EPH – International Journal of Mathematics and Statistics
ISSN:2208-2212
https://eijms.com/index.php/ms/issue/archive

SMALL-SAMPLE LOGISTIC REGRESSION WITH PENALIZED LIKELIHOOD AND PRINCIPAL COMPONENT ANALYSIS

[*]Nisha Khah
[*]Researcher, Human Genetics, Manipal School of life sciences, Manipal, Karnataka
Email: nishakhah66@gmail.com
Corresponding author: nishakhah66@gmail.com

A R T I C L E I N F O

Received: 03/12/2025
Revised: 10/03/2026
Accepted: 15/04/2026
Published: 29/04/2026

DOI: https://doi.org/10.69980/jpc4eb82
Volume:12, Issue:02, April, 2026

Abstract

This study examined the applicability of the principal component analysis (PCA) and Firth penalized logistic regression (FPLR) approach to modelling binary outcomes with small samples. An open-access dataset of 23 older adults and their demographic, health-related, and eight predictor variables from questionnaires was used to complete a secondary quantitative analysis. The binary outcome consisted of 11 observations in class 0 and 12 observations in class 1. The eight variables from the questionnaire were standardised and then subjected to PCA, the Kaiser-Meyer-Olkin statistic was used to assess sampling adequacy. PCs 1 and 2 were kept for further Firth logistic regression. The PCA analysis revealed that PC1 and PC2 accounted for 47.71% and 19.46% of the variance, respectively, with a total of 67.18% variance explained. In the penalized regression model, PC1 was found to be associated with the binary outcome (OR = 16.68, $p < 0.001$), while PC2 was not (OR = 0.77, $p = 0.792$). For the model, in-sample accuracy, sensitivity, specificity, precision and F1-score were all 100% while the AUC was 1.000. The results are preliminary due to the very limited sample size and lack of external validation, however. In conclusion, the PCA–Firth framework was a parsimonious exploratory framework for small-sample binary classification problems in which predictors are correlated.

Keywords: Principal component analysis; Firth logistic regression; penalized likelihood; small-sample modelling; binary classification

1. Introduction

Logistic regression is among the most frequently used statistical models for predicting binary outcomes because the results from logistic regression models are interpretable regression relationships between the predictor variables and the probability of the event. It's especially useful in biomedical, behavioral and social-science studies where the outcomes are often dichotomous. In case the available sample is small compared to the number of predictive factors, maximum likelihood logistic regression is problematic however. Small sample size results in variable regression parameters, an inflated standard error, a skewed probability estimate, and poor generalizability [1,2]. Sometimes the datasets are sparse, and this leads to complete or quasi-complete separation, where the predictors almost perfectly distinguish between the outcome groups, which makes the conventional maximum-likelihood estimates very large or even makes them non-convergent [3,4]. The difficulties can be lessened by increasing the sample size, but this is not always possible, especially in the case of special or exploratory studies [5]. Therefore, the modelling approaches used need to be different when using data with fewer samples.

To overcome problems of overfitting and coefficient estimation in the presence of small samples, sparse events, or separation, penalized likelihood methods have been proposed for conventional regression. Firth logistic regression is useful since it penalizes the likelihood function in a way that reduces the bias, and thus yields finite coefficient estimates even in the absence of separation. This method can be used to achieve better estimation when the number of events is small, and when maximum-likelihood estimates are not valid [3]. However, while penalisation leads to better predictive performance, especially in small datasets, it does not always result in the desired performance, and it is important to validate the models [6,7]. When sample sizes are approaching the minimum acceptable size, the tuning, estimation and prediction results also vary significantly [8]. Therefore, transparent reporting and careful assessment is required when building small-sample prediction models [9–11].

