MethodsX 14 (2025) 103186

Contents lists available at ScienceDirect

MethodsX

journal homepage: www.elsevier.com/locate/methodsx

Risk factor analysis for stunting incidence using sparse categorical principal component logistic regression ☆,☆☆

Anna Islamiyati [a,*], Muhammad Nur [b], Abdul Salam [c], Wan Zuki Azman Wan Muhamad [d], Dwi Auliyah [a]

[a] Department of Statistics, Faculty of Mathematical and Natural Sciences, Hasanuddin University, Makassar 90245, Indonesia
[b] Department of Mathematics, Faculty of Mathematical and Natural Sciences, Hasanuddin University, Makassar 90245, Indonesia
[c] Department of Nutrition, Faculty of Public Health, Hasanuddin University, Makassar 90245, Indonesia
[d] Institute of Engineering Mathematics, Universiti Malaysia Perlis, Arau 02600, Malaysia

ARTICLE INFO

Method name:
Sparse Categorical Principal Component Logistic Regression

Keywords:
Binary logistic
Categorical predictor
Multicollinearity
Sparse Categorical PCA

ABSTRACT

The risk factors for stunting incidence involve categorical data in both the response and predictor variables. Therefore, we developed a sparse categorical principal component logistic regression model capable of handling data with multicollinearity. The parameters of the sparse categorical principal component logistic regression model were estimated using the maximum likelihood method and the Newton-Raphson iterative approach. The analysis yielded a likelihood ratio value of 144.81 and a chi-square statistic value of 11.07, indicating that all factors included in the model are statistically significant. The results highlight that medical history, inadequate complementary feeding, formula feeding, lack of complementary feeding programs, and lack of iron supplementation for mothers are highly associated with the risk of stunting in toddlers. This emphasizes the need for attention to maternal nutrition from pregnancy through breastfeeding, as well as the nutrition of the toddler. Some important points proposed in this method are:

- Stunting data consists of categorical variables containing multicollinearity.
- The method applied is sparse logistic regression combined with categorical principal component analysis.
- Analysis of risk factors for stunting in toddlers is based on the child's own condition, as well as parental factors, namely age, education, and intake of additional food and supplementary tablets during pregnancy.

Specifications table

Subject area:	Mathematics and Statistics
More specific subject area:	Statistics modelling
Name of your method:	Sparse Categorical Principal Component Logistic Regression
Name and reference of original method:	None
Resource availability:	Software R

☆ **Related research article:** Detecting Age Prone to Growth Retardation in Children Through a Bi-Response Nonparametric Regression Model with a Penalized Spline Estimator.

☆☆ **For a published article:** A. Islamiyati, A. Kalondeng, M. Zakir, S. Djibe, U. Sari, Original Article, Iranian Journal of Nursing and Midwifery Research 29, no. 5 (2024): 549–554, https://doi.org/10.4103/ijnmr.ijnmr_342_22.

* Corresponding author.
E-mail address: annaislamiyati@unhas.ac.id (A. Islamiyati).

https://doi.org/10.1016/j.mex.2025.103186
Received 6 December 2024; Accepted 26 January 2025
Available online 27 January 2025

Background

Stunting is a chronic growth disorder in children caused by prolonged malnutrition and frequent infections, particularly during the critical 1000 days of life. This condition is characterized by a height below the age-standard reference, reflecting the accumulated negative effects of nutrient deficiencies and repeated infections, such as diarrhea and respiratory tract infections, which can hinder nutrient absorption [1–2]. Limited access to economic resources, low maternal education levels, and poor access to quality healthcare services further exacerbate the risk of stunting in children [3]. In addition to internal factors, environmental conditions such as clean water and adequate sanitation are crucial, as unhealthy environments can serve as sources of infections that impair child development [4]. Various studies emphasize that improving maternal education and providing adequate nutritional interventions during pregnancy and childhood can significantly reduce stunting rates, with environmental improvements being a key preventive step [5–6].

Many researchers have studied stunting data, for example, a study of variations in weight and height growth in children in rural South Africa using the SITAR model [6]. Analysis of stunting by age in the South Asian Region [7] and detection of child growth retardation in Indonesia using penalized spline regression [8]. However, stunting data often involves categorical data on response and predictor variables, requiring a more precise statistical approach to analyzing it. Logistic regression is a commonly used statistical method to predict categorical response variables based on one or more predictor variables. Specifically, binary logistic regression is frequently used to analyze the relationship between predictors and a dichotomous response or a categorical variable with two levels [9–10]. In contrast, ordinal logistic regression is used for data with ordered categories [11]. One of the critical assumptions in logistic regression is the absence of multicollinearity among predictors, as multicollinearity can distort the accuracy of the regression coefficients, leading to incorrect interpretations, particularly with odds ratios [12]. This issue frequently arises in high-dimensional datasets, where strong correlations between predictors are challenging to avoid and can affect the stability of the model [13]. Principal component analysis has long been used as a dimensionality reduction technique to address multicollinearity, transforming the predictor variables into uncorrelated principal components [14]. With principal component analysis, high multicollinearity data can be converted into principal components free of correlation and subsequently used as predictors in logistic regression.

In high-dimensional datasets, such as those used in genomics or pattern recognition applications, multicollinearity can significantly impact the interpretation and stability of the model. Classic principal component analysis has limitations, as the principal components produced are linear combinations of all variables, making the component loadings difficult to interpret [15]. One approach to address this issue is sparse principal component analysis, which introduces a sparsity penalty to reduce dimensionality and select relevant features for logistic regression models. Sparse principal component analysis offers an alternative by combining dimensionality reduction and sparsity, ensuring that only significant variables are included in each principal component while diminishing the weights of less relevant variables to zero [16]. Sparse principal component analysis has proven to be more effective for high-dimensional data modeling, as it produces principal components that are more interpretable for categorical responses, facilitating better insights into the underlying patterns in the data [17].

