjee's correlation matches to the $\zeta_n^{(k-j+1)}$. Then,

go to the next step.

.
.
.

4.k: If $p.value^{(k)} > \alpha$, then stop. If not, enter all the covariates into the general linear model, and then stop.

In the above algorithm, $\zeta_n(X_j, Y)$s, $j = 1, 2, \cdots, k$, are calculated by using the equation (2). We have presented a theorem which are used to calculate the significance levels in Step 3. More details of Step 3 are given in the upcoming theorem. In this paper, we have assumed that the response variable is continuous. If the response variable wasn't necessarily continuous, then Theorem 2.1 of Chatterjee (2021) and also our proposed theorem will be valid yet, but the variance parameter of Normal distribution will be different (see Chatterjee (2021)). In Step 4, the sequential tests are applied to select appropriate variables, where the covariates with higher dependency on the response variable eventuate higher amounts of p.value. Our proposed theorem guarantee that the level of all combinations of tests in the proposed algorithm be at most α. In the above algorithm, variable selection happens in Step 4. Indeed, the following hypotheses are tested:

$$H_0^{(j)} : \zeta_n^{(k-j+1)} = 0 \quad vs \quad H_1^{(j)} : \zeta_n^{(k-j+1)} \neq 0, \qquad j = 1, 2, \cdots . \tag{5}$$

The null hypothesis of the above equation, which is based on the ordered Chatterjee's correlation, implies that the certain covariate is independent of the response variable. The following test statistics are used to test the above hypotheses:

$$T_j = \frac{\zeta_n^{(k-j+1)}}{\sqrt{\frac{2n}{5}}}, \qquad j = 1, 2, \cdots . \tag{6}$$

The null hypotheses will be rejected if $|T_j| > c_j$, where c_j is the critical point. The coming theorem describes the properties of our proposed algorithm.

Theorem 1 *The test statistics of equation* (6) *under hypotheses* (5) *follow* $N(0, 1)$ *distribution for large sample sizes. Moreover, the level of any combinations of tests in the proposed algorithm is at most* α.

Proof It is the direct result of Theorem 2.1 of Chatterjee (2021) and Theorem 2 of Holm (1979). The Normality comes from Theorem 2.1 of Chatterjee (2021) and the level of tests comes from Theorem 2 of Holm (1979). Where, in Theorem 2 of Holm (1979), one can use similar arguments by considering $c_j s = 1$, $j = 1, 2, \cdots, k$, and conclude that the level of tests in the algorithm is at most α. □

Theorem 1 provides that the level of any combinations of tests be at most α in the algorithm. Furthermore, it gives a more powerful scheme in comparison to the

traditional Bonferroni scheme. Also, in Theorem 1, a good approximation of the test statistics distribution extremely depends on the large sample sizes. The following flowchart, given in Fig. 1 is designed for variable selection based on the new algorithm.

After selecting the appropriate covariates, we can derive the optimal model in a pre-assumed class of general linear models by the AIC or BIC criterion. To calculate AIC and BIC in any cases, the following formulas are used:

$$\begin{aligned} AIC_j &= nln\left(\frac{SSE}{n}\right) + 2p, \quad j = 1, 2, \cdots, m, \\ BIC_j &= nln\left(\frac{SSE}{n}\right) + pln(n), \quad j = 1, 2, \cdots, m, \end{aligned} \tag{7}$$

where, p is the number of parameters in each model and m is the number of general linear models in the assumed class. Moreover, $g(\cdot)$s are known functions in the assumed class. For more details about the model selection criteria, see also Montgomery et al. (2021). In the next section, we just simulate the experiments to select the covariates. However, in real data studies, it is necessary to consider a class of $g(\cdot)$ functions to find the optimal general linear model.

3 Simulation study

In this section, we have considered some separate cases to study the properties of our proposed algorithm to see its performances in different situations. However, throughout the studies, we just thought about the variable selection part of the problem.

In continuation of this section, we will consider different functional cases of the general linear model. In any study, we have repeated the simulation study 5000 times with R software. In any study, we have supposed the level of whole tests in algorithm to be $\alpha = 0.05$, and to see the sensitivity of models under changing measurement errors, we have considered different variances to ϵ. Moreover, to see the sample size effects in studies, we have considered different sample sizes. Also, for a better understanding of different cases of simulation studies, we have started with the simplest model, and we have continued to study the most complex general linear models. In any studies, the empirical Coverage Probability (CP) of accepting any covariates in the general linear model has been calculated based on the proposed algorithm for different values of σ^2 and n.

