Democratic and People's Republic of Algeria
University of Abderrahmane MIRA- Bejaia
Faculty of exact sciences

Department of Operational research

Dissertation submitted for the obtention of a masters degree
in Applied Mathematics

Speciality: Data science and Decision Support

Logistic Regression Models: Theory and Applications

Presented by :
Njagi Kenedy Fundi

Supervised by : Dr S. Amroun

Defended on 30 /06/2025 , before the jury composed of :

M[r] N. Zougab	Professor	President	UAMB - Bejaia.
M[rs] K. Bouchebah	M.C. classe/ B	Examiner	UAMB - Bejaia
M[rs] L. Djeroud	M.C. classe/ A	Examiner	UAMB - Bejaia.

Academic year 2024 – 2025

Dedication

With great love, I dedicate this research to my late dad, beloved mom, twin , siblings, relatives and most specifically the family of Hon Genesio Njagi Mugo, Mr. John Muchiri, Rev Canon Difatha Nyaga and Mr Mutero for their overwhelming and great support throughout my academic journey.

To my best friends (Zidi Ziane, Anis Ikhlef), friends and classmates, your immeasurable support, love, care and kindness greatly contributed to my success.

May God bless you always.

Njagi Kenedy

Acknowledgments

This thesis reflects nothing but the climax of hard work, support and intellectual growth and hence I would wish to express my deepest gratitude to everyone that contributed to its success.

First and foremost, I am grateful to the Almighty God for the gift of life, good health and strength to achieve this great milestone.

Secondly, I extend my heartfelt gratitude and appreciation to my supervisor **Dr. Sonia Amroun** for her great supervision, guidance and unwavering support throughout the writing of this thesis. Her dedication, patience, time and intellectual assistance have been instrumental in shaping and improving the quality of this work.

Thirdly, I am profoundly grateful to the entire teaching staff at the University of Abderrahmane Mira - Bejaia for their dedication, insight, and unwavering support throughout my academic journey. Their passion for knowledge and excellence has left a lasting impact on me, and I deeply appreciate the role they have played in shaping my academic and professional development.

In particular, I would like to extend my heartfelt thanks to two exceptional individuals -**Dr. Larbi Asli** and **Dr. Djabri**. Beyond being outstanding educators, they have been remarkable mentors and friends. Their encouragement, support, and invaluable words of advice have inspired me to persevere and aim higher, especially in challenging moments. I am truly honored to have learned under their guidance.

I also wish to express my sincere appreciation to my beloved family and friends, especially the members of *Living Hope*, for their constant prayers, emotional support, and unwavering belief in my potential. Their presence and encouragement have been pillars of strength and motivation throughout this journey.

Finally, I acknowledge the contributions of fellow researchers and scholars whose work has laid the foundation for this study. Their tireless pursuit of knowledge continues to inspire and challenge the academic community to think critically, explore deeper, and contribute meaningfully to our fields.

To each and every one of you - thank you.

Table of Contents

List of figures

List of tables

General introduction

In the contemporary data-centric era, most institutions and businesses carry out vast research before taking decisions that could have an impact on their businesses in one way or the other. Most businesses are therefore required to make binary decision-making as to whether to adopt certain measures or not. A good example could be in the health sector where a lot of research is required before carrying out diagnosis of various diseases. The generalised linear models plays part among the various research models as it extends the ordinary linear regression to accommodate non-normal response variables (e.g., binary, count, or skewed data). [18]

Common models present in the generalised linear models include linear regression (Gaussian, identity link), logistic regression (Binomial, logit link), and Poisson regression (Poisson, log link) [11]. The use of logistic regression techniques is one of the methods that is employed in various sectors to classify, predict and explain several aspects[12]. What distinguishes the binary logistic regression from other generalised linear models is that the outcome has two levels. Example (presence or absence)[13]. It provides a straightforward way to model the decision boundaries hence making it possible to establish clear thresholds for action.

Within this context, the logistic regression techniques have been employed as it offers vigorous interpretability of the model, with the coefficients clearly depicting the degree of influence each predictor has on the probability of the outcome. The choice of applying the logistic regression model is evidently supported by its strong explanatory power, its computational efficiency and its simplicity to model binary and multiple outcomes hence making it ideal for industrial analysis.

This research therefore cites an application of binary logistic regression to predict the likelihood of the patients being diagnosed with diabetes based on clinical symptoms and other demographic factors. For our case the modelling is based on the diagnosis of the presence and absence of diabetes as a function of clinical symptoms as variables.

Similarly, multinomial logistic regression model has been employed in the prediction of the

choice of educational programs from the social economic status of the students and writing scores. It has been used to classify students' choice preferences from a set of three programs- General, academic, and vocational. The predictor social economic status used has three levels ie low, high and middle.

Finally, the ordinal logistic regression model has also been utilized to identify the risk levels of maternal mortality during pregnancy categorising them in three levels- **high,low and moderate risks.** The results are obtained by studying the key factors like age, blood pressure blood sugar levels and body tempareture.

By identifying the key factors and quantifying their impact, this analysis aims to enhance clinical decision-making and support in the different sectors that support classification of data and predictions.

The main research question driving this study is: ***How efficient is the logistic regression model in predicting and classifying the outcomes based on a given set of predictors?***

In addition, this central question is accompanied by the following secondary questions.

1. Which key factors greatly exposes and encourages the risk of an event happening?
2. How can the odds ratios derived from logistic regression coefficients provide actionable insights into the relationship between predictors and the possible outcomes?

This research emphasizes the central role of logistic regression as both a predictive and explanatory tool for various sectors. By transforming the collected data into actionable risk probabilities, the logistic regression model supports researchers in making timely and evidence-based decisions. The results of this research have implications not only for the health sector but also other institutions that heavily rely on the decision support systems.

To carry out the application of the logistic regression, our research is hence divided into three chapters.

Chapter one focuses on the introduction of linear models. It explores on the notion of linear models covering the simple linear regression models and multiple linear regression models which have link in the logistic regression practicability.

Chapter two discusses the logistic regression model into details. It covers the various tests and estimations carried out while carrying out logistic regression model.

Finally, the last chapter focuses on the application of the logistic regression using the real data set. It outlines the model's performance and the empirical results drawn from the study case.

1

Introduction to linear models

Linear models can be defined as statistical tools aimed at describing and understanding the relationship between the dependent and independent variables [22]. They specifically depict the changes on the response variables(known as dependent) while more independent variables (also known as predictors) vary. These models are employed in different domains and disciplines including economics among others and to whose purpose is mainly to :

- Predict future occurrences given a known dataset.
- Show correlations between variables using linear equations.
- Assess the impacts of an outcome given the different independent variables and finally
- Establish the insights into patterns in data.

For the linear regression model to be valid, it is assumed that the model is free from outliers, the data points are independent and that the distribution of these residuals should be normal with a mean of zero and a constant variance [7]

In the analysis of linear models, there exist the **simple linear regression models** and the **multiple linear regression models** which are the two main categories of linear models.

1.1 Simple linear regression models

Simple linear regression can be defined as a model with a single independent variable x (also called explanatory variable) that has a relationship and one dependent variable y (also called the response variable) that is a straight line[24].

For example by collecting data on the heights of trees and fitting it on a simple linear regression model, you can predict the number of leaves on the tree based on the height. In this case the height of the trees is considered the independent variable while the number of the leaves present on the trees is the dependent variable. This approach helps us understand the relationship and the impact of the tree sizes to the corresponding number of leaves on the tree.

The simple regression model can be expressed as :

$$y = \beta_0 + \beta_1 x + \varepsilon_i. \tag{1.1}$$

where:

- y is the dependent variable,
- x is the independent variable or explanatory variable,
- β_0 is the y-intercept term,
- β_1 is the slope parameter,
- ε_i is the random error component.

The parameters β_0 and β_1 are usually called the **regression coefficients.** The random error component ε_i in the model accounts for the failure of data to align itself on the straight line. This error clearly demonstrates and give the differences between the observed value of y and its actual true value. [21] The differences in the models could be as a result of inherent randomness of the model or the variables were qualitative.

From the model, it is assumed that the errors are uncorrelated which means that the value of an error is independent to the value of any other error. Other assumption made on the error term is that it is identically distributed and has a mean of zero that is $E(\varepsilon_i) = 0$ and a constant variance $Var(\varepsilon_i)=\sigma^2$. Later, we will additionally assume that ε_i is normally distributed.

The explanatory variables are viewed to be controlled by the analyst and hence considered to be non-stochastic while the dependent variable is viewed as a random variable. This means that there is a probability distribution for y given any possible value of x[4].

The mean of this distribution is therefore

$$E(y|x) = \beta_0 + \beta_1 x. \tag{1.2}$$

While the variance is given by

$$Var(y|x) = Var(\beta_0 + \beta_1 x + \varepsilon_i) = \sigma^2. \tag{1.3}$$

1.1.1 Estimation of parameters

The parameters β_0 and β_1 are unknown and therefore there's great need to have them estimated. In order to have these values determined, the **least square method** is therefore used[**\cite**]. This involves finding the sum of the squares of the differences between the observations y_i and the regression line where its minimum. While Eq.(1) is a simple regression model we can find terms of the n pairs of data (y_i, x_i) with $(i = 1, 2, \ldots, n)$ in order to proceed.

Thus the least square criterion is given as follows

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2 \tag{1.4}$$

The overall solution is therefore obtained as follows:

$$\begin{cases} \hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \\ \\ \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}. \end{cases} \tag{1.5}$$

where:

$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$ representing the mean of y_i ; and

$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$ representing the mean of x_i.

1.1.2 Properties of $\hat{\beta}_0$ and $\hat{\beta}_1$

In the simple linear regression realm the intercepts are crucial for understanding the relationship between the predictor and the response variables. Examining the statistical properties of these parameters, we gain deeper insights into their behaviour and reliability in the models. In order to find a foundational understanding of their central tendencies and dispersion, we

derive their expected values and variances. From the line of best fit represented as follows;

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x. \tag{1.6}$$

The variance is hence given by:

$$V(Y_i) = \sigma^2.$$

while the expected value is given by;

$$E(Y_i) = \beta_0 + \beta_1 x.$$

Expected value of $\hat{\beta}_0$ and $\hat{\beta}_1$

The expected value is mainly calculated in order to understand the central tendency of the coefficients. By assessing the $E(\hat{\beta_0})$ and $E(\hat{\beta_1})$ we can tell on average if our estimates align with the true population parameters. The subsequent derivations will reveal that under the standard assumptions of simple linear regression, both the estimated intercept and slope are unbiased of their respective true values.

Given:

$$\begin{aligned} E(\hat{\beta}_0) &= E(\bar{Y} - \beta_1 \bar{x}) \\ &= E(\bar{Y}) - \bar{x}E(\beta_1) \\ &= \beta_0 + \beta_1 \bar{x} - \bar{x}\beta_1 \\ &= \beta_0 \end{aligned}$$

The expected value for the estimator β_1 is hence calculated as follows:

$$\begin{aligned} E(\hat{\beta}_1) &= E\left(\frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right) \\ &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})E(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\ &= \frac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sum_{i=1}^{n}(x_i - \bar{x})\beta_1(x_i - \bar{x}) \\ &= \beta_1 \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\ &= \beta_1 \end{aligned}$$

The estimators β_0 and β_1 are therefore **non-biased**[23].

Variance of $\hat{\beta}_0$ and $\hat{\beta}_1$

Beyond calculations of the expected values, it is equally important to quantify the variability associated with the estimated coefficients. These variations provides a measure of the precision of the estimates and they indicate how they vary across different samples. These derivations moreover provide a tool for constructing confidence intervals and conducting hypothesis tests. The variance of $\hat{\beta}_1$ is hence given by:

$$Var(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

while the variance of $\hat{\beta}_0$ is presented as follows

$$Var(\hat{\beta}_0) = \frac{\sigma^2 \sum_{i=1}^{n} x_i^2}{n \sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

Therefore the distribution of β_0 and β_1 are represented as:

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right).$$

$$\hat{\beta}_0 \sim N\left(\beta_0, \frac{\sigma^2 \sum_{i=1}^{n} x_i^2}{n \sum_{i=1}^{n}(x_i - \bar{x})^2}\right).$$

1.1.3 Confidence intervals for β_0 and β_1

The confidence intervals in simple linear regression are conducted in order to provide a range of plausible values for the true population parameters. These calculations occurs in two different dimensions with assumptions on the variance of the error term as known and unknown which differs from each other with every scenario leading to a different distributional assumptions for our test statistics.

1^{st} case: Known variance (σ^2)

When the variance of the error term is assumed to be known, the construction of the confidence intervals β_0 and β_1 relies on the properties of the standard normal distribution. This assumption allows for straight forward calculations based on the known variance and the sampling distributions of our estimators.

The confidence interval for the Slope (β_1) is given by:

$$\beta_1 \in \left[\hat{\beta}_1 \pm z_{\alpha/2} \cdot \sqrt{Var(\hat{\beta}_1)}\right];$$

where;

$$Var(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

The confidence Interval for the Intercept (β_0) is given by:

$$\beta_0 \in \left[\hat{\beta}_0 \pm z_{\alpha/2} \cdot \sqrt{Var(\hat{\beta}_0)}\right];$$

where;

$$Var(\hat{\beta}_0) = \frac{\sigma^2 \sum_{i=1}^{n} x_i^2}{n \sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

2^{nd} case: Unknown variance (σ^2)

In most scenarios the true variance error is typically unknown and need to be estimated from the data provided. The construction uses the t-distribution accounting for the added uncertainty introduced by estimating the error variance.