An alternative approach to dealing with multiple correlated predictors is called principal component analysis (PCA). The original variables are converted to a much smaller number of uncorrelated variables that preserve much of the information in the original variables [12]. PCA reduces the dimensionality and thus the complexity of the model, lessens multicollinearity and reduces the number of parameters fed into logistic regression. When the number of predictors is large compared to the number of observations, feature engineering and systematic variable transformation also enhance the predictive modelling efficiency [13]. Other dimension-reduction methods specifically developed for binary classification have shown that they can be used to efficiently represent predictive information with transformed components combined with logistic regression [14].

Penalized regression and dimensionality-reduction methods are two methods that are very well-known in their respective applications, but they have not been applied together in very small data sets. Their ability to provide stable performance, avoid overfitting and avoid overestimation of performance is particularly sensitive in small samples, highlighting the need for methodological design and validation [15]. Moreover, the interpretation of discrimination measures, e.g. receiver operating characteristic curves, should be considered together with other measures of predictive performance and not in isolation [16]. A joint PCA-Firth method then proves to be a promising method in that it is not only capable of dimensionality reduction of predictors, but also of correcting for small-sample estimation bias.

This study therefore attempts to reduce the number of correlated predictors into a smaller number of informative principal components by applying PCA fitting a Firth penalized logistic regression model with these reduced predictors and compare the statistical and predictive performance of the resultant model. The study aims to analyze the ability of this combined framework to deliver stable, parsimonious and interpretable binary classification results in the context of severe small-sample settings.

2. Materials and Methods

2.1 Study Design and Dataset

The design of this study was secondary quantitative using an open-access dataset of 23 older adults. Demographic, health-related, and questionnaire variables were included in the dataset. The study outcome was the binary variable "class_escore" and eight items of the questionnaire were chosen to be reduced in dimensions and modelled [17].

2.2 Outcome Variable

The dependent variable class_escore was binary coded (0, 1). The outcome distribution was nearly balanced with 11 observations in class 0 and 12 observations in class 1. The purpose of the modelling was to estimate the likelihood of class 1 as compared to class 0.

2.3 Predictor Variables

Eight questionnaire variables were included as the initial predictors: P_1_DF, P_2_DF, P_3_PS, P_4_PS, P_5_PSI, P_6_RS, P_7_MA, and P_8_MA. These variables represented multiple quality-of-life-related dimensions and were used as the input variables for PCA rather than being entered individually into the logistic regression model.

2.4 Data Preprocessing

The data were analysed for missing or incorrect values, value coding and reasonable ranges. No missing observations were identified. To put each questionnaire variable on a similar scale, the data were standardized before PCA. To determine dimensionality reduction suitability their distributions and pairwise correlations were also examined.

2.5 Principal Component Analysis

To reduce dimensionality and to minimise multicollinearity, PCA was carried out on the 8 standardised variables of questionnaire. Kaiser-Meyer-Olkin statistics were used to evaluate the sampling adequacy. Examination of eigenvalues, explained variance, scree pattern and component loadings. Thus, following the analytical approach, the first two principal components (PC1 and PC2) were considered for regression modelling.

2.6 Penalized Logistic Regression

The class_escore was used as the response variable, and PC1 and PC2 as the explanatory variables, in a Firth penalized logistic regression. The model was class_escore~ PC1+PC2. Since there were only 23 observations in the sample, there was a high risk of small-sample bias, unstable estimates and separation so Firth regression was used. In such situations, the penalized likelihood method gives more stable estimates of the coefficients.

2.7 Model Evaluation

The results of the model were evaluated with regression coefficients, odds ratio, 95% confidence intervals, and p-values. The predicted probabilities were transformed to binary classifications with a 0.50 threshold. The performance of prediction was determined by accuracy, sensitivity, specificity, precision, F1 score and ROC-AUC. These measures of performance were regarded as exploratory due to the very small sample.

