KINGDOM OF SAUDI ARABIA
Northern Border University (NBU)
Journal of the North for Basic & Applied Sciences(JNBAS)
p - ISSN : 1658 -7022 / **e-ISSN:** 1658 - 7014
www.nbu.edu.sa
s.journal@nbu.edu.sa

Statistical Modeling of Childhood Diarrhea Using Penalized Regression and Bootstrap Techniques in High-Dimensional Survey Data

Ali Satty
Mathematics Department, College of Science, Northern Border University, Saudi Arabia

(**Received:** 27th September 2025; **Accepted:** 14th October 2025)

Abstract

High-dimensional survey data, such as those from the Multiple Indicator Cluster Survey (MICS), pose challenges for traditional statistical analysis due to multicollinearity and correlated variables. The objective of this study is to demonstrate the application of penalized regression (PR), specifically the Least Absolute Shrinkage and Selection Operator (LASSO), combined with post-selection inference and bootstrap validation for analyzing childhood diarrhea. Data from 8,923 children were analyzed to assess predictors of diarrhea. LASSO and Ridge regression models were fitted, and post-selection bootstrap resampling (R = 1,000) was conducted. LASSO identified child age, region, unimproved water source, and child weight as the strongest predictors, while Ridge regression retained all variables with similar predictive performance (AUC ~0.61). The findings highlight the value of PR methods for variable selection and model stability in high-dimensional survey data. In conclusion, combining LASSO with post-selection inference and bootstrap validation provides a practical framework for statistical analysis of complex survey data, with applicability beyond diarrhea research.

Keywords: Childhood diarrhea, Penalized regression (PR), Least absolute shrinkage and selection operator (LASSO), High-dimensional survey data, Post-selection inference.

(*) **Corresponding Author**:
Ali Satty
Mathematics Department, College of Science, Northern Border University, Saudi Arabia
E-mail: alisatty1981@gmail.com; ali.hassan@nbu.edu.sa

DOI: 10.12816/0062289

1. Introduction

Childhood diarrhea remains a leading cause of morbidity and mortality among children under five, particularly in low- and middle-income countries (Walker et al., 2013; Rego et al., 2022). Despite advancements in water, sanitation, and hygiene, the burden of diarrheal diseases persists due to a complex interplay of demographic, environmental, and socioeconomic factors (Black et al., 2010). Large-scale surveys such as the Multiple Indicator Cluster Surveys (MICS) provide valuable data on child health and household characteristics, offering opportunities to study these factors (UNICEF, 2020). However, these datasets often include a large number of correlated variables, posing challenges for traditional statistical models. Standard logistic regression can produce unstable estimates in the presence of multicollinearity or high-dimensional data, limiting reproducibility and interpretability (Zou and Hastie, 2015).

Previous studies analyzing childhood diarrhea disease have primarily relied on standard logistic regression or generalized linear models to identify risk factors (Satty et al., 2024; Black et al., 2010). While informative, these approaches often fail to efficiently handle correlated predictors or select the most relevant variables when the number of predictors is large. PR methods, such as LASSO and Ridge regression, have been increasingly applied in epidemiological and public health settings to address high-dimensional data, perform variable selection, and improve model interpretability (Tibshirani, 1996; Zou and Hastie, 2005; Friedman et al., 2010). Nevertheless, most applications focus on predictive performance rather than demonstrating reproducible post-selection inference or integrating robust bootstrap validation (Lee et al., 2016; Kammer et al., 2022; Wong et al., 2023), leaving a methodological gap in epidemiological research using survey data.