3.1 General linear model with no covariates in the model

In most regression studies, maybe, there is no relationship between the response variable and the covariates. For this reason, we consider a case in which there are no covariates in the model. Therefore, consider the Model (1), where, $g(\mathbf{x}) = 1$, $\beta = 1$ and $\epsilon \sim N(0, \sigma^2)$ with known values of σ^2. Furthermore, suppose we have three covariates in hand that are candidates to enter into the model where $X_1 \sim N(0, 1)$,

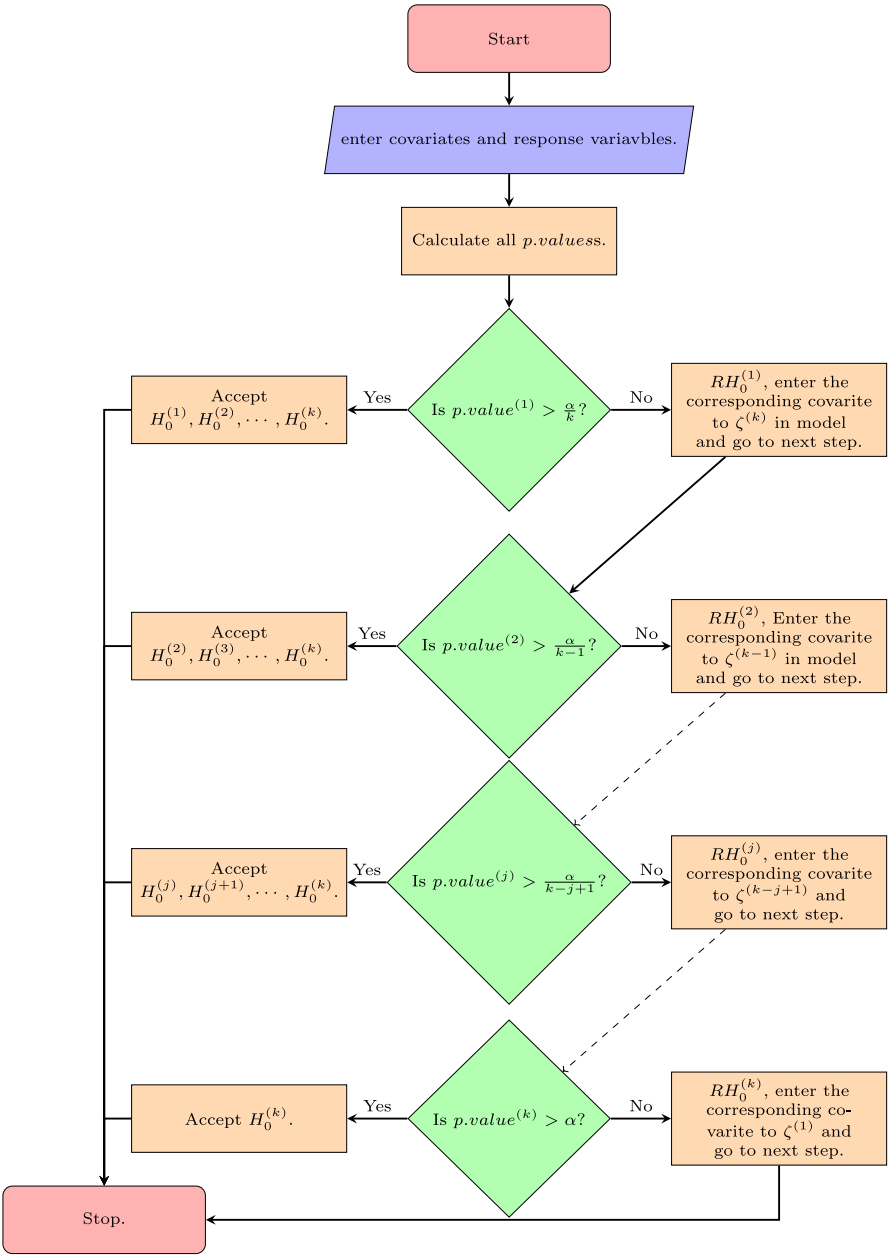

Fig. 1 Variable selection flowchart for glm based on the new algorithm

Table 1 Empirical coverage probability of entering any covariates into the general linear model when there are no covariates in the model with three candidate covariates for different values of σ^2 and n based on the 5000 repetitions of the Monte Carlo algorithm

n	σ^2	X_1	X_2	X_3
100	0.01	0.0156	0.0158	0.0122
	0.10	0.0116	0.0146	0.0154
	0.25	0.0152	0.0156	0.0134
	0.50	0.0134	0.0160	0.0176
	1.00	0.0156	0.0170	0.0152
200	0.01	0.0146	0.0140	0.0166
	0.10	0.0158	0.0164	0.0176
	0.25	0.0172	0.0150	0.0168
	0.50	0.0178	0.0158	0.0128
	1.00	0.0162	0.0160	0.0196
300	0.01	0.0138	0.0142	0.0180
	0.10	0.0152	0.0170	0.0174
	0.25	0.0134	0.0150	0.0170
	0.50	0.0162	0.0156	0.0152
	1.00	0.0154	0.0184	0.0162

$X_2 \sim U(0, 3)$ and $X_3 \sim N(0, 1)$. In this case, Model (1) is equivalent to:

$$y = 1 + \epsilon, \tag{8}$$

where is the simplest general linear model with no covariates. The results of this study are given in Table 1 for various sample sizes and different values of σ^2.

