Vol.4, No.7, Page 2398- 2409, 2026 https://doi.org/10.59628/jast.v4i7. 2876 ISSN 2958-9568 JAST

SUN PUBLISHER

مجلة جامعة صنعاء للعلوم التطبيقية والتكنولوجيا

Sana'a University Journal of Applied Sciences and Technology

https://journals.su.edu.ye/index.php/jast/

Comparison of Logistic Regression and Discriminant Analysis for Predicting Anemia Among Pregnant Women in Sana'a, Yemen

Amtalteef M.A. Al-Hamzi [1]* and Amani M.A. Al-Hamzi[2]

[1]**Department of Mathematical, Faculty of Education, Sana'a University, Sana'a, Yemen,**
[2]**Department of Pediatric, Faculty of Medicine, Sana'a University, Sana'a, Yemen**

*Corresponding author: alhmzi2018@gmail.com.

Abstract

This study aimed to identify the primary factors influencing anemia among pregnant women and to compare the classification efficiency of **Logistic Regression LR** and **Discriminant Analysis** DA. The study utilized a random sample of **(361) pregnant women** visiting Al-Sabeen Hospital during the period (2023-2024). Independent variables included demographic characteristics, reproductive history, and health habits. Results indicated that nutritional **Supplement Intake** was the most influential predictor for classifying anemia in both models, followed by **Body Missed Index BMI**, **Smoling, Age, Family size**, **N. of pregnancy**, and **N. of Births**. Conversely, education level and household income showed no statistically significant impact. In terms of performance, **LR** demonstrated superior classification accuracy at **88.1%**, compared to **86.9%** for **DA**. These findings confirm the high predictive and discriminatory capacity of the models in identifying anemic cases within the clinical women.

Article Info

Keywords:
Anemia, Pregnancy, Logistic Regression LR, Discriminant Analysis DA, Discriminative Functions DF, Classification.

Article History:
Received: 22-February-2026,
Revised: 16-April-2026,
Accepted: 10-May-2026,
Published: 28 July 2026.

1. INTRODUCTION

Anemia in pregnant women remains a critical global health challenge, particularly in developing economies. The World Health Organization (WHO) defines maternal anemia as a hemoglobin (Hb) concentration below 11.0 g/dL during the first and third trimesters and below 10.5 g/dL during the second trimester [1]. Globally, nearly half a billion women of reproductive age are affected by this disease. Estimates from 2011 indicate that approximately 38% of pregnant women (4.32 million) aged 15– 49 years suffer from anemia, with nearly 20% of maternal mortality cases attributed to its complications [2]. Anemia in pregnant women is clinically classified into three levels: severe (< 7.0 g/dL), moderate (7.0–9.9 g/dL), and mild (10.0–10.9 g/dL) [3]. Statistical analysis is a vital component of scientific research in the health sciences, where researchers often encounter numerous intertwined variables. Medical studies emphasize a strong link between diagnostic aspects and statistical applications, which accelerate disease diagnosis. Multivariate methods, specifically discriminant analysis (DA) and binary logistic regression (LR), are effective tools for classifying individuals into groups based on predictive variables. For statistical modeling using Binary Logistic Regression and Two-Group **DA**, these clinical classifications were collapsed into a single dichotomous category. The dependent variable was operationalized as follows anemic (**Hb < 11.0 g/dL)** and non-anemic (**Hb $\geq$ 11.0 g/dL)** based on several predictors: **(Age, Education Level, Family Income, Family size, Age at Marriage, Smoking, N. of Pregnancy, N. of Births, N. of Miscarriages, Taking dietary Supplement during Pregnancy, weekly consumption of Vegetables, Fruits, meat and Tea Consumption), and Body Mass Index (MBI)**. This methodological approach ensured a robust analysis of the predictive factors influencing the presence or absence of the condition.

1.1. Research Problem

The primary challenge in managing maternal anemia lies in the **multi-factor complexity** of its determinants, which complicates early and accurate diagnosis. While clinical thresholds are well established, there is a persistent gap in **statistically classifying and predicting** of at-risk cases based on overlapping biological and socioeconomic variables. This study addresses this gap by applying **LR** and **DA** to develop a high-precision predictive model that identifies the most significant risk factors, thereby enhancing clinical decision-making and improving maternal outcomes.

1.2. Research Objectives

1. **Identify Predictive Factors:**
To determine the significant variables influencing the classification of pregnant women (anemic vs. non-anemic).
2. **Model Comparison:**
To evaluate and compare the performance of **LR** and **DA** to identify the superior model for accurate predictive classification (Hit Ratio).

1.3. Significance of the Study

1. **Methodological:**
Comparing classification techniques to identify the most accurate model for biostatistical applications.
2. **Clinical:**
This model helps healthcare providers prioritize interventions based on significant risk factors, ultimately improving maternal and neonatal outcomes.

1.4. Research Methodology

This study adopted a **Descriptive-Analytical approach** combined with a **Survey method** for data collection. The methodological procedures were as follows.

- **Data Collection Tool:** A structured questionnaire covering various social, health, and demographic variables related to anemia was developed.
- **Population and Sample:** The study targeted pregnant women visiting **Al-Sabeen General Hospital for Maternity and Childhood**.
- **Timeframe:** The fieldwork was conducted between **October 2023 and March 2024**.

· **Ethical Considerations:** This study was conducted in accordance with the required ethical standards and received official clearance under the approval number **[Ref. No. 1056].** Informed consent was obtained from all participants prior to data collection. Stringent measures were implemented to ensure data anonymity and confidentiality, and all information was utilized strictly for academic purposes.