The current study aims to demonstrate the methodological application of PR in high-dimensional survey data, using childhood diarrhea as a case study. LASSO regression was employed to identify the most informative predictors, with Ridge regression used for comparative performance evaluation. To ensure robust and unbiased inference, post-LASSO analysis with bootstrap resampling was conducted to estimate adjusted odds ratios (ORs) and confidence intervals (CIs). The significance of this study lies in highlighting reproducibility, interpretable effect estimation, and methodological transparency, rather than solely identifying risk factors. By combining PR with post-selection inference and bootstrap validation, this study provides a practical, reproducible, and interpretable framework for modern epidemiological research. Although applied here to childhood diarrhea, the methodology is broadly applicable to other public health outcomes with high-dimensional data, offering a roadmap for improving statistical rigor and interpretability. Overall, this approach illustrates how advanced statistical techniques can be leveraged for robust, reproducible, and interpretable analyses of large-scale survey datasets.

2. Methods

2.1 Mathematical formulation of the PR analysis

Let $Y_i \in \{0,1\}$ denote the binary outcome for child i (1 = diarrhea, 0 = no diarrhea), and let $X_i=(X_{i1},X_{i2},\ldots,X_{ip})$ represent the p explanatory variables including child, maternal, household, and environmental characteristics. The PR model is defined as:

$$P(Y_i = 1|X_i) = \pi_i = \frac{\exp\left(\beta_0 + \sum_{j=1}^{p} \beta_j X_{ij}\right)}{1 + \exp\left(\beta_0 + \sum_{j=1}^{p} \beta_j X_{ij}\right)},$$

where β_0 is the intercept and β_j are the regression coefficients. The LASSO estimator is obtained by maximizing the penalized log-likelihood:

$$\hat{\beta}^{LASSO} = \arg max \left\{ \iota(\beta) - \lambda \sum_{j=1}^{p} |\beta_j| \right\},$$

where $\iota(\beta)$ is the log-likelihood function for the logistic model, $\lambda>0$ is the tuning parameter controlling the penalty, and $\sum_{1}^{p}=_1|\beta_j|$ is the L_1-norm penalty that shrinks some coefficients exactly to zero. Similarly, the Ridge estimator minimizes the L_2-penalized log-likelihood:

$$\hat{\beta}^{Ridge} = \arg max \left\{ \iota(\beta) - \lambda \sum_{j=1}^{p} \beta_j^2 \right\},$$

where the L_2-norm penalty shrinks coefficients toward zero without eliminating any variable. For post-selecting inference, for variables selected by LASSO ($\beta_j^{LASSO} \neq 0$), adjusted ORs were estimated using standard logistic regression on the selected subset:

$$OR_j = exp\left(\hat{\beta}_j^{post-LASSO}\right), \; j \in S,$$

where $S=\{j:\beta_j^{LASSO} \neq 0\}$ and $\beta_j^{post\text{-}LASSO}$) are coefficients from the refitted logistic model. Bootstrap resampling (R=1,000) was used to obtain 95% CIs for these ORs.

2.2 Study data and variable description

The study analyzed data from the 2018–2019 MICS conducted in the Central African Republic (CAR) (ICASEES, 2021). The binary outcome variable was childhood diarrhea (1 = yes, 0 = no), based on a sample of 8,923 children. Explanatory variables covered child, maternal, and household/environmental characteristics. Child-level factors included sex, current breastfeeding status, weight, and age grouped into five categories (0–11, 12–23, 24–35, 36–47, and 48–59 months). Maternal education was classified as none/preschool, primary, or secondary and above. Household and environmental variables included area of residence (urban/rural), region (1–7), wealth index (poorest to richest quintiles), source of drinking water (improved/unimproved), sanitation facility (improved/unimproved), and handwashing facilities (observed/unobserved).