From Table 1, one can conclude that the chance of entering any covariate into the model is very low. In this case, we can say that no covariates are inserted into the model as we expected, and the probability of inserting all covariates into the model is approximately 0.05. This probability comes from the assumed level for the tests in the algorithm, i.e. $\alpha = 0.05$. In this study, the effect of sample sizes and measurement errors are not very obvious because there are no covariates in the general linear model. The effects of n and σ^2 will be seen more obviously in the next studies by the existence of covariates in the general linear model. However, the good performance of our proposed algorithm is obvious and the empirical coverage probability of entering the candidate covariates into the model is very low.

3.2 Multiple linear model with two significant covariates in the model

To see the good performance of our proposed algorithm in the multiple regression model, which is the special case of the general linear model, Model (1) is simulated. Where in Model (1), $g(\mathbf{x}) = (x_1, x_2)$, $\beta = (1, 1)$ and $\epsilon \sim N(0, \sigma^2)$ where σ^2 is a known value. Furthermore, $X_1 \sim N(0, 1)$, $X_2 \sim U(0, 3)$, and the variable X_3 which is not in the model, follows from $N(0, 1)$. In this case, Model (1) is equivalent to:

$$y = x_1 + x_2 + \epsilon, \tag{9}$$

Table 2 Empirical coverage probability of entering any covariates into the general linear model when there are two covariates in the model with three candidate covariates for different values of σ^2 and n based on the 5000 repetitions of the Monte Carlo algorithm

n	σ^2	X_1	X_2	X_3
100	0.01	0.9954	0.9628	0.0490
	0.10	0.9908	0.9382	0.0430
	0.25	0.9790	0.8862	0.0404
	0.50	0.9366	0.7950	0.0426
	1.00	0.7938	0.5916	0.0322
200	0.01	1.0000	0.9998	0.0464
	0.10	1.0000	0.9990	0.0470
	0.25	0.9998	0.9940	0.0448
	0.50	0.9986	0.9782	0.0496
	1.00	0.9782	0.9000	0.0422
300	0.01	1.0000	1.0000	0.0496
	0.10	1.0000	1.0000	0.0520
	0.25	1.0000	0.9996	0.0476
	0.50	1.0000	0.9982	0.0450
	1.00	0.9986	0.9750	0.0518

where is the multiple regression model with two covariates. The following table shows the results of simulation studies under the mentioned assumptions.

From Table 2, as we expected, the probability of entering X_1 and X_2 into the general linear model is very high, and the probability of entering X_3 into the model is very low. Moreover, by increasing the variance of measurement errors, the chance of entering X_1 and X_2 into the model has decreased. On the other hand, by increasing the sample sizes, the chance of entering X_1 and X_2 into the general linear model has increased. However, the empirical coverage probability of entering X_3 into the model is not affected much by changing the values of σ^2 and n. This comes from the fact that X_3 is not in the assumed model. The results of this study show the good performance of our proposed algorithm.

3.3 General linear model with two significant covariates in the model

In this section, we simulate the general linear model by two covariates in the model, and in addition, we assume that there exist 4 extra covariates where these extra covariates don't affect the model. Therefore, we simulate Model (1) where in Model (1) $g(\mathbf{x}) = (x_1, x_2^2)$, $\beta = (1, 1)$ and $\epsilon \sim N(0, \sigma^2)$ where σ^2 is a known value. Also, $X_1 \sim U(0, 3)$ and $X_2 \sim N(0, 1)$. Moreover, we assume that the extra covariates in this study, X_3, X_4, X_5 and X_6 follow from $\sim N(0, 1)$. In this case, Model (1) is equivalent to:

$$y = x_1 + x_2^2 + \epsilon. \tag{10}$$

This general linear model is equivalent to a polynomial regression model. The results of this study are given in Table 3 by the mentioned assumptions.

Table 3 Empirical coverage probability of entering any covariates into the general linear model when there are two covariates in the model with six candidate covariates for different values of σ^2 and n based on the 5000 repetitions of the Monte Carlo algorithm

n	σ^2	X_1	X_2	X_3	X_4	X_5	X_6
100	0.01	0.9822	0.9806	0.0128	0.0122	0.0090	0.0101
	0.10	0.9822	0.9812	0.0114	0.0134	0.0116	0.0130
	0.25	0.9684	0.9738	0.0098	0.0116	0.0122	0.0112
	0.50	0.9018	0.9446	0.0104	0.0106	0.0106	0.0108
	1.00	0.4690	0.7796	0.0120	0.0104	0.0074	0.0088
200	0.01	1.0000	0.9996	0.0112	0.0124	0.0098	0.0114
	0.10	1.0000	1.0000	0.0108	0.0128	0.0140	0.0114
	0.25	0.9994	1.0000	0.0146	0.0116	0.0104	0.0120
	0.50	0.9968	0.9998	0.0132	0.0106	0.0114	0.0114
	1.00	0.8142	0.9838	0.0124	0.0118	0.0126	0.0122
300	0.01	1.0000	1.0000	0.0120	0.0132	0.0146	0.0134
	0.10	1.0000	1.0000	0.0104	0.0130	0.0130	0.0108
	0.25	1.0000	1.0000	0.0128	0.0104	0.0152	0.0110
	0.50	0.9996	1.0000	0.0126	0.0118	0.0134	0.0100
	1.00	0.9436	0.9998	0.0112	0.0088	0.0120	0.0112

Similar conclusions to the previous study are obtained in this study. Where, the chance of entering existing covariates into the assumed model is very high and the chance of the covariates that are not in the assumed model is very low for entering into the model. Also, by increasing the number of candidate covariates in comparison to the last study, the effects of sample sizes and measurement errors are very obvious. Figures 2 and 3 show the empirical coverage probability of entering X_1 and X_2 into the general linear model under changing σ^2 and n, respectively.