The estimate of residual variance ($\hat{\sigma}^2$) is then provided as follows:

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2.$$

- **Confidence interval for the intercept (β_0)**

we compute standard errors of estimators to obtain

$$SE(\hat{\beta}_0) = \sqrt{\frac{\hat{\sigma}^2 \sum_{i=1}^{n} x_i^2}{n \sum_{i=1}^{n} (x_i - \bar{x})^2}};$$

we then use compute the critical value read from the student's distribution table using $n - 2$ degrees of freedom $t_{\alpha/2,n-2}$;

Combining the two equation computed above the confidence interval for β_0 is then computed as follows:

$$\beta_0 \in \left[\hat{\beta}_0 \pm t_{\alpha/2,n-2} \cdot \sqrt{\frac{\hat{\sigma}^2 \sum_{i=1}^{n} x_i^2}{n \sum_{i=1}^{n} (x_i - \bar{x})^2}} \right];$$

- **Confidence interval for the intercept (β_1)**

The standard error for the β_1 is given as follows:

$$SE(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}};$$

combining the standard error with the critical value formula based on the student distribution with $n-2$ degrees of freedom $t_{\alpha/2,n-2}$, we therefore obtain the formula for the confidence interval as:

$$\beta_1 \in \left[\hat{\beta}_1 \pm t_{\alpha/2,n-2} \cdot \sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}} \right].$$

1.1.4 Hypothesis testing for the simple linear regression estimators

In simple linear regression, we test two main hypotheses:

Null hypothesis (H_0): Tests whether the model as a whole is significant and that there's no relationship between the X and Y.

$$H_0 : \beta_1 = 0$$

.

Alternative hypothesis (H_1): Tests whether the slope (β_1) is significantly different from zero and that there exist a relationship between X and Y.

$$H_1 : \beta_1 \neq 0.$$

Test for β_0

The tests are based on a student distribution with $n-2$ degrees of freedom on H_0:

$$\mathrm{T}_{\beta_0} = \frac{|\hat{\beta}_0 - \beta_0|}{\sqrt{\frac{\hat{\sigma}^2 \sum_{i=1}^{n} x_i^2}{n \sum_{i=1}^{n} (x_i - \bar{x})^2}}} \sim t_{n-2};$$

We then compare the value of T_{β_0} with $t_{(n-2,\frac{\alpha}{2})}$:

If $\mathrm{T}_{\beta_0} > t_{(n-2,\frac{\alpha}{2})}$ we reject the **null hypothesis** i.e $H_0 : \beta_0 = 0$;

If $\mathrm{T}_{\beta_0} < t_{(n-2,\frac{\alpha}{2})}$ we accept the **alternative hypothesis** i.e $H_0 : \beta_0 \neq 0$;

Test for β_1

This test is based on the statistic distribution below:

$$\mathrm{T}_{\beta_1} = \frac{|\hat{\beta}_1 - \beta_0|}{\sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}} \sim t_{n-2};$$

T_{β_1} follows the student distribution with $n-2$ degrees of freedom on H_0.

We then compare the value of T_{β_1} with $t_{(n-2,\frac{\alpha}{2})}$:

If $\mathrm{T}_{\beta_1} > t_{(n-2,\frac{\alpha}{2})}$ we reject the **null hypothesis** i.e $H_0 : \beta_1 = 0$;

If $\mathrm{T}_{\beta_1} < t_{(n-2,\frac{\alpha}{2})}$ we accept the **alternative hypothesis** i.e $H_0 : \beta_1 \neq 0$.

1.1.5 Quality of adjustment

Coefficient of determination R^2

The most common metric of evaluating the goodness of fit is the R-squared. The analysis of the goodness fit helps us determine whether the relationship between the predictor variables and

the response variables are statistically significant. This metric represents the proportion of the variance of the response variable and the independent variable in the model. It is represented as follows:

$$R^2 = \frac{SSE}{SST} = \frac{\sum_{i=1}^{n} (\hat{Y}_i - \bar{Y}_i)^2}{\sum_{i=1}^{n} (Y_i - \bar{Y}_i)^2}.$$

Similarly the formula can be represented as:

$$R^2 = 1 - \frac{SSR}{SST} = 1 - \frac{\sum_{i=1}^{n} (\bar{Y}_i - \hat{Y}_i)^2}{\sum_{i=1}^{n} (Y_i - \bar{Y}_i)^2}.$$

where:
SSR is the sum of the square of residuals.
SST is the total sum of squares.
SSE is the explained sum of the squares.

Properties of R^2

From the mathematical formulation given above, we formulate the following interval,

$$0 \leq R^2 \leq 1$$

.

If $R^2 = 1$ means that the points of the sample are perfectly aligned in the line of best fit therefore confirming that the model is excellent.

If $R^2 = 0$ implies that the model doesn't align well with the sampled data and that there's no linear relationship.

Correlation coefficient r

The correlation coefficient measures the strength and direction of the linear relationship between two variables. The most commonly used correlation coefficient is **Pearson's correlation coefficient**, denoted by r.

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}} = \frac{COV(X,Y)}{\sqrt{Var(X)} \cdot \sqrt{Var(Y)}}.$$

Properties of r

The Pearson's correlation coefficient varies between the interval -1 and 1 represented as follows.

$$-1 \leq r \leq 1.$$

Remark: It is worth noting that the square of correlation coefficient equals the coefficient of determination:

$$R^2 = r^2.$$

If $r = 1$ it means that the points of the sample are perfectly aligned in the line of best on a positive direction therefore confirming that the model is excellent.

If $r = -1$ implies that the model perfectly aligns well with the sampled data but on the opposite direction.

If $r = 0$ implies that there exist no linear correlation with the sampled data in the model.

1.2 Multiple linear regression models

Multiple linear regression extends simple linear regression to include more than one explanatory variable. It estimates the relationship between a quantitative dependent variable and two or more independent variables using a straight line. The equation for multiple linear regression has the same form as that for simple linear regression but has more terms. It is however one of the most widely used statistical techniques for predictive modelling, hypothesis testing, and understanding complex relationships in data.

Model equation

In the multiple linear regression model, the relationship between the dependent variable Y (the response) and the independent variables $X_1, X_2, \ldots, X_p$ (the predictors) is expressed as

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon.$$

Here, β_0 represents the intercept, a constant term that captures the expected value of Y when all predictors are zero, while $\beta_1, \beta_2, \ldots, \beta_p$ are the coefficients associated with each predictor, quantifying their individual contributions to the response. The random error term, ϵ, accounts for the variability in Y not explained by the predictors and is assumed to follow a normal distribution with a mean of zero and constant variance σ^2, denoted as $\epsilon \sim N(0, \sigma^2)$. This assumption

ensures that the errors are symmetrically distributed around zero and have consistent variability, which is critical for the validity of statistical inferences drawn from the model. Together, these components define the structure of the linear regression framework, enabling the estimation and interpretation of relationships between predictors and the response. The matrix of the multiple linear regression is therefore represented as follows:

$$\mathbf{X} = \begin{bmatrix} 1 & X_{11} & X_{12} & \dots & X_{1p} \\ 1 & X_{21} & X_{22} & \dots & X_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & X_{n2} & \dots & X_{np} \end{bmatrix}; \quad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}; \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}; \quad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}.$$

The model is therefore written as:

$$Y = X\beta + \epsilon.$$

In the context of linear regression, the model is defined by several key components: the response vector $\mathbf{Y}$, an $n \times 1$ vector representing the observed outcomes; the design matrix $\mathbf{X}$, an $n \times (p + 1)$ matrix containing the predictor variables (including an intercept term), which is assumed to have full column rank to ensure the necessary algebraic conditions are met; the coefficient vector $\boldsymbol{\beta}$, a $(p+1)\times 1$ vector that parametrizes the relationship between the predictors and the response; and the error vector $\boldsymbol{\varepsilon}$, an $n \times 1$ vector with $\mathbb{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$ (indicating zero mean) and $\mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}_n$ (indicating constant variance σ^2 and uncorrelated errors).

1.2.1 Estimation of parameters

The estimation of errors is determined using the **least square estimate method**. It involves finding the regression function where the error term is minimal.
Therefore this can be expressed as:

$$\sum_{i=1}^{n} \epsilon_i^2 = \epsilon^\top \epsilon = ||Y - X\beta||^2 = (Y - \beta X)^\top (Y - \beta X).$$

Deriving the equation further, we obtain the estimated value of β as:

$$\hat{\beta} = (X^\top X)^{-1} X^\top Y.$$

The estimated value of Y from the expression are therefore represented as follows:

$$\hat{Y} = X\hat{\beta} = X(X^\top X)^{-1}X^\top Y.$$

1.2.2 Properties of $\hat{\beta}$

- $\hat{\beta}$ is an unbiased estimator of β and hence the expected value of $\hat{\beta}$ is equivalent to the corresponding value of β.

  This can be represented as follows:

$$\begin{aligned} E(\hat{\beta}) &= E((X^\top X)^{-1}X^\top Y) \\ &= (X^\top X)^{-1}X^\top E(Y) \\ &= (X^\top X)^{-1}X^\top X\beta \\ &= \beta \end{aligned}$$

  Thus, $\hat{\boldsymbol{\beta}}$ is an unbiased estimator for $\boldsymbol{\beta}$.

- Using linearity of expectation and $\mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}_n$:

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top)\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}.$$

  Since $\mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top) = \sigma^2\mathbf{I}_n$:

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top(\sigma^2\mathbf{I}_n)\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}.$$

  Simplify further:

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1} = \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}.$$

1.2.3 Variance and the confidence interval of the estimator $\hat{\beta}_j$

The variance of the j-th estimator $\hat{\beta}_j$ is the j-th diagonal element of $\mathrm{Var}(\hat{\boldsymbol{\beta}})$:

$$\mathrm{Var}(\hat{\beta}_j) = \sigma^2\left[(\mathbf{X}^\top\mathbf{X})^{-1}\right]_{jj}.$$

Case 1: Known variance σ^2

Distribution of $\hat{\boldsymbol{\beta}}$

Under normality:

$$\hat{\boldsymbol{\beta}} \sim \mathcal{N}\left(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}\right).$$

For the j-th coefficient β_j:

$$\hat{\beta}_j \sim \mathcal{N}\left(\beta_j, \sigma^2[(\mathbf{X}^\top\mathbf{X})^{-1}]_{jj}\right),$$

where $[(\mathbf{X}^\top\mathbf{X})^{-1}]_{jj}$ is the j-th diagonal element of $(\mathbf{X}^\top\mathbf{X})^{-1}$.

Standardization and confidence interval

Define the standard error (SE) of $\hat{\beta}_j$:

$$\text{SE}(\hat{\beta}_j) = \sigma\sqrt{[(\mathbf{X}^\top\mathbf{X})^{-1}]_{jj}}.$$

Standardizing gives:

$$Z_j = \frac{\hat{\beta}_j - \beta_j}{\text{SE}(\hat{\beta}_j)} \sim \mathcal{N}(0,1).$$

The $(1-\alpha)$ confidence interval (CI) for β_j is:

$$\beta_j \in \left[\hat{\beta}_j \pm z_{\alpha/2} \cdot \sigma\sqrt{[(\mathbf{X}^\top\mathbf{X})^{-1}]_{jj}}\right],$$

where $z_{\alpha/2}$ is the critical value from the standard normal distribution.

Case 2: Unknown variance σ^2

Estimating σ^2

The sum of the square of residuals (SSR) estimates σ^2:

$$s^2 = \frac{\text{SSR}}{n-p-1} = \frac{(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})^\top(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})}{n-p-1}.$$

The distribution of s^2 is:

$$(n-p-1)\frac{s^2}{\sigma^2} \sim \chi^2_{n-p-1}.$$

Student's t-Distribution

The standardized estimator follows the student's distribution with $n-p-1$ degrees of freedom and it becomes:

$$T_j = \frac{\hat{\beta}_j - \beta_j}{s\sqrt{[(\mathbf{X}^\top\mathbf{X})^{-1}]_{jj}}} \sim t_{n-p-1}.$$

This follows because:

- $\hat{\beta}_j$ is normal,
- s^2 is chi-squared,
- $\hat{\beta}_j$ and s^2 are independent.

Confidence intervals

The $(1-\alpha)$ confidence interval from a case where the variance is unknown for β_j is hence given by:

$$\beta_j \in \left[\hat{\beta}_j \pm t_{\alpha/2,n-p-1} \cdot s\sqrt{[(\mathbf{X}^\top\mathbf{X})^{-1}]_{jj}}\right],$$

The $t_{\alpha/2,n-p-1}$ is calculated as the critical value derived from the t-distribution table with $n-p-1$ as the degrees of freedom.

In statistical practice, the concept of known variance is largely theoretical, whereas unknown variance is more common and standard. To address the uncertainty in estimating the variance (σ^2), the t-distribution is used, as it provides a way to account for this variability. A critical assumption in this context is the normality of errors, which ensures the validity of the results. For large sample sizes (n), the t-distribution approximates the normal distribution ($\mathcal{N}$), making it easier to work with. Additionally, it is essential to ensure that the design matrix $\mathbf{X}$ has full column rank, as this guarantees the existence of the inverse $(\mathbf{X}^\top\mathbf{X})^{-1}$, which is crucial for parameter estimation.

1.2.4 Tests on the β_j parameter

In multiple linear regression analysis, testing the parameters of the model typically involves assessing the statistical significance of individual coefficients and the collective contribution of predictor variables. For individual parameters, the null hypothesis $H_0 : \beta_j = 0$ tests whether a specific predictor X_j has no linear relationship with the response variable Y, conditional on other predictors. This is evaluated using a t-test, where the test statistic $t = \frac{\hat{\beta}_j}{\mathrm{SE}(\hat{\beta}_j)}$ follows a t-distribution under H_0, with degrees of freedom $n-p-1$ (where n is the sample size and p is the number of predictors). The standard error $\mathrm{SE}(\hat{\beta}_j)$ is derived from the estimated error variance $\hat{\sigma}^2$, which incorporates the residual sum of squares.

