Linear Statistical Models Project

Statistical Modelling of CDC Diabetes Health Indicators Dataset

Swapnamoy Samadder BS2133
Deeptendra Banerjee BS2120
Rudrashis Bardhan BS2118

December 2023

1 Introduction

Diabetes is a globally prevalent chronic disease, imposing a substantial societal and economic burden. It hinders effective glucose regulation, leading to complications like heart disease and kidney issues. Mitigation strategies, including lifestyle changes and medical interventions, are crucial. Predictive models for diabetes risk play a vital role in global public health. With millions affected worldwide, awareness is imperative. The prevalence of diabetes varies across regions, influenced by socio-economic factors. The global economic impact is significant, emphasizing the urgent need for comprehensive strategies to address this widespread health challenge.

The Behavioral Risk Factor Surveillance System (BRFSS) is an annual health-related telephone survey administered by the CDC (Centers for Disease Control and Prevention, the national public health agency of the United States) since 1984. It gathers information from over 400,000 Americans regarding health-related risk behaviors, chronic conditions, and preventative service utilization.

This project seeks to create a good predictive model using the said dataset.

2 Data Description

The data has been taken from https://www.kaggle.com/code/alexteboul/diabetes-health-indicators-dataset-notebook. Our dataset consists of the following indicators:

Variable	*Description*
Numeric Variables	
MentHlth	Number of days mental health not good in the last 30 days
PhysHlth	Number of days physical health not good in the last 30 days
BMI	Body Mass Index
Categorical Variables	
Diabetes_binary	0 indicates absence, 1 indicates presence of diabetes
HighBP	0 indicates absence, 1 indicates presence of high blood pressure
HighChol	0 indicates absence, 1 indicates presence of high cholesterol
CholCheck	0 indicates no cholesterol check, 1 indicates cholesterol check
Smoker	0 indicates non-smoker, 1 indicates smoker
Stroke	0 indicates no stroke history, 1 indicates history of stroke
HeartDiseaseorAttack	0 indicates no history of heart disease or attack, 1 indicates history
PhysActivity	0 indicates low physical activity, 1 indicates high physical activity
Fruits	0 indicates low fruit consumption, 1 indicates high fruit consumption
Veggies	0 indicates low vegetable consumption, 1 indicates high vegetable consumption
HvyAlcoholConsump	0 indicates no heavy alcohol consumption, 1 indicates heavy alcohol consumption
AnyHealthcare	0 indicates no healthcare, 1 indicates healthcare received
NoDocbcCost	0 indicates having healthcare coverage, 1 indicates no healthcare coverage
DiffWalk	0 indicates no difficulty walking, 1 indicates difficulty walking
Sex	0 indicates male, 1 indicates female
Age	1: 18-24 years; 2: 25-29 years; 3: 30-34 years; 4: 35-39 years 5: 40-44 years 6: 45-49 years; 7: 50-54 years; 8: 55-59 years; 9: 60-64 years; 10: 65-69 years 11: 70-74 years; 12: 75-79 years; 13: 80 years or older
GenHlth	1: Excellent; 2: Very Good; 3: Good 4: Fair; 5: Poor
Education	1: Never attended school or kindergarten; 2: Grades 1 through 8; 3: Grades 9 through 11 4: Grade 12 or GED; 5: College 1 year to 3 years; 6: College 4 years or more

Variable	*Description*
Income	1.$< \$10,000$; 2. $\$10,000$-$\$15,000$; 3. $\$15,000$-$\$20,000$; 4.$\$20,000$-$\$25,000$ 5. $\$25,000$-$\$35,000$; 6. $\$35,000$-$\$50,000$; 7. $\$50,000$-$\$75,000$; 8. $\geq \$75,000$

3 Exploratory Data Analysis

3.1 Correlation Heatmap

We plot the correlation heatmap for the numeric variables. As depicted in Fig.1, it is evident that there is minimal correlation among the numeric covariates.