Figure 2 shows the decreasing behavior of empirical coverage probability under increasing σ^2, and Fig. 3 shows the increasing behavior of empirical coverage probability under increasing sample sizes. The results of this study show good performances of our proposed algorithm.

3.4 More complex general linear model with three significant covariates in the model

In This case, we consider a more complex general linear model with a large number of candidate covariates. In model (1), suppose $g(\mathbf{x}) = (x_1, x_2^3, e^{x_3})$, $\beta = (2.75, 2.5, 1.5)$, $\epsilon \sim N(0, \sigma^2)$ with known values of σ^2, $X_1 \sim U(-3, 3)$, $X_2 \sim N(0, 1)$, and $X_3 \sim Exp(1)$. Also, there exist some covariates that do not have a rule in the model where $X_4, X_5, X_6 \sim N(0, 1)$, $X_7 \sim U(-1, 1)$, $X_8 \sim N(1, 1)$, $X_9 \sim Exp(0.5)$, and $X_{10} \sim Beta(2, 2)$. In this case, Model (1) is equivalent to:

$$y = 2.75x_1 + 2.5x_2^3 + 1.5e^{x_3} + \epsilon. \tag{11}$$

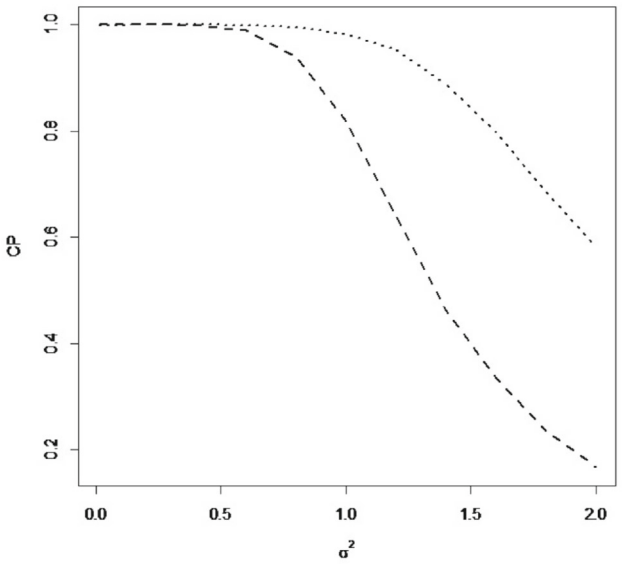

Fig. 2 Empirical coverage probability of existing covariates in the general linear model for entering into the model for different values of σ^2 where the dashed curve shows the CP of X_1 and the dotted curve shows the CP of X_2

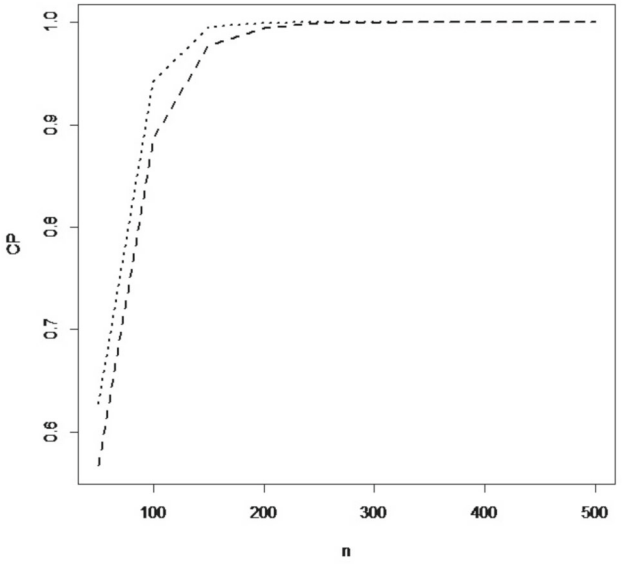

Fig. 3 Empirical Coverage probability of existing covariates in the general linear model for entering into the model for different values of n where the dashed curve shows the CP of X_1 and the dotted curve shows the CP of X_2

This model is more complex than the previously studied model, and there are 10 candidate covariates to enter into the model. The results of this study are given in Table 4.

