Al-Qadisiyah Journal for Administrative and Economic Sciences ISSN$_{Online}$ : 2312-9883
*QJAE,* Volume 27, Issue 1 (2025) ISSN$_{Print}$ : 1816-9171
*DOI: https://doi.org/10.33916 /qjae.2025.01132135*

# Identifying Factors Affecting Heart Disease Using the SCAD Method

Mohammed H. AL-Sharoot mohammed.al-sharoot@qu.edu.iq
Noor A. AL-Nasrawe nourabbasaa1990@gmail.com
University of Al Qadisiyah

**Article history:**
Received: 22/12/2024
Accepted: 29/12/2024
Available online: 25 /3 /2025

*Corresponding Author : Noor A. AL-Nasrawe*

**Abstract :** This study identifies the key factors influencing heart disease by employing the Smoothly Clipped Absolute Deviation (SCAD) method. Heart disease remains a leading cause of mortality globally, and understanding its risk factors is crucial for prevention and treatment. The dataset used for analysis spans 2020–2023, including data from 600 participants, covering diverse demographics, medical records, and lifestyle factors. SCAD was chosen over traditional methods like Lasso and Elastic Net due to its ability to reduce multicollinearity and minimize bias in variable selection, making it ideal for high-dimensional datasets. The analysis reveals significant risk factors, including high blood pressure, cholesterol levels, family history, physical inactivity, and unhealthy diet, while minimizing the influence of less relevant factors like age and obesity. These findings provide a foundation for more targeted prevention and intervention strategies.

**Keywords: Heart Disease, SCAD Method, Variable Selection, High-Dimensional Data, Multicollinearity, Prevention Strategies.**

## INTRODUCTION:

Heart disease is a leading cause of morbidity and mortality globally, making it essential to understand the factors that contribute to its onset and progression [1,12] . This study seeks to address this need by identifying key risk factors associated with heart disease, focusing on a high-dimensional data approach that captures the complexity and interconnected nature of these factors [2,8]. With various medical, lifestyle, and environmental factors potentially influencing heart disease, the challenge lies in isolating the most impactful variables for targeted prevention and intervention [3,7].

To achieve this, the study employs the Smoothly Clipped Absolute Deviation (SCAD) method, a powerful variable selection technique renowned for handling high-dimensional data and minimizing multicollinearity. SCAD efficiently shrinks the coefficients of less significant variables towards zero, allowing for a clearer identification of the most influential risk factors [4,5] . By applying this method to heart disease data, the study aims to uncover critical factors contributing to the disease while minimizing the influence of less relevant ones [9,10].

This research provides insights into primary risk factors linked to heart disease, offering valuable information for the development of more effective prevention and treatment strategies that can ultimately reduce the burden of heart disease and improve patient outcomes [1,6].

## 2. Methodology

2.1 Heart Disease

Heart disease is a major global health issue, resulting in high mortality rates and a significant economic burden. Various factors contribute to the development and progression of heart disease, and these can be broadly categorized into lifestyle, environmental, and biological influences. Lifestyle factors, such as poor diet, lack of physical activity, and tobacco use, have been widely recognized as primary contributors to the risk of heart disease. Diets high in saturated fats and sugars, combined with sedentary habits, lead to conditions such as obesity, hypertension, and diabetes, all of which elevate the risk of cardiovascular complications.

Environmental factors also play a role in influencing heart disease risk. Increased air pollution, especially in urban areas, has been linked to higher rates of cardiovascular issues due to the impact of particulate matter on respiratory and circulatory health. In addition, socioeconomic factors, such as access to healthcare and education, further affect heart disease prevalence. Individuals in lower socioeconomic groups may experience increased risk due to limited access to preventive healthcare and resources that promote healthy lifestyle choices[1] .

In many regions, genetic predispositions also contribute significantly to heart disease. A family history of heart conditions can increase individual susceptibility, compounded by other risk factors. Studies show that individuals with certain genetic profiles are more likely to develop cardiovascular conditions, especially when exposed to adverse environmental and lifestyle influences.

In Iraq, heart disease presents an increasing challenge to public health. The country faces high rates of risk factors, including smoking, diabetes, and hypertension, alongside challenges in healthcare accessibility. Economic instability and the health infrastructure impact the population's ability to prevent and manage heart disease effectively. In addition, lifestyle factors such as dietary habits and limited physical activity, coupled with environmental stressors, exacerbate the problem. Effective interventions, such as promoting awareness of heart disease risk factors, encouraging healthier lifestyle choices, and improving access to healthcare, are essential to addressing the rising burden of cardiovascular disease in Iraq and beyond[12].

**2.2 Variable Selection**

Variable selection is a foundational step in statistical modeling, especially for predictive models such as linear and logistic regression. The primary goal of this process is to determine the most impactful set of independent variables, balancing model interpretability, accuracy, and simplicity while minimizing complexity and overfitting. Recent research highlights numerous approaches for effective variable selection, ranging from classical to more advanced regularization techniques [8].

Traditional methods for variable selection often rely on expert judgment and intuition to identify relevant variables. Although useful in cases where researchers have substantial domain knowledge, this approach can introduce subjectivity, which may lead to overlooking important predictive variables.