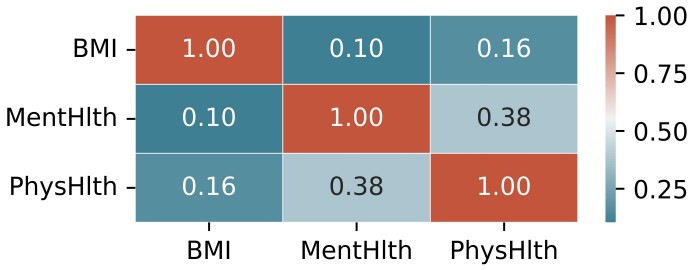

Figure 1: Correlation Heatmap

3.2 χ^2 test

We perform a χ^2 test for association between each categorical covariate and the variable *Diabetes_binary*. The results indicate that the p-values in all instances are approximately 0, suggesting a significant correlation.

4 Logistic Regression

Due to the binary nature of our response variable, we employ logistic regression as a suitable modeling technique. Logistic regression assumes that the responses, denoted as Y_i for the i^{th} observation, follow independent Bernoulli random variables with a parameter p_i.

$$Y_i \sim \text{Bern}(p_i) \implies E(Y_i|X_i = x_i) = p_i = P(Y_i = 1|X_i = x_i)$$

The basic idea of logistic regression is to use the mechanism already developed for linear regression by modeling the p_i's as a function of the covariates. Since p_i falls within the interval (0,1), a direct consideration of such a function is not feasible. Instead, we establish a "link" from (0,1) to the real numbers $\mathbb{R}$. This is known as a **link function**, essentially a bijective function from $[0,1] \to \mathbb{R}$. For Logistic Regression, the chosen link function is $\ln\left(\frac{p}{1-p}\right)$. Thus, our model takes the form:

$$\ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots \beta_k x_{ik} + \varepsilon$$

Here, there are a total of k covariates, and the task is to estimate the parameters $\beta_0, \beta_1, \ldots, \beta_k$.

4.1 Binary Cross Entropy

For a set of binary variables $(y_1, y_2 \cdots, y_N)$ and corresponding 'probabilities of occurrences' $(p_1, p_2, \cdots, p_N)$, Binary Cross Entropy (BCE) is defined as

$$\frac{-1}{N}\sum_{i=1}^{N} y_i \log(p_i) + (1-y_i)\log(1-p_i)$$

In our specific context, the choice of the BCE loss function is motivated by two key considerations:

(i) It strongly penalizes any misclassification errors in the estimates derived from our regression model.

(ii) BCE, being equivalent to $-N \times$ log-likelihood, ensures that minimizing the BCE loss is equivalent to maximizing the likelihood function.

4.2 Preliminary Model fitting

Prior to model fitting, the dataset was partitioned into training and testing sets at a ratio of 4:1. Subsequently, categorical variables were replaced with their respective dummy variables. Following this preprocessing step, the model was trained using the training set and evaluated on the testing set. The outcomes are presented in the tabulated results found in Table 5.1.1

Table 4.2.1: Confusion Matrix

	Predicted Negative	Predicted Positive
Actual Negative	5145	1945
Actual Positive	1606	5443

Accuracy: 74.89%

BIC: 58215.713790518625

4.3 Selecting Sub-Models

With our dataset containing 45 covariates, obtained after substituting dummy variables, we embark on model selection from submodels derived from the complete model. This undertaking employs the Bayesian Information Criterion (BIC) and systematically removes covariates through a backward pass. Our findings are encapsulated in Fig. 2 and Fig. 3.