3. Results

3.1 Descriptive Characteristics

All 23 participants were included in the analysis as there were no missing observations found. The mean age of the participants was 68.00 ± 4.08 years (range: 60 - 74 years). The majority of observations were coded as 0 for sex (87.0%) and 26.1% as diabetic, and 60.9% as osteoporosis. The binary outcome was nearly equally divided, with 11 (47.8%) participants in the group class_escore = 0 and 12 (52.2%) in the group class_escore = 1. It was a very small data set, though, as can be seen in Table 1; it resulted in a nearly balanced binary response.

Table 1. Demographic, clinical, and outcome characteristics of participants

Characteristic	Category/Statistic	n (%) or Mean ± SD
Age, years	Mean ± SD	68.00 ± 4.08
Age, years	Range	60–74
Sex	Code 0	20 (87.0%)
	Code 1	3 (13.0%)
Diabetes	No (0)	17 (73.9%)
	Yes (1)	6 (26.1%)
Osteoporosis	No (0)	9 (39.1%)
	Yes (1)	14 (60.9%)
Education	Code 0	4 (17.4%)
	Code 1	12 (52.2%)
	Code 2	7 (30.4%)
class_escore	0	11 (47.8%)
	1	12 (52.2%)

The balanced class pattern is illustrated in Figure 1, with only a one-participant difference between the two outcome groups.

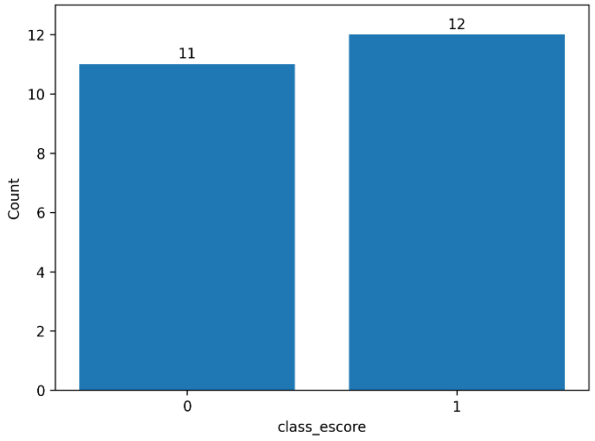


Figure 1. Binary outcome distribution

3.2 Distribution of Predictor Variables

There was middle-range variability of the eight predictor questionnaires among the participants. Mean scores ranged from 3.00 for P_7_MA to 3.87 for P_5_PSI. Most of the median values were 4, with the exception of P 2 DF and P 7 MA which had median 3. The observed ranges were typically in the 1 to 5 range as shown in Table 2, which is large enough to allow multivariate dimensionality reduction.

Table 2. Descriptive statistics of questionnaire predictors

Variable	Mean	SD	Median	Minimum	Maximum
P_1_DF	3.74	0.86	4.00	1	5
P_2_DF	3.13	1.10	3.00	1	5
P_3_PS	3.65	0.94	4.00	2	5
P_4_PS	3.70	1.15	4.00	1	5
P_5_PSI	3.87	1.22	4.00	1	5
P_6_RS	3.83	1.27	4.00	1	5
P_7_MA	3.00	1.00	3.00	1	5
P_8_MA	3.57	1.24	4.00	1	5

3.3 Correlation and PCA Suitability

The questionnaire variables demonstrated several moderate-to-strong positive correlations. The strongest relationships occurred between P_4_PS and P_6_RS (r = 0.776), P_3_PS and P_6_RS (r = 0.714), and P_5_PSI and P_6_RS (r = 0.633). In contrast, P_8_MA showed generally weak correlations with the remaining variables and a negative association with P_2_DF (r = −0.291). These relationships are illustrated in Figure 2.