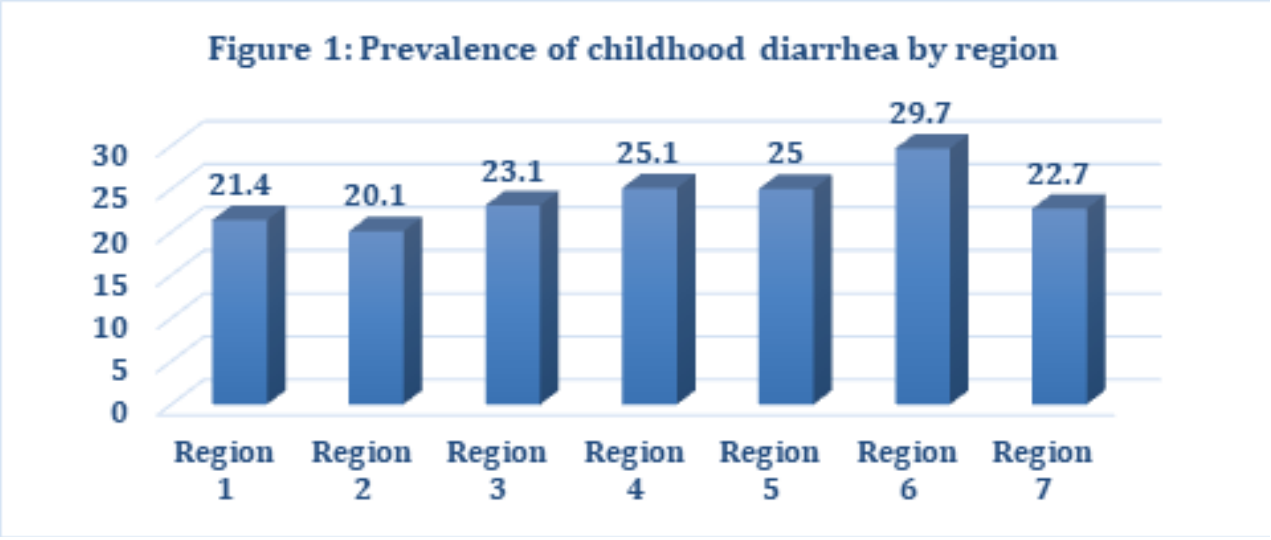
2.3 The PR analysis and post-selection inference

Data analysis was conducted using R (version 4.5.1). Predictor variables were preprocessed by dummy coding categorical variables, and continuous variables were standardized where appropriate. LASSO and Ridge regression models were fitted using the glmnet package. The optimal penalty parameters (λ) for LASSO and Ridge were selected via 10-fold cross-validation, minimizing the binomial deviance. The final LASSO and Ridge models were fitted using the λ values corresponding to the minimum cross-validated deviance (λ_LASSO = 0.0046, λ_Ridge = 0.0028). Variables retained by LASSO (non-zero coefficients) were used for post-selection inference. Specifically, a standard logistic regression was refitted on the LASSO-selected predictors, and bootstrap resampling (R = 1,000, set.seed = 123) was performed to obtain 95% CIs for adjusted ORs. Model performance was evaluated using area under the receiver operating characteristic curve (AUC), accuracy, sensitivity, and specificity. Sensitivity was particularly emphasized due to the imbalance in diarrhea prevalence, with fewer positive cases relative to non-cases. This combined approach allowed robust identification of key predictors while ensuring reproducible and interpretable effect estimates in a high-dimensional survey dataset.

3. Results:

3.1 Population characteristics and diarrhea prevalence

The analysis included 8,923 children under five (4,379 boys and 4,544 girls), with diarrhea slightly more common among boys (25.1%) than girls (23.4%). Prevalence peaked at ages 6–23 months (33.6%) and was lowest at 48–59 months (15.1%). Children ever breastfed showed higher prevalence (25.4%) than those never breastfed (20.0%). By maternal education, the highest prevalence was among children of mothers with fundamental 1 education (26.7%), compared to 21.6% with no education. Wealth differences were modest (22.6%–25.2%). Prevalence was slightly higher in rural (24.5%) than urban (23.5%) areas, and across regions ranged from 21.3% in Region 2 to 29.7% in Region 6. Children from households with unimproved water sources (25.8%) and unimproved sanitation (24.4%) had higher prevalence than those with improved facilities, and diarrhea was also more common in households with observed handwashing places (25.7%) than unobserved (23.6%). Figure 1 shows childhood diarrhea prevalence in the CAR by region, highlighting notable regional disparities.