From Table 4, the sample size effects are very obvious. By increasing the sample sizes from 100 to 300, the changes in the empirical coverage probability criterion are meaningful. This comes from the fact that in Theorem 2.1 of Chatterjee (2021) and also, in our proposed theorem, the distribution of tests statistics depend on the sample sizes. However, the results are very satisfactory. Where, the existing covariate in the general linear model enter into the model by the high probability. Also, the covariates that don't have a rule in the model are rejected by the high probability.

3.5 Variable selection with high-dimensional covariates

In this subsection, we examine a model similar to the one used in the previous study, which includes three covariates. To investigate the impact of the number of candidate covariates (k) on variable selection using the proposed algorithm, we extend the model to include 20 covariates. In model (1), it is assumed that $g(\mathbf{x}) = (x_1, x_2^2 + e^{(1-x_3)})$, $\beta = (2, 0.5)$, $\epsilon \sim N(0, \sigma^2)$ with known values of σ^2. Additionally, we simulate the covariates from the following distributions:

$$
\begin{aligned}
&X_1, X_2, X_3, X_4, X_5 \sim N(1, 5),\\
&X_6, X_7, X_8, X_9, X_{10} \sim Exp(0.5),\\
&X_{11}, X_{12}, X_{13}, X_{14}, X_{15} \sim U(-1, 1),\\
&X_{16}, X_{17}, X_{18}, X_{19}, X_{20} \sim Gamma(5, 1).
\end{aligned}
$$

In this case, Model (1) is equivalent to:

$$y = 2x_1 + 0.5(x_2^2 + e^{(1-x_3)}) + \epsilon. \tag{12}$$

In equation (12), the general linear model includes 3 covariates and is characterized by 2 parameters. This model allows us to evaluate how the proposed algorithm performs in a higher-dimensional setting and assess its sensitivity to the number of candidate predictors. The results of this study are presented in Table 5 for various values of n and σ^2.

The results of this study align closely with those of the previous study. The effect of sample size on variable selection remains evident. More importantly, increasing the number of covariates does not significantly impact the results. Additionally, a key finding is that covariates which do not contribute to the model (i.e., irrelevant variables) are rejected with high probability across different values of k. Moreover, in this study as well as in previous studies, the rejection of irrelevant covariates with high probability is independent of both the number of covariates and the sample size.

Table 4 Empirical coverage probability of entering any covariates into the general linear model when there are three covariates in the model with ten candidate covariates for different values of σ^2 and n based on the 5000 repetitions of the Monte Carlo algorithm

n	σ^2	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
100	0.01	0.7626	0.6234	0.6620	0.0048	0.0048	0.0052	0.0048	0.0058	0.0064	0.0058
	0.10	0.7568	0.6122	0.6642	0.0070	0.0060	0.0060	0.0066	0.0066	0.0066	0.0046
	0.25	0.7526	0.5982	0.6568	0.0056	0.0072	0.0040	0.0050	0.0040	0.0038	0.0046
	0.50	0.7580	0.6086	0.6674	0.0046	0.0056	0.0056	0.0044	0.0036	0.0036	0.0044
	1.00	0.7240	0.5928	0.6560	0.0036	0.0042	0.0052	0.0036	0.0064	0.0060	0.0040
200	0.01	0.9772	0.9310	0.9488	0.0068	0.0060	0.0052	0.0068	0.0054	0.0044	0.0084
	0.10	0.9752	0.9232	0.9502	0.0066	0.0090	0.0052	0.0052	0.0058	0.0058	0.0066
	0.25	0.9778	0.9326	0.9526	0.0072	0.0058	0.0052	0.0074	0.0074	0.0052	0.0062
	0.50	0.9768	0.9368	0.9472	0.0086	0.0070	0.0086	0.0050	0.0070	0.0052	0.0078
	1.00	0.9682	0.9238	0.9460	0.0060	0.0058	0.0054	0.0056	0.0054	0.0082	0.0072
300	0.01	0.9990	0.9906	0.9956	0.0068	0.0062	0.0058	0.0064	0.0064	0.0078	0.0062
	0.10	0.9986	0.9916	0.9942	0.0060	0.0086	0.0080	0.0076	0.0072	0.0052	0.0068
	0.25	0.9990	0.9898	0.9948	0.0082	0.0080	0.0072	0.0072	0.0060	0.0048	0.0050
	0.50	0.9986	0.9894	0.9956	0.0072	0.0072	0.0090	0.0070	0.0066	0.0074	0.0058
	1.00	0.9978	0.9890	0.9942	0.0076	0.0082	0.0078	0.0078	0.0058	0.0072	0.0064