- **Statistical Analysis:** Data were analyzed using **BLR** and **DA** via **SPSS v.27** software.
- These statistical techniques were used to evaluate predictive accuracy and determine the optimal model for estimating the probability of infection based on clinically validated independent variables. All statistical tests were performed at a significance level ( ).

1.5. Inclusion- exclusion criteria

All pregnant women with known hemoglobin levels were included in this study, while those with unknown hemoglobin levels were excluded.

2. RELATED WORK

Al-Gazzar (2012) compared LDA and Logistic Regression using simulated data across various sample sizes. While Logistic Regression exhibited slightly higher classification accuracy, both models demonstrated a high degree of coefficient similarity and comparable performance in terms of Sensitivity, Specificity, and AUCcriteria [4]. Al-Nuwairi (2013) aimed to distinguish between diabetic patients with and without renal failure using Discriminant Analysis. This study identified the most significant factors contributing to the differentiation between the two groups. These findings led to the development of a discriminant function capable of classifying patients with a high classification accuracy of 91% [5]. "Suliman (2015) aimed to compare Discriminant Analysis, Logistic Regression, and Neural Network models in classifying observations. This study utilized independent variables affecting household income adequacy, including family size, housing tenure, and the presence of children in university. The results indicate that the discriminant function is statistically significant for only two variables: family size and housing tenure. Furthermore, the study concluded that Neural Networks outperformed both the Logistic Regression model and Discriminant Analysis in terms of classification efficiency [6]. Eshtiwi (2020) conducted a comparative study between Logistic Regression and Discriminant Analysis using a sample of 199 male and female students from schools in Al-Khums Municipality, Libya, including both infected and non-infected cases of Helicobacter Pylori. The study concluded that the proposed models yielded identical results regarding the statistical significance and importance of the independent variables. However, the **LR** method demonstrated a higher classification accuracy than **DF** [7]. Al-Nabawy (2023) conducted a comparative study between Discriminant Analysis and Logistic Regression to determine the most effective predictive method under various experimental conditions using Monte Carlo simulations for data generation. A logistic model capable of differentiating between the two methods was developed, achieving a remarkably high classification

accuracy of 98.5% [8]. Tirore (2024) investigated anemia determinants among pregnant women in sub-Saharan Africa, reporting a high prevalence of 51.26%. The study found that higher educational attainment and advanced age were protective factors, whereas multiparity (multiple births within 5 years), advanced gestational age (2nd and 3rd trimesters), and extreme poverty were significant risk factors for anemia susceptibility [9]. According to Tettegah (2024), the prevalence of pregnancy anemia in the district was 78.5%. The study found that 92% and 8% of pregnant women had excellent and good knowledge of anemia in pregnancy, respectively. It also identified several barriers to maintaining an appropriate hemoglobin level during pregnancy, such as long distances to healthcare facilities, non-intake of antimalarial drugs, and lack of nutritious meals. Finally, the study found that low education level, number of pregnancies, and number of children a woman had were significant determinants of anemia during pregnancy in the district [10]. Hua (2025) found that validated logistic regression models optimize diagnostic accuracy and risk estimation, provided that data integrity and rigorous assumption testing are maintained to prevent misclassification. By utilizing visual analytics, such as violin plots, and integrating clinical biomarkers, these models transform complex variables into actionable, clinically significant inferences [11].

3. THEORETICAL PART

Logistic Regression is a pivotal tool for analyzing categorical dependent variables with quantitative or mixed independent predictors. It is categorized into Simple or Multiple Binary Logistic models and Multinomial Logistic Regression If dependent variable contains more than two unorganized choices. Its primary advantage over Discriminant Analysis lies in its **flexibility,** as it does not assume normality for independent variables, making it highly effective in medical and economic research.

3.1. Logistic Regression LR Model

The LR provides each predictor with an independent coefficient for its contribution to the change in a dependent variable. The dependent variable Y will take a value of 1 if the response is "yes," and 0 if it is "no." The predicted probabilities were formulated using the natural logarithm of the odds ratio (OR) [12].

$$\ln\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_m X_{im} \quad (1)$$

where $\ln\left(\frac{P}{1-P}\right)$ is the natural logarithm of the odds of outcome Y (Y is a binary outcome). By exponentiating both sides, we convert the log odds into odds, turning the logarithmic expression into an exponential one. Then, we solve the probability by rearranging the equation to isolate the probability term. This process leads to the logistic function (Eq. 2), which expresses the probability as a function of the exponentiated combination of the predictors [13, 14].

$$P(Y) = \pi(X_i) = \frac{\exp\left(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_m X_{im}\right)}{1 + \exp\left(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_m X_{im}\right)} \quad (2)$$

3.1.1. Statistical Assumptions for LR

The robustness of the logistic regression analysis is contingent on several fundamental assumptions. First, the dependent variable was **dichotomous**, representing a binary outcome. Second, the model assumes **independence of observations**, requiring that data points are not correlated or derived from repeated measures. Third, the **absence of multicollinearity** among independent variables is essential to prevent the inflation of standard errors and to ensure the precision of the estimated coefficients. Finally, a **sufficiently large sample size** is required to achieve parameter stability and provide the necessary statistical power for reliable inference**.**

3.2. Maximum Likelihood Estimation (MLE)

The parameters of the **LR** model were estimated using the pseudo-maximum likelihood or weighted maximum likelihood method. The pseudo-log-likelihood function is approximately the likelihood function of a finite population based on the observed samples and known sampling weights. For the binary **LR** model, the likelihood function is given by

$$f(X,\beta) = L(\beta_0,\beta_1,\ldots,\beta_k) = \prod_{i=1}^{n} f(x_i,\beta) = P_i^{y_i}(1-P_i)^{1-y_i} \quad (3)$$