3.2 Factors retained in the LASSO model

The LASSO-PR analysis (Table 1) identified several factors associated with childhood diarrhea, while other variables were not selected by the model, with their coefficients shrunk to zero. In Table 2, variables marked as "No" under the LASSO selection column indicate that their coefficients were shrunk exactly to zero by the LASSO penalty and thus excluded from the post-selection model. Variables marked as "Yes" were retained with non-zero coefficients and included in the post-LASSO inference. Age showed the strongest associations, with children aged 12–23 months ($\beta = 0.76$) and 24–35 months ($\beta = 0.53$) having the highest positive coefficients compared to the reference group, and smaller effects observed for children aged 36–47 months ($\beta = 0.14$) and 48–59 months ($\beta = 0.15$). Male sex ($\beta = 0.09$), current breastfeeding ($\beta = 0.17$), urban residence ($\beta = 0.08$), unimproved water source ($\beta = 0.18$), and higher child weight ($\beta = 0.17$) were also positively associated with diarrhea. In contrast, maternal education of preschool or none ($\beta = -0.29$), unobserved handwashing facilities ($\beta = -0.22$), and certain wealth index categories showed negative associations, suggesting reduced risk. Regional differences were evident, with progressively larger coefficients in regions 4 ($\beta = 0.18$), 5 ($\beta = 0.27$), and 6 ($\beta = 0.41$), indicating region 6 as the area with the strongest effect. Some variables, such as maternal education at secondary or higher ($\beta = 0.00$) and region 7 ($\beta = 0.00$), were not selected by LASSO, suggesting minimal contribution to predicting diarrhea. Overall, Table 2 highlights child age, geographic region, and water source as the most influential predictors, with other factors contributing more moderately.

Table 1: Variables retained in the LASSO model:

Factor	Coefficient (β)	Selected (LASSO)
Sex: male	0.089	Yes
Age:12-23	0.755	Yes
Age: 24-35	0.530	Yes
Age: 36-47	0.135	Yes
Age: 48-59	0.147	Yes
Breastfeeding: yes	0.170	Yes
Maternal education: preschool or none	-0.286	Yes
Maternal education: secondary or higher	0.000	No
Area of residence: urban	0.080	Yes
Region: region 2	-0.111	Yes
Region: region 3	0.019	Yes
Region: region 4	0.180	Yes
Region: region 5	0.269	Yes
Region: region 6	0.406	Yes
Region: region 7	0.000	No
Wealth index: poor	-0.042	Yes
Wealth index: poorest	0.034	Yes
Wealth index: rich	-0.117	Yes
Wealth index: richest	-0.044	Yes
Source of drinking water: unimproved	0.177	Yes
Sanitation status: unimproved	0.031	Yes
Handwashing facilities: unobserved	-0.222	Yes
Child weight	0.172	Yes

3.3 Model performance metrics: LASSO vs. Ridge

The 10-fold cross-validation results (Table 2) indicate that LASSO and Ridge regression models performed similarly in terms of discrimination and overall accuracy. LASSO achieved a slightly higher AUC (0.612) compared to Ridge (0.609), while both models had comparable accuracy (~76%). However, both models exhibited extremely high specificity (close to 1.0) and very low sensitivity, suggesting that while non-cases were correctly classified, true cases of diarrhea were poorly identified. Despite comparable predictive performance, LASSO was preferred because it performs both shrinkage and variable selection, shrinking some coefficients exactly to zero. This property improves interpretability by highlighting the most relevant predictors of childhood diarrhea and simplifying the model.