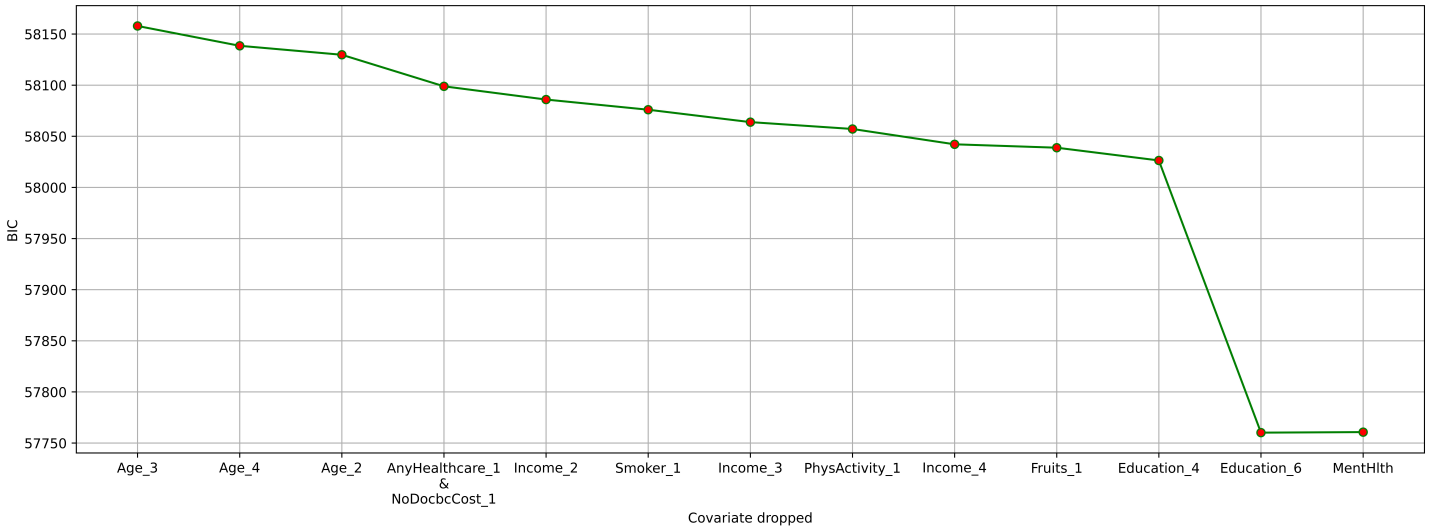

Figure 2: Covariates dropped vs BIC

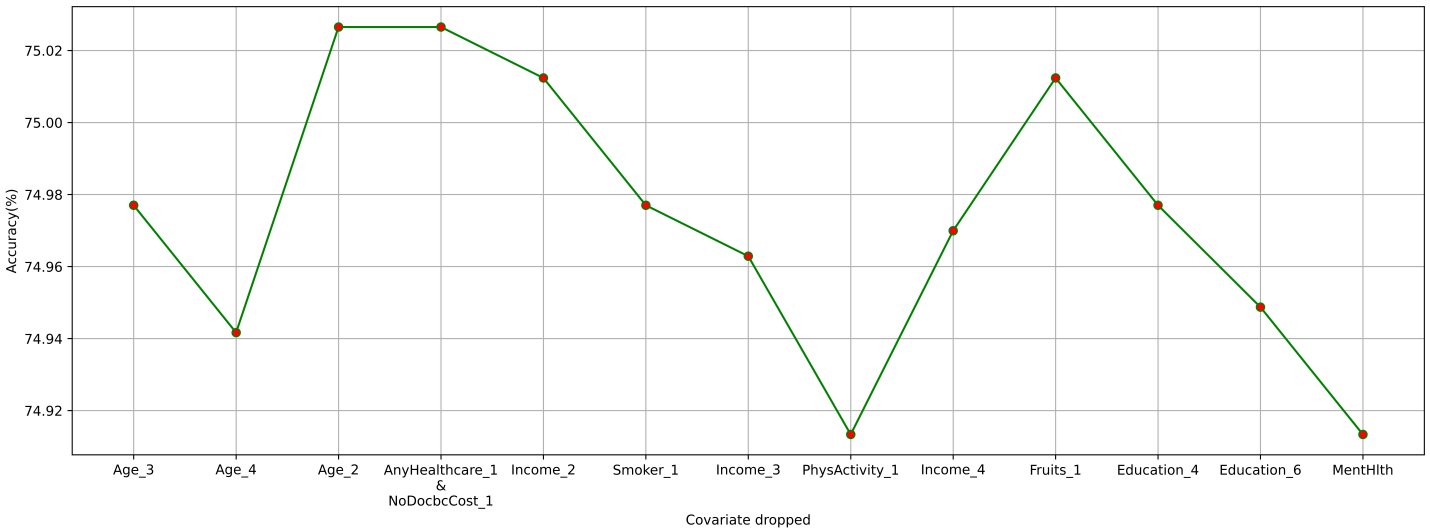

Figure 3: Covariates dropped vs Accuracy

From Fig.3, it is clear that each of the obtained submodels, perform better than the original model. Therefore, we obtain a model with improved accuracy compared to the initial model, yet with a reduced number of covariates.

5 Principal Components Regression

In the context of a generalized linear model using the columns of the data matrix X as covariates, denoted by $g(Y) = X\beta + \varepsilon$, the corresponding Principal Components Regression (PCR) model is represented as $g(Y) = U\theta + \varepsilon$. Here, the singular value decomposition of X is expressed as $X = U\Delta V^T$, and the coefficient vector θ is defined as $\Delta V^T \beta$.

Advantages of PCR:

(i) **Orthogonality and Reduction of Covariates:** The efficacy of PCR lies in the orthogonal nature of the columns of U and the decreasing variation along each component in the order of columns of U.

(ii) **Dimension Reduction:** When seeking to diminish the number of covariates, one can selectively drop columns in the order they appear in U. This reduction is performed while assessing if the removal of a column enhances the predictive power of the model.

(iii) **Multicollinearity Mitigation:** The orthogonal columns of U eliminate any potential issues of multicollinearity.

Although we do achieve a sub-model with enhanced accuracy, we aim to explore the possibility of obtaining a model with similar accuracy while further reducing the number of covariates. Consequently, the adoption of Principal Components Regression (PCR) represents a strategic approach to formulate a logistic model that maintains comparable accuracy and predictive power, yet utilizes an even smaller number of covariates.

5.1 Model fitting

Initially, we conduct Singular Value Decomposition (SVD) on our covariate matrix, and the resulting singular values are illustrated in Fig.4. The observed pattern raises concerns about potential multicollinearity.

Subsequently, we initiate our model selection process through a forward pass method, employing accuracy as the criterion. This entails sequentially adding principal components to our submodel. If the inclusion of a component leads to an improvement in predictive accuracy, we retain that predictor; otherwise, we omit the corresponding component. Following the incorporation of the first principal component, we achieve an accuracy of 64.5%.