The essence of the **(ML)** principle is to find the estimator $\hat{\beta}$ that maximizes the **Likelihood Function**. In other words, the MLE for parameter β is the specific value $\hat{\beta}$ that satisfies the following relationship [15, 16]:

$$f(X,\hat{\beta}) \geq f(X,\beta) \quad (4)$$

3.3. Evaluating the performance of the models

3.3.1. Wald Statistic

Proposed by Wald (1943), this statistic is used to evaluate the statistical significance of each estimated parameter (individual coefficients) obtained using the MLE method. It was specifically designed to test the following null hypothesis:

$$H_0 : \beta_i = 0 \quad \text{vs.} \quad H_1 : \beta_i \neq 0 \quad (5)$$

The Wald statistic is formulated as follows: [17, 18]

$$W = \left[\frac{\hat{\beta}_i}{\text{S.E.}(\hat{\beta}_i)}\right]^2 \quad (6)$$

$\hat{\beta}$ Estimated Coefficient (S.E.($\hat{\beta}_i$)) Standard Error.

3.3.2. Hosmer & Lemeshow (H & L) Goodness estimating equations (GEE)

The **(H & L) statistic**, which follows a **Chi-square distribution**, is utilized to evaluate the **Goodness-of-Fit** of the model. It assesses whether the model adequately represents the data by analyzing discrepancies between **observed and predicted values**. In this test, a **p-value greater than 0.05** indicates that there is no significant difference between the observed and predicted frequencies, suggesting that the model fits the data well [14, 19].

3.3.3. Determination factor R^2

These two statistics $R^2_{\text{Nagelkerke}}$ and $R^2_{\text{Cox-Snell}}$ are utilized to evaluate the **explanatory power** of the logistic regression model regarding the relationship between the independent variables and the binary response variable. Since the $R^2_{\text{Nagelkerke}}$ ranges from **0 to 1**, it is considered more reliable and interpretable than the $R^2_{\text{Cox-Snell}}$ whose upper limit typically does not reach 1. Both are defined by the following equations [20]:

$$R^2_{\text{Cox-Snell}} = 1 - \left[\frac{L_0}{L_1}\right]^{(n/2)} \quad (7)$$

$$R^2_{\text{Nagelkerke}} = \frac{R^2_{\text{Cox-Snell}}}{1 - (L_0)^{(2/n)}} \quad (8)$$

L_0 : The Likelihood Function for the Intercept-Only Model step 0.

L_1 : Likelihood function for the full model step 1. n: Sample size for LR model.

3.4. Odds Ratio (OR)

The Odds Ratio **OR** is a measure of association used to quantify the relationship between an independent variable (exposure) and a binary outcome. It represents how the odds of the outcome change with the presence (or increase) of a predictor variable.

1. OR = 1 indicates no association between the exposure and the outcome.

2. OR > 1 indicates a positive association, meaning the exposure is associated with higher odds of the outcome (risk factor).

3. An OR < 1 indicates a negative (inverse) association, meaning that the exposure is associated with lower odds of the outcome (protective factor).

In practical terms, the OR indicates how many times the odds of the outcome increase or decrease in one group compared to a reference group. The further the OR is from 1, the stronger is the association. Additionally, the interpretation should always consider the confidence interval (CI):

If the CI includes 1 $\rightarrow$ the association is not statistically significant.

If the CI does not include 1, the association is statistically significant [21].

4. DISCRIMINANT ANALYSIS DA

Discriminant Analysis is a powerful multivariate statistical technique used to classify observations into distinct, predefined groups based on a set of independent predictor variables. Its primary objective is to develop a Discriminant Function—a linear combination of predictors—that maximizes the variation between groups while minimizing the variation within groups**.** Assumptions and Applications of **DA**. To ensure the validity of **DA**, the following assumptions must be met:

(1) Multivariate Normality across groups

(2) Homogeneity of variance-covariance matrices, **typically assessed** using Box's M test [22].

(3) Sufficient and random sample size.

(4) Absence of multicollinearity among predictors.

While DA can be applied to multiple categories, the current study utilizes Two-Group (Binary) DA, as the population is divided into two distinct groups. The model employs Linear Discriminant Functions (e.g., Fisher's function) to achieve optimal group separation, a method widely applied across diverse scientific fields.

4.1. Linear Discriminant Function (LDF)

Assuming that two random samples are drawn from two normally distributed populations with sizes n_1, n_2 and characterized by m independent variables (x_1, x_2,...,x_m). Given the respective mean vectors $\bar{X}_1$ and $\bar{X}_2$, and a pooled variance-covariance matrix S a linear combination of observations can be constructed to distinguish between the two groups [23]. The Linear Discriminant Function (LDF) is formulated as follows:

$$Y = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m \quad (9)$$

Y: Binary dependent variable discminant value

β: Discriminatory function coefficients

4.2. Estimation of the DF Parameters

The estimation of the Linear DF aims to find the coefficient vector (β) that maximizes the ratio of between-group variance to within-group variance.

First: Data Preparation and Group Mean Vectors: Calculate the Mean Vector($\bar{X}_1$ and $\bar{X}_2$) for each group across all independent variables.

Second: Calculating Sum of Squares and Cross-Products: Compute the **Within-Group Matrices (w_1 and w_2)** and the **Between-Group Matrix** (M). These matrices quantify the dispersion of data points around their respective group means and the distance between the groups.