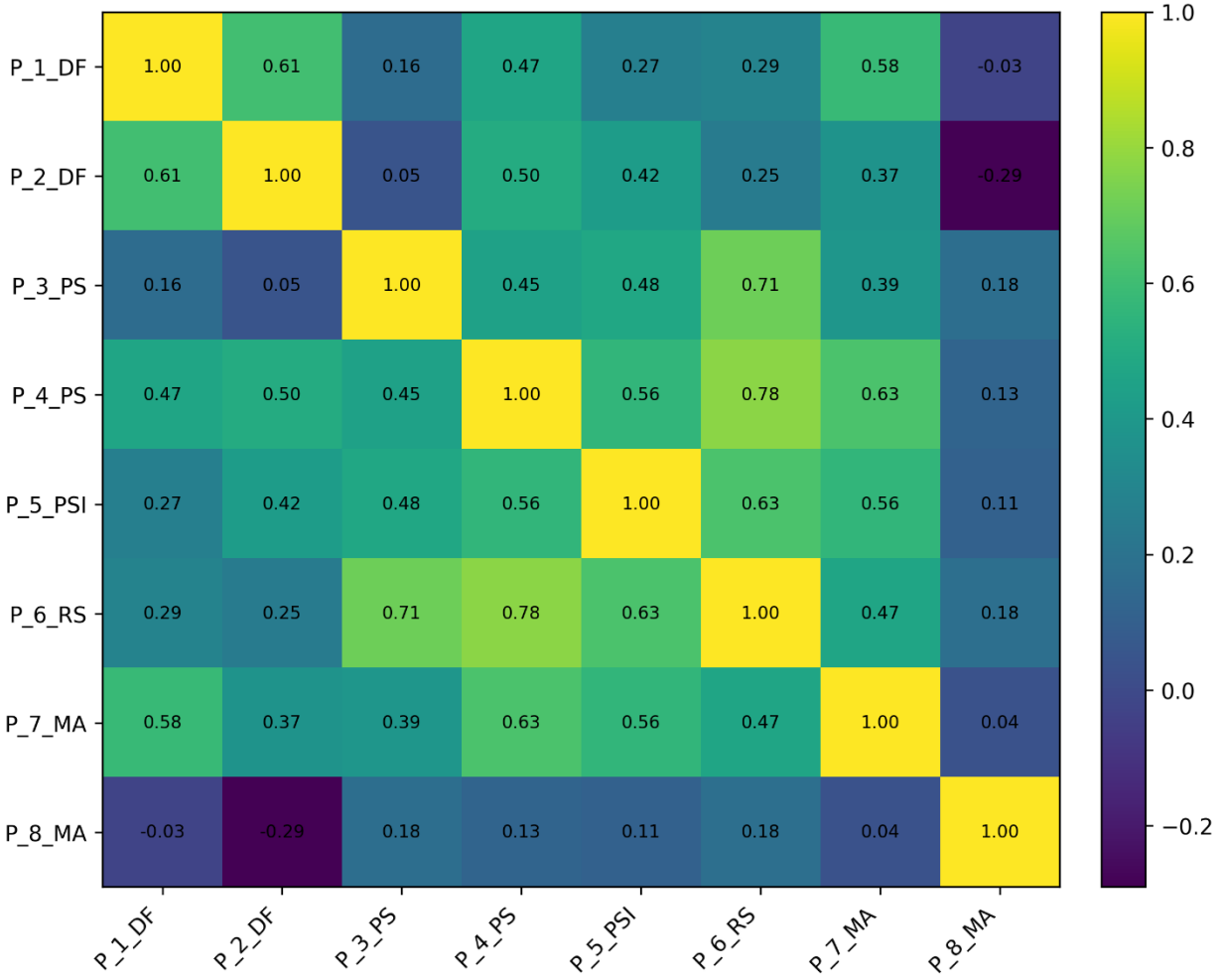


Figure 2. Correlation matrix of the eight questionnaire predictors

The Kaiser-Meyer-Olkin statistic overall was 0.581, which is borderline sampling adequacy. This value was somewhat

lower than the norm of 0.60 and thus was not very supportive of latent factor analysis. However, the observed correlation structure was consistent with the use of PCA as an exploratory dimensionality-reduction technique, consistent with the small-sample analytical framework.