Decision rules

Decisions made to declare whether the predictor variables are relevant and appropriate to the model are mostly guided by the testing of the statistical hypothesis. This involves the tests made on each of the regression coefficients β_j. The tests are therefore:

1. **Null hypothesis**:

$$H_0 : \beta_j = 0.$$

   This test implies that the predictor X_j has no significant linear relationship with the dependent variable Y.
   On the contrary we have;

2. **Alternative hypothesis**:

$$H_1 : \beta_j \neq 0.$$

   This test clearly signifies that the regression coefficients have meaningful contributions to the model

t-test

This tests is based on the *t-statistic* which following a student distribution with $n - p - 1$ degrees of freedom where n is the sample size and p is the number of predictors represented as follows:

$$t = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)} \sim \text{student distribution } (n - p - 1)$$

where $\hat{\beta}_j$ is the estimator and $SE(\hat{\beta}_j)$ is the standard error

The decision on whether to reject or accept the tested hypothesis involves the comparison of the calculated t-value with the critical value read from the student's distribution table, or alternatively comparing the corresponding p-values with a pre-specified significance level α (normally 0.05).

If the absolute value ($|t|$) is greater than the critical value ($t_{critical}$) or else the $p - value$ is less than the pre-specified α value, the null hypothesis is rejected, indicating that the predictor variable significantly in one way or the other affects the response variable Y [19].

Fisher-test

This test is based on the assessment of the significance of the model as a whole. It is based on the assumption that all the regression coefficients except the intercept are equal to zero as demonstrated below.

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0.$$

On the other hand the alternative hypothesis given as follows;

$$H_1 : \exists\, j \in \{1, 2, \ldots, k\} \text{ such that } \beta_j \neq 0.$$

The F-statistic is therefore mathematically computed as:

$$F = \frac{MSE}{MSR} = \frac{SSE/p}{SSR/(n-p-1)} \sim \text{Fisher distribution } F(p, n-p-1)$$

Where **SSR** is the sum of squares residuals, **SSE** is the explained sum of squares, **MSR** is the mean squared for residuals, **MSE** is the explained(regression) mean squared, **p** represents the number of predictor variables while **n** is the total number of observations.

The results are then calculated and compared using the critical value calculated from the Fisher's distribution table. If the calculated F-statistic is found having exceeded the critical value , or else if the p-value is less than the given α value, the null hypothesis is rejected, indicating that the model provides a better fit than a model with no predictors. [17].

1.2.5 Quality of adjustment

The quality of adjustment in multiple linear regression involves evaluating the variability in the response variable. This can be applied using the coefficient of determination R^2, defined as

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}.$$

whereby **SSR** represents the residual sum of squares and **SST** defining the total sum of squares. While R^2 increases with each added predictor, in some scenarios it can be misleading mostly with models with many variables. A better alternative is the use of adjusted R^2, given by

$$R^2_{adj} = 1 - \frac{\frac{SSE}{(n-p-1)}}{\frac{SST}{n-1}};$$

, which penalizes unnecessary predictors and provides a more reliable measure of model fit. [9].

It is worth noting that adjusted R^2_{adj} increases only if the new predictor improves the model more than what would be expected by chance.

The coefficient of determination R^2 ranges from 0 to 1. whereby a range closer to 1 indicates a better fit.

1.3 Conclusion

Chapter one laid the foundational concepts of linear models, emphasizing their significance in statistical modeling and data analysis. It began with an overview of the simple linear regression model, focusing on how a continuous response variable can be predicted using one or more explanatory variables. The chapter covered key assumptions underlying linear models, such as linearity, independence, homoscedasticity, and normality of errors, which are essential for valid inference. Estimation techniques, particularly the method of least squares, were introduced to determine model parameters. Furthermore, the chapter discussed the evaluation of model fit using R-squared, residual analysis, and significance testing through the F-test and t-tests. It concluded with an introduction to multiple linear regression, highlighting how it extends the simple linear model to accommodate multiple predictors, allowing for more complex and realistic modeling of relationships within data.

Linear Models to Generalized Linear Models

While linear regression models are foundational tools for examining relationships between continuous dependent variables and one or more predictors, they rely on assumptions such as normality of errors, homoscedasticity, and linearity. However, in many real-world applications, especially in fields like healthcare, social sciences, and economics, the response variable is not continuous but categorical, count-based, or otherwise non-normally distributed. To address such limitations, **Generalized Linear Models (GLMs)** extend the linear modeling framework by allowing for a flexible range of distributions for the response variable and by introducing a link function that connects the expected value of the response to the linear predictor. One important member of the GLM family is **logistic regression**, which is specifically designed for modeling binary outcomes. Thus, GLMs provide a coherent and unified framework that accommodates both linear regression and logistic regression as special cases, bridging classical linear models and modern statistical modeling [8]

2

Logistic regression models

Introduction

Logistic regression models can be defined as statistical models which describe the relationship between a qualitative dependent variable and an independent variable. [20] The logistic regression model has been made the standard method of analysis in many fields over the years.[14]. It involves the study of the effects of predictor variables on outcomes. Normally, the results produces a two-sided outcome whereby in this case the model is known as a **binary logistic model**. In the case where only one predictor variable is used in the model it is then referred to as a **simple logistic regression** while where multiple predictors are used its referred as **multiple or multivariable logistic regression.**[5]. The difference between the logistic and linear regression models is reflected both in the choice of a parametric model and in the assumptions. [14]What distinguishes logistic regression from the regression model is that the outcome is binary or categorical (e.g., success/failure, yes/no) by linking the probability of an event to predictors via the **logit function**.

2.1 Binary logistic regression

2.1.1 Mathematical formulation

While working with the binary outcome data, the conditional mean expressed as $E(Y|X)$ is taken to be greater than or equal to zero and less than or equal to one. ie $[0 \leq E(Y|x) \leq 1]$. As the conditional mean gets closer to zero or one, the smaller change in the $E(Y|X)$ per unit change in X becomes progressively smaller . As a result, the curve is said to be **S-shaped** ressembling a cumulative distribution of a random variable.

The logit transformation is applied which helps us to establish a relationship between the logistic distribution and the linear regression model properties. This is because the logit transformation is linear in its parameters, may be continous, and may range from $-\infty$ to $+\infty$ depending on the range of X.

To simplify the notation, the quantity $E(Y|X = x)$ is used in representation of the conditional mean of Y given x when the logistic regression model is used. As a result the logistic regression model is represented as follows:

$$E(Y|X = x) = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}}. \tag{2.1}$$

In order to have a clear and concise interpretation of our conditional mean used in the logistic distribution, the logit transformation is used which enables us have a clear picture of the linear relationship of the model. The logit transformation is hence represented as follows:

$$g(E(Y|X = x)) = ln\left(\frac{E(Y|X)}{1 - E(Y|X)}\right) = ln\left[\frac{\frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}}{1 - \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}}\right] = \beta_0 + \beta_1 x \tag{2.2}$$

Several distribution functions have been proposed for the use in the analysis of a binary outcome variable primarily because the logistic distribution is extremely flexible and easily used as a function and that it lends itself a clinically meaningful interpretation.[14]

Apart from the logit transformation g(x) providing many desirable properties of a linear regression model the second important difference between the linear and logistic regression model concerns the conditional distribution of the outcome variable.

Consider a collection of several predictor variables denoted by the vector $X' = (X_1, X_2, ..., X_p)$. The conditional probability that the outcome is present is denoted by $\pi(x_i)= P(Y = 1|X = x_i)$ with $x_i = (x_{i1}, x_{i2}, ..., x_{ip})$ the logit of the multiple regression model is hence expressed by the below given equation;

$$g(\pi(x_i)) = log\left[\frac{\pi(x_i)}{1-\pi(x_i)}\right] = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}. \tag{2.3}$$

With respect to the logistic regression, each β component represents a separate coefficient. Each coefficient assumes that when it is interpreted, the other predictors are held as constants.

In this case the probability of success is

$$P(Y = 1 \mid X = x_i) = \pi(x_i) = \frac{e^{\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\cdots+\beta_p x_{ip}}}{1 + e^{\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\cdots+\beta_p x_{ip}}}. \tag{2.4}$$

and therefore the logistic model in multiple predictors generalizes to:

$$\log \frac{P(Y = 1 \mid X_1, \ldots, X_p)}{1 - P(Y = 1 \mid X_1, \ldots, X_p)} = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p. \tag{2.5}$$

This formulation is the same in principle as the single-predictor case, but now X is a vector and there are p slope coefficients.

2.1.2 Estimation of parameters

The binary response logistic regression is based on the bernoulli probability distribution which is expressed with ones and zeros. The density function for the pair (x_i, y_i) is hence represented as

$$f(y_i, x_i) = \pi(x_i)^{y_i}[1 - \pi(x_i)]^{1-y_i} \tag{2.6}$$

Maximum likelihood estimation

As the observations are assumed to be independent, the coefficients β are estimated using maximum likelihood estimation (MLE), which maximizes the likelihood function given by:

$$L(\beta) = \prod_{i=1}^{n} \pi(x_i)^{y_i}[1 - \pi(x_i)]^{1-y_i} \tag{2.7}$$

While working with log of the equation makes the mathematical calculation easier, the expression of the log-likelihood is hence defined as

$$\ell(\beta) = \log[L(\beta] = \sum_{i=1}^{n} [y_i \log(\pi(x_i)) + (1 - y_i) \log(1 - \pi(x_i))] \tag{2.8}$$

Expressing the log likelihood $\ell(\beta)$ in exponential form is equally important so that we can extract from it the link function as well as the variance and the mean of the bernoulli distribution. In this case it is given as

$$\log\left[\frac{\pi(x_i)}{1-\pi(x_i)}\right]$$

The fitted (predicted) value of the logistic regression is based on the link function represented by,

$$g(\pi(x_i)) = \log\left(\frac{\pi(x_i)}{1-\pi(x_i)}\right) = \eta_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$$

To get the fitted probability (i.e., the expected value of Y_i) from the linear predictor, we apply the inverse of the logit link:

$$\hat{\pi}(x_i) = \frac{1}{1+e^{-\hat{\eta}_i}} = \frac{1}{1+e^{-(\hat{\beta}_0+\hat{\beta}_1 x_{i1}+\cdots+\hat{\beta}_k x_{ik})}}$$

This expression gives the fitted value $\hat{\pi}(x_i)$, the predicted probability of success for observation i.

Newton-Raphson algorithm

To find the maximum likelihood estimates (MLEs) of parameters in models where analytical solutions are not available we employ the Newton-Raphson algorithm as an iterative optimization method.

It is particularly common in logistic regression, where the log-likelihood equations are nonlinear. The method uses both the gradient and the hessian matrix of the log-likelihood function to make updates, which allows for faster and more accurate convergence compared to gradient descent. Mathematically, the update step is given by:

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} - \left[\nabla^2 \ell(\boldsymbol{\beta}^{(t)})\right]^{-1} \nabla \ell(\boldsymbol{\beta}^{(t)}),$$

where $\nabla \ell(\boldsymbol{\beta})$ is the score function and $\nabla^2 \ell(\boldsymbol{\beta})$ is the observed information matrix.[1].

The **Gradient (score function):** is given by

$$\nabla \ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} (y_i - \pi(\mathbf{x}_i))\mathbf{x}_i.$$

In order to estimate coefficients, the following steps are followed

1. Initialize $\boldsymbol{\beta}^{(0)}$.
2. Compute $\nabla \ell(\boldsymbol{\beta}^{(t)})$ and $\nabla^2 \ell(\boldsymbol{\beta})$.
3. Update using:

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} - \left[\nabla^2 \ell(\boldsymbol{\beta}^{(t)})\right]^{-1} \nabla \ell(\boldsymbol{\beta}^{(t)}),$$

4. Repeat until:

$$\|\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^{(t)}\| < \epsilon.$$

2.1.3 Model interpretation

Odds in logistic regression

The **odds** of an event represent the ratio of the probability that the event occurs to the probability that it does not occur:

$$\text{Odds} = \frac{P(\text{Event})}{1 - P(\text{Event})} = \frac{\pi(x_i)}{1 - \pi(x_i)}$$

Example, if the probability of an event is 0.75, the odds are:

$$\text{Odds} = \frac{0.75}{0.25} = 3 \quad \text{(or "3 to 1").}$$

Log-odds (logit)

Logistic regression models the **log-odds** (logit) of the probability of the outcome as a linear combination of predictors:

$$\log\left(\frac{\pi(x_i)}{1 - \pi(x_i)}\right) = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip}$$

with $\pi(x_i)$ representing the probability of the outcome represented as

$$\pi(x_i) = \frac{\exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})}{1 + \exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})}.$$

, p representing the number of predictors and X_i representing the predictor variables.[15]
This formulation forces $0 \leq p \leq 1$, and the quantity $\frac{\pi(x_i)}{1-\pi(x_i)}$ is the **odds** of the vent.

Odds ratio

In binary logistic regression, the **Odds Ratio (OR)** quantifies the relationship between the odds of the occurring outcome and a predictor variable. It serves as a crucial measure for interpreting the effect size of independent variables on a binary dependent variable.

For example, if one group has probability p_A of $Y = 1$ and another group has probability p_B, then the odds ratio comparing group A to group B is

$$\text{OR} = \frac{p_A/(1-p_A)}{p_B/(1-p_B)}.$$

Equivalently, it is the ratio of the odds in A to the odds in B. An OR greater than 1 means the odds are higher in group A (a positive association), while an OR less than 1 means they are lower (a negative association).

If one predictor x_j increases by one unit (holding other covariates fixed), the new odds become $\exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_j(x_j + 1) + \cdots + \beta_p x_p)$. Dividing the new odds by the original odds gives

$$\frac{\text{Odds}(x_j + 1)}{\text{Odds}(x_j)} = \frac{\exp(\beta_0 + \cdots + \beta_j(x_j + 1) + \cdots)}{\exp(\beta_0 + \cdots + \beta_j x_j + \cdots)} = \exp(\beta_j).$$

Thus a one-unit increase in x_j multiplies the odds by $\exp(\beta_j)$. In particular, for a binary indicator X the odds ratio for $X = 1$ versus $X = 0$ is

$$\exp(\beta_j) = \frac{\dfrac{P(Y = 1 \mid X = 1, \mathbf{Z})}{P(Y = 0 \mid X = 1, \mathbf{Z})}}{\dfrac{P(Y = 1 \mid X = 0, \mathbf{Z})}{P(Y = 0 \mid X = 0, \mathbf{Z})}},$$

meaning e^{β_j} is the conditional OR for X (given other covariates $\mathbf{Z}$).