Third: Estimating the Pooled Covariance Matrix (S): To ensure a stable linear estimation, the group covariance matrices are combined into a single **Pooled Variance-Covariance Matrix** (S), assuming that the groups have equal dispersions (homoscedasticity):

$$\underline{S} = \frac{(n_1-1)\underline{S}^{(1)} + (n_2-1)\underline{S}^{(2)}}{n_1+n_2-2} \tag{10}$$

$\underline{S}$: Variance and covariance matrices within the two groups

Fourth: Computing Discriminant Coefficients (Weights)

The coefficient vector (β) is estimated by multiplying the inverse of the pooled covariance matrix ($S^{-}1$) by the difference between the two mean vectors. Thus, the estimated linear LDF is given as the equation [24]:

$$\hat{Y}_i = \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \cdots + \hat{\beta}_m X_{mi} \quad ; \quad i = 1,2,\ldots,n \tag{11}$$

4.3. Selection of Predictors in D A

The selection of the most significant independent variables was based on their ability to maximize group separation. The process follows the rigorous steps described below.

4.3.1. Assessing Overall Significance (Wilks' Lambda)

Wilks' Lambda (Λ) was used to test the null hypothesis that the means of the groups were equal across all predictors. It represents the ratio of the within-group sum of squares to the total sum of squares [25].

$$\begin{aligned} H_0 &: \underline{\mu_1} = \underline{\mu_2} = \cdots = \underline{\mu_m} = 0 \\ \text{VS} \quad H_1 &: \underline{\mu_1} \neq \underline{\mu_2} \neq \cdots \neq \underline{\mu_m} \neq 0. \end{aligned} \tag{12}$$

where the calculated value of (Λ) as:

$$\Lambda = \prod_{i=1}^{m} \frac{1}{1+\lambda_i} \tag{13}$$

Where: λ_i is the eigenvalues of all variables, and m is the number of variables. A value of (Λ) near **0** indicates high group discrimination, whereas a value near **1** suggests no difference [26].

4.3.2. Individual Variable Testing (F-to-Enter)

Each variable was evaluated based on its **Partial F-statistic**. This measures the change in Wilks' Lambda when a specific variable is added to the model:

$$F_{\text{cal}} = \frac{n-k-m}{k-1} \cdot \frac{1-\Lambda}{\Lambda} \tag{14}$$

Criterion: A variable is included if its F-value is greater than a specified threshold, $F_{\text{cal}} > F_{\text{tab}}$,

where n is the sample size, m is the number of variables, and k is the number of groups [27, 28].

4.4. Classification and the Optimal Cut-off Value

To categorize observations into their respective groups, the model establishes a decision rule based on **group centroids** [29]. First, the centroid for each group is calculated by substituting the group mean vectors into the discriminant function as follows:

$$Y_1 = \sum_{i=1}^{m} \beta_i \bar{X}_{1i} \quad \text{And} \quad Y_2 = \sum_{i=1}^{m} \beta_i \bar{X}_{2i}$$

where Y_1 and Y_2 are the discriminant scores (centroids) for Groups 1 and 2, respectively.

The **optimal cut-off point** (Y_{critical}), which serves as the decision boundary, is determined as the weighted average of the two centroids:

$$Y_{\text{critical}} = \frac{n_1 Y_1 + n_2 Y_2}{n_1+n_2} \tag{15}$$

Decision Rule: For any new observation with a discriminant score Y_{Obs}

1. If $Y_{Obs} < Y_{critical}$, the observation is classified into Group 2.

2. If $Y_{Obs} \geq Y_{critical}$ the observation is classified into Group 1 [30, 31].

5. METHODOLOGY AND STATISTICAL ANALYSIS

A descriptive-analytical design was employed to investigate the predictors of anemia among a random sample of pregnant women (n=361) at Al-Sabeen Maternity and Childhood Hospital in Yemen. The participants were categorized according to a binary dependent variable into two groups: anemic (n=210) and non-anemic (n=151). A comparative study was conducted between the BLR and DA for data analysis.

Table (1) presents the independent variables incorporated into the study, classified into quantitative and categorical variables, along with their respective identification codes. For statistical analysis, categorical variables were dummy-coded as follows: Educational Level (1 = illiterate, 2 = read and write, 3 = secondary, 4 = university) and Smoking Status (1 = smoker, 0 = non-smoker). Regarding dietary behaviors, Supplement Intake was coded as (1= intake, 0 = no intake) and Post-meal Tea Consumption as (0 = no consumption, 1 = consumption). Additionally, the weekly frequency of fruit, vegetable, and

Table 1. Symbol and names of the independent variables nominated for the study

Symbol	Variables	Symbol	Variables
X_1	Age	X_8	N. of births
X_2	Education level	X_9	N. of miscarriages
X_3	Family income in R	X_{10}	Supplement Intake during pregnancy
X_4	Family size	X_{11}	N. of eat fruits a week
X_5	Age at marriage	X_{12}	N. of eat vegetables a week
X_6	Smoking	X_{13}	Tea consumption immediately after eating
X_7	N. of pregnancies	X_{14}	N. of eat meat of all kinds that per week
X_{15}	Body Mass Index (BMI)		

meat consumption was measured using a four-point ordinal scale (1 = never, 2 = once, 3 = twice, 4 = three times or more). The binary dependent variable was categorized based on hemoglobin **(Hb)** concentrations, where a value of '1' was assigned to 'Anemic' and '0' to 'non-anemic.

5.1. Logistic Analysis

Binary Logistic Regression BLR was employed to identify statistically significant predictors influencing the prevalence of anemia among pregnant women.

5.1.1. Chi-square Test for Model Significance

Table 2. Omnibus Tests of Model Coefficients

		Chi square	d.f.	Sig.
Step 1	Step	301.108	15	0.000
	Block	301.108	15	0.000
	Model	301.108	15	0.000

The results presented in Table (2) show that the chi-square test results are statistically significant (p=0.000 < 0.05), indicating that the predictive model is a significant improvement over the null model.