As the need for models capable of handling categorical predictors increases, such as in the case of risk factors for stunting, one promising approach is categorical principal component analysis, introduced by Kemalbay and Korkmazoglu [18]. Given that sparse principal component analysis can enhance the interpretability and variable selection in binary logistic regression models [19], developing a sparse principal component analysis logistic regression model for categorical predictors offers a valuable solution for analyzing risk factors for stunting. This method leverages the sparsity constraint to select significant variables while reducing dimensionality, thus improving the stability and interpretability of the model when applied to high-dimensional categorical data. This approach could provide more accurate insights into the factors contributing to stunting by focusing on the most relevant predictors.

Method details

Categorical principal component analysis

One of the violations of assumptions that often occurs in data is the occurrence of multicollinearity, namely a high correlation between predictors. In such cases, principal component analysis is an old statistical technique that can be used by linearly transforming the original set of variables into a smaller set of uncorrelated variables, which can still represent the information of the original set of variables. These new variables are called principal components. Let $\boldsymbol{R}$ represent the correlation matrix of predictor variables $\boldsymbol{X} = [\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_p]$, and the pairs of eigenvalues and eigenvectors are $(\lambda_1, \boldsymbol{v}_1), (\lambda_2, \boldsymbol{v}_2), \ldots, (\lambda_p, \boldsymbol{v}_p)$ where p is the number of predictor variables and $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$. The following Eq. (1) can define the principal components formed as linear combinations.

$$\boldsymbol{Z}_r = \boldsymbol{v'}_r \boldsymbol{r} \boldsymbol{X} = v_{r1}\boldsymbol{X}_1 + v_{r2}\boldsymbol{X}_2 + \ldots + v_{rp}\boldsymbol{X}_p,\ r = 1, 2, \ldots,\ p \tag{1}$$

Eq. (1) can be expressed in matrix form as in Eq. (2).

$$\boldsymbol{v'}_r \boldsymbol{X} = \begin{bmatrix} v_{11} & v_{21} & \cdots & v_{p1} \\ v_{12} & v_{22} & \cdots & v_{p2} \\ \vdots & \vdots & \ddots & \vdots \\ v_{1p} & v_{2p} & \cdots & v_{pp} \end{bmatrix} \begin{bmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \\ \vdots \\ \boldsymbol{X}_p \end{bmatrix} \tag{2}$$

where $\boldsymbol{Z}_r$ represents the r-th principal component, r is the index for the principal components from 1 to p with $r \leq p$. The criterion for selecting principal components is based on the cumulative proportion of the total variance that can be explained by the selected components, with a minimum threshold of 80% explained variance [20].

Next, categorical principal component analysis was developed, a method for dealing with multicollinearity in categorical data by applying optimal scaling, which converts categorical labels into numeric values while maximizing the variance between variables [21]. There are n individuals with p predictor variables, where each variable is defined by $\boldsymbol{X}_j$ for $j = 1, 2, \ldots, p$. If $\boldsymbol{X}_j$ is a nominal or ordinal scale, then the optimal linear transformation is observed for each score by converting them into categorical quantifications. The variable $\boldsymbol{X}_j$ resulting from quantification is denoted by $\boldsymbol{X}^*$.

Let $\boldsymbol{C}_j$ be a matrix of size $v_j \; x \; 1$, where v_j is the number of categories for each variable $\boldsymbol{X}_j$, and the values of $\boldsymbol{C}_j$ are consecutive integers. To find the solution for $\boldsymbol{X}_j$ and the values of $\boldsymbol{C}_j$, the problem can be formulated by minimizing the function in Eq. (3) through the alternating least square algorithm. It includes the least squares approximation of each parameter by updating each parameter matrix [22].

$$\sigma(\bar{\boldsymbol{X}}; \boldsymbol{C}_j) = \frac{1}{n} \sum_{j=1}^{p} \frac{1}{j} \left\{ tr\left(\bar{\boldsymbol{X}} - \boldsymbol{G}_j \boldsymbol{C}_j\right)^T \left(\bar{\boldsymbol{X}} - \boldsymbol{G}_j \boldsymbol{C}_j\right) \right\} \tag{3}$$

where $\boldsymbol{G}_j$ is the indicator matrix of $\boldsymbol{X}_j$ which has a value of 1 if the value of $\boldsymbol{X}_j$ is included in the category c_j, and a value of 0 if the value of $\boldsymbol{X}_j$ is not included in the category c_j. The optimal value of $\boldsymbol{C}_j$ resulting from quantification can be denoted by $\boldsymbol{C}_j^*$. Next, the optimal quantification process is obtained through Eq. (4), namely:

$$\left(\boldsymbol{G}_j^T \boldsymbol{G}_j\right)^{-1} \boldsymbol{G}_j^T \bar{\boldsymbol{X}}^{(t-1)} = \boldsymbol{C}_j^{(t)} \tag{4}$$

Where t is the iteration. The value of $\boldsymbol{C}_j^{(t)}$ at $t = 1$ is obtained based on $\bar{\boldsymbol{X}}^{(0)}$. The variable $\bar{\boldsymbol{X}}^{(0)}$ is the average of $\boldsymbol{X}_j^{(0)}$, $j = 1, 2, \ldots, p$. After that, we determine the value of $\boldsymbol{X}_j^{(t)}$ based on Eq. (5).

$$\boldsymbol{X}_j^{(t)} = \boldsymbol{G}_j \boldsymbol{C}_j^{(t)} \tag{5}$$

The iteration process is stopped when $\sigma(\bar{\boldsymbol{X}}^{(t-2)}; \boldsymbol{C}_j^{(t-1)}) - \sigma(\bar{\boldsymbol{X}}^{(t-1)}; \boldsymbol{C}_j^{(t)}) < \epsilon$, $\epsilon = 10^{-6}$. The value of $\boldsymbol{C}_j^*$ and $\boldsymbol{X}^*$ are obtained when the iteration process is stopped.