Example

In practice, the odds ratio is calculated by exponentiating the estimated coefficient. For example, if $\beta_j = 0.5$ then $\text{OR} = e^{0.5} \approx 1.65$, meaning a one-unit increase in X_j raises the odds by 65%.[2]

Conversely, if $\beta_j = -0.45$ then $\text{OR} = e^{-0.45} \approx 0.64$, meaning the odds are multiplied by 0.64 (a 36% reduction). For a two-group comparison (e.g. group A vs. group B) one computes

$$\text{OR} = \frac{p_A/(1-p_A)}{p_B/(1-p_B)},$$

where p_A and p_B are the event probabilities in each group.

*Interpretation of the odds ratio In general, the odds ratio quantifies the direction and strength of association.
Therefore, OR = 1 indicates no association (equal odds), OR > 1 indicates increased odds of the event, and OR < 1 indicates decreased odds.

For example, $\exp(-0.45) \approx 0.64$ would imply the odds decrease by 36% per unit of the predictor. More generally, an OR of 2.5 means the odds are 2.5 times higher, and an OR of 0.5

means the odds are half as large, compared to the reference.

Standard errors

The **standard error** measures the variability of the estimated coefficients.[16]
In logistic regression, it is derived from the **Fisher information matrix**, the inverse of the observed information matrix $\mathcal{I}(\beta)$.

The variance-covariance matrix of the MLEs $\hat{\beta}$ is:

$$\text{Var}(\hat{\beta}) = \left(X^T W X\right)^{-1}$$

where: X denotes the design matrix and W is a diagonal matrix with elements $w_i = \pi_i(1 - \pi_i)$, derived from the second derivative (Hessian) of the log-likelihood function.

Each diagonal element of this matrix gives the variance $\text{Var}(\hat{\beta}_j)$, and its square root is the **standard error**:

$$SE(\hat{\beta}_j) = \sqrt{[(X^T W X)^{-1}]_{jj}}$$

2.1.4 Tests on the parameters β_j

Z-Statistic

The z statistics can be simply defined as the ratio of a coefficient to its standard error.

To test whether a particular coefficient β_j is significantly different from zero, we use the **z-statistic**, which under large samples is approximately normally distributed due to the maximum likelihood estimator's (MLE) asymptotic normality.

$$z_j = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)}$$

When the model records a larger value of z, it typically indicates that a predictor significantly contributes to the model.
Under the null hypothesis $H_0 : \beta_j = 0$, the z-statistic follows a standard normal distribution $N(0, 1)$.
Note:If we take the square root of **Wald test** W, we get the z-statistic.This statistic follows a chi-squared distribution with 1 degree of freedom.

P-values

A p-value is a two-tail test of the z statistic. It is used to test the null hypothesis given that the associated coefficient value is 0. The **p-value** tests the significance of individual predictors.

Given the z-statistic for each coefficient $\hat{\beta}_j$, the p-value is computed as:

$$p\text{-value} = 2 \cdot P(Z > |z_j|) = 2 \cdot (1 - \Phi(|z_j|))$$

where Φ is the cumulative distribution function (CDF) of the standard normal distribution.

A small p-value (typically <0.05) indicates that we reject the null hypothesis $H_0 : \beta_j = 0$, implying the predictor x_j has a statistically significant relationship with the response variable.

Confidence intervals

The $(1 - \alpha)$ **confidence interval** for a logistic regression coefficient $\hat{\beta}_j$ is:

$$\beta_j \in [\hat{\beta}_j \pm z_{\alpha/2} \cdot SE(\hat{\beta}_j)].$$

For a $1 - \alpha$ confidence level, $z_{\alpha/2}$. This interval estimates the range within which the true coefficient β_j lies with $1 - \alpha$ confidence.

However, since logistic regression deals with **log-odds**, we often exponentiate the interval to obtain the **odds ratio** confidence interval:

$$\left(e^{\hat{\beta}_j - z_{\alpha/2} \cdot SE(\hat{\beta}_j)},\ e^{\hat{\beta}_j + z_{\alpha/2} \cdot SE(\hat{\beta}_j)}\right)$$

This transformed interval is more interpretable: it gives the multiplicative effect on the odds of $Y = 1$ per unit increase in x_j.

Example derivation (simplified)

Assume a logistic model with a single predictor x:

$$\log\left(\frac{\pi(x_i)}{1 - \pi(x_i)}\right) = \beta_0 + \beta_1 x$$

After estimating the model:

$$\hat{\beta}_1 = 0.7, \quad SE(\hat{\beta}_1) = 0.2$$

Z-statistic:

$$z = \frac{0.7}{0.2} = 3.5$$

P-value:

$$p = 2 \cdot (1 - \Phi(3.5)) \approx 2 \cdot (1 - 0.99977) = 0.00046$$

95% Confidence interval for β_1:

$$0.7 \pm 1.96 \cdot 0.2 = (0.312, 1.088)$$

Odds Ratio interval:

$$(e^{0.312}, e^{1.088}) = (1.37, 2.97)$$

This tells us a unit increase in x multiplies the odds by between 1.37 and 2.97.

2.1.5 Model evaluation:Testing and fitting a logistic regression

Pearson *Chi2* goodness of fit

This method of model fitting is mainly used to assess the goodness-of-fit of a logistic regression model.

The test statistic is defined as:

$$X^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\pi}_i)^2}{\hat{\pi}_i(1 - \hat{\pi}_i)};$$

n is the number of runs (observations) in y_i is the number of successes of individual i,$\hat{\pi}_i$ is the predicted probability of success in of individual i.

The test is carried out using $n-p-1$ degrees of freedom where p is the number of parameters in the logistic regression model.

There is evidence of poor fit if either of the ratio of the test statistic to its degrees of freedom differs significantly from 1:

$$\frac{X^2}{df} = \frac{X^2}{n-k-1}.$$

Similarly if the p-value associated with the Chi-square statistic is less than the chosen significance level α (commonly 0.05) it is evident that the model is poorly fitted. If the design contains few or no replicates, the Pearson Chi-Square test may not be appropriate. In such cases, the **Hosmer-Lemeshow test** is recommended as an alternative for assessing the model's goodness-of-fit.

Hosmer-Lemeshow statistics

The Hosmer Lemeshow test is a statistical test used in testing the for goodness of fit and calibration for logistic regression models. It is used frequently in risk prediction models. The test assesses whether or not the observed event rates match expected probabilities in subgroups of the model population.

The test works by dividing the dataset into G groups based on predicted probabilities created by sorting the dataset in terms of the predicted probabilities.

This statistic normally follows a Chi-square distribution with $(G-2)$ degrees of freedom and it is represented as follows

$$\chi^2 = \sum_{g=1}^{G} \left(\frac{(O_{1g} - E_{1g})^2}{E_{1g}} + \frac{(O_{0g} - E_{0g})^2}{E_{0g}} \right).$$

with O_{1g} representing observed number of success events ($y = 1$) in group g , E_{1g} showing the expected number of events in group g , O_{0g}: observed number of failed events ($y = 0$) in group g and E_{0g} representing the expected number of failed events in group g

The statistics is then computed and compared with the Chi-square distribution where by a high p-value (> 0.05) indicates good fit while a low p-value suggests a poor fit.

The advantage behind this test is that it provides a single, easily interpretable value that can be used to assess fit. However its greatest limitation could be that in the process of grouping, we may lack an important deviation from fit due to a small number of individual data points.

Confusion matrix

The confusion matrix is a fundamental tool for evaluating the performance of classification models such as logistic regression. It evaluates the number of correct and incorrect predictions made by the model, compared to the actual outcomes. It consists of four components:

- **True Positives (TP)**: Model correctly predicts class 1.
- **True Negatives (TN)**: Model correctly predicts class 0.
- **False Positives (FP)**: Model predicts class 1 but actual is class 0.
- **False Negatives (FN)**: Model predicts class 0 but actual is class 1

For binary classification, the confusion matrix is represented by the following form:

	Predicted: 1	**Predicted: 0**
Actual: 1	True Positive (TP)	False Negative (FN)
Actual: 0	False Positive (FP)	True Negative (TN)

Several important performance metrics can be computed from the confusion matrix:

- **Accuracy:** Overall correctness of the model

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}.$$

- **Precision:** Correct positive predictions among all predicted positives

$$\text{Precision} = \frac{TP}{TP + FP}.$$

- **Recall (Sensitivity):** Correct positive predictions among all actual positives

$$\text{Recall} = \frac{TP}{TP + FN}.$$

- **Specificity:** Correct negative predictions among all actual negatives

$$\text{Specificity} = \frac{TN}{TN + FP}.$$

- **F1 Score:** Harmonic mean of precision and recall

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

S-S plot

An S-S plot is a graph of the full range of sensitivity and specificity values that occur for cut point values ranging from 0 and 1. The point whereby the sensitivity and specificity values intersects indicates the point at which the two statistics are closest in value. The practical implementation of the S-S plot, a very low threshold indicates that you get all the positive outcomes but misclassifying many healthy individuals in a health case.

If in case you use a very high threshold, you correctly classify healthy people, but miss the actual positive cases.

ROC analysis

The ROC curve (Receiver Operator Characteristic curve) can help to decide which value of the threshold is best. The ROC curve shows the variation of the true positive and false-positive rates at different threshold values. When using the ROC analysis, it is advisable to look at both the ROC statistic as well at the top of the sensitivity versus one minus the specificity.[16]

A model with no predictive power represents a slope of 1. While interpreting the ROC values, values ranging from 0.5 to 0.65 are interpreted to be having a little predictive power whereas values ranging from 0.65 to 0.80 have moderate predictive values. Most logistic models fit within this range. On the other hand, values ranging from 0.8 and less than 0.9 are regarded as having a strong predictive power whereas values greater than 0.9 records the highest amount of predictive power. To show that the model did well in the prediction of the outcomes, the Area

Under the Curve (AUC) is used is which measures the power of the model to predict outcomes. It is simply the area under the ROC curve.

2.2 Multinomial logistic regression

Multinomial logistic regression is a statistical technique used to predict a categorical dependent variable with more than two categories, given one or more independent variables. It's an extension of binary logistic regression, which is used for predicting a variable with two categories.

The main use case is when your outcome variable has more than two possible values, and those values are not ordered.

The most common approach to multinomial logistic regression is the **baseline-category logit model**. This is where one category is selected as the reference (or baseline), and the model estimates the log odds of being in each of the other categories relative to the baseline.

Example, let $Y \in \{0, 1, \ldots, J\}$ be the categorical dependent variable with $J + 1$ categories, and $x = (x_1, x_2, \ldots, x_k)$ be a vector of independent variables. Choose category 0 as the baseline. Then, for each category $j = 1, 2, \ldots, J$ we can have:

$$\log\left(\frac{P(Y = j \mid x)}{P(Y = 0 \mid x)}\right) = \beta_{j0} + \beta_{j1}x_1 + \beta_{j2}x_2 + \ldots + \beta_{jk}x_k;$$

This implies:

$$P(Y = j \mid x) = \frac{e^{\beta_j^T x}}{1 + \sum_{l=1}^{J} e^{\beta_l^T x}} \quad \text{for } j = 1, \ldots, J;$$

$$P(Y = 0 \mid x) = \frac{1}{1 + \sum_{l=1}^{J} e^{\beta_l^T x}};$$

where β_j is taken to be the vector of coefficients for category j.

2.2.1 Estimation and interpretation of β_j

The parameters β_j are estimated using maximum likelihood estimation (MLE). The log-likelihood function is:

$$\ell(\beta) = \sum_{i=1}^{n} \sum_{j=0}^{J} y_{ij} \log P(Y_i = j \mid x_i).$$

where $y_{ij} = 1$ if the observed outcome for observation i is category j, otherwise 0.

Each coefficient β_{jk} reflects the effect of variable x_k on the log-odds of being in category j relative to the baseline category 0. A positive β_{jk} implies that higher values of x_k increase the log-odds of being in category j versus category 0, while a negative β_{jk} implies the opposite.

The odds

The odds of being in category j relative to the baseline category K is

$$\frac{P(Y = j \mid x)}{P(Y = K \mid x)} = \exp\left(\beta_{j0} + \beta_{j1}x_1 + \cdots + \beta_{jp}x_p\right) = e^{\beta_j^T x}$$

.

The odds ratio for a unit increase in x_k is $e^{\beta_{jk}}$, holding other variables constant.

2.2.2 Model evaluation

Since multinomial logistic regression is not linear, the traditional R^2 does not apply. Instead, several pseudo-R^2 measures are used. McFadden's pseudo -R^2 is defined as:

$$R^2 = 1 - \frac{\ell(\text{full model})}{\ell(\text{null model})}.$$

It ranges from 0 to 1, where values greater than 0.2 are considered decent. Other variants include Cox & Snell and Nagelkerke's pseudo-R^2, which are adjusted for sample size and interpretability.

The likelihood ratio test evaluates the significance of predictors:

$$\text{Deviance} = -2(\ell(\text{restricted model}) - \ell(\text{full model})).$$

This statistic is compared against a chi-square distribution with degrees of freedom equal to the difference in number of parameters between models.

Although multinomial logistic regression is probabilistic, predicted class labels can be obtained using the maximum predicted probability for each observation. A confusion matrix summarizes the true classes versus predicted classes and helps compute accuracy, precision, recall, and F1 score for each category.

2.3 Ordinal logistic regression

Ordinal logistic regression is a statistical technique used to predict a dependent variable that has an ordered scale, like a likert scale. It's an extension of binary logistic regression and

multiple linear regression, allowing you to model the relationship between one or more independent variables and an ordinal outcome. For example (strongly disagree, disagree, neutral, agree,strongly agree)

2.3.1 Model formulation and estimation of parameters

Ordinal logistic regression models parameters by estimating regression coefficients and thresholds also known as cut points). These cut points define the boundaries in which the categories are separated. On the other hand the regression coefficients are used to indicate the predictor variables influence the odds of falling into or below a particular category.