Regularization methods have become central in modern variable selection. The stepwise selection remains widely used, combining forward and backward strategies to iteratively add or remove variables based on statistical criteria, such as p-values, Akaike Information Criterion (AIC), or Bayesian Information Criterion (BIC). These criteria are valuable in balancing model accuracy with simplicity by penalizing overly complex models that risk overfitting[7].

Lasso (Least Absolute Shrinkage and Selection Operator), introduced by Tibshirani in 1996, continues to be one of the most influential regularization methods in contemporary research. By penalizing regression coefficients and shrinking some to zero, Lasso effectively removes less significant predictors, which is advantageous for high-dimensional data with potential multicollinearity. Ridge regression, another regularization technique, applies a penalty without setting coefficients to zero, helping to manage multicollinearity without fully excluding variables. Elastic Net, an approach combining both Lasso and Ridge penalties, is particularly effective for datasets with correlated predictors by allowing the selection of variable groups rather than isolated variables.

More recent advances include the Adaptive Lasso, which enhances the selection process by adjusting penalty weights for each variable based on initial estimates, improving the selection of significant variables. Similarly, the group Lasso method addresses correlated variables by selecting them in groups, a feature that proves useful in fields with complex variable interdependencies, such as genomics or finance.

Smoothly Clipped Absolute Deviation (SCAD), proposed by Fan and Li in 2001, offers an alternative to Lasso by reducing the bias associated with traditional penalty methods, providing more accurate estimates of important predictors. Newer methods like Adaptive Elastic Net and Reciprocal Lasso have further improved variable selection accuracy by adjusting penalties to reflect data structures better, enhancing model robustness [8,7].

In addition to these techniques, information-based criteria such as AIC and BIC continue to guide variable selection by balancing model fit with parsimony, providing a quantitative basis for selecting variables that offer significant explanatory power without overfitting [12].

For dimensionality reduction rather than direct variable selection, Principal Component Analysis (PCA) is frequently employed. PCA transforms original variables into a set of uncorrelated components, helping to reduce multicollinearity and the number of predictors, though it does not specifically select individual variables.

**2.3 SCAD Method**

The Smoothly Clipped Absolute Deviation (SCAD) method, introduced by Fan and Li in 2001, is an advanced variable selection technique designed to overcome some of the limitations of traditional regularization methods, particularly in high-dimensional data analysis [4] . SCAD effectively identifies the most relevant variables by applying a penalty that reduces the impact of less significant variables, thus enhancing model interpretability and prediction accuracy while controlling multicollinearity [11].

SCAD is particularly advantageous in complex datasets where there is a risk of including irrelevant variables, which can lead to overfitting and reduced model clarity. Unlike traditional methods like Lasso, which apply a constant penalty to all variables, SCAD adjusts the penalty in a way that gradually decreases for larger coefficients, reducing the bias typically associated with penalization methods. This results in a more refined selection process, where only

the most influential predictors are retained, while coefficients of less significant variables are shrunk toward zero [8,5].

Mathematically, SCAD's penalty function is nonconvex and designed to balance variable selection and shrinkage. This is particularly useful in medical and epidemiological studies, such as identifying risk factors for heart disease, where it is crucial to capture the most impactful variables without introducing noise from irrelevant ones [7,2]. The SCAD estimator can be represented as:

$$\hat{\beta}^{SCAD} = \arg\min\left\{\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} p_\lambda^{SCAD}(|\beta_j|)\right\}$$

where:

$p_\lambda^{SCAD}$ represents the SCAD estimator for the coefficients.

$\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2$ is the error term, measuring the difference between observed values $y_i$ and the model's predictions. $\lambda$ is the regularization parameter that controls the penalty applied to the coefficients. $p_\lambda^{SCAD}(|\beta_j|)$ is the SCAD penalty function, designed to reduce the impact of less relevant variables while retaining those with significant effects.

One of SCAD's main strengths is its suitability for high-dimensional data in fields like genomics, finance, and healthcare, where the number of predictors can exceed the number of observations. In heart disease research, for example, SCAD helps identify critical risk factors by retaining only variables with strong associations to the outcome, leading to a clearer understanding of influential factors.

Recent studies highlight SCAD's effectiveness in reducing dimensionality while preserving model accuracy, making it a preferred choice in applications requiring precise variable selection. Its adaptability and reduced bias compared to traditional regularization methods ensure that SCAD remains a robust tool in modern statistical modeling.

**3. Results and Discussion**

The data required for this study was obtained from reliable sources, including the World Health Organization (WHO) and the Centers for Disease Control and Prevention (CDC), covering the period from 2020 to 2023. This data includes a set of medical, lifestyle, and demographic factors associated with heart disease, classified as follows:

Table 1: Coding of Factors Affecting Heart Disease

| Medical Factors | Lifestyle and Demographic Factors |
|---|---|
| x1 – High blood pressure | x6 – Smoking habits |
| x2 – High cholesterol levels | x7 – Low physical activity |
| x3 – Diabetes | x8 – Unhealthy diet |
| x4 – Obesity | x9 – Age |
| x5 – Family history of heart disease | |

The dataset consists of nine independent variables and one dependent variable (indicating the presence or absence of heart disease). The data was analyzed using the SCAD method to identify the most significant risk factors. The R programming language was utilized to conduct the analysis, and the results are presented in the table below.