In categorical principal component analysis, the correlation matrix $\boldsymbol{R}$ is computed among the quantified variables $\boldsymbol{X}^*$ as in Eq. (6).

$$\boldsymbol{R} = \frac{1}{n-1} \boldsymbol{X}^{*T} \boldsymbol{X}^* \tag{6}$$

with $\lambda_j = \lambda_1, \lambda_2, \ldots, \lambda_p$ as the eigenvalues of the correlation matrix $\boldsymbol{R}$ and $\boldsymbol{a}_j = (a_1, a_2, \ldots, a_p)^T$ as the eigenvector corresponding to the eigenvalue λ_j. If the eigenvalues λ_j are ordered from the largest to the smallest, then the ordering of the principal components corresponds to λ_j. This study uses a minimum cumulative variance proportion of 80% for forming categorical principal components. If the principal components for the quantified predictor variables are expressed as $\boldsymbol{Z}_k$ formed are $k = 1, 2, \ldots, r$, then $\boldsymbol{Z}_k$ can be expressed by the following Eq. (7).

$$\boldsymbol{Z}_k = \boldsymbol{a'}_k \boldsymbol{X}^* = a_{k(1)} \boldsymbol{X}_1^* + a_{k(2)} \boldsymbol{X}_2^* + \ldots + a_{k(p)} \boldsymbol{X}_p^*, \; k = 1, 2, \ldots, r \tag{7}$$

Sparse principal component analysis

Sparse principal component analysis is an extension of the principal component analysis method that can eliminate ineffective variables from the principal components by reducing the loading values to zero [16]. The sparse principal component analysis uses the elastic net, developed from the least absolute shrinkage and selection operator method, to produce modified principal components from sparse loading values by combining L_1-norm and squared L_2-norm constraints on the elastic net estimator δ [23]. The L_1-Norm constraint can result in a simpler model by shrinking some δ values to exactly zero, while the squared L_2-Norm constraint generates a model that does not select variables but enhances the grouping and shrinking effects of δ [24].

The sparse principal component analysis algorithm for reducing data dimensionality using the naive elastic net estimator is formulated with the following steps [25]:

1. Let $\boldsymbol{A} = [\boldsymbol{A}_1, \ldots, \boldsymbol{A}_r]$, representing the loading values of each principal component.
2. The naive elastic net estimator for $k = 1, 2, \ldots, r$ is calculated by the following Eq. (8).

$$\widehat{\boldsymbol{\delta}_k} = \arg\min_{\boldsymbol{\delta}_k} \left\{ \left(\boldsymbol{A}_k - \boldsymbol{\delta}_k\right)^T \boldsymbol{X}^{*T} \boldsymbol{X}^* \left(\boldsymbol{A}_k - \boldsymbol{\delta}_k\right) + \ell_2 \|\boldsymbol{\delta}_k\|_2^2 + \ell_1 \|\boldsymbol{\delta}_k\|_1 \right\} \tag{8}$$

where $\|\boldsymbol{\delta}_k\|_2^2 = \sum_{j=1}^{p} \delta_{jk}^2$, $\|\boldsymbol{\delta}_k\|_1 = \sum_{j=1}^{p} |\delta_{jk}|$, ℓ_1 and ℓ_2 are positive numbers, tuning parameters determined by considering the parsimony principle.

3. For each δ obtained from step 2, calculate the SVD of $\boldsymbol{X}^{*T} \boldsymbol{X}^* \boldsymbol{B}$ with $\boldsymbol{X}^{*T} \boldsymbol{X}^* \boldsymbol{B} = \boldsymbol{U} \boldsymbol{D} \boldsymbol{V}^T$, to obtain $\boldsymbol{A} = \boldsymbol{U} \boldsymbol{V}^T$.
4. Repeat steps 2–3 until δ converges.
5. Perform normalization: $\hat{v}_k = \frac{\hat{\delta}_k}{\|\hat{\delta}_k\|}; k = 1, \ldots, r$.

Next, the optimal sparse principal component $\boldsymbol{Z}_k^*$ can be obtained based on the values of $\hat{v}_k$ which is the result of normalizing the values of $\boldsymbol{\delta}_k$ by choosing the values of ℓ_1 and ℓ_2 to obtain a simple model with a minimum cumulative variance proportion of 80%. The sparse categorical principal component is shown in Eq. (9).

$$\boldsymbol{Z}_k^* = \hat{v}_k \boldsymbol{X}^* = \hat{v}_{k(1)} \boldsymbol{X}_1^* + \hat{v}_{k(2)} \boldsymbol{X}_2^* + \ldots + \hat{v}_{k(p)} \boldsymbol{X}_p^* \tag{9}$$

Principal component logistic analysis

Principal component logistic regression aims to improve parameter estimation in logistic regression models with multicollinearity among predictor variables by using the principal components of these predictors. The equation for the principal component logistic regression model is shown in Eq. (10) [25].

$$\pi(Z) = \frac{\exp\{\beta_0 + \sum_{k=1}^{r}\sum_{j=1}^{p} Z_k v_{jk}\beta_j\}}{1+\exp\{\beta_0 + \sum_{k=1}^{r}\sum_{j=1}^{p} Z_k v_{jk}\beta_j\}} = \frac{\exp\{\beta_0 + \sum_{k=1}^{r} Z_k\gamma_k\}}{1+\exp\{\beta_0 + \sum_{k=1}^{r} Z_k\gamma_k\}} \tag{10}$$

where $\pi(Z)$ is the probability of success and $\beta_0,\ \gamma_1,\ \ldots,\ \gamma_r$ are the regression parameters. The principal component logistic regression model can be expressed in terms of the linear relationship between log odds and predictor variables through the logit transformation $g = \ln(\frac{\pi(Z)}{1-\pi(Z)})$. The model is shown in Eq. (11).

$$g = \beta_0 + \sum_{k=1}^{r} \boldsymbol{Z_k}\gamma_k + \varepsilon \tag{11}$$

where g is the probability of a successful outcome when $Y = 1$, and γ_k is the logistic regression coefficient based on the principal components formed.