3.4 Principal Component Analysis Results

PCA produced eigenvalues of 3.817, 1.557, 0.848, 0.566, 0.508, 0.412, 0.210, and 0.081 for PC1 through PC8, respectively. PC1 accounted for 47.71% of the total variance and PC2 accounted for 19.46% which made cumulative explained variance 67.18%. Smaller amounts of variance were accounted for by the other components individually. The scree and cumulative variance pattern is shown in Figure 3.

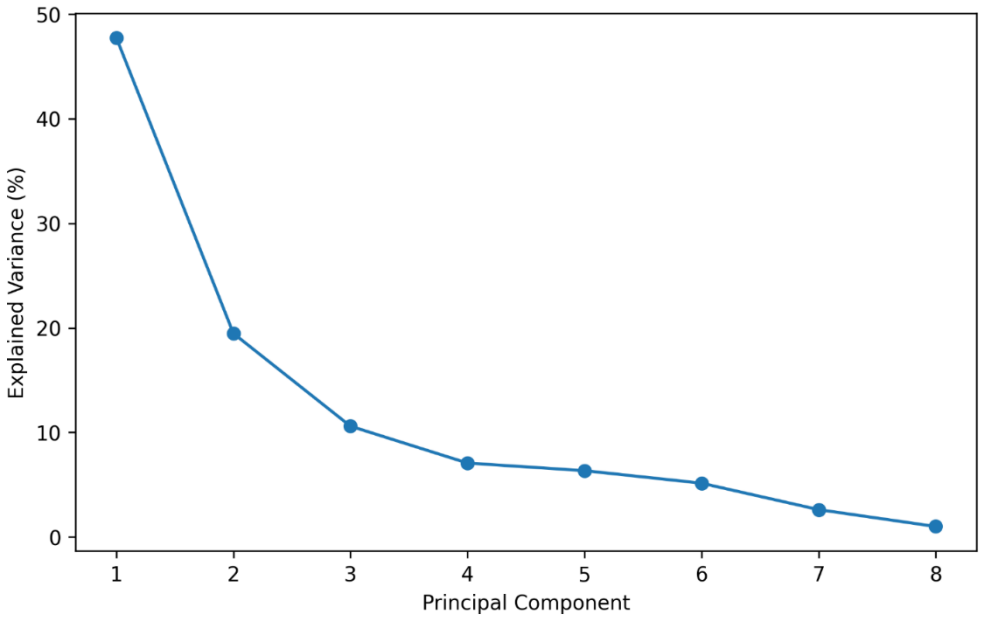


Figure 3. PCA scree plot and cumulative explained variance

PC1 showed positive contributions from nearly all questionnaire variables, with the largest loadings observed for P_4_PS (0.447), P_6_RS (0.424), P_7_MA (0.401), and P_5_PSI (0.396). PC2 was primarily characterized by positive loadings for P_2_DF (0.538) and P_1_DF (0.396), together with negative loadings for P_8_MA (−0.537) and P_3_PS (−0.403). The component-loading structure is summarized in Table 3.

Table 3. Loadings of questionnaire variables on PC1 and PC2

Variable	PC1 loading	PC2 loading
P_1_DF	0.321	0.396
P_2_DF	0.298	0.538
P_3_PS	0.328	−0.403
P_4_PS	0.447	−0.013
P_5_PSI	0.396	−0.095
P_6_RS	0.424	−0.290
P_7_MA	0.401	0.102
P_8_MA	0.055	−0.537

3.5 Firth Logistic Regression Results

The Firth penalized logistic regression model included PC1 and PC2 as predictors of class_escore. PC1 was strongly and positively correlated with the likelihood of class 1 membership as shown in Table 4. This had a coefficient of 2.814 and an odds ratio of 16.68. Because of the small sample size, the profile-likelihood 95% confidence interval for the association was very large (2.59–8191.73); however, the association was still statistically significant ($p < 0.001$). There was a small negative coefficient for PC2 which was not significant (OR = 0.77, $p = 0.792$).