Table 2: Performance metrics of Ridge and LASSO models

Model	λ	AUC	Accuracy	Sensitivity	Specificity
Ridge	0.0028	0.609	0.764	0.0001	1.000
LASSO	0.0046	0.612	0.764	0.0001	0.999

3.4 Adjusted ORs from post-LASSO model

Post-LASSO analysis was conducted to estimate the adjusted *OR*s for the variables retained in the LASSO model (Table 3). Male children (*OR* = 1.093; 95% CI: 1.000-1.211) and those aged 12–23 months (*OR* = 2.128; 95% CI: 1.797-2.605) or 24–35 months (*OR* = 1.699; 95% CI: 1.416-2.077) had significantly higher odds of diarrhea. Breastfeeding was also associated with increased odds (*OR* = 1.186; 95% CI: 1.038-1.362). Maternal education had a protective effect, with children of mothers with no or preschool education less likely to experience diarrhea (*OR* = 0.751; 95% CI: 0.674-0.842). Regional disparities were evident, with children in Regions 4, 5, and 6 facing significantly higher odds, while no differences were observed by urban–rural residence or wealth quintiles. Environmental risk factors included unimproved water sources (*OR* = 1.193; 95% CI: 1.077-1.340), whereas unimproved sanitation was not significant. Interestingly, lack of observed handwashing facilities was protective (*OR* = 0.801; 95% CI: 0.715-0.884). Higher child weight was also associated with greater odds of diarrhea (*OR* = 1.188; 95% CI: 1.066-1.305). Overall, age, region, unimproved water, and child weight emerged as the most important predictors of childhood diarrhea.

Table 3: Post-LASSO results for childhood diarrhea

Variable	Adjusted odds ratios (*OR*s)	CI Lower	CI Upper
Sex: male	1.093	1.000	1.211
Age:12-23	2.128	1.797	2.605
Age: 24-35	1.699	1.416	2.077
Age: 36-47	1.145	0.985	1.409
Age: 48-59	1.158	1.000	1.389
Breastfeeding: yes	1.186	1.038	1.362
Maternal education: preschool or none	0.751	0.674	0.842
Area of residence: urban	1.084	0.974	1.258
Region: region 2	0.895	0.757	1.033
Region: region 3	1.019	0.878	1.238
Region: region 4	1.197	1.013	1.422
Region: region 5	1.309	1.076	1.578
Region: region 6	1.501	1.263	1.847
Wealth index: poor	0.959	0.825	1.091
Wealth index: poorest	1.035	0.908	1.196
Wealth index: rich	0.889	0.754	1.001
Wealth index: richest	0.956	0.764	1.095
Source of drinking water: unimproved	1.193	1.077	1.340
Sanitation status: unimproved	1.032	0.920	1.177
Handwashing facilities: unobserved	0.801	0.715	0.884
Child weight	1.188	1.066	1.305

4. Discussion

In this study, penalized regression (PR) techniques, including LASSO and Ridge models, were applied to high-dimensional survey data to provide a reproducible and interpretable framework for variable selection. By simultaneously performing shrinkage and selection, LASSO identifies a concise set of predictors while controlling model complexity, whereas Ridge regression retains all variables but provides comparable predictive performance, illustrating the trade-off between sparsity and variance reduction (Zou and Hastie, 2005; Friedman et al., 2010). In the present analysis of childhood diarrhea, both LASSO and Ridge regression effectively accommodated correlated predictors through shrinkage (Pavlou and Ambler, 2015; Kipruto and Wang, 2025). LASSO highlighted child age, region, unimproved drinking water, and greater child weight as primary risk factors, while factors such as maternal education, urban residence, and unobserved handwashing facilities showed weaker or protective associations. Post-selection inference with 1,000 bootstrap resamples confirmed these findings, and Ridge regression results mirrored LASSO's patterns, demonstrating the reliability and stability of the selected predictors (Bainter and Binns, 2023).