Parameter estimation of the sparse categorical principal component logistic regression model

The model obtained for sparse categorical principal component logistic regression analysis is shown in Eq. (12).

$$g = \beta_0 + \sum_{k=1}^{r} \boldsymbol{Z_k^*}\gamma_k + \varepsilon \tag{12}$$

where g represents the probability of success for $Y = 1$, β_0 is the intercept or the value of g when all $\boldsymbol{Z_k^*}$ are zero, $\boldsymbol{Z_k^*}$ is the optimal sparse principal component, and $\boldsymbol{\gamma_k}$ is the regression coefficient that measures the effect of $\boldsymbol{Z_k^*}$ on the probability of g. The parameters $(\beta_0, \gamma_1, \ldots, \gamma_r)$ are estimated using the maximum likelihood estimation method by maximizing the likelihood function in Eq. (13).

$$f\left((\beta_0,\gamma_1,\gamma_2,\ldots,\gamma_r);Y_i\right) = \prod_{i=1}^{n} f(Y_i) = \prod_{i=1}^{n}\left(\frac{\pi(Z_i^*)}{1-\pi(Z_i^*)}\right)^{Y_i}\left(1-\pi(Z_i^*)\right) \tag{13}$$

Therefore, we obtain the log-likelihood function as in Eq. (14).

$$\begin{aligned}\ln L\left((\beta_0,\gamma_1,\gamma_2,\ldots,\gamma_r);Y_i\right) &= ln\prod_{i=1}^{n}\left(\frac{\pi(Z_i^*)}{1-\pi(Z_i^*)}\right)^{Y_i}\left(\frac{\pi(Z_i^*)}{1-\pi(Z_i^*)}\right)\Bigg)\\ &= \sum_{i=1}^{n}\left[Y_i\left(\beta_0+\sum_{k=1}^{r}Z_k^*\gamma_k\right) - ln\left(1+exp\left(\beta_0+\sum_{k=1}^{r}Z_k^*\gamma_k\right)\right)\right]\end{aligned} \tag{14}$$

Next, Eq. (14) is differentiated for its parameters, shown in Eq. (15-17).

$$\frac{\partial ln\left((\beta_0,\gamma_1,\gamma_2,\ldots,\gamma_r);Y_i\right)}{\partial\beta_0} = \sum_{i=1}^{n}\left[Y_i - \frac{exp\left(\beta_0+\sum_{k=1}^{r}Z_k^*\gamma_k\right)}{1+exp\left(\beta_0+\sum_{k=1}^{r}Z_k^*\gamma_k\right)}\right] \tag{15}$$

$$\frac{\partial ln\left((\beta_0,\gamma_1,\gamma_2,\ldots,\gamma_r);Y_i\right)}{\partial\gamma_1} = \sum_{i=1}^{n}\left[Y_iZ_{i1}^* - Z_{i1}^*\frac{exp\left(\beta_0+\sum_{k=1}^{r}Z_k^*\gamma_k\right)}{1+exp\left(\beta_0+\sum_{k=1}^{r}Z_k^*\gamma_k\right)}\right] \tag{16}$$

$$\frac{\partial ln\left((\beta_0,\gamma_1,\gamma_2,\ldots,\gamma_r);Y_i\right)}{\partial\gamma_r} = \sum_{i=1}^{n}\left[Y_iZ_{ir}^* - Z_{ir}^*\frac{exp\left(\beta_0+\sum_{k=1}^{r}Z_k^*\gamma_k\right)}{1+exp\left(\beta_0+\sum_{k=1}^{r}Z_k^*\gamma_k\right)}\right] \tag{17}$$

Since the first derivatives of the log-likelihood function do not provide a solution, the Newton-Raphson iteration is used for parameter estimation $(\beta_0, \gamma_1, \ldots, \gamma_r)$. The Newton-Raphson method requires both the first and second derivatives of the log-likelihood function. The second derivative of the log-likelihood function for $(\beta_0, \gamma_1, \ldots, \gamma_r)$ can be seen in Eq. (18-20).

$$\frac{\partial^2 \ln L\left((\beta_0,\gamma_1,\ldots,\gamma_r);Y_i\right)}{\partial^2\beta_0} = -\sum_{i=1}^{n}\left[\frac{exp\left(\beta_0+\sum_{k=1}^{r}Z_k^*\gamma_k\right)}{1+exp\left(\beta_0+\sum_{k=1}^{r}Z_k^*\gamma_k\right)}\right] \tag{18}$$

$$\frac{\partial^2 \ln L\left((\beta_0,\gamma_1,\ldots,\gamma_r);Y_i\right)}{\partial\beta_0\partial\gamma_l} = -\sum_{i=1}^{n}\left[Z_{il}^*\frac{exp\left(\beta_0+\sum_{k=1}^{r}Z_k^*\gamma_k\right)}{\left(1+exp\left(\beta_0+\sum_{k=1}^{r}Z_k^*\gamma_k\right)\right)^2}\right] \tag{19}$$

$$\frac{\partial^2 \ln L\left((\beta_0,\gamma_1,\ldots,\gamma_r);Y_i\right)}{\partial\gamma_l\partial\gamma_k} = -\sum_{i=1}^{n}\left[Z_{il}^*Z_{ik}^*\frac{exp\left(\beta_0+\sum_{k=1}^{r}Z_k^*\gamma_k\right)}{\left(1+exp\left(\beta_0+\sum_{k=1}^{r}Z_k^*\gamma_k\right)\right)^2}\right] \tag{20}$$

for $l, k = 1, 2, \ldots, r$.

Table 1
Frequency distribution of stunting occurrence.