The maximum likelihood estimation (MLE) is the most common method for finding these parameter values.The likelihood function is built on the cumulative probabilities of the response categories.

The likelihood function is thus represented as follows:

$$L(\theta) = \prod_{i=1}^{n} \prod_{j=1}^{J} P(Y_i = j \mid X_i)^{I(Y_i=j)},$$

with $I(Y_i = j)$ being the indicator function that is 1 when the observation belongs to category j and 0 otherwise.

The log-likelihood function for the ordinal logistic regression model is given by:

$$\ell(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \sum_{i=1}^{n} \sum_{j=1}^{J} I(Y_i = j) \log \left(P(Y_i = j \mid \mathbf{x}_i) \right)$$

The probability that an observation i lies on category j are computed as:

$$P(Y_i = j \mid \mathbf{x}_i) = \begin{cases} \dfrac{1}{1 + e^{-(\alpha_1 - \mathbf{x}_i^\top \boldsymbol{\beta})}}, & \text{if } j = 1 \\ \dfrac{1}{1 + e^{-(\alpha_j - \mathbf{x}_i^\top \boldsymbol{\beta})}} - \dfrac{1}{1 + e^{-(\alpha_{j-1} - \mathbf{x}_i^\top \boldsymbol{\beta})}}, & \text{if } 2 \leq j \leq J - 1 \\ 1 - \dfrac{1}{1 + e^{-(\alpha_{J-1} - \mathbf{x}_i^\top \boldsymbol{\beta})}}, & \text{if } j = J \end{cases}$$

The most widely used formulations for ordinal logistic regression modelling is the **proportional odds model**, also known as the **cumulative logit model**. This model is used to compare the probability of an equal or smaller response, $Y \leq k$ to the probability of a larger response, $Y > k$.

The calculations of the **odds** are hence based on these cumulative probabilities.

Cumulative odds

For a response variable Y with ordered categories $1, 2, \ldots, K$, the **cumulative odds** of being in category k or lower (i.e., $Y \leq k$) versus being in a higher category (i.e., $Y > k$) with X being the given predictors is defined as:

$$\text{Odds}(Y \leq k \mid X) = \frac{P(Y \leq k \mid X)}{P(Y > k \mid X)}.$$

Modelling the odds

The **proportional odds model** links the cumulative odds to the predictors as:

$$C_k(x) = \ln\left(\frac{P(Y \leq k \mid X)}{P(Y > k \mid X)}\right) = \alpha_k - X^\top \beta$$

where α_k is the intercept (threshold) for category k, X is vector of predictor variables and β represents the vector of regression coefficients (common across all categories)

To obtain the cumulative odds from the proportional odds model the exponentiate our function from both sides to obtain:

$$\frac{P(Y \leq k \mid X)}{P(Y > k \mid X)} = \exp(\alpha_k - X^\top \beta).$$

Example:
Let Y represent educational attainment, with ordered categories:

$$Y = \begin{cases} 1 & \text{Low} \\ 2 & \text{Medium} \\ 3 & \text{High} \end{cases}$$

Let x be the number of years of study. The proportional odds model specifies that the cumulative logit of being in category k or below is:

$$\log\left(\frac{P(Y \leq k)}{P(Y > k)}\right) = \alpha_k - \beta x, \quad k = 1, 2.$$

This gives us the following system of equations:

$$\log\left(\frac{P(Y \leq 1)}{P(Y > 1)}\right) = \alpha_1 - \beta x.$$
$$\log\left(\frac{P(Y \leq 2)}{P(Y > 2)}\right) = \alpha_2 - \beta x.$$

The ordinal logistic regression assumes that the relationship between each pair of outcome groups is the same which means that the effect of the predictors is consistent across all thresholds of the ordinal outcome.

Interpreting the odds ratio

The **odds ratio (OR)** measures the change in odds for a one-unit increase in a predictor x_j, holding other variables constant.

Let β_j be the coefficient of predictor x_j. Then the odds ratio is given by:

$$\text{OR} = \exp(\beta_j).$$

In the interpretation of results, if $\beta_j > 0$ and the odds ratio is greater than one i.e (OR > 1) means that the odds of being in a **higher category** increase as x_j increases. On the other hand given the $\beta_j < 0$ and the odds ratio is less than one (i.e OR < 1) indicates that the odds of being in a **lower category** increase as x_j increases. Consequently if $\beta_j = 0$ and the odds ratio is equal to 1 (OR $= 1$) means that there's no effect.

2.3.2 Evaluating the model's performance

Confusion matrix

A confusion matrix gives the report for the model's perfomance by simply comparing the actual categories of the model against its predicted categories in the data. The perfomance metrics like the accuracy, precision, recall, specificity and F1 score are then calculated to give the overall perfomance of the model.

Pseudo R-Squared (McFadden's R^2)

Pseudo R-squared values are used to assess the fit of models in generalized linear models (GLMs), including logistic regression, when there's no traditional R-squared like in ordinary least squares (OLS) regression. They represent the proportion of variance explained by the model, but are based on log-likelihoods rather than sums of squares. It is represented as follows:

$$R^2{}_{McFadden} = 1 - \frac{logL_{Model}}{logL_{null}}.$$

Values ranging from 0.2 are considered decent with 0.4 being considered as an excellent fit.

Brant test

The Brant test of proportional odds is used in ordinal logistic regression, specifically to assess the proportional odds assumption. It simply compares their coefficients to check whether there is violation of the proportional odds assumption that the coefficients for the predictor variables are the same across all categories of the outcome variable. In simple terms the effect of a predictor is constant across all logits. It is represented as follows:

$$\log\left(\frac{P(Y \leq j)}{P(Y > j)}\right) = \alpha_j - \beta^\top X.$$

whereby the same β is used for all j, and the Brant Test checks if this is violated.

2.4 Conclusion

Chapter two focused on logistic regression, a statistical method tailored for modeling categorical outcome variables, especially binary and multinomial responses. The chapter began by presenting the limitations of linear regression in handling binary outcomes, leading to the justification and formulation of the logistic regression model using the logit link function. It detailed the interpretation of model coefficients in terms of odds and odds ratios, providing a more intuitive understanding of predictor effects. Estimation of parameters was addressed through the maximum likelihood estimation (MLE) method, given the non-linearity of the model. The chapter also covered model diagnostics and performance evaluation, including goodness-of-fit measures like the deviance, classification tables, and ROC curves. Additionally, it extended logistic regression to handle multinomial and ordinal responses, thus broadening its applicability to a wider range of categorical data analysis problems. This chapter provided essential tools for analyzing categorical outcomes and interpreting results in various practical contexts.

3

Applications of logistic regression

Introduction

This chapter aims to investigate the steps and techniques used in the study. It concentrates on the application of the logistic regression technics employed to real datasets that lead to the identification of the factors influencing the different outcomes. It describes the models used, practical analysis and goes ahead to discuss and interpret the results obtained and suggest the recommendations found from the findings.

3.1 Application on a binary logistic regression

The data utilized in this work was collected using direct questionnaires from the patients of **Sylhet Diabetes Hospital in Sylhet Bangladesh**[10]. It represents a sample space of five hundred and twenty instances of patients records. It examines two classes of the target variable (positive and negative) with whether the patient recorded diabetic positive or rather diabetic negative as driven by the predictors. The predictor variables include age, gender, polyuria, polydipsia, sudden weight loss, weakness, polyphagia, genital thrush, visual blurring, itching, irritability, delayed healing, partial paresis, muscle stiffness, alopecia and obesity. The numerical variables such as age are continuous while the other predictor variables are categorical

taking binary responses. The figure 3.1 shows a list of the predictor variables used in the model.

ge	Gender	Polyuria	Polydipsia	sudden weight loss	weakness	Polyphagia	Genital thrush	visual blurring	Itching	Irritability	delayed healing	partial paresis	muscle stiffness	Alopecia	Obesity	clas
40	Male	No	Yes	No	Yes	No	No	No	Yes	No	Yes	No	Yes	Yes	Yes	Positiv
58	Male	No	No	No	Yes	No	No	Yes	No	No	No	Yes	No	Yes	No	Positiv
41	Male	Yes	No	No	Yes	Yes	No	No	Yes	No	Yes	No	Yes	Yes	No	Positiv
45	Male	No	No	Yes	Yes	Yes	Yes	No	Yes	No	Yes	No	No	No	No	Positiv
60	Male	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Positiv

Figure 3.1 – A diagram showing the respective dataset columns

3.1.1 Model fitting

A binary logistic regression model was constructed with `Class` as the dependent variable alongside other predictor variables. The logistic regression equation is given by:

$$\log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 \cdot \text{Age} + \beta_2 \cdot \text{Gender} + \beta_3 \cdot \text{Polyuria} + \beta_4 \cdot \text{Polydipsia} + \cdots \tag{3.1}$$

The model was fitted on the training data, and the estimated coefficients and odds ratios were obtained (see table 3.1).

Interpretation of the summarized table presented in 3.1

Given the logistic regression model as follows:

$$\log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 \cdot \text{Age} + \beta_2 \cdot \text{Gender} + \beta_3 \cdot \text{Polyuria} + \cdots + \beta_{16}\text{Obesity}$$

where the $\pi(x)$ is indicating the predicted probability of class = 1.ie

$$\pi(x) = P(\text{Class} = 1 | \text{Age,Gender,Polyuria,...})$$

Coefficients

The model estimates the **log-odds** from the logistic regression represented by the column **Coef.** The coefficient β_j in this case gives more information of how much log-odds of the

Table 3.1 – Logistic Regression Results: Coefficients, Odds Ratios, and Confidence Intervals

Variable	Coef.	OR	SE	z	p-value	95% CI Lower	95% CI Upper
Intercept	2.7466	15.5895	1.0755	2.5538	0.0107	1.8940	128.3146
Age	-0.0512	0.9511	0.0254	-2.0174	0.0437	0.9040	0.9985
Gender	-4.3512	0.0129	0.5982	-7.2739	0.0001	0.0040	0.0416
Polyuria	4.4395	84.7359	0.7053	6.2948	0.0001	21.2687	337.5939
Polydipsia	5.0704	159.2446	0.8289	6.1172	0.0001	31.3704	808.3679
Sudden weight loss	0.1903	1.2096	0.5477	0.3475	0.7282	0.4135	3.5385
Weakness	0.8171	2.2638	0.5368	1.5221	0.1279	0.7905	6.4830
Polyphagia	1.1936	3.2948	0.5335	2.2375	0.0255	1.1596	9.3881
Genital thrush	1.8637	6.4472	0.5533	3.3682	0.0008	2.1797	19.0701
Visual blurring	0.9159	2.4999	0.6512	1.4064	0.1596	0.6973	8.9551
Itching	-2.8029	0.0606	0.6727	-4.1663	0.0001	0.0162	0.2266
Irritability	2.3407	10.3888	0.5905	3.9638	0.0001	3.2652	33.0539
Delayed healing	-0.9364	0.3920	0.5510	-1.6997	0.0893	0.2299	1.9870
Partial paresis	1.1593	3.1877	0.5248	2.2089	0.0278	1.1396	8.9167
Muscle stiffness	-0.7288	0.4826	0.5808	-1.2561	0.2091	0.1548	1.5044
Alopecia	0.1504	1.1623	0.6201	0.2425	0.8084	0.3447	3.9185
Obesity	-0.2890	0.7499	0.5443	-0.5390	0.5945	0.2577	2.1768

outcome change with one unit increase in any of the predictor variables X_j holding all other variables constant.

For example : The coefficient of the diabetes predictor Polyuria is given by: **4.4395** while the coefficient for Gender given as **-4.3512**. This indicates that testing Polyuria positive increases the log-odds of diabetes by **4.4395** while on the other hand being male (coded as 1) decreases the log-odds of contracting diabetes by 4.3512.

Odds Ratio

While **Odds** can be defined as a way of expressing the likelihood of an event happening compared to it not happening, the **Odds Ratio** can therefore be defined as a measure used to compare the odds of an event happening in one group to the odds of it happening in another group.

Focusing the outcomes from the model using an example of the predictor Polyuria. The coefficient of this predictor is given as 4.4395 which represent the amount of the rate of change in our class when our predictor moves from 0 to 1 in value.

Exponentiating our logit function $log(\frac{\pi(x_i)}{1-\pi(x_i)0})$ we simply obtain the odds as $\frac{\pi(x_i)}{1-\pi(x_i)}$. **Example** suppose that the probability of testing polyuria positive was 0.65, then the Odds could be calculated as:

$$\text{Odds} = \frac{\text{Probability of polyuria positive } (\pi(x_i))}{\text{Probability of polyuria negative (1-}\pi(x_i))} = \frac{0.65}{1-0.65} = 1.85714$$

The odds ratio of polyuria = 1 is the ratio of the odds of polyuria = 1 to the odds of polyuria = 0.

Thus, for each predictor the odds ratio is constituted as $OR_i = e^{\beta_j}$ From our case the odds ratio for polyuria is $e^{4.4395} \approx 84.7359$.

This results indicate that patients with polyuria have 84.7359 times higher odds of having diabetes compared to patients testing polyuria negative. Similarly considering the predictor Age with the Odds ratio as **0.9511**. This suggests that a one year increase in age of a patient reduces the odds of testing diabetes positive by 4.89% (since $1 - 0.9511 = 0.0489$), holding all other factors constant.

When the ratio of the odds equals to **1**, it indicates that the predictor doesn't change the odds and hence no effect whereas when **OR** > 1 demonstrates that the predictor increases the odds while when **OR**< 1 shows that the predictor decreases the odds. From our model, its enough to judge that all the variables increases the odds with Polydipsia (159.2446) recording the highest odds ratio value and gender (0.0129) recording the lowest odds ratio of testing diabetes positive respectively.