The study incorporated extensive internal validation to enhance reproducibility and assess model reliability. Ten-fold cross-validation selected $\lambda = 0.0006376$ for LASSO and $\lambda = 0.0028$ for Ridge, and both penalties produced similar predictive performance, indicating comparable discrimination but very low sensitivity, likely due to severe class imbalance (Saito and Rehmsmeier, 2015). To mitigate bias in penalized coefficient estimates, we refitted the LASSO-selected predictors and applied bootstrap resampling, which provided less-biased effect estimates and valid 95% CIs. This procedure confirmed the stability of variable selection and the consistency of effect magnitudes across resamples (Robert, 1996).

Although LASSO and Ridge produced nearly identical AUC metrics (~0.61), LASSO was preferred for its interpretability, setting unimportant coefficients to zero. However, these coefficients are shrinkage estimates optimized for prediction; their magnitudes should not be interpreted as unbiased causal effects. For formal inference on effect sizes, post-selection procedures, such as refitting an unpenalized model on LASSO-selected variables or using post-selection inference frameworks, are recommended to reduce bias and obtain valid CIs (Taylor & Tibshirani, 2016; Meinshausen and Bühlmann, 2010; Lee et al., 2016; Chernozhukov et al., 2015).

The observed extremely low sensitivity combined with very high specificity underscores practical challenges in class-imbalanced data. Decision thresholds tuned for overall accuracy/AUC may fail to detect positive cases. In such settings, precision–recall analysis (PR-AUC), threshold optimization, cost-sensitive learning, resampling strategies (e.g., oversampling, SMOTE), or class-weighted penalties can provide more informative evaluation metrics and improve positive-case detection (Saito and Rehmsmeier, 2015).

In summary, this study illustrates that PR provides a robust, reproducible, and flexible framework for modeling complex disease outcomes. By combining variable selection, shrinkage, bootstrap validation, and post-selection inference, researchers can identify key predictors, quantify uncertainty, and account for overfitting and multicollinearity. While the application here focused on childhood diarrhea, these methods are broadly applicable across epidemiological and public health research where high-dimensional, correlated predictors are common.

5. Conclusion

This study demonstrates that application of penalized regression analysis, particularly LASSO combined with post-selection inference and bootstrap validation, provides a robust and reproducible framework for analyzing high-dimensional survey data. This application effectively identified child age, region, unimproved water sources, and child weight as the strongest predictors of childhood diarrhea, while accounting for multicollinearity and class imbalance.

Strengths of this study include the use of advanced variable selection methods, internal validation through cross-validation and bootstrap resampling, and the generation of interpretable effect estimates via post-LASSO inference. These methods improve reproducibility and allow for a compact, stable model, which is particularly valuable when working with complex survey data. Limitations include the low sensitivity observed in the predictive models due to class imbalance, and the inherent bias in penalized coefficients that requires post-selection adjustment. Additionally, the analysis is cross-sectional and observational, limiting causal interpretation of the associations.

Future directions include refining methods for handling imbalanced outcomes, applying PR techniques to other public health datasets, incorporating external validation to assess model generalizability, and combining these approaches with causal inference methods to strengthen policy-relevant recommendations.

6. Ethical considerations

This study used de-identified, secondary data obtained from MICS. The data are publicly available upon request through the UNICEF MICS website (https://mics.unicef.org/). Ethical approval for the original data collection was obtained by the implementing agencies, and informed consent was provided by all participants.

The use of de-identified secondary data in this analysis posed no risk to participant privacy or confidentiality, and all analyses adhered to ethical standards for research using human subjects.

7. Conflict of interest:

The author declares that there are no conflicts of interest.