Stunting Status	Frequency	Percentage (%)
Non-stunted	251	84.8
Stunted	45	15.2
Total	296	100

After obtaining the first and second derivatives of the log-likelihood function concerning $(\beta_0, \gamma_1, \ldots, \gamma_r)$, the parameter estimates $(\hat{\beta}_0, \hat{\gamma}_1, \ldots, \hat{\gamma}_r)$ are generally obtained using the Newton-Raphson iteration as in Eq. (21).

$$\begin{bmatrix} \hat{\beta}_{0(t+1)} \\ \hat{\gamma}_{1(t+1)} \\ \vdots \\ \hat{\gamma}_{r(t+1)} \end{bmatrix} = \begin{bmatrix} \hat{\beta}_{0(t)} \\ \hat{\gamma}_{1(t)} \\ \vdots \\ \hat{\gamma}_{r(t)} \end{bmatrix} - \boldsymbol{D}_2{}^{-1}\boldsymbol{D}_1 \tag{21}$$

Where $\boldsymbol{D}_2$ is the matrix of second derivatives of the log-likelihood function concerning $(\beta_0, \gamma_1, \ldots, \gamma_r)$, and $\boldsymbol{D}_1$ is the matrix of first derivatives of the log-likelihood function concerning $(\beta_0, \gamma_1, \ldots, \gamma_r)$.

The estimated model for sparse categorical principal component logistic regression analysis can be seen in Eq. (22).

$$g = \hat{\beta}_0 + \boldsymbol{Z}_1^*\hat{\gamma}_1 + \boldsymbol{Z}_2^*\hat{\gamma}_2 + \ldots + \boldsymbol{Z}_3^*\hat{\gamma}_3 \tag{22}$$

In Eq. (22), the regressed variables are the optimal sparse principal component variables formed from the sparse categorical principal component analysis. Therefore, Eq. (22) needs to be transformed back into the original X variables so that the response variable Y can be clearly expressed in terms of the predictor variable X.

Method validation

Data description

This study uses data on stunting occurrences in Maginti District in 2023 from the West Muna District Health Office, Southeast Sulawesi Province, Indonesia. The categorical predictor variables in this study are risk factors that may lead to stunting. The outcome variable for stunting consists of 0 (non-stunted) and 1 (stunted). The predictor variables include gender, birth height, birth weight, history of illness, age at the introduction of solid food, formula milk consumption, food during pregnancy, mother received iron tablets, father and mother's age, father and mother's highest education level, and the drinking water consumed. This study uses R software in analyzing data.

Based on the data, the frequency distribution of stunting occurrences among the respondents is shown in Table 1. There were 296 respondents; 84.8% (251) were not identified as stunted, while 15.2% (45) were identified as stunted.

Table 2 shows that from 15.2% of respondents who were indicated as stunted, the incidence in girls and boys is the same, namely 7.8% and 7.4%. Furthermore, the high percentage of stunting in children who have low height and birth weight is 14.5% and 12.8%, respectively. Furthermore, 12.2% of children have a history of illness, 11.8% receive additional food that is not according to their age, and 11.5% consume formula milk. For the child's mother, 7.8% do not receive additional food, and 8.8% do not receive iron tablets. Furthermore, for the age of the parents, the highest percentage of children indicated as stunting have fathers who are over 40 years old, namely 6.1%. As for the mother's age, the highest is between the ages of 21 – 35 years, namely around 7.1%. For the last education of the parents, the highest percentage occurs in children with father and mother education at the high school level, namely 6.7% and 6.4%, respectively.

Risk factor analysis for stunting occurrence

In the data analysis process, the use of categorical principal component analysis begins with the formation of optimal quantification of each category of predictor variables. It is known that $\boldsymbol{X}^*$ is the variable of the optimal quantification result obtained based on the value of $\boldsymbol{C}_j^*$ from the iteration process as in Eq. (5). The values for each variable $\boldsymbol{X}^*$ are shown in Table 3.

Next, through categorical principal component analysis, five optimal principal components, $\boldsymbol{Z}_k$, $k = 1, 2, \ldots, 5$, were obtained, which collectively explain a cumulative variance percentage of 85.44%. The loading values $\boldsymbol{A}_{jk}$, which represent the correlation between $\boldsymbol{Z}_k$ and $\boldsymbol{X}_j^*$, are shown in Table 4.

Table 4 needs to be simplified using sparse principal component analysis so that each $\boldsymbol{Z}_k$ consists only of a few $\boldsymbol{X}_j^*$ variables that contribute significantly to the optimal principal component $\boldsymbol{Z}_k$. The initial analysis steps to obtain the sparse loading values $\boldsymbol{A}$ involve selecting the best model $\widehat{\delta_k}$ by choosing the values of ℓ_1 and ℓ_2 to achieve a simple model with a cumulative variance proportion of at least 80%. Therefore, in this study, ℓ_1=0.009 and ℓ_2=0.0001 were used. This is because both values have provided a cumulative variance proportion that has exceeded 80% which is 81.8%. Next, the resulting sparse loading values are as in Table 5.

The obtained optimal sparse loading values are simpler because each $\boldsymbol{Z}_k^*$ consists only of a few $\boldsymbol{X}_j^*$ variables that are considered to contribute significantly to the formation of $\boldsymbol{Z}_k^*$, while $\boldsymbol{X}_j^*$ variables that are deemed not to contribute significantly are ignored, as indicated by the weights $\boldsymbol{A}_{jk}$ being zero. As shown in Table 5, there is also an effect of grouping $\boldsymbol{X}_j^*$ variables with high correlation

Table 2
Frequency distribution of each predictor variable based on stunting incidence.