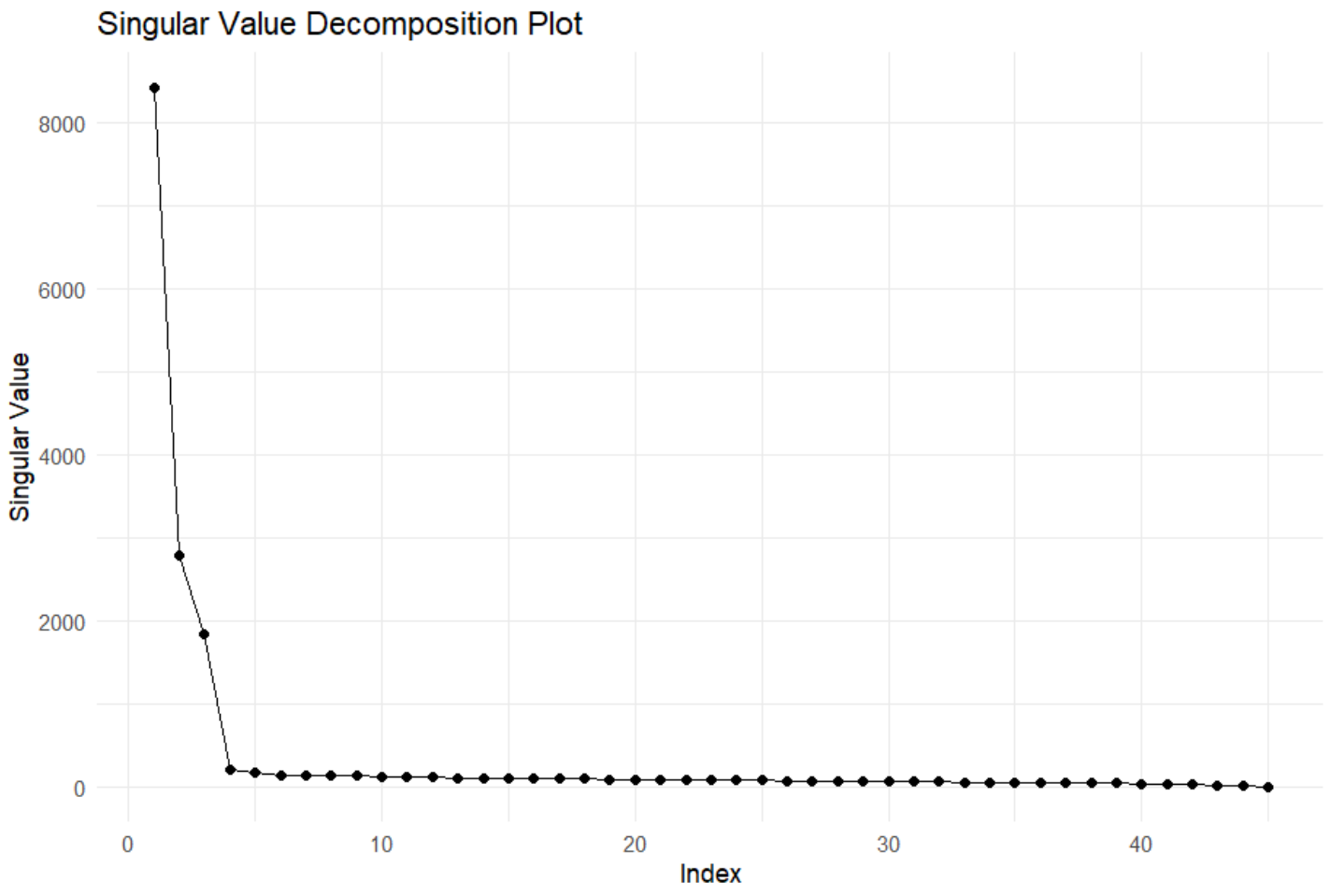

Figure 4: Singular Values

Table 5.1.1: Confusion Matrix

	Predicted Negative	Predicted Positive
Actual Negative	4570	2510
Actual Positive	2509	4550

Accuracy: 64.50%

We examine the plot illustrating the relationship between the component number and the corresponding accuracies, as depicted in Fig. 5. Notably, our analysis reveals the utilization of only 14 principal components, and within this selection, the submodel incorporating the first seven (from the chosen list) achieves an accuracy surpassing 70%. This marks an improvement in terms of reducing the number of covariates. Initially, the increase in accuracy is substantial, but within the middle range (components 11-36), the rate of improvement becomes notably slower. Interestingly, the last 50% of components contribute only a modest increase in the model accuracy, amounting to 2.5%.