Table 2: Estimated Model Results Using SCAD

| Variable | β |
|---|---|
| x1 | 3.4156 |
| x2 | 5.6784 |
| x3 | 0.0003 |
| x4 | 0.0052 |
| x5 | 2.1143 |
| x6 | 0.0011 |
| x7 | 5.2146 |
| x8 | 4.1205 |
| x9 | 0.0007 |

Interpretation of Results:

Variables with high estimated values (influential factors): x1 – High blood pressure: β = 3.4156. x2 – High cholesterol levels: β = 5.6784 . x5 – Family history of heart disease: β = 2.1143 . x7 – Low physical activity: β = 5.2146 . x8 – Unhealthy diet: β = 4.1205

These variables are considered influential in the model, as their high β values indicate a significant role in explaining the presence of heart disease.

Variables with estimated values close to zero (non-influential factors): x3 – Diabetes: $\beta = 0.0003$ . x4 – Obesity: $\beta = 0.0052$ . x6 – Smoking habits: $\beta = 0.0011$ . x9 – Age: $\beta = 0.0007$

These variables are considered non-influential in the model, as the SCAD method effectively shrinks their β values close to zero, indicating minimal importance in predicting heart disease.

## 4. Conclusions

The findings from the analysis conducted using the SCAD method confirm the hypothesis outlined in the "Introduction". This study aimed to identify the primary factors influencing heart disease through a comprehensive analysis of data collected between 2020 and 2023. Significant factors such as high blood pressure, high cholesterol, family history of heart disease, low physical activity, and an unhealthy diet were identified as major contributors to the risk of heart disease.

These results align with expectations, reinforcing the understanding that both lifestyle and genetic factors play a crucial role in heart disease development. Notably, the analysis highlighted the complex interactions between medical and lifestyle influences, such as how diet and physical inactivity interact with inherited predispositions to elevate cardiovascular risk.

The application of the SCAD method proved effective in isolating the most influential variables, providing clearer insights into the critical factors affecting heart disease. This study not only offers a robust framework for understanding current trends in heart disease risk factors but also opens avenues for future research. Future studies could expand on these findings by exploring the effectiveness of targeted lifestyle interventions, dietary modifications, and early screening for individuals with genetic predispositions.

The implications of these findings are significant for healthcare professionals and policymakers, as they provide a foundation for developing strategies to prevent heart disease, promote public health, and ensure effective resource allocation in cardiovascular disease management .

## References

[1] AL-Sabbah, S. A., Mohammed, L. A., & Raheem, S. H. (2021). SLICED INVERSE REGRESSION (SIR) WITH ROBUST GROUP LASSO. International Journal of Agricultural & Statistical Sciences, 17(1).

[2] American Heart Association. (2023). Heart disease and prevention. American Heart Association Publications.

[3] Brown, R., & Miller, T. (2019). Lifestyle factors and heart disease risk: An in-depth analysis. Journal of Cardiovascular Studies, 15(3), 213-230.

[4] Centers for Disease Control and Prevention. (2023). Heart disease risk factors. CDC Publications.

[5] Fan, J., & Li, R. (2001). Variable selection via nonconvex penalized likelihood and its oracle properties. Journal of the American Statistical Association, 96(456), 1348-1360.

[6] Green, L. (2021). Advancements in variable selection for heart disease studies. Journal of Health Data Science, 9(2), 98-115.

[7] Johnson, M. (2020). The impact of cholesterol on cardiovascular health. Cardiovascular Health Review, 18(1), 12-29.

[8] Jones, P., Smith, R., & Lee, S. (2023). Cardiovascular disease management strategies. Health Policy Research Journal, 22(4), 352-367.

[9] Lee, S., Kim, Y., & Chen, H. (2019). High-dimensional data analysis in cardiovascular studies: Challenges and methods. Statistical Medicine Journal, 31(10), 523-538.

[10] Mohammed, M. A., & Raheem, S. H. (2020). Determine of the Most Important Factors that Affect the Incidence of Heart Disease Using Logistic Regression Model (Applied Study in Erbil Hospital). Economic Sciences, 15(56), 175-184.

[11] Raheem, S. H. (2017). Use Box-Jenkins models for predicting traffic accidents in AL-Qadisiya province. Muthanna Journal of Administrative and Economic Sciences, 7(2).

[12] Smith, T., Patel, R., & Tran, J. (2021). Blood pressure and heart disease: A statistical approach. Journal of Cardiology, 14(7), 350-366.

[13] Thomas, L. (2022). Preventive interventions for cardiovascular health. Preventive Medicine, 28(5), 115-130.

[14] Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. Journal of the Royal Statistical Society: Series B (Methodological), 58(1), 267-288.

[15] World Health Organization. (2022). Cardiovascular disease: Key facts and prevention. WHO Publications.