Table 4. Firth penalized logistic regression results

Predictor	Coefficient	SE	Odds Ratio	95% CI for OR	p-value
Intercept	−0.639	0.974	0.53	0.02–3.14	0.486
PC1	2.814	1.252	16.68	2.59–8191.73	<0.001
PC2	−0.256	0.913	0.77	0.05–11.36	0.792

These findings indicate that the primary multidimensional pattern represented by PC1 carried most of the discriminatory information, whereas PC2 made little independent contribution to the binary outcome.

3.6 Classification Performance

The fitted Firth model classified all 23 observations correctly when a probability threshold of 0.50 was used. There were 11 true negatives and 12 true positives, and no false positives or false negatives. Thus, the in-sample accuracy, sensitivity, specificity, precision and F1Score were 100% each.

These results show complete separation of the two classes observed after the PCA-based reduction, but should be interpreted with caution. Only 23 observations were analysed with no independent validation sample. Furthermore, class_escore has a good alignment with the global score derived from the questionnaire, so it is structurally similar to the global score items of the questionnaire used to build PC1 and PC2. The best classification result is therefore a measure of in-sample discrimination, and not external validation of predictive accuracy.

3.7 ROC Analysis

An AUC of 1.000 was obtained for receiver operating characteristic analysis, which means there was complete discrimination between the two observed outcome categories within the analyzed sample. The ROC curve as shown in Figure 4 moved towards the upper-left corner, without the error.

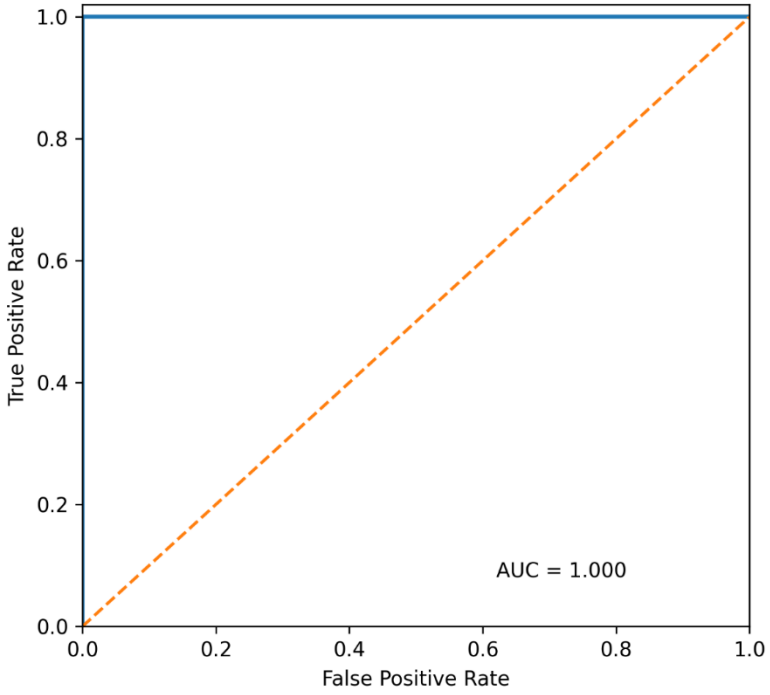


Figure 4. ROC curve for the PCA–Firth logistic regression model

The AUC result was also in line with the results obtained in the confusion matrix, but it was calculated from the training data set that is the same small data set used to build the models and, therefore, should not be interpreted as a perfect performance of the model in new participants. The results generally suggest that PC1 was the major component related to class_escore, PC2 offered little extra information and the PCA–Firth approach had a parsimonious description of the binary outcome under extreme small-sample situations.