Its worth noting that when carrying out multiple logistic regression, the effect of each predictor is adjusted for the others.This makes the estimates cleaner and more accurate.

Standard error

The standard error is a statistical tool derived from the inverse Fisher information matrix used to demonstrate the degree of uncertainty (variability). It shows how much the coefficient β might vary if we repeated the study many times. It is denoted by the formula:

$$SE(\hat{\beta}_j) = \sqrt{var(\hat{\beta}_j)}$$

A smaller value of the standard errors indicates a more precise estimate and vice versa. Our predictor variable age has the least standard error of **0.0254** while polydipsia had the largest standard error of **0.8289** hence making age more reliable.

3.1.2 Tests for significance

P- Values

A p-value is a statistical tool used to test if a predictor variable has a significant effect on the outcome. It helps us in making a decision as to whether to keep or drop a predictor from the model. It helps us understand which variables actually matter for predicting the outcome. It is normally derived from the standard normal distribution whereby it is based on the tests for null

hypothesis:i.e $H_0 : \beta_j = 0$ with a low p-value ($i.e < 0.05$) simply signifying that the predictors have statistically significant effects on the model while p-values ($i.e > 0.05$) are not significant. Moreover the wald-statistic represented as by the column z can be used to depict the factors that are significant. A large $|z|$ value depicts a stronger evidence that the predictor is significant.

Partial paresis, age, gender, polyuria, polydipsia, itching, irritability, genital thrush, polyphagia were the predictors with the significant effect on the models. From the evaluation of the model, they recorded a p-value of less than < 0.05 hence confirming their great contribution in the models. While other predictors would in some way influence the model, the most basic and determinants of whether the patient is diabetic or will be diabetic are the above mentioned predictors.

The table 3.2 below represents a summary of the predictors with significance on the model based on the p-values.

Variable	**Coef.**	**OR**	**SE**	**z**	**p-value**	**Significance**
Intercept	2.7466	15.5895	1.0755	2.5538	0.0107	✓
Age	-0.0512	0.9511	0.0254	-2.0174	0.0437	✓
Gender	-4.3512	0.0129	0.5982	-7.2739	0.0001	✓
Polyuria	4.4395	84.7359	0.7053	6.2948	0.0001	✓
Polydipsia	5.0704	159.2446	0.8289	6.1172	0.0001	✓
Sudden weight loss	0.1903	1.2096	0.5477	0.3475	0.7282	
Weakness	0.8171	2.2638	0.5368	1.5221	0.1279	
Polyphagia	1.1936	3.2948	0.5335	2.2375	0.0255	✓
Genital thrush	1.8637	6.4472	0.5533	3.3682	0.0008	✓
Visual blurring	0.9159	2.4999	0.6512	1.4064	0.1596	
Itching	-2.8029	0.0606	0.6727	-4.1663	0.0001	✓
Irritability	2.3407	10.3888	0.5905	3.9638	0.0001	✓
Delayed healing	-0.9364	0.3920	0.5510	-1.6997	0.0893	
Partial paresis	1.1593	3.1877	0.5248	2.2089	0.0278	✓
Muscle stiffness	-0.7288	0.4826	0.5808	-1.2561	0.2091	
Alopecia	0.1504	1.1623	0.6201	0.2425	0.8084	
Obesity	-0.2890	0.7499	0.5443	-0.5390	0.5945	

Table 3.2 – A summary logistic regression results table representing the predictors with significance on the model using the P-Values

Confidence intervals

The confidence interval is calculated in order to assess the uncertainty of the estimate for the coefficient $\hat{\beta}_j$. It is calculated using the predictors coefficient ($\hat{\beta}_j$), standard error and the $z_{\alpha/2}$ which represents the critical value from the standard normal distribution for a two-tailed test (e.g., 1.96 for 95% confidence level).

$$\text{CI}_{\text{log-odds}} = \left[\hat{\beta}_j \pm z_{\alpha/2} \cdot \text{SE}(\hat{\beta}_j)\right]$$

Example at 95% confidence level:

$$z_{0.025} = 1.96$$

To interpret the effect in terms of **odds ratio** rather than log-odds, we exponentiate the bounds of the confidence interval:

$$\text{CI}_{\text{OR}} = \left(e^{\hat{\beta}_j - 1.96 \cdot \text{SE}},\ e^{\hat{\beta}_j + 1.96 \cdot \text{SE}}\right).$$

This gives a range within which the true odds ratio lies with 95% confidence.

For example, using the predictor variable **polyphagia** we have the following values

$$\hat{\beta}_j = 1.1937, \quad \text{SE} = 0.5335.$$

Computing the 95% confidence intervals for log-odds we obtain

$$\text{CI}_{\text{log}} = 1.1937 \pm 1.96 \cdot 0.5335 = (0.1471,\ 2.2403).$$

We exponentiate the upper bound and lower bounds to get the confidence intervals for the odds ratio:

$$\text{CI}_{\text{OR}} = \left(e^{0.1471},\ e^{2.2403}\right) = (1.1596,\ 9.3881).$$

from the confidence interval (1.1596, 9.3881) the interval does not include hence the predictor is **statistically significant**. Confidence intervals with one included in their intervals indicates that they are statistically insignificant.

The summary table 3.3 below shows the factors that are significant basing on the confidence interval estimates:

Under the significant column the mark(✓) indicates the values that are statistically significant at 95% confidence (CI does not include 1).

3.1.3 Model evaluation

Pearson Chi-square goodness of fit

This test is usually carried out to determine if observed data from a sample aligns with a specific theoretical distribution. Its from the results that we get that enables us judge how effective

Variable	Coef	SE	z	OR(e^{β})	95% CI OR	Signif.
const	2.7466	1.0755	2.5538	15.5895	(1.8940 , 128.3146)	✓
Age	-0.0512	0.0254	-2.0174	0.9511	(0.9040 , 0.9985)	✓
Gender	-4.3512	0.5982	-7.2739	0.0129	(0.0040 , 0.0416)	✓
Polyuria	4.4395	0.7053	6.2948	84.7359	(21.2687 , 337.5939)	✓
Polydipsia	5.0704	0.8289	6.1172	159.2446	(31.3704 , 808.3679)	✓
Suddenweight loss	0.1903	0.5477	0.3475	1.2096	(0.4135 , 3.5385)	
Weakness	0.8171	0.5368	1.5221	2.2638	(0.7905 , 6.4830)	
Polyphagia	1.1938	0.5335	2.2375	3.2949	(1.1596 , 9.3881)	✓
Genital thrush	1.8637	0.5533	3.3682	6.4472	(2.1797 , 19.0701)	✓
Visual blurring	0.9159	0.6512	1.4064	2.4999	(0.6973 , 8.9551)	
Itching	-2.8029	0.6727	-4.1663	0.0606	(0.0162 , 0.2266)	✓
Irritability	2.3407	0.5961	3.9638	10.3888	(3.2652 , 33.0539)	✓
Delayed healing	-0.3916	0.5375	-0.7287	0.6763	(0.2299 , 1.9870)	
Partial paresis	1.1593	0.5248	2.2089	3.1872	(1.1396 , 8.9167)	✓
Muscle stiffness	-0.7288	0.5808	-1.2561	0.4826	(0.1548 , 1.5044)	
Alopecia	0.1504	0.6201	0.2425	1.1623	(0.3447 , 3.9185)	
Obesity	-0.2890	0.5443	-0.5400	0.7499	(0.2577 , 2.1768)	

Table 3.3 – A summary logistic regression results table representing the predictors with significance on the model using the 95% CI

and fit the model is. This fitting method is majorly based on the test of the two hypothesis

H_0 : The observed frequencies follow the expected distribution.

H_1 : The observed frequencies do not follow the expected distribution.

Our model produces a Chi-square statistic of **175.4591** with **503** degrees of freedom derived from the 520 instances minus the 17 considered predictors and a p-value: **1.0000**. The p-value being greater than the standard significance level of 0.05 signifies that the observed frequencies follow the chi-square distribution and that the observed data fits the expected distribution extremely well.

This results well explains that the observed frequencies are highly consistent with the expected frequencies under the theoretical distribution being tested. No statistically significant deviation was found and hence proving a good fit.

Hosmer-Lemeshow statistics

The model provided a Hosmer-Lemeshow test statistic of **1.6085** and a corresponding p-value of **0.9908**. We accept the null hypothesis ie (H_0 : The model fits the data well) since the p-value is greater than the 0.05 significance level proving that there is no significant difference

between the actual outcomes and predicted probabilities and hence the logistic regression model fits the data well.

Confusion matrix

Figure 3.2 represents how the model classified the true positive values, true negative, false positive and false negative values. A corresponding confusion matrix table was therefore drafted from the figure and presented as follows: see table 3.4.

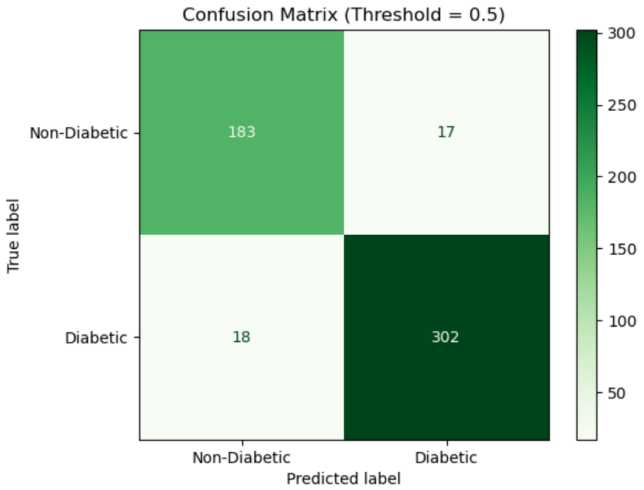


Figure 3.2 – Confusion matrix

	Predicted Negative	**Predicted Positive**
Actual Negative	183	17
Actual Positive	18	302

Table 3.4 – Confusion Matrix

Classification metrics

The table 3.5 below shows the classification performance metrics that the model produced.

Metric	**Value**
Accuracy	0.933
Precision	0.947
Recall (Sensitivity)	0.944
Specificity	0.915
F1 Score	0.945

Table 3.5 – Classification performance metrics

An accuracy score of **0.933** signifies that the model measured an overall correctness of the model with a 93.3% accuracy in the total predictions confirming that the model performs very well overall. A precision score of **0.947** depicts that the predicted positive cases were actually positive with 94.7% precision. The model identifies most of the actual positive cases were corrrectly identified with a 94.4% sensitivity score. This score is very important as in some occasions a missing a positive case is could be harmful and dangerous.

The specificity score measured that 91.5% of the negative cases were correctly identified and labelled as negative successfully. This specificity score is considered vital since the incorrectly classified positive cases needs to be minimized for better results. Finally the F1 Score metric of 94.5% demonstrates that there is a strong balance between precision and sensitivity. It is useful when you want a single metric that considers both false positives and false negatives.

Based on the tests from our data set, the sensitivity, specificity and high precision scores gives a direct confirmation that our model had a balanced performance. On the other hand , while the F1 Score of 94.5% which is closer to 100% showing the high reliability hence the model being termed as robust.

These outcomes suggest that the model is suitable for practical use, especially in domains where both types of errors (false positives and false negatives) are critical.

Sensitivity- Specificity plot

The figure 3.3 below shows a Sensitivity(true positive rate) - Specificity (true negative rate) plot for our dataset showing the variation across different threshold values used to classify the diabetes positive case.

The Sensitivity curve represented by the (Green Line) starts at a high rate ie ($\approx$ 1) when threshold is low (near 0) indicating that almost all positives are correctly identified (few false negatives). The curve decreases as the threshold increases because the model becomes stricter in classifying a patient as diabetic, so you miss more true cases.

Similarly, the Specificity curve (Red Line) begins low when the threshold is low signifying that many false positives-non-diabetics were wrongly flagged. The curve then increases as the threshold increases and eventually reaches $\approx$ near threshold 1 where (almost all true negatives are correctly classified).

From this plot, the two curves intersect at around 0.6 threshold value.At this point Sensitivity and specificity are both $\approx$ 0.9, indicating strong performance of the model

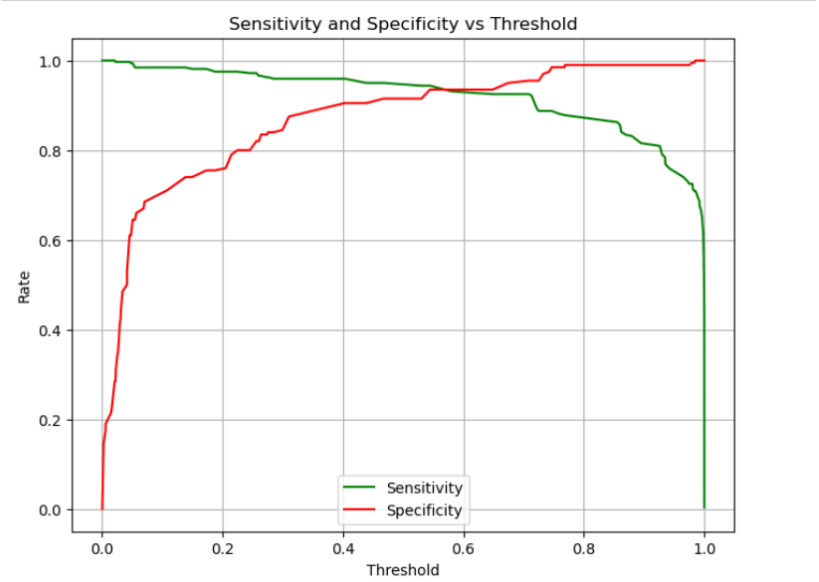


Figure 3.3 – S-S plot

ROC-AUC

The figure 3.4 below shows a Receiver Operator Characteristic curve (ROC) drawn from our data. It clearly shows the classifier's performance across all the classification thresholds.