5.1.2. Goodness -of Fit of the model (H & L) Test)

Table 3. Test of Goodness-of-Fit

Step	Chi square	d.f.	Sig.
1	8.029	8	0.431
Source SPSS Version 27			

The Hosmer-Lemeshow test results indicate a good model fit, as the (p-value= 0.431) is greater than 0.05 (p>0.05). suggesting no significant difference between the observed and predicted classifications. study data, as further corroborated by the results in Table (4).

5.1.3. Model Explanatory Power and R-Square

The results in Table (5) illustrate the explanatory power of the estimated regression models. The $R^2_{\text{Cox and Snel}}$ were calculated at 56.6%, while the $R^2_{\text{Nagelkerke}}$ reached 76.1%. These findings indicate that the independent variables incorporated into the model accounted for approximately 76.1% of the variance in the dependent variable. These values signify a robust predictive capability and a high degree of fit between the model and data.

5.1.4. Variable Significance and LR Equation

Table (6) presents the Wald test results used to prioritize the independent variables based on their classification efficiency. Variables exhibiting statistical significance at the level of (α = 0.05) were ranked by their Wald statistics in descending order: parity, **Supplements Intak** ranked first, Wald = 60.6), BMI (37.62), smoking (16.43), and age (14.35). Other significant predictors included number of births (10.77), number of pregnancies (8.921), and family size (8.32). Non-significant variables were excluded from the final models. In addition, binary logistic regression analysis revealed that dietary supplementation was a **strong protective factor** against the risk of anemia among pregnant women **(OR = 0.037; 95% CI: 0.016–0.085)**. This indicates that the odds of developing anemia were reduced by **96.3%** in women who adhered to supplementation compared to those who did not. These findings underscore the **pivotal role** of nutritional supplements in maintaining optimal hemoglobin levels throughout the gestational period, also, (BMI) was also significantly associated with anemia among pregnant women **(OR = 0.629, 95% CI: 0.542–0.730)**. Since the odds ratio was less than one, BMI had a protective effect, where an increase in BMI was associated with a **37.1%** reduction in the odds of anemia. The confidence interval does not include 1, which confirms the statistical significance of this association.

The resulting **LR** equation is expressed as:

$$P(Y=1) = \frac{1}{1+\exp(} 0.45X_{15} - 2.53X_{10} - 1.69X_6 - 0.17X_1 - 0.36X_4 - 0.66X_7 + 0.63X_8)$$

Table 4. Observed and expected values of the logistic model

	Anemic		Not anemic		
	Observed value	Expected	Observed value	Expected	Total
1	36	35.840	0	0.160	36
2	36	35.168	0	0.832	36
3	28	31.698	8	4.302	36
4	27	24.707	9	11.293	36
5	14	13.253	22	22.747	36
6	6	6.617	30	29.383	36
7	4	2.367	32	33.633	36
8	0	0.999	36	35.001	36
9	0	0.313	36	35.687	36
10	0	0.038	37	36.962	37
Source: SPSS Version 27					

Table 5. Explanatory power of the estimated regression model

Step	Log Likelihood	Cox and Snell R Square	Nagelkerke R Square
1	189.667	0.566	0.761
Source: SPSS Version 27			

5.1.5. Classification Accuracy and Predictive Power

The classification matrix in Table (7) demonstrates the predictive accuracy of the model. The correct classification rate for the anemic subjects was 86.1% (sensitivity), whereas the rate for the non-anemic group was 89.5% (specificity). The model achieved an overall classification accuracy of 88.1%. These metrics demonstrated a robust discriminatory capacity and confirmed that the model adequately represented the empirical data.

5.2. Discriminatory analysis DA

5.2.1. Checking the conditions for applying DA

1. Assumptions: Normality Test.

The Kolmogorov-Smirnov test results indicated that most of the study's independent variables followed a normal distribution. Although certain variables, namely, educational level, family income, number of miscarriages, weekly consumption frequency of meat, and post-meal tea consumption, deviated from normality, the data were approximated to a normal distribution. This approximation is justified by the Central Limit Theorem (CLT), given that the sample size exceeds the threshold of 30 observations, which ensures the robustness of the subsequent parametric analysis.

2.Homogeneity of Covariance Matrices (Box's M Test)

The Box's M test resulted in a significance level of (P=0.08 > 0.05), confirming the homogeneity of the variance-covariance matrices across groups. This result satisfies the core assumption required for utilizing the Linear DF in the analysis.

3.Multicollinearity Test

Table (10) shows that all Variance Inflation Factor (VIF) values for the explanatory variables are less than 5, indicating the absence of a multicollinearity problem among these variables. indicates the absence of a multicollinearity problem among these variables.

5.2.2. Variable Significance (Wilks' Lambda)

The relative importance of each predictor was assessed using Wilks' Lambda and approximate F-values, with the results summarized in Table (11).

As illustrated in Table (11), **Supplement Intake** X_{10} **, BMI** X_{15} **Smoking** X_6 **,Age** X_1 **, N. of Pregnancies** X_7 **, Family Size** X_4 **,** and **N. of Births** X_8 emerged as the primary predictors in the discriminant model, yielding the most substantial **F-ratios** and minimum **Wilks' Lambda** coefficients. These variables were statistically significant at the **0.05 level**. In contrast, the remaining variables failed to meet the significance threshold (p > 0.05) and were omitted from the final function.

5.2.3. Significance of the DF

As indicated in Table (12), the DF is statistically significant, the Wilks' Lambda value was 0.476, χ^2=264.83) at a significance level of (P < 0.001), so that the null hypothesis is rejected, confirming the model's robust discriminatory power in group classification.