4. Discussion

The current study assessed a combined principal component analysis and Firth penalized logistic regression approach for a very small binary-outcome dataset (N = 23). The eight questionnaire predictors were reduced to two principal components by PCA, which accounted for about 67% of the total variance. PC1 explained the highest amount of variance and was statistically significant with the binary outcome, while PC2 was not statistically significant. For the Firth model, discrimination was 100% in-sample, but this should be viewed with some caution due to the very small dataset size and the absence of external validation. However, limited sample sizes are known to reduce prediction-model stability even

when apparently strong discrimination is observed [18].

The use of PCA was beneficial as it reduced the dimensionality of the questionnaire, which comprised 8 correlated variables, into fewer uncorrelated components to make the following regression model more tractable. This method is especially helpful in scenarios with limited observations and a relatively small number of predictors [19,20]. PCA also offers a systematic approach to finding dominant patterns in multivariate data, that retain a minimum amount of redundancy among predictors [21]. However, PCA solutions of very small samples must be interpreted with care as component stability and loading patterns may be quite different across samples [22,23].

Firth penalized logistic regression was used because 23 observations were available and there was a significant problem of coefficient instability, separation under maximum likelihood estimation, and sparse-data bias. Firth regression is a penalised logistic regression that decreases the bias associated with first order estimation; it yields finite and more stable parameters when standard logistic regression is not successful. It has been used in applied health data contexts to show its usefulness in sparse or imbalanced analytical situations [24]. Penalized regression methods are also commonly used when information is less than complete for unrestricted multivariable modelling [25].

A combination of PCA and Firth logistic regression produced two alternative regularizations. The original predictor space was contractable by PCA, and small-sample bias in the estimation of binary outcomes was reduced by Firth regression. This approach resulted in a model of only two principal components instead of eight correlated questionnaire variables. When there are many more predictors than observations, this helps to make the model more efficient [26]. But in lower dimensions careful validation and uncertainty assessment are still necessary.

The results obtained here support the earlier studies demonstrating that penalized logistic regression proves to be beneficial in small or sparse data sets, though the models are still dependent on the sample size [25]. The dimensionality-reduction component is also aligned with research which underscored the effectiveness of PCA in reducing the number of correlated predictors [20,21]. However, changes in the data's characteristics also lead to lower out-of-sample accuracy, so the in-sample accuracy observed should not be used as a measure of out-of-sample accuracy [27].

The PCA–Firth approach could be applied in exploratory biomedical, behavioral and social-science research with a binary outcome and small numbers of observations. This yields a simpler and more interpretable model after regression, while reducing the dimensionality of predictors and avoiding overparameterization. However, sample size requirements and external validation principles should also be taken into account in model evaluation and not just apparent development-sample accuracy [15].

The main constraint was the very limited number of observations (n=23). The small KMO value also suggests that the structure of the PCA is tentative and exploratory in nature. Data were obtained from one dataset only, and no independent validation sample was found. Hence, coefficient estimates, component loadings, and discrimination measures may be unstable and the external generalizability is limited. The study therefore offers methodological evidence instead of definitive predictive validation.

The PCA-Firth model should be validated with larger independent datasets, and test for the stability of the retained components across populations. Comparisons with other methods for logistic regression, such as conventional logistic regression, ridge logistic regression, and LASSO logistic regression should be made to see if Firth penalization provides better small-sample performance. Other methods of dimensionality reduction should be explored as well, especially when predictor relationships are not linear or PCA assumptions are not strong. Larger validation samples are particularly important for reliable discrimination, calibration and generalizability.