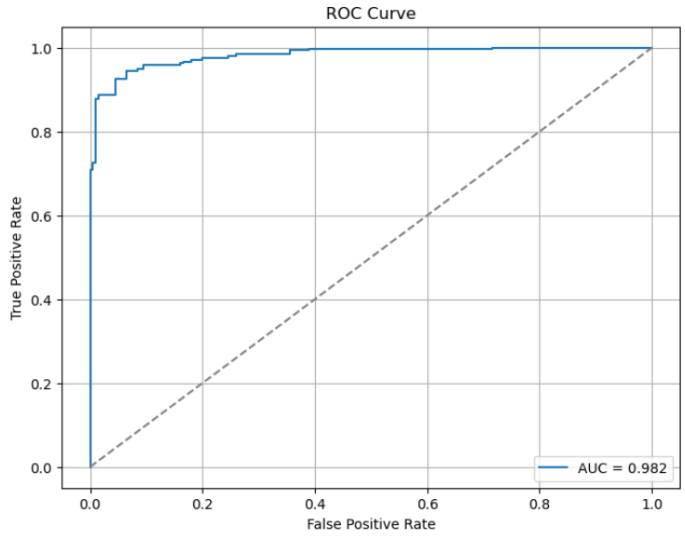


Figure 3.4 – ROC curve

The ROC curve shown in figure 3.4 presents the trade-off between sensitivity and specificity for varying classification thresholds. The corresponding AUC (Area Under the Curve) value of 0.97 indicates excellent model performance, suggesting that the logistic regression model has a very high ability to distinguish between diabetic and non-diabetic patients. The curve's

Metric	Value / Interpretation
ROC Curve	Plots True Positive Rate (Sensitivity) vs. False Positive Rate
Shape of ROC	The curve hugs the top-left corner, indicating high model performance
AUC (Area Under Curve)	0.97
AUC Interpretation	Excellent discrimination ability (97% chance of correct classification)
Best Threshold (Optional)	Based on Youden's Index (can be computed for threshold selection)
Model Strength	High Sensitivity and Specificity across most thresholds

Table 3.6 – ROC Curve and AUC Summary for the logistic regression model

proximity to the top-left corner confirms a strong predictive capability. For optimal threshold selection, Youden's Index could be computed to identify the point that maximizes the sum of sensitivity and specificity.

3.2 Application on a multinomial logistic regression

While multinomial logistic regression analysis is based on a more than two unordered outcomes, in our study we are going keenly carry out multinomial logistic regression analysis in which students are to make choices of their study program based on several factors. In our case, it involves students joining college who make program choices among three sets of programs available in the institution namely general program, vocational program and academic program. Their choices could be modelled using associated factors like the writing score and the social economic status.

For our case we are going to use the hsbdemo data which can obtained from the site below *https://stats.idre.ucla.edu/stat/data/hsbdemo.*[6]

The dataset contains data for two hundred students with id, gender, school type, program, read, write, maths scores, science scores, social status, honours, awards, cid, prog encoded and social economic status represented in three levels (middle,low and high) and writing scores as the main predictor variables. The dependent variable is the program type with three outcomes (general, vocational and academic)

The figure 3.5 below shows a quick representation of the data.

3.2.1 Model presentation

Unlike the binary logistic regression case where the outcome is binary, the multiple logistic regression predicts the variables with more than two categories. The model is lied on the baseline-category approach where one category is selected as the baseline and then the log-odds are estimated for the other categories relative to the baseline. For our case the model is formulated with **General** and **ses high** as the baseline category in predicting the students program

	id	female	ses	schtyp	prog	read	write	math	science	socst	honors	awards	cid	prog_encoded
0	45	female	low	public	vocation	34	35	41	29	26	not enrolled	0	1	2
1	108	male	middle	public	general	34	33	41	36	36	not enrolled	0	1	1
2	15	male	high	public	vocation	39	39	44	26	42	not enrolled	0	1	2
3	67	male	low	public	vocation	37	37	42	33	32	not enrolled	0	1	2
4	153	male	middle	public	vocation	39	31	40	39	51	not enrolled	0	1	2
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
195	100	female	high	public	academic	63	65	71	69	71	enrolled	5	20	0
196	143	male	middle	public	vocation	63	63	75	72	66	enrolled	4	20	2
197	68	male	middle	public	academic	73	67	71	63	66	enrolled	7	20	0
198	57	female	middle	public	academic	71	65	72	66	56	enrolled	5	20	0
199	132	male	middle	public	academic	73	62	73	69	66	enrolled	3	20	0

Figure 3.5 – multinomial logistic regression data

choices (General, Academic, Vocational) basing on their writing scores and socio-economic status. The model formulation is hence provided by the below equation:

$$log\left(\frac{P(Y=j)}{P(Y=0)}\right) = \beta_{j0} + \beta_{j1} \cdot \text{write} + \beta_{j2} \cdot \text{ses low} + \beta_{j3} \cdot \text{ses middle}$$

where j takes the 2 categories ie Academic and Vocational.

For example, using the results derived from table 3.7, computing the log odds of vocational program category using the baseline category general can be calculated as:

$$\log\left(\frac{P(Y=2)}{P(Y=0)}\right) = 4.2355 - 0.1136 \cdot \text{write} + 0.9827 \cdot \text{ses low} + 1.2741 \cdot \text{ses middle}$$

Table 3.7 below shows the output generated when the multinomial logistic regression was carried out with General and high social economic status as the baseline categories.

Category	Variable	Coef.	Std. Err.	z	P>|z|	CI -Log odds[0.025, 0.975]
Academic	Intercept	1.6894	1.227	1.377	0.169	[-0.715, 4.094]
Academic	write	-0.0579	0.021	-2.706	0.007	[-0.100, -0.016]
Academic	ses_low	1.1628	0.514	2.261	0.024	[0.155, 2.171]
Academic	ses_middle	0.6295	0.465	1.354	0.176	[-0.282, 1.541]
Vocational	Intercept	4.2355	1.205	3.516	0.000	[1.874, 6.597]
Vocational	write	-0.1136	0.022	-5.113	0.000	[-0.157, -0.070]
Vocational	ses_low	0.9827	0.596	1.650	0.099	[-0.185, 2.150]
Vocational	ses_middle	1.2741	0.511	2.493	0.013	[0.272, 2.276]

Table 3.7 – Multinomial logistic regression coefficients relative to general program summary table

3.2.2 Interpretation of results

The individual coefficients

With reference to the academic category the writing score has a coefficient of **-0.0579** whereas basing on the vocational category the coefficient to the writing score is given as **-0.1136**. This signifies that as writing scores increases the probability of choosing both academic or vocational over general decreases significantly. This conclusion is based on the negative coefficient scores. In simple terms the outcome claims that students with higher writing scores are more likely to be enrolled to in general programs.

On the other hand, drawing observations from the social economic status with the reference category ses-high, the low social economic status significantly increases odds of choosing Academic as while the middle economic status has no effect on the model. Similarly middle social economic status significantly increases the odds of Vocational as the ($p = 0.013$). This means that students from lower social economic backgrounds are more likely to be in either Academic or Vocational programs relative to General while a higher social economically status students tend to be enrolled in General programs.The positive intercepts reflect the baseline log-odds when predictors are zero.

The figure 3.6 below shows a plot showing the relationship between the predicted probabilities, the writing scores and social economic status.

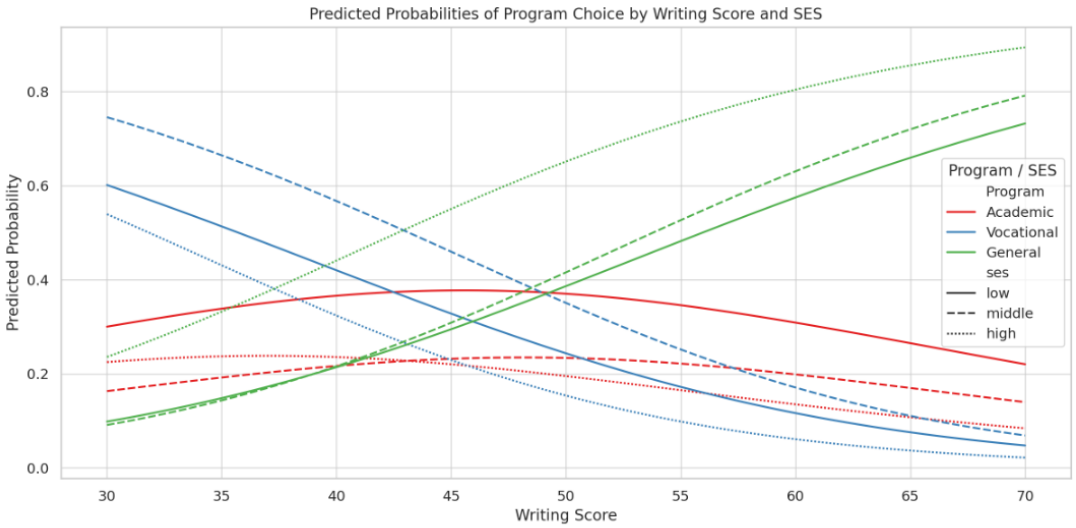


Figure 3.6 – Graph showing the relationship of the predicted probabilities to the writing scores and social economic statuses

Interpretation of the curves

As writing scores increase, the demand for General program becomes more likely especially for students with high social economic status as represented by the green curves. The blue curves represents the Vocational programs. It is clear that the Vocational probability drops sharply with higher writing scores across all social economic levels and finally the Academic program stays moderate with highest for low social economic status, but still decreases as writing score increases.

P-values

We check on the corresponding p-values in order to have an understanding of which predictors have influence in the model. Starting with the academic category, it is evident that all the variables are statistically significant except the middle social economic status which recorded a higher p-value score of 0.176 which is greater than the threshold p-value of 0.05 therefore confirming that it is not significant and hence it does not influence on our model.

Likewise, putting our focus on the vocational category, the lower social economic status recorded a slightly higher p-value of **0.099** which is greater than our fixed threshold value of **0.05**. While the other factors have a lower p-value, this proves that the lower social economic status is statistically insignificant and hence it does not affect the model whatsoever.

The odds Ratio

We then exponentiate coefficients of the independent variables to obtain the odds ratios:

$$\text{OR}_{write} = exp(-0.1136) \approx 0.8925$$

Confidence intervals formulation

In multinomial logistic regression, confidence intervals (CI) for regression coefficients provide a range within which the true population coefficient is likely to fall with a specified level of confidence (usually 95%). In the log-odds confidence interval interpretation, if 0 is within the interval, the effect is considered not statistically significant while the absence of zero in the interval signifies that the effect is statistically significant.

Concentrating on the confidence intervals drawn from the academic program the writing scores and low social economic status ie **([-0.100, -0.016] and [0.155, 2.171])** respectively shows a significant effect while the middle social economic status was not significant. ie (ie(**[-0.282, 1.541]**)). CI = [-0.100, -0.016]

Further interpretations made on the vocational category indicates that the writing scores and ses-middle are statistically significant while the ses-low is not.

3.2.3 Model evaluation

McFadden's pseudo R^2

The pseudo R^2 value derived from the model was **0.1182** (refer to appendix 3.15)suggesting that the model explains about 11.8% of the variance in program choice. This value shows a low explanatory power and hence calling up for the model to be improved.

Findings

The graphs shown in figure 3.6 clearly shows that lower SES students opt more for Vocational or Academic programs. It is also evident that high writing scores do not align with academic and vocational programs probably could be due to other factors.

Confusion matrix

The table 3.2.3 below designed matrix was derived from the results of the model.

	Predicted: Academic	**Predicted: General**	**Predicted: Vocation**
Actual: Academic	92	4	9
Actual: General	27	7	11
Actual: Vocation	23	4	23

Confusion matrix table

The confusion matrix interpretations given by the model are as follows: **Academic program**

- True Positives (TP) = 92
- False Negatives (FN) = 4 + 9 = 13
- False Positives (FP) = 27 + 23 = 50

Vocational program

- True Positives (TP) = 23
- False Negatives (FN) = 4 + 23 = 27
- False Positives (FP) = 9 + 11 = 20

The classification metrics were carried out and represented in the table 3.8

deriving conclusions based on the confusion matrix, the model gives a high recall record of 88% and a F1-score of 74% in the academic program which indicates that most students who choose academic are correctly predicted whereas the model fails to identify students enrolled under the general program as evident in the recall percentage i.e 16% and F1 score of

Class	Precision	Recall	F1-Score	Support
Academic	0.65	0.88	0.74	105
General	0.47	0.16	0.23	45
Vocation	0.53	0.46	0.49	50
Accuracy		0.61		200
Macro Avg	0.55	0.50	0.49	200
Weighted Avg	0.58	0.61	0.57	200

Table 3.8 – Classification report for multinomial logistic regression

0.23 indicating some potential model biasness. On the other hand, under the vocational program a precision score of **53.5%**,recall of **46%** and an F1 score of 0.494 shows a moderate performance with balanced precision and recall of the model.

Overall performance

The model records an accuracy level of **61%** and an average F1-score of **0.49** signifying that the model needs more improvement simply by trying different algorithms for classification or else adding more features.

3.3 Application on an ordinal logistic regression

In this section, we are going to see how effective the ordinal logistic regression technics can be employed in the study for identifying the risk intensity levels for maternal mortality during pregnancy in the rural areas of Bangladesh considering a given set of factors. This data was collected from different hospitals, community clinics and maternal health cares through the Internet of Things(IoT) see [3].

The predicted risks were categorised into three categories ie **high risk, low risk and moderate risk** given the factors as: Age in years when a woman is pregnant, upper value of Blood pressure in mmHg (SystolicBP), lower value of blood pressure in mmHg (DiastolicBP), blood glucose levels in terms of a molar concentration (ie BS), body temperature (BodyTemp) and a normal resting heart rate (HeartRate) . All these are the responsible and significant risk factors for maternal mortality.

One thousand and fourteen observations were made and figure 3.7 belows shows a quick representation of the data.