5.2.4. Canonical DF Coefficients

Table (13) ranks the predictors by their relative impact based on unstandardized coefficients. These weights define the DF model for group separation as follows:

$$\hat{Y} = -4.116 + 0.183X_{15} + 1.794X_{10} + 0.717X_6 - 0.043X_1 + 0.297X_7 - 0.063X_4 + 0.253X_8$$

Table 6. Variables in the Equation

Symbol	β	S.E.	Wald	d.f.	Sig	Exp(β)	95% C.I. EXP(β)	
							Lower	Upper
X_1	0.181	0.048	14.354	1	0.000	1.198	1.091	1.316
X_2	-0.087	0.214	0.164	1	0.685	0.917	0.603	1.394
X_3	0.000	0.000	0.227	1	0.634	1.000	1.000	1.000
X_4	0.317	0.110	8.307	1	0.004	1.373	1.107	1.703
X_5	-0.040	0.078	0.263	1	0.608	0.961	0.824	1.120
X_6	1.664	0.411	16.431	1	0.000	5.282	2.362	11.811
X_7	0.848	0.284	8.921	1	0.003	2.336	1.339	4.075
X_8	-1.005	0.306	10.770	1	0.001	0.366	0.201	0.667
X_9	-0.517	0.325	2.531	1	0.112	0.596	0.315	1.127
X_{10}	-3.295	0.423	60.555	1	0.000	0.037	0.016	0.085
X_{11}	-0.490	0.283	3.094	1	0.079	0.607	0.349	1.059
X_{12}	0.712	0.307	5.393	1	0.020	2.039	1.118	3.718
X_{13}	-0.368	0.420	0.766	1	0.381	0.692	0.304	1.579
X_{14}	-0.039	0.292	0.018	1	0.894	0.962	0.543	1.704
X_{15}	-0.463	0.076	37.617	1	0.000	0.629	0.542	0.730
Constant	7.922	2.712	8.534	1	0.003	2756.5	–	–
Source: SPSS Version 27								

Table 7. Model Classification Predicted Efficiency

Observed		Predicted		
		Anemic	Non–anemic	Total
Count	Anemic	130	21	151
	Non–anemic	22	188	210
%	Anemic	86.1	13.9	100
	Non–anemic	10.5	89.5	100
Overall Percentage		**88.1**		
Source: Researcher's analysis				

5.2.5. Classification Accuracy and Predictive Power

Table (14) presents the classification matrix used to evaluate the predictive validity of the model. The correct classification rate for anemic pregnant women was 85.4% (with a 14.6% error rate), whereas the accuracy for the non-anemic group reached 88.1% (with a 11.9% error rate). Overall, the discriminant function achieved a total classification accuracy of 86.9%, demonstrating its effectiveness in predicting group membership based on specified predictors.

Table (15) presents a comparative analysis of the classification accuracy between Binary LR and Linear **DA**. The results demonstrated that **LR** outperformed **DA** in terms of predictive capacity and overall classification hit rate (1.2%).

5.2.6. Model Validation

The **Hold-out** method was used as an internal validation technique by splitting the dataset into training and testing subsets to evaluate the predictive performance of the model and ensure the robustness of the comparative analysis between **LR** and Linear **DA.**

The results in Table (16) indicate that both models demonstrate strong and comparable predictive performance on the validation set. The **LR** model achieved a slightly higher overall accuracy (87.7%) than the DF model (85.9%), suggesting a marginal advantage in overall classification correctness. In terms of sensitivity, **DA** (87.6%) slightly outperformed **LR** (85.5%), indicating a better ability to correctly identify positive cases. However, **LR** showed higher specificity (89.7% vs. 83.8%), reflecting its superior performance in correctly classifying negative cases. Overall, the differences between the two models were relatively small, implying that both approaches provided robust and consistent classification performance. Therefore, the choice between them may depend on the study objective: prioritizing sensitivity favors **DA**, whereas prioritizing specificity and overall accuracy favors **LR**.

6. RESULTS AND DISCUSSION

First: Methodological Consistency and Variable Ranking both **LR** and **DA** yielded consistent results in identifying and ranking the key predictors of anemia among pregnant women. The variables were ranked in descending order of their statistical contribution as follows: **Supplement Intake, MBI**, **Smoking, Age**, number of **Pregnancies, Family Size**, and **N.** of births. This alignment between the two statistical approaches reinforces the reliability and robustness of the identified risk factors.

Second: Comparative Analysis and Model Selec-

Table 8. Kolmogorov-Smirnov Test for Normality Assessment

Variables	Statistical	Sig	Variables	Statistical	Sig
Age	0.984	0.077	N. of miscarriage	1.537	0.026
Education level	0.909	0.025	Supplement Intake	4.625	0.140
Family income	0.642	0.008	N. of eat fruits	1.709	0.081
family size	1.052	0.201	N. of eat vegetables	1.674	0.074
Age at marriage	1.307	0.056	Tea Consumption	1.235	0.001
Smoking	1.391	0.320	N. of eat meat	1.023	0.015
N. of pregnancies	0.672	0.063	BMI	2.34	0.130
N. of births	0.971	0.705			

Source: Researcher's analysis

Table 9. Results of Box's M test for the homogeneity of variance test

	F			
Box's M	**Approx**	$\mathbf{df_1}$	$\mathbf{df_2}$	**Sig**
87.646	1.609	15	416708.42	0.08

Source: Researcher's analysis

Table 10. Variance Inflation Factor (VIF)