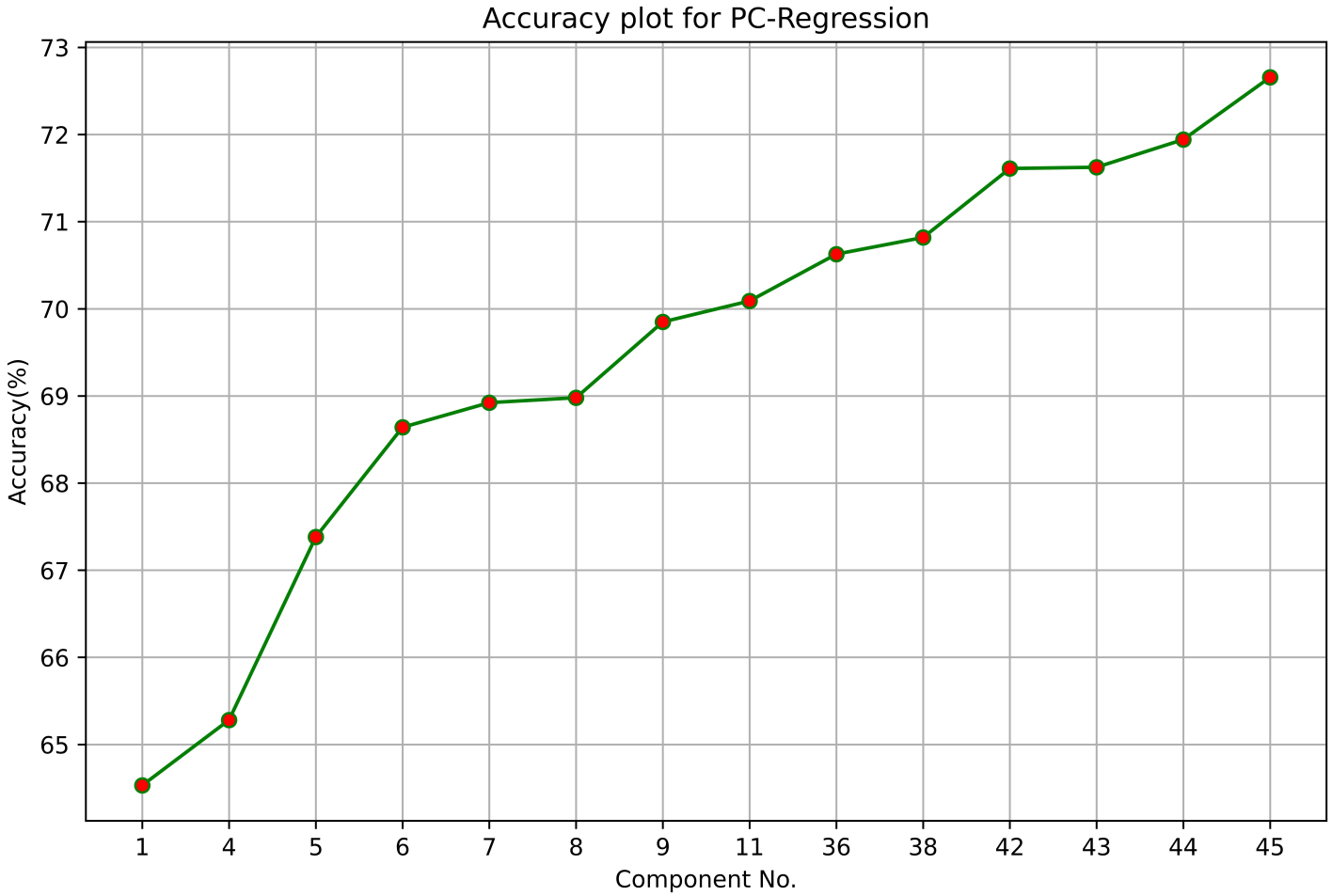

Figure 5: Accuracy in PCR

6 Roles

(i) **Swapnamoy Samadder:** Programming and Theory of PC Regression

(ii) **Deeptendra Banerjee:** Programming and Theory of Logistic Regression

(iii) **Rudrashis Bardhan:** EDA and Assistance in programming (SVD of design matrix)

7 References

- **Exponential Scheduler:** https://pytorch.org/docs/stable/generated/torch.optim.lr_scheduler.ExponentialLR.html

- **Binary Cross Entropy:** https://towardsdatascience.com/binary-cross-entropy-and-logistic-regression-bf7098e75559
- **ADAM Optimizer:** https://machinelearningmastery.com/adam-optimization-algorithm-for-deep-learning/