	Age	SystolicBP	DiastolicBP	BS	BodyTemp	HeartRate	RiskLevel
0	25	130	80	15.0	98.0	86	high risk
1	35	140	90	13.0	98.0	70	high risk
2	29	90	70	8.0	100.0	80	high risk
3	30	140	85	7.0	98.0	70	high risk
4	35	120	60	6.1	98.0	76	low risk
...	...	...	...	...	...	...	...
1009	22	120	60	15.0	98.0	80	high risk
1010	55	120	90	18.0	98.0	60	high risk
1011	35	85	60	19.0	98.0	86	high risk
1012	43	120	90	18.0	98.0	70	high risk
1013	32	120	65	6.0	101.0	76	mid risk

1014 rows × 7 columns

Figure 3.7 – Ordinal logistic regression data

3.3.1 Model formulation

The model estimates the coefficients associated with ordinal logistic regression which helps us in the formation the model's equation. The model's equation also referred to as the **cumulative logit model** focuses on the probability of an observation falling into a category or below, rather than modelling each category independently. For response categories $j = 1, 2$ having three separate categories, the model estimates

$$log\left(\frac{P(Y \leq j)}{P(Y > j)}\right) = \theta_j - (\beta_1 \cdot Age + \beta_1 \cdot SystolicBP + \cdots + \beta_5 \cdot HeartRate)$$

Since we have three categories (low, mid and high risks), we shall therefore have θ_1 and θ_2 to represent the threshold points also known as the cut points used to separate the three categories.

A summary table 3.9 was obtained from our dataset.

Variable	Coef.	Std. Error	z	P> |z|	Odds Ratio	[0.025	0.975]
Age	-0.0098	0.0061	-1.5979	0.1101	9.9020e-01	9.7840e-01	1.0220e+00
SystolicBP	0.0489	0.0061	8.0077	0.0000	1.0501e+00	1.0376e+00	1.0628e+00
DiastolicBP	-0.0019	0.0075	-0.2512	0.8016	9.9810e-01	9.8350e-01	1.0129e+00
BS	0.4554	0.0365	12.4601	0.0000	1.5767e+00	1.4677e+00	1.6938e+00
BodyTemp	0.4881	0.0537	9.0873	0.0000	1.6292e+00	1.4664e+00	1.8100e+00
HeartRate	0.0418	0.0092	4.5293	0.0000	1.0427e+00	1.0240e+00	1.0617e+00
low/mid	59.4307	5.5213	10.7638	0.0000	6.4625e+25	1.2920e+21	3.2373e+30
mid/high	0.8173	0.0508	16.0882	0.0000	2.2644e+00	2.0498e+00	2.5014e+00

Table 3.9 – Ordinal logistic regression statistical summary

3.3.2 Interpretation of results

Coefficients

The coefficients in a model greatly tells how much the predictor factors influences the odds of moving to a higher category. Positive coefficients increases the odds of being at a higher risk while negative values pushes towards a lower risk likelihood.

Our model records the coefficients of Age and Diastolic blood pressure as negative values (ie -0.0098 **and** -0.0019 **respectively**). This signifies that a unit change in age and diastolic blood pressure reduces the likelihood of maternal mortality. Similarly, the upper value of Blood pressure in mmHg (SystolicBP), blood glucose levels (BS), body temperature (BodyTemp) and a normal resting heart rate (HeartRate) recorded a positive coefficient value (ie **0.0489, 0.4554, 0.4881, 0.0418 respectively** . This can be interpreted as, a unit increase in the predictors value increases the risk of mortality.

P- values

This indicates the significance of the model predictors. From the summary table 3.9 presented above, it is evident that all other predictors except Age and DiastolicBP have a significant influence on the model as their p-values are below the fixed p-value of 0.05. It is necessary to note that the p-values are based on the tests for **hypothesis testing** with the null hypothesis H_0 standing on the point that the predictors have no significant influence in the model. A lower value than the fixed p-value leads to the rejection of the null hypothesis as demonstrated by the predictors Age and DiastolicBP.

Odds Ratio

While the odds ratio (OR) in ordinal logistic regression quantifies the change in the odds of moving from one ordinal level to the next for a one-unit increase in a predictor variable, holding

other variables constant, our model gave the Odds ratio for age and diastolic blood pressure as less than 1 ($OR < 1$) suggesting that the two predictors have a greater likelihood of moving towards the lower risk levels while the others whose Odds Ratio is greater than 1 ($OR > 1$) indicates a greater likelihood of moving towards the higher ordinal levels.

Thresholds (Intercepts)

The threshold value represents the cut-points between categories on the latent variable scale. **59.4307** is the threshold for the log-odds of being at a low risk versus higher while **0.8173** is the threshold for the log-odds of being at mid risk or low versus high risk.

3.3.3 Model evaluation

Confusion matrix

Consider the confusion matrix represented by the figure 3.8

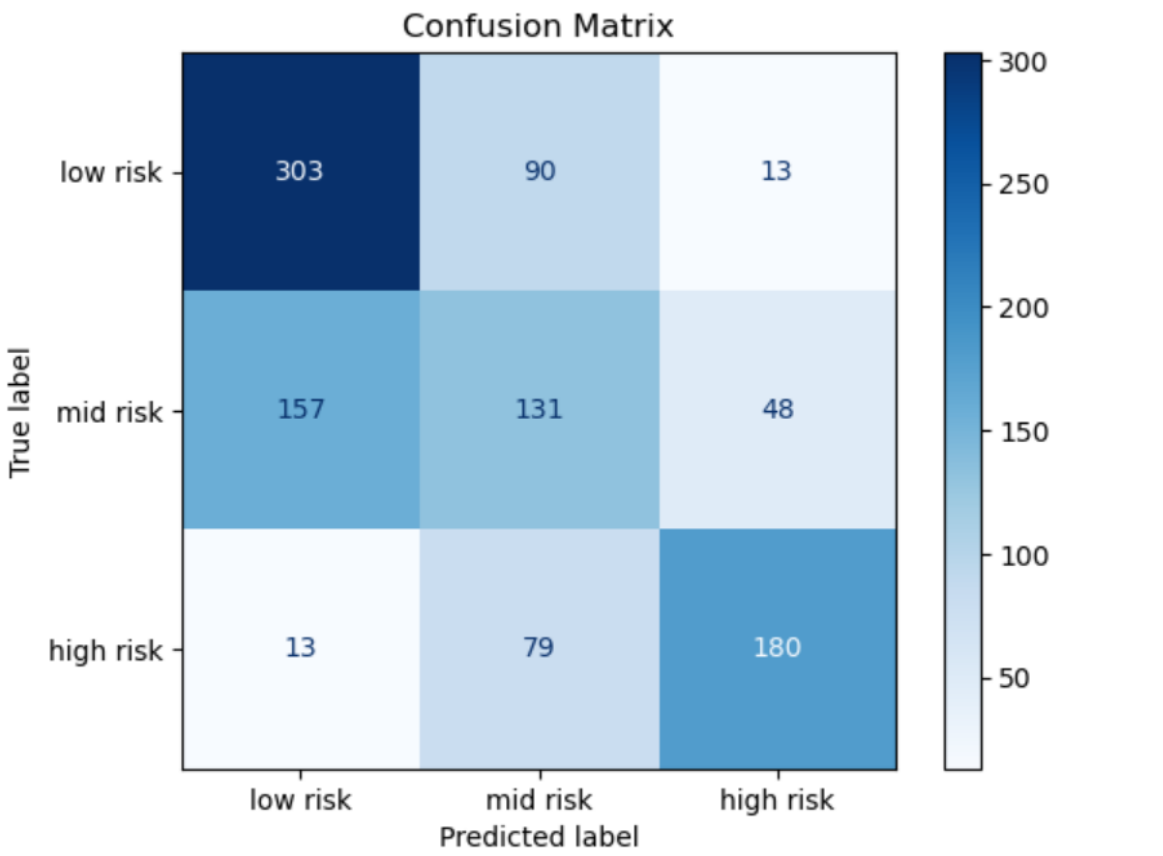


Figure 3.8 – Maternal mortality confusion matrix

From the confusion matrix represented in figure 3.8 the following performance metrics were extracted:

Metric	Value
Accuracy	0.606
Precision	0.608
Recall (Sensitivity)	0.599
Specificity	0.771
F1 Score	0.601

Table 3.10 – Classification performance metrics

The model produced an accuracy score of **60.6%** , precision score of **60.8%** , recall of **59.9%**, F1 Score: **60.1%** and a specificity score of **77.1%**. This indicates that the model performance was moderate but not strong as it misses many true cases drawing conclusions from the accuracy and recall scores. Focusing on the F1-scores, we can as well conclude that the model was not yet reliable for critical tasks and hence required room for improvement.

Pseudo R-Squared (McFadden's R^2)

The model produces a Pseudo R-Squared value of **0.267** indicating it being a moderate fit for the dataset.

Brant test

Consider the figure 3.9 representing the analysis of the brant test.

```
--- Brant Test Approximation ---
             Threshold 1 Coef  Threshold 2 Coef  Difference
Intercept          -58.870790        -71.211782   12.340992
Age                 -0.005745         -0.016681    0.010936
SystolicBP           0.060863          0.028424    0.032439
DiastolicBP         -0.026531          0.046220   -0.072751
BS                   0.495207          0.496913   -0.001706
BodyTemp             0.489291          0.561051   -0.071761
HeartRate            0.033632          0.051243   -0.017611

--- McFadden's Pseudo R^2 ---
Pseudo R-squared: 0.267
```

Figure 3.9 – Showing the coefficients of the brant test

The figure 3.9 clearly shows that the coefficients differ notably and therefore the proportional odds assumption may be violated for those predictors. In this case, SystolicBP and DiastolicBP show larger deviations suggesting partial violation of the assumption. Variables like BS behave consistently assumption likely holds for them.

Using the Brant test it is hence clear that there is a violation of the proportional odds assumption hence making the model less performing.

3.4 Conclusion

The binary logistic regression model depicted the strong explanatory power by clearly examining the two target classes. The accuracy scores and other model evaluation metrics gave high records clearly showing the great performance of the model.

Likewise, the multinomial logistic regression model identifies clear relationships between socio-economic status, writing ability, and program selection. It reveals potential policy concerns about how academic placement correlates with socioeconomic and skill factors. Further improvements could involve interaction terms and additional predictors to increase explanatory power.

Finally, the ordinal logistic regression gave moderate results as seen in the brant test, the confusion matrix and pseudo R-squared. The results concluded that model was not reliable for critical tasks and hence required more improvement.

General conclusion

The prediction modelling technics to support clinical decision-making will forever remain useful in the various sectors in the world. While this thesis focused on the use of logistic regression tools for predicting the likelihood of events in the education and health care system, it has been observed that the model could be a well suit capable of assisting professionals in making timely and accurate diagnostic decisions due to its robustness, interpretability and efficiency.

The study demonstrated the high performance scores of binary logistic regression translating it to its efficiency and satisfaction. The multinomial logistic regression revealed the significant statistical relations between the academic choices the students made in reference to the social economic factors and writing scores. Finally the ordinal logistic regression permitted us to predict the risk levels in maternal mortality and gave a moderate performance scores maybe simply because of possible partial violations of the hypothesis of the constant proportions.

The study moreover enabled us have responses to our research questions. It confirmed how logistic regression classified the outcomes basing on the predictors and factors presented. The results obtained carry a great practical impact mostly on the various sectors it is applied. This helps in guiding the key decision makers in the various domains come up with insights from the provided predictor variables that could be of great help both in the education and health sector

Our work was thus structured into three chapters. The first chapter presented the statistical grounds for understanding the logistic regression while the second chapter dived into the mathematical formulation and technics of evaluation in the logistic regression. Finally the third chapter presented the application of logistic regression on real datasets, performed the rigorous model evaluation providing a better understanding of how each predictor contributed to the model's performance.

Although the model demonstrated a great practical approach in the diverse sectors, it was noted that the model's performance heavily depended on the quality and relevance of the predictors used. Similarly, the model was based on the assumption that there is a linear relationship between the independent variables and the log-odds of the dependent variable. This poses a great problem where there is complex or non-linear relationship. As a result of the findings and

limitations identified in the study, the following recommendations were proposed:

- Use of advanced and complex models like random forests, support vector machines, and neural networks could be explored to ensure a higher accuracy, especially when non-linear interactions are suspected.
- Incorporating more diverse predictors. While there could be a lot of factors that could be considered in the diagnosis of various diseases, future research should put into account more and diverse predictors like clinical and genetic predictors in order to improve the model's predictive power and uncover more relationships.
- Employing comparative studies between different classification algorithms aimed at optimizing the predictive performance of the real data sets.

In conclusion this research work gave light on the power of logistic regression technics on the predictive modelling in both social and medical fields. Its high efficiency and high explanatory power makes it ideal for decision makers working with diverse datasets. Similarly this thesis positions itself as a significant contribution to the applied statistical methodologies and opens opportunities for more analysis integrating the advantage of its complexity and precision.

Appendix

In [11]:
```
# Libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import statsmodels.api as sm
from sklearn.model_selection import train_test_split
from sklearn.metrics import (
    confusion_matrix, accuracy_score, precision_score, recall_score,
    f1_score, roc_curve, roc_auc_score
)
from scipy.stats import chi2
from statsmodels.stats.outliers_influence import variance_inflation_factor
```

In [12]:
```
# Load Dataset
df= pd.read_csv("diabetes_data_upload.csv")
```

In [13]:
```
# Inspect data
print(df.head())
```

```
   Age Gender Polyuria Polydipsia sudden weight loss weakness Polyphagia  \
0   40   Male       No         Yes                 No      Yes         No
1   58   Male       No          No                 No      Yes         No
2   41   Male      Yes          No                 No      Yes        Yes
3   45   Male       No          No                Yes      Yes        Yes
4   60   Male      Yes         Yes                Yes      Yes        Yes

  Genital thrush visual blurring Itching Irritability delayed healing  \
0             No              No     Yes           No              Yes
1             No             Yes      No           No               No
2             No              No     Yes           No              Yes
3            Yes              No     Yes           No              Yes
4             No             Yes     Yes          Yes              