Symbol	VIF	Symbol	VIF
X_1	1.238	X_9	2.127
X_2	1.322	X_{10}	1.174
X_3	1.358	X_{11}	1.451
X_4	1.189	X_{12}	1.686
X_5	1.132	X_{13}	1.161
X_6	1.168	X_{14}	1.430
X_7	4.048	X_{15}	1.205
X_8	2.794		

Source: Researcher's analysis

tion A comparison between the two statistical models reveals that the **LR model** outperforms the **DA** in terms of predictive accuracy. The **LR** achieved a higher overall classification rate of **88.1%** compared to **86.9%** for the **DA**. Furthermore, **LR** demonstrated superior sensitivity (**86.1%**) in correctly identifying cases of anemia. Consequently, **LR** was the preferred model for this study because of its higher discriminatory power and robustness. The **LR** model's superiority observed in this study is consistent with the findings of Al-Gazzar (2012) and Eshtiwi (2020). Furthermore, the high classification accuracy achieved in our results reinforces the findings of Al-Nuwairi (2013) regarding the effectiveness of statistical modeling in clinical differentiation. Our findings are consistent with those of a study [31] that identified smaller family size as a key protective factor against maternal anemia. Furthermore, a short inter-pregnancy interval was significantly associated with an elevated risk of anemia in mothers. Non-adherence to Iron and Folic Acid (IFA) supplementation significantly increased the risk of anemia **[RR: 1.53; 95% CI: 1.30–1.81]**, underscoring the critical impact of nutritional and reproductive factors on maternal health.

7. RECOMMENDATIONS

1.EnhancingNutritional Supplementation Programs: Prioritize and intensify awareness campaigns regarding adherence to Iron and Folic Acid (IFA) supplementation, as it is the most significant protective factor against maternal anemia.

2. Monitoring Maternal Body Mass Index (BMI): Integrate regular nutritional assessment and BMI monitoring into prenatal care protocols to ensure healthy weight gain, which significantly reduces the probability of developing anemia.

3. Implementing Smoking Cessation Initiatives: Develop targeted prevention programs to educate pregnant women on the risks of smoking, given its critical role as a major risk factor for maternal anemia.

4. Methodological Adoption: We recommend the adoption of **Logistic Regression** as a robust diagnostic and predictive tool in clinical settings because of its superior classification accuracy (**88.1%).**

Table 11. Significant differences among the means the groups

Variables	Symbol	Wilkes's Lambda	F	Sig
Age	X_1	0.920	31.32	0.000
Education level	X_2	0.993	2.707	0.101
Family income	X_3	1.000	0.179	0.672
Family size	X_4	0.938	23.66	0.000
Age at marriage	X_5	0.994	2.301	0.130
Smoking	X_6	0.837	69.65	0.000
N. of pregnancies	X_7	0.929	27.41	0.000
N. of births	X_8	0.981	6.953	0.009
N. of miscarriage	X_9	0.993	2.645	0.105
Supplement Intake	X_{10}	0.603	236.13	0.000
N. of eat fruits	X_{11}	0.995	1.642	0.201
N. of eat vegetables	X_{12}	0.997	1.029	0.311
Tea Consumption	X_{13}	0.999	0.231	0.631
N. of eat meat	X_{14}	9.992	2.835	0.093
BMI	X_{15}	0.729	133.1	0.000
Source: Researcher's analysis				

Table 12. Wilks's Lambda test for DF significance

Test of Function	Wilkes' Lambda statistic	Chi square	d.f.	Sig
1	0.476	264.83	7	< 0.001
Source: SPSS version: 27				

Table 13. Linear discriminant function coefficients

Variables	Symbol	$\hat{\beta}$
Constant		-4.161
Supplement Intake	X_{10}	1.794
BMI	X_{15}	0.183
Smoking	X_6	-0.717
Age	X_1	-0.043
N. of pregnancies	X_7	0.297
Family Size	X_4	-0.064
N. of births	X_8	0.253
Source: Researcher's analysis		

Table 14. Model Classification Predicted Efficiency

Observed		Predicted		
		Anemic	Non-anemic	Total
Count	Anemic	129	22	151
	Non-anemic	25	185	210
%	Anemic	85.4	14.6	100
	Non-anemic	11.9	88.1	100
Overall %		86.9		
Source: Researcher's analysis				

Table 15. Comparison of classification methods

Method				
	Logistic regression		Discriminatory Analysis	
%	correct classification	Error classification	correct classification	Error classification
Anamae	86.1	13.9	85.4	14.6
Non–Anamae	89.5	10.5	88.1	11.9
Overall %	88.1		86.9	

Source: Researcher's analysis

Table 16. Comparative Analysis of Predictive Accuracy: LR vs. DA (Validation Set, N=253)

Statistical efficiency index	Logistic Regression LR	Discriminatory Analysis DA
Accuracy	87.7%	85.9%
Sensitivity	85.5%	87.6%
Specificity	89.7%	83.8%

Source: Researcher's analysis

REFERENCES

[1] World Health Organization, *WHO Guideline on Use of Ferritin Concentrations to Assess Iron Status in Individuals and Populations*. Geneva: World Health Organization, 2020.

[2] G. A. Stevens, M. M. Finucane, and N. I. M. S. G. (Anemia), "Global, regional, and national trends in hemoglobin concentration and prevalence of total and severe anemia in children and pregnant and non-pregnant women for 1995–2011: A systematic analysis of population-representative data," *The Lancet Glob. Health*, vol. 1, no. 1, pp. 16–25, 2013.

[3] World Health Organization, *Prevention and Management of Severe Anemia in Pregnancy*. Geneva: World Health Organization, 1993.

[4] M. A. Al-Gazzar, "A comparative study between linear discriminant analysis and multiple logistic regression in classification and prediction," M.S. thesis, Al-Azhar University, Faculty of Economics and Administrative Sciences, Gaza, 2012.