Table 5 Empirical coverage probability of entering any covariates into the general linear model when there are three covariates in the model with 20 candidate covariates for different values of σ^2 and n based on the 5000 repetitions of the Monte Carlo algorithm

n	σ^2	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
100	0.01	0.9216	0.2518	0.4802	0.0016	0.0030	0.0032	0.0030	0.0018	0.0026	0.0022
	0.10	0.9208	0.2428	0.4684	0.0026	0.0012	0.0024	0.0026	0.0024	0.0022	0.0022
	0.25	0.9124	0.2392	0.4622	0.0028	0.0022	0.0018	0.0026	0.0028	0.0020	0.0016
	0.50	0.9110	0.2370	0.4752	0.0020	0.0014	0.0020	0.0014	0.0018	0.0028	0.0020
	1.00	0.8954	0.2232	0.4742	0.0018	0.0026	0.0024	0.0034	0.0014	0.0012	0.0018
200	0.01	0.9984	0.5912	0.8514	0.0024	0.0030	0.0024	0.0026	0.0032	0.0030	0.0020
	0.10	0.9980	0.6044	0.8594	0.0032	0.0026	0.0026	0.0024	0.0010	0.0020	0.0022
	0.25	0.9986	0.6016	0.8482	0.0026	0.0044	0.0028	0.0032	0.0030	0.0026	0.0022
	0.50	0.9976	0.5836	0.8560	0.0026	0.0026	0.0012	0.0030	0.0014	0.0028	0.0028
	1.00	0.9992	0.5588	0.8438	0.0026	0.0024	0.0020	0.0024	0.0020	0.0018	0.0030
300	0.01	1.0000	0.8250	0.9700	0.0022	0.0026	0.0030	0.0034	0.0018	0.0030	0.0020
	0.10	1.0000	0.8388	0.9696	0.0026	0.0024	0.0028	0.0028	0.0020	0.0032	0.0022
	0.25	1.0000	0.8284	0.9730	0.0032	0.0024	0.0018	0.0038	0.0018	0.0032	0.0032
	0.50	1.0000	0.8232	0.9734	0.0020	0.0016	0.0028	0.0022	0.0020	0.0032	0.0040
	1.00	1.0000	0.8126	0.9666	0.0016	0.0026	0.0012	0.0018	0.0016	0.0030	0.0036

Table 5 continued

n	σ^2	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
n	σ^2	X_{11}	X_{12}	X_{13}	X_{14}	X_{15}	X_{16}	X_{17}	X_{18}	X_{19}	X_{20}
100		0.0022	0.0024	0.0028	0.0020	0.0016	0.0022	0.0012	0.0028	0.0022	0.0026
	0.10	0.0018	0.0022	0.0012	0.0036	0.0018	0.0026	0.0022	0.0018	0.0020	0.0022
	0.25	0.0022	0.0026	0.0028	0.0018	0.0018	0.0022	0.0030	0.0026	0.0022	0.0022
	0.50	0.0028	0.0016	0.0026	0.0014	0.0028	0.0030	0.0016	0.0016	0.0014	0.0018
	1.00	0.0018	0.0032	0.0022	0.0028	0.0018	0.0024	0.0020	0.0014	0.0026	0.0016
200	0.01	0.0024	0.0020	0.0030	0.0038	0.0036	0.0036	0.0034	0.0010	0.0036	0.0044
	0.10	0.0022	0.0028	0.0026	0.0034	0.0036	0.0028	0.0028	0.0034	0.0032	0.0030
	0.25	0.0018	0.0034	0.0024	0.0034	0.0022	0.0030	0.0036	0.0028	0.0024	0.0032
	0.50	0.0028	0.0032	0.0022	0.0030	0.0022	0.0024	0.0036	0.0030	0.0020	0.0040
	1.00	0.0026	0.0014	0.0024	0.0022	0.0038	0.0020	0.0028	0.0018	0.0040	0.0024
300	0.01	0.0016	0.0028	0.0028	0.0022	0.0032	0.0034	0.0020	0.0030	0.0028	0.0028
	0.10	0.0022	0.0024	0.0026	0.0024	0.0036	0.0030	0.0022	0.0036	0.0028	0.0020
	0.25	0.0014	0.0038	0.0024	0.0026	0.0038	0.0032	0.0030	0.0028	0.0032	0.0014
	0.50	0.0028	0.0042	0.0028	0.0020	0.0018	0.0026	0.0028	0.0036	0.0010	0.0024
	1.00	0.0026	0.0036	0.0036	0.0032	0.0012	0.0036	0.0022	0.0022	0.0032	0.0024

4 Real data studies

In this section, we analyze two real datasets. In the first study, we apply our method to the"Academic Achievements" dataset, and in the second study, we apply it to the"Wavelength Reflection"dataset. For both studies, we begin by selecting the appropriate covariates for the model using the proposed algorithm. Next, we fit several general linear models and identify the optimal model based on the BIC. The final model is selected using the BIC criterion, as discussed in Sect. 2.

4.1 Study1: academic achievements

Consider the data set,"Effects of Students Background on Academic Achievements"used by the UCLA group to introduce structural equation modeling. This data set and more details about the data are given on their site by address "https://stats.oarc.ucla.edu/r/seminars/rsem/". Moreover, the direct link to access to the data is "https://stats.idre.ucla.edu/wp-content/uploads/2021/02/worland5.csv". This data set contains observations of 9 continuous variables with no ties. Also, the sample size is $n = 500$. The variables are motiv(Y):Motivation; harm(X_1): Harmony; stabi(X_2): Stability; ppsych(X_3):Negative Parental Psychology; ses(X_4): Socioeconomic Status; verbal(X_5): Verbal IQ; read(X_6): Reading; arith(X_7): Arithmetic and spell(X_8): Spelling. As determined in the definition of variables, we consider the motiv variable as the response variable, and the other variables are considered as the candidate covariates that can enter into the general linear model.