Variable	Category	Stunting Status	
		Non-stunted	Stunted
Gender (X_1)	Female	50.3%	7.8%
	Male	34.5%	7.4%
Birth Height (X_2)	Abnormal (Male <48 cm, Female <47.3 cm)	17.2%	14.5%
	Normal (Male ≥48 cm, Female ≥47.3 cm)	67.6%	0.7%
Birth Weight (X_3)	Abnormal (<2.5 kg)	19.3%	12.8%
	Normal (≥2.5 kg)	65.5%	2.4%
History of Illness (X_4)	No	52.7%	3%
	Yes	32.1%	12.2%
Age of Introduction to Solid Foods (X_5)	Appropriate (6 months)	47%	3.4%
	Inappropriate (<6 months & >6 months)	37.8%	11.8%
Formula Milk Consumption (X_6)	Yes	65.5%	11.5%
	No	19.3%	3.7%
Mother Received Additional Food During Pregnancy (X_7)	Yes	57.4%	7.4%
	No	27.4%	7.8%
Mother Received Iron Tablets (X_8)	Yes	57.4%	6.4%
	No	27.4%	8.8%
Father's Age (X_9)	<25 years	6.7%	5.1%
	25 – 40 years	48.6%	4.1%
	>40 years	29.4%	6.1%
Mother's Age (X_{10})	<21 years	7.4%	6.4%
	21 – 35 years	59.5%	7.1%
	>35 years	17.9%	1.7%
Father's Highest Education (X_{11})	Elementary School	4.1%	1.4%
	Middle School	7.4%	5.4%
	High School	47.6%	6.7%
	University	25.7%	1.7%
Mother's Last Education Level (X_{12})	Elementary School	3%	1.7%
	Middle School	8.8%	6.1%
	High School	51.7%	6.4%
	University	21.3%	1%
Type of Drinking Water Used (X_{13})	Packaged Gallon	3.4%	0%
	Refillable Gallon	70.6%	5.4%
	Boiled Water	10.8%	9.8%

in each $\boldsymbol{Z}_{\boldsymbol{k}}^{*}$, meaning that the formation of the optimal sparse principal component $\boldsymbol{Z}_{\boldsymbol{k}}^{*}$ can identify groups of $\boldsymbol{X}_{\boldsymbol{j}}^{*}$ variables that have a similar impact on data variation.

After the optimal sparse principal components $\boldsymbol{Z}_{\boldsymbol{k}}^{*}$ are formed, we can obtain the sparse binary logistic regression estimates from categorical principal component analysis on the stunting risk event data, as shown in Table 6.

Based on Table 6, we can express the sparse categorical principal component logistic regression analysis model for the stunting risk event data as following Eq. (21).

$$g = -3.361 - 2.137\boldsymbol{Z}_1{}^{*} + 0.791\boldsymbol{Z}_2{}^{*} + 0.484\boldsymbol{Z}_3{}^{*} + 1.014\boldsymbol{Z}_4{}^{*} - 0.150\boldsymbol{Z}_5{}^{*} \tag{21a}$$

The feasibility test of the relationship model for the stunting risk factors, as shown in Eq. (21), was carried out using the likelihood ratio statistical test, namely G^2 at a significance level of $\alpha = 0.05$. We obtained a value of $G^2 = 144.81$ and $\chi^2 = 11.07$, which means that the sparse category principal component logistic regression model is worthy of explaining the relationship model of stunting risk factors with all the predictors studied. Furthermore, the level of accuracy of the model obtained based on the confusion matrix results in Table 7 is around 88.85%, and the sensitivity level is around 75.55%. This value is higher than the results of the categorical principal component logistic regression model without sparse, with an accuracy of around 78.71% and a sensitivity of around 57.77%. The sparse categorical principal component logistic regression model provides better results. more accurate compared to the non-sparse model.

These results indicate that the model can explain the relationship between the response and all the predictors studied. The next step is to transform Eq. (21) back into the X variables based on the relationship between each Z_k^{*} and the corresponding categorical X_j^{*}. Table 5 shows that the response variable Y can be clearly expressed by the predictor variable X. The model describing the relationship between stunting events in Maginti District and the risk factors that may cause stunting events is shown in Eq. (22).

$$\begin{aligned} g \;=\; & -3.361 - 0.097X_{1.2} - 0.404X_{2.2} - 0.378X_{3.2} - 0.022X_{4.2} + 0.046X_{5.2} - 0.149X_{6.2} \;- 0.162X_{7.2} - 0.133X_{8.2} - 0.005X_{9.2} \\ & - 0.005X_{9.3} - 0.004X_{10.2} - 0.005X_{10.3} \;+ 0.101X_{11.2} - 0.066X_{11.3} - 0.036X_{11.4} + 0.107X_{12.2} - 0.069X_{12.3} \\ & - 0.037X_{12.4} - 0.107X_{13.2}) + 0.070X_{13.3} \end{aligned} \tag{22a}$$

Based on the estimated model for stunting risk factors obtained through sparse categorical principal component analysis logistic regression, several factors have more than a one-time higher risk of affecting stunting events. These factors include a history of

Table 3
Optimal category quantification C_j^* for each category in X^*.

Variable	Category	Optimal Quantification
Gender (X_1)	Female	−0.049
	Male	0.068
Birth Height (X_2)	Abnormal (Male <48 cm, Female <47.3 cm)	−0.085
	Normal (Male ≥48 cm, Female ≥47.3 cm)	0.039
Birth Weight (X_3)	Abnormal (<2.5 kg)	−0.084
	Normal (≥2.5 kg)	0.039
History of Illness (X_4)	No	−0.051
	Yes	0.065
Age of Introduction to Solid Foods (X_5)	Appropriate (6 months)	−0.057
	Inappropriate (<6 months & >6 months)	0.058
Formula Milk Consumption (X_6)	Yes	−0.031
	No	0.106
Mother Received Additional Food During Pregnancy (X_7)	Yes	−0.042
	No	0.079
Mother Received Iron Tablets (X_8)	Yes	−0.043
	No	0.077
Father's Age (X_9)	<25 years	−0.158
	25 – 40 years	0.021
	>40 years	0.021
Mother's Age (X_{10})	<21 years	−0.145
	21 – 35 years	0.002
	>35 years	0.029
Father's Highest Education (X_{11})	Elementary School	−0.157
	Middle School	−0.102
	High School	0.018
	University	0.042
Mother's Last Education Level (X_{12})	Elementary School	−0.145
	Middle School	−0.106
	High School	0.023
	University	0.041
Type of Drinking Water Used (X_{13})	Packaged Gallon	−0.048
	Refillable Gallon	−0.028
	Boiled Water	0.113

Table 4
Optimal loading value.