[5] F. M. Al-Nuwairi, "Using the linear discriminant function to distinguish between diabetic patients with and without renal failure," M.S. thesis, University of Gezira, Faculty of Economics and Rural Development, Sudan, 2013.

[6] A. A. F. Suliman, "A comparison between discriminant analysis, logistic model, and neural network models in classifying observations," M.S. thesis, Sudan University of Science and Technology, College of Graduate Studies, Sudan, 2015.

[7] M. Eshtiwi and M. Al-Fiqi, "The optimal model for classifying helicobacter pylori infection among schoolchildren: A comparison between logistic regression and discriminant analysis," *Int. J. Adm. Sci.*, vol. 27, no. 27, pp. 1–20, 2020.

[8] A. Al-Banawi and Z. Bani Ata, "Comparison between the effectiveness of discriminant analysis and logistic regression in classifying cases into their groups," *Jordanian Educ. J.*, vol. 8, no. 3, pp. 314–337, 2023.

[9] L. L. Tirore et al., "Determinants of severity levels of anemia among pregnant women in sub-saharan africa: Multilevel analysis," *Front. Glob. Women's Health*, vol. 5, p. 1 367 426, 2024. DOI: 10.3389/fgwh.2024.1367426.

[10] E. Tettegah, T. Hormenu, and N. Ebu-Enyan, "Risk factors associated with anemia among pregnant women in the adaklu district, ghana," *Front. Glob. Women's Health*, vol. 4, p. 1 140 867, 2024. DOI: 10.3389/fgwh.2023.1140867.

[11] Y. Hua, T. S. Stead, A. George, and L. Ganti, "Clinical risk prediction with logistic regression: Best practices, validation techniques, and applications in medical research," *Acad. Med. & Surg.*, 2025. DOI: 10.62186/001c.131964.

[12] A. C. Rencher, *Methods of Multivariate Analysis*, 2nd ed. John Wiley & Sons, 2002.

[13] D. Dey, M. S. Haque, et al., "The proper application of logistic regression model in complex survey data: A systematic review," *BMC Med. Res. Methodol.*, vol. 25, p. 15, 2025. DOI: 10.1186/s12874-024-02454-5.

[14] K. L. Wuensch, *Binary logistic regression with spss*, https://core.ecu.edu/psyc/wuenschk/MV/MultReg/Logistic-SPSS.pdf, Accessed online, 2014.

[15] A. El-Habi and M. El-Jazzar, "Comparative study between linear discriminant analysis and multinomial logistic regression," *An-Najah Univ. J. for Res. (Humanities)*, vol. 28, no. 6, pp. 1526–1548, 2014.

[16] D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, *Applied Logistic Regression*, 3rd ed. New York: John Wiley & Sons, 2013.

[17] K. S. Taha and D. J. Hassan, "Using the logistic model to study the most significant factors affecting diabetes according to disease type," *Qalaai Zanist Sci. J.*, vol. 5, no. 4, pp. 622–643, 2020.

[18] J. Al-Sayyad, *Statistical Inference*. Riyadh, Saudi Arabia: Dar Al-Mareekh Publishing, 1993.

[19] K. F. Hamad, "Applying logistic regression model to study some statistical measurements in distinguishing shapes of mass in medical images," B.Sc. Thesis, Salahaddin University, Department of Statistical Sciences, 2016.

[20] A. Sliman, "Comparison between discriminant analysis, binary logistic regression and neural networks for classification observation," PhD Thesis, Sudan, 2015.

[21] J. F. Hair, R. E. Anderson, R. L. Tatham, and W. C. Black, *Multivariate Data Analysis*, 7th ed. New York: Maxwell International, 2009.

[22] K. I. Mouloud, "Using discriminant analysis to diagnose the most important factors influencing the clinical classification of heart disease," Master's Thesis in Statistical Sciences, Salahaddin University, College of Administration and Economics, Erbil, 2000.

[23] M. Jouda, *Basic Statistical Analysis Using SPSS*, 1st ed. Amman, Jordan: Dar Wael for Publishing, 2007.

[24] A. C. Rencher, *Methods of Multivariate Analysis*, 2nd ed. Brigham Young University: John Wiley & Sons, 2002.

[25] F. Al-Jaouni and A. Ghanem, "Multivariate statistical analysis (discriminant analysis) in characterizing and distributing households within the socio-economic structure," *Damascus Univ. J. for Econ. Leg. Sci.*, vol. 23, no. 2, 2007.

[26] M. A. Suleiman, "Comparison between discriminant analysis, the binary logistic model, and neural network models in classifying observations," PhD Thesis, Sudan University of Science and Technology, Sudan, 2015.

[27] R. D. Bock, *Multivariate Statistical Methods in Behavioral Research*. USA: McGraw-Hill, Inc., 1975.

[28] D. F. Morrison, *Multivariate Statistical Methods*, 3rd ed. New York: McGraw-Hill Book Company, 1976.

[29] A. M. Muhammad, "Using the linear discriminant function to distinguish children with and without diabetes," Master's Thesis, University of Gezira Wad Madani, Sudan, 2023.

[30] R. A. Johnson and D. W. Wichern, *Applied Multivariate Statistical Analysis*. Upper Saddle River, NJ: Prentice-Hall, 2002.

[31] T. G. Geta, S. Gebremedhin, and A. O. Omigbodun, "Prevalence and predictors of anemia among pregnant women in ethiopia: Systematic review and meta-analysis," *PLOS ONE*, vol. 17, no. 6, e0267005, 2022. DOI: 10.1371/journal.pone.0267005.