In the first step, the appropriate covariates are selected by the proposed algorithm in level $\alpha = 0.05$. The significant levels of any covariates for entering into the model are 0.0000, 5.1026×10^{-10}, 0.2672, 0.8984, 0.0469, 1.4647×10^{-8}, 0.0000 and 7.2875×10^{-13}, respectively. Therefore, by the proposed algorithm, one can conclude that X_1, X_2, X_6, X_7, and X_8 are related to the response variable but the variables X_3, X_4, and X_5 are not related to the response variable. We assume the following models as the assumed class and then, we choose the final optimal general linear model in this class.

$$
\begin{aligned}
&\text{model 1: } y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_6 + \beta_4 x_7 + \beta_5 x_8.\\
&\text{model 2: } y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_6 + \beta_4 x_7 + \beta_5 x_8.\\
&\text{model 3: } y = \beta_0 + \beta_1 (x_1 + x_2) + \beta_2 (x_6 + x_7 + x_8).\\
&\text{model 4: } y = \beta_1 (x_1 + x_2) + \beta_2 (x_6 + x_7 + x_8).\\
&\text{model 5: } y = \beta_1 (x_1 + x_2)^2 + \beta_2 (x_6 + x_7 + x_8)^2.\\
&\text{model 6: } y = \beta_1 (x_1 + x_2 + x_6 + x_7 + x_8).\\
&\text{model 7: } y = \beta_0 + \beta_1 (x_1 + x_2 + x_6 + x_7 + x_8).\\
&\text{model 8: } y = \beta_1 e^{(x_1 + x_2)} + \beta_2 (x_1 + x_2 + x_6 + x_7 + x_8).\\
&\text{model 9: } y = \beta_1 \tfrac{x_1 + x_2}{x_6 + x_7 + x_8}.\\
&\text{model 10: } y = \beta_1 (x_1 + x_2) \times (x_6 + x_7 + x_8).
\end{aligned}
$$

To select the final general linear model, we use the BIC criterion. The BIC of different models are 1719.308, 1725.522, 1762.158, 1768.587, 2309.688, 1842.385, 1848.600, 1848.456, 2307.274 and 2306.358, respectively. Based on the BIC criterion, one can choose the first model as the final model among the assumed class of the

general linear models. Therefore, the final model for data will be as follows:

$$y = 0.5256x_1 + 0.1518x_2 - 0.1080x_6 + 0.2269x_7 + 0.2264x_8. \tag{13}$$

In the final model, the covariates X_3, X_4 and X_5 are omitted from inferences by the proposed algorithm.

4.2 Study 2: wavelength reflection

In agriculture, waves are sent into the soil to measure the concentration of soil minerals. In this study, we used data collected from Ardabil County, Iran, by researchers at the University of Mohaghegh Ardabili. Waves with wavelengths ranging from 580 to 680 nm, at intervals of 5 nm, were transmitted to the soil at various locations in Ardabil. The reflection rates of these waves were then recorded. Additionally, the pH of the soil was measured at each location.

The objective of this study is to construct a general linear model for soil pH (the response variable) using the reflection rates of 21 different wavelengths as covariates. However, preliminary analysis revealed that the covariates are highly correlated, leading to a multicollinearity issue. To address this problem, we applied Principal Component Regression (PCR), which avoids multicollinearity by transforming the covariates into uncorrelated principal components.

In the initial analysis, we observed that the cumulative proportion of variance explained by the principal components reached approximately 1 by the 9th component. Therefore, we used the first 9 principal components to transform the covariates. As a result, the final dataset consists of one response variable (soil pH) and 9 rotated covariates derived from the original 21 wavelength reflection rates.

By applying our proposed algorithm at level of $\alpha = 0.05$, the p-values for entering covariates into the model were 0.0132, 0.1175, 0.9460, 0.1479, 0.8656, 0.04868, 0.07560, 0.8062, and 0.6784, respectively. Therefore, by our algorithm, no covariates were included in the final model. This result aligns with the findings of Ben-Dor and Banin (1995), who demonstrated that soil pH does not emit or reflect specific wavelengths. This results suggesting that wavelength-based measurements may not be suitable for predicting soil pH directly. However, when we reduce the confidence level of the tests in the algorithm to 88% ($\alpha = 0.12$), the first transformed covariate enters the model. This result suggests that, at a lower confidence level, the relationship between the first principal component (pc_1) and soil pH becomes statistically significant. We assume the following models as the assumed class and then, we choose the final optimal general linear model in this class.

model 1: $y = \beta_0 + \beta_1 pc_1$.
model 2: $y = \beta_1 pc_1$.
model 3: $y = \beta_0 + \beta_1 pc_1^2$.
model 4: $y = \beta_1 pc_1^2$.
model 5: $y = \beta_1 pc_1^3$.
model 6: $y = \beta_0 + \beta_1 pc_1 + \beta_2 pc_1^2$.

model 7: $y = \beta_1 pc_1 + \beta_2 pc_1^2$.
model 8: $y = \beta_1 pc_1 + \beta_2 pc_1^2 + \beta_3 pc_1^3$.
model 9: $y = \beta_0 + \beta_1 e^{pc_1}$.
model 10: $y = e^{\beta_0 + \beta_1 pc_1}$.