5. Conclusion

This study demonstrated the applicability of a combined principal component analysis and Firth penalized logistic regression framework for binary outcome modelling under severe small-sample conditions. With just 23 observations, PCA was able to compress eight correlated predictor questions in a dataset to two principal components which captured about 67% of the total variance, simplifying the structure of the predictor variables and decreasing multicollinearity. PC1 was significantly associated with the binary outcome among the retained components, but PC2 did not provide much independent information. The use of Firth logistic regression was an appropriate substitute for the traditional maximum-likelihood estimation since it minimized the chances of small-sample bias, unstable coefficients and separation. The model had a 100 per cent in-sample classification, and an AUC of 1.00, but these results are to be used with care since the estimates are made with the same very small dataset on which the model was developed. The borderline KMO value and a lack of an external validation sample are further constraints to the generalizability of the findings. In general, the PCAFirth method provides a parsimonious and statistically suitable exploratory plan for small biomedical, behavioral and social-science data with predictors correlated with one another, and binary outcomes. Future investigations need to verify the stability of the retained components and predictive performance with larger, independent datasets, and comparisons to other penalized regression methods.

References

1. Intergovernmental Panel on Climate Change (IPCC). (2023). Climate Change 2023: Synthesis Report. Geneva: IPCC.
2. World Meteorological Organization (WMO). (2024). State of the Global Climate 2023. Geneva: WMO.
3. India Meteorological Department (IMD). (2024). Annual Climate Summary of India. New Delhi: Ministry of Earth Sciences.
4. Kumar, V., Jain, S. K., & Singh, Y. (2022). Analysis of rainfall variability and trends over India. Theoretical and Applied Climatology, 147(1–2), 341–356.
5. Dash, S. K., Jenamani, R. K., Kalsi, S. R., & Panda, S. K. (2021). Some evidence of climate change in twentieth-century

India. Climatic Change, 85(3–4), 299–321.
6. Goswami, B. N., Venugopal, V., Sengupta, D., Madhusoodanan, M. S., & Xavier, P. K. (2006). Increasing trend of extreme rain events over India. Science, 314(5804), 1442–1445.
7. Rajeevan, M., Bhate, J., & Jaswal, A. K. (2008). Analysis of variability and trends of extreme rainfall events over India. International Journal of Climatology, 28(14), 1837–1847.
8. Guhathakurta, P., Sreejith, O. P., & Menon, P. A. (2011). Impact of climate change on extreme rainfall events and flood risk in India. Journal of Earth System Science, 120(3), 359–373.
9. Pai, D. S., Sridhar, L., Rajeevan, M., Sreejith, O. P., Satbhai, N. S., & Mukhopadhyay, B. (2014). Development of a new high spatial resolution rainfall dataset for India. Mausam, 65(1), 1–18.
10. Krishnan, R., Shrestha, A. B., Ren, G., et al. (2020). Climate Change and Land: IPCC Special Report. Cambridge University Press.
11. Singh, D., Tsiang, M., Rajaratnam, B., & Diffenbaugh, N. S. (2014). Observed changes in extreme wet and dry spells during the South Asian monsoon season. Nature Climate Change, 4(6), 456–461.
12. Mishra, V., Kumar, D., & Ganguly, A. R. (2020). Climate change and hydrologic extremes in India. Journal of Hydrology, 589, 125–142.
13. Jain, S. K., Kumar, V., & Saharia, M. (2013). Analysis of rainfall and temperature trends in northeast India. International Journal of Climatology, 33(4), 968–978.
14. Chattopadhyay, S., & Edwards, D. R. (2016). Long-term trends in precipitation and temperature in eastern India. Atmospheric Research, 178–179, 146–157.
15. Pattanaik, D. R., Mohanty, U. C., & Kumar, A. (2019). Climate variability and changing monsoon characteristics over eastern India. Meteorological Applications, 26(3), 450–462.

EPH – International Journal of Mathematics and Statistics
https://eijms.com/index.php/ms/index

This work is licensed under a Creative Commons Attribution 4.0 International License. The authors retain ownership of the copyright for their article,You are free to: Share — copy and redistribute the material in any medium or format for any purpose, even commercially. To see the complete license contents,

please visit **http://creativecommons.org/licenses/by/4.0/.**