Loading value	Z_1^*	Z_2^*	Z_3^*	Z_4^*	Z_5^*
X_1^*	0.488	0.208	0.495	0.338	−0.407
X_2^*	−0.716	−0.321	0.333	−0.215	−0.358
X_3^*	−0.741	−0.365	0.281	−0.199	−0.320
X_4^*	0.729	0.412	0.276	0.280	−0.028
X_5^*	0.694	0.414	0.238	0.169	0.039
X_6^*	0.629	0.384	−0.004	−0.405	−0.231
X_7^*	0.795	0.241	−0.212	−0.374	−0.175
X_8^*	0.776	0.125	−0.198	−0.334	−0.152
X_9^*	−0.528	0.619	0.325	−0.234	0.315
X_{10}^*	−0.515	0.623	0.341	−0.312	0.207
X_{11}^*	−0.632	0.530	−0.298	0.202	−0.135
X_{12}^*	−0.664	0.621	−0.191	0.156	−0.157
X_{13}^*	0.621	−0.526	0.315	−0.048	0.298

illness, inadequate complementary feeding, formula milk consumption, and the lack of supplementary food and iron tablets for the mother. Regarding birth height and weight, we found that respondents with normal birth height have a 0.667 times lower risk of experiencing stunting compared to those with low birth height. Similarly, respondents with normal birth weight have a 0.685 times lower risk of experiencing stunting compared to those with low birth weight. Birth height and weight also play a role, with toddlers born with normal height and weight being at a lower risk of stunting than those with low birth height and weight. This highlights the importance of monitoring the health of both toddlers and mothers during pregnancy, mainly by providing additional nutrition and supplements for mothers.

The results of this study have provided knowledge related to several factors that have a major influence on the indication of stunting in toddlers through a statistical inferential approach. Child factors, such as birth weight and height, medical history, breast milk intake, and age of additional food provision, are the dominant factors causing more significant stunting. Previous studies have

Table 5
Optimal sparse loading value.

Loading value	Z_1^*	Z_2^*	Z_3^*	Z_4^*	Z_5^*
X_1^*	0.060	0	−0.670	0	0
X_2^*	0.695	0	0	0	0
X_3^*	0.650	0	0	0	0
X_4^*	−0.046	0	−0.527	0	0
X_5^*	−0.146	0	−0.419	0	0
X_6^*	0	0	0	−0.561	0
X_7^*	0	0	0	−0.607	0
X_8^*	0	0	0	−0.500	0
X_9^*	0	0	0	0	0.708
X_{10}^*	0	0	0	0	0.665
X_{11}^*	0	−0.538	0	0	0
X_{12}^*	0	−0.568	0	0	0
X_{13}^*	0	0.570	0	0	0

Table 6
Estimation results.

Variable	Coef. Estimation
Constant	−3.361
Z_1^*	−2.137
Z_2^*	0.791
Z_3^*	0.484
Z_4^*	1.014
Z_5^*	−0.150

Table 7
Confusion matrix.

Actual	Predicted with sparse		Predicted without sparse	
	1	0	1	0
1	34	11	26	19
0	22	229	44	207

found that various diseases, such as diarrhea, respiratory tract infections, and fever, contribute to child growth [26]. However, maternal factors also significantly influence the growth and development of their children, including nutritional intake and iron tablets during pregnancy. For the case of stunting, this study can be further developed by adding data sets and variables that have not been studied, such as childcare factors, economy, environment, and other factors. The variety of factors that can influence the risk of stunting means that method development can be carried out in further studies, for example, mixed effects derived from predictors.

Limitations

The method applies to categorical data in both the response and its predictors. In addition, a high correlation occurs between predictors, which is called multicollinearity.

Ethics statements

This research involved human subjects, and in the data collection process, it has gone through Research Permit No. 11,043/UN4.11.7/PT.01.04/2023 from the Department of Statistics.

Supplementary material *and/or* additional information [OPTIONAL]

None.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Anna Islamiyati: Conceptualization, Methodology, Formal analysis, Writing – original draft. **Muhammad Nur:** Formal analysis. **Abdul Salam:** Investigation. **Wan Zuki Azman Wan Muhamad:** Software. **Dwi Auliyah:** Data curation, Project administration.

Data availability

Data will be made available on request.

Acknowledgments

This work was funded by a National Competitive Basic Research grant from the Directorate of Research, Technology, and Community Service of the Ministry of Education, Culture, Research, and Technology with research contract No: 050/E5/PG.02.00.PL/2024.