8. References:

1. *Bainter, S. A., McCauley, T. G., Fahmy, M. M., Goodman, Z. T., Kupis, L. B., & Rao, J. S. (2023). Comparing Bayesian variable selection to Lasso approaches for applications in psychology. Psychometrika, 88(3), 1032–1055. https://doi.org/10.1007/s11336-023-09914-9*
2. *Black, R. E., et al. (2010). Global, regional, and national causes of child mortality in 2008: A systematic analysis. The Lancet, 375(9730), 1969–1987. https://doi.org/10.1016/S0140-6736(10)60549-1*
3. *Friedman, J. H., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software, 33(1), 1–22. https://doi.org/10.18637/jss.v033.i01*
4. *ICASEES. (2021). MICS6-RCA Multiple Indicator Cluster Survey 2018–2019, final report of survey results. Retrieved from https://mics.unicef.org/surveys*
5. *Kammer, M., et al. (2022). Evaluating methods for Lasso selective inference in biomedical research: A comparative simulation study. BMC Medical Research Methodology. https://doi.org/10.1186/s12874-022-01681-y*
6. *Kipruto, E., & Wang, Y. (2025). Evaluating prediction performance: A simulation study comparing penalized regression methods. Applied Sciences, 15(13), 7443. https://doi.org/10.3390/app15137443*
7. *Lee, J. D., Sun, D. L., Sun, Y., & Taylor, J. E. (2016). Exact post-selection inference, with application to the Lasso. The Annals of Statistics, 44(3), 907–927. https://doi.org/10.1214/15-AOS1371*
8. *Pavlou, M., Ambler, G., Seaman, S., De Iorio, M., & Omar, R. Z. (2016). Review and evaluation of penalised regression methods for risk prediction in low-dimensional data with few events. Statistics in Medicine, 35(7), 1159–1177. https://doi.org/10.1002/sim.6782*
9. *Rego, R., Watson, S., Gill, P., & Lilford, R. (2022). The impact of diarrhoea measurement methods for under 5s in low- and middle-income countries on estimated diarrhoea rates at the population level: A systematic review and meta-analysis of methodological and primary empirical studies. Tropical Medicine & International Health, 27(4), 347–368. https://doi.org/10.1111/tmi.13739*
10. *Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS ONE. https://doi.org/10.1371/journal.pone.0118432*
11. *Satty, A., Salih, M., Abdalla, F. A., Mahmoud, A. F. A., Gumma, E. A. E., Saad Mohamed Khamis, G., Adam, A. M. A., Hassaballa, A. A., Hamed, O. M. A., & Mohammed, Z. M. S. (2024). Statistical analysis of factors associated with diarrhea in Yemeni children under five: Insights from the 2022–2023 Multiple Indicator Cluster Survey. Journal of Epidemiology and Global Health, 14(3), 1043–1051. https://doi.org/10.1007/s44197-024-00253-1*
12. *Taylor, J., & Tibshirani, R. (2018). Post-selection inference for ℓ1-penalized likelihood models. Canadian Journal of Statistics, 46(1), 41–61. https://doi.org/10.1002/cjs.11313*
13. *Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. Journal of the Royal Statistical Society: Series B (Methodological), 58(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x*
14. *UNICEF. (2020). Multiple Indicator Cluster Surveys (MICS) - UNICEF. https://mics.unicef.org/*
15. *Walker, C. L. F., Rudan, I., Liu, L., Nair, H., Theodoratou, E., Bhutta, Z. A., O'Brien, K. L., Campbell, H., & Black, R. E. (2013). Global burden of childhood pneumonia and diarrhoea. The Lancet, 381(9875), 1405–1416. https://doi.org/10.1016/S0140-6736(13)60222-6*
16. *Wong, A., Kramer, S. C., Piccininni, M., Rohmann, J. L., Kurth, T., Escolano, S., Grittner, U., & Domenech de Cellès, M. (2023). Using LASSO regression to estimate the population-level impact of pneumococcal conjugate vaccines. American Journal of Epidemiology, 192(7), 1166–1180. https://doi.org/10.1093/aje/kwad061*
17. *Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67(2), 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x*