The BIC of different models are -134.3546, -17.7761, -147.3464, 84.6675, 131.6210, -152.0495, -92.7720, -141.5162, -112.2814, and -336.8436, respectively. Based on the BIC criterion, one can choose the 10th model as the final model among the assumed class of the general linear models. Therefore, the final model for data will be as follows:

$$y = e^{1.1817 - 0.1160 pc_1}. \tag{14}$$

5 Discussion

In our studies, we have considered that all the variables are continuous and there are no ties among the observations in any variables. However, one can consider some more complex cases. In general, The empirical distribution of $\zeta_n(\cdot, \cdot)$ is Normal yet, but its variance is different. Considering the more complex cases may be our next study.

In real data study, we first choose the appropriate covariates based on the proposed algorithm. Then, we picked the best model among the assumed class of general linear models. It is necessary to emphasize that selected covariates are unique by the algorithm. But, the fitted model may changes by changing the class of candidate general linear models and used criteria. Therefore, fitting an appropriate general linear model depends on the experiences of researchers about the experiments and data.

Remark that we used the empirical distribution of Chatterjee's correlation to select the appropriate covariates. This criterion extremely depends on the sample sizes, especially when the number of candidate covariates is high. Therefore, in using our proposed algorithm, determining the appropriate sample size is essential. However, in simulation studies, we have observed that irrelevant covariates associated with the response variable are removed from the model with a high probability, even in cases with low sample sizes.

Data Availability Dataset 1 is directly linked within the manuscript text. Dataset 2 is available via the following Google Drive link: https://drive.google.com/file/d/13IbTXtbvRtO76nBmubZzmBIz3NLq3afg/view.

Declarations

Conflict of interest We declare that there are no Conflict of interest in this article.

References

Azadkia, M., & Chatterjee, S. (2021). A simple measure of conditional dependence. *Ann. Stat., 49*(6), 3070–3102.

Bahari, F., Parsi, S., & Ganjali, M. (2021). Goodness of fit test for general linear model with nonignorable missing on response variable. *Adv. Stat. Anal., 105*, 163–196.

Ben-Dor, E., & Banin, A. (1995). Near-infrared analysis as a rapid method to simultaneously evaluate several soil properties. *SSSAJ, 59*(2), 364–372.

Chatterjee,S. (2024). A survey of some recent developments in measures of association, Probab. Stoch. Proc.: A Volume in Honour of Rajeeva L. Karandikar , 109–128.

Chatterjee, S. (2021). A new coefficient of correlation. *J. Am. Stat. Assoc., 116*(536), 2009–2022.

Creemers, A., Aerts, M., Hens, H., & Molenbergh, G. (2011). A nonparametric approach to weighted estimating equations for regression analysis with missing covariates. *Comput. Stat. Data. Anal., 56*, 100–113.

Efroymson,M.A. (1960), Multiple regression analysis, Mathematical methods for digital computers.

Fuchs, S., & Wang, Y. (2024). Hierarchical variable clustering based on the predictive strength between random vectors. *Int. J. Approx. Reason., 170*, Article 109185.

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *J. Mach. Learn. Res., 3*, 1157–1182.

Hardle, W., Mammen, E., & Muller, M. (1998). Testing parametric versus semiparametric modeling in generalized linear models. *J. Am. Stat. Assoc., 93*(444), 1461–1474.

Holm, S. (1979). A simple sequentially rejective multiple test procedure, Scand. J. Stat, pp. 65–70.

Kutner, H. M., Nachtsheim, J. C. J., Neter, J., & William, L. (2005). *Applied linear statistical models*. USA: McGraw-Hill.

Lin, Z., & Han, F. (2023). On boosting the power of Chatterjee's rank correlation. *Biometrika., 110*(2), 283–299.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2021). *Introduction to linear regression analysis*. John Wiley & Sons.

Pei, J., Zhu, F., & Li, Q. (2024). Diagnostic checks in time series models based on a new correlation coefficient of residuals. *J. Appl. Stat., 51*(12), 2402–2419.

Shi, H., Drton, M., & Han, F. (2022). On the power of Chatterjee's rank correlation. *Biometrika, 109*(2), 317–333.

Thottolil, R., Kumar, U., & Chakraborty, T. (2023). Prediction of transportation index for urban patterns in small and medium-sized Indian cities using hybrid RidgeGAN model. *Sci. Rep., 13*(1), 21863.

Zhu, L., & Chu, H. (2005) Testing the adequacy for a general linear errors-in-variables model, Stat. Sin. 1049–1068.