References

[1] U. Khadija, S. Mahmood, A. Ainee, M.Y. Quddoos, H. Ahmad, A. Khadija, S.M. Zahra, A. Hussain, Nutritional health status: association of stunted and wasted children and their mothers, BMC Pediatr 22 (1) (2022) 255, doi:10.1186/s12887-022-03309-y.
[2] N. Linawati, Relationship Between Low Birth Weight and Infectious Diseases with Stunting in Children Aged 4 to 5 Years, Indonesian, J. Multidiscipl. Sci. 1 (9) (2022) 1020–1030, doi:10.55324/ijoms.v1i9.143.
[3] O.R. Katoch, Urgent call to address child malnutrition: a matter of life and future, Nutrition 116 (2023), doi:10.1016/j.nut.2023.112236.
[4] D. Vilcins, P. Sly, P. Jagals, Environmental risk factors associated with child stunting: a systematic review of the literature, Ann. Glob. Health 84 (4) (2018) 551, doi:10.29024/aogh.2361.
[5] G. Rezaeizadeh, M.A. Mansournia, A. Keshtkar, Z. Farahani, F. Zarepour, M. Sharafkhah, Maternal education and its influence on child growth and nutritional status during the first two years of life: a systematic review and meta-analysis, eClinicalMedicine 71 (2024) 102574, doi:10.1016/j.eclinm.2024.102574.
[6] C. Nembidzane, M. Lesaoana, K.D. Monyeki, A. Boateng, P.J. Makgae, Using the sitar method to estimate age at peak height velocity of children in rural south africa: ellisras longitudinal study, Children. (Basel) 7 (17) (2020) 1–9, doi:10.3390/children7030017.
[7] S. Thurstans, C. Opondo, A. Seal, J. Wells, T. Khara, C. Dolan, A. Briend, M. Myatt, M. Garenne, R. Sear, M. Kerac, Boys are more likely to be undernourished than girls: a systematic review and metaanalysis of sex differences in undernutrition, BMJ Glob. Health 5 (12) (2020) 1–17, doi:10.1136/bmjgh-2020-004030.
[8] A. Islamiyati, A. Kalondeng, M. Zakir, S. Djibe, U. Sari, Detecting age prone to growth retardation in children through a bi-response nonparametric regression model with a penalized spline estimator, Iran. J. Nurs. Midwifery Res. 29 (5) (2024) 549–554, doi:10.4103/ijnmr.ijnmr_342_22.
[9] P. Schober, T.R. Vetter, Sample Size and Power in Clinical Research, Anesth. Analg. 129 (2) (2019), doi:10.1213/ANE.0000000000004316.
[10] A. Islamiyati, M.Zakir Anisa, U. Sari, D.S. Salam, The use of the binary spline logistic regression model on the nutritional status data of children, Commun. Mathem. Biol. Neurosci. 37 (2023) 1–11, doi:10.28919/cmbn/7935.
[11] S. Arifin, A. Islamiyati, E.T. Herdiani, Ability of ordinal spline logistic regression model in the classification of nutritional status data, Commun. Mathem. Biol. Neurosci. 83 (2023) 1–11, doi:10.28919/cmbn/8072.
[12] O.B. Emine, D. Franklin, Multicollinearity in logistic regression models, Anesth. Analg. 133 (2) (2021) 362–365, doi:10.1213/ANE.0000000000005593.
[13] A. Agarwal, D. Shah, D. Shen, D. Song, On robustness of principal component regression, J. Am. Stat. Assoc 116 (536) (2021) 1731–1745, doi:10.1080/01621459.2021.1928513.
[14] A. Islamiyati, A. Kalondeng, N. Sunusi, M. Zakir, A.K. Amir, Biresponse nonparametric regression model in principal component analysis with truncated spline estimator, J. King Saud. Univer. 34 (2022) 101892, doi:10.1016/j.jksus.2022.101892.
[15] I.T. Jolliffe, J. Cadima, Principal component analysis: a review and recent developments, Philosoph. Transac. A 374 (20150202) (2016) 1–16, doi:10.1098/rsta.2015.0202.
[16] S. Kawano, H. Fujisawa, T. Takada, T. Shiroishi, Sparse principal component regression with adaptive loading, Comput. Stat. Data. Anal 89 (2015) 192–203 September, doi:10.1016/j.csda.2015.03.016.
[17] E.N. Erichson, P. Zheng, K. Manohar, S.L. Brunton, J.N. Kutz, A.Y. Aravkin, Sparse principal component analysis via variable projection, SIAM. J. Appl. Math 80 (2) (2020), doi:10.1137/18M1211350.
[18] G. Kemalbay, Ö.B. Korkmazoğlu, Categorical principal component logistic regression: a case study for housing loan approval, Procedia-Social Behav. Sci. 109 (2) (2014) 730–736, doi:10.1016/j.sbspro.2013.12.537.
[19] H. Zou, L. Xue, A selective overview of sparse principal component analysis, Proc. IEEE 106 (8) (2018) 1311–1320, doi:10.1109/JPROC.2018.2846588.
[20] M. Greenacre, P.J.F. Groenen, T. Hastie, A.L. D'Enza, A. Markos, E. Tuzhilina, Principal component analysis, Nat. Rev. Method. Primers 2 (100) (2022), doi:10.1038/s43586-022-00184-w.
[21] H. Abou-Senna, E. Radwan, H.T. Abdelwahab, Categorical principal component analysis (CATPCA) of pedestrian crashes in central Florida, J. Transpor. Safety Secur. 14 (11) (2021) 1–23, doi:10.1080/19439962.2021.1988788.
[22] Q. Zhang, S. Lu, L. Xie, W.H. Xu, H.Y. Su, Dynamic fault detection and diagnosis of industrial alkaline water electrolyzer process with variational Bayesian dictionary learning, Int. J. Hydrogen Energy 71 (2024) 1492–1506, doi:10.1016/j.ijhydene.2023.03.373.
[23] S. Kumar, P. Sarkar, Oja's algorithm for sparse PCA, arXiv (2024). http://arxiv.org/abs/2402.07240.
[24] P.J. Pan, C.H. Lee, N.W. Hsu, T.L. Sun, Combining principal component analysis and logistic regression for multifactorial fall risk prediction among community-dwelling older adults, Geriatr. Nurs. (Minneap) 57 (2024) 208–216, doi:10.1016/j.gerinurse.2024.04.021.
[25] J. Wu, Y. Asar, Efficiency of the principal component liu-type estimator in logistic regression model, Revstat-Statist. J. 18 (3) (2017) 1–16, doi:10.57805/revstat.v18i3.304.
[26] A. Santosa, E.N. Arif, D.A. Ghoni, Effect of maternal and child factors on stunting: partial least squares structural equational modeling, Clin. Exp. Pediatr 65 (2) (2022) 90–97, doi:10.3345/cep.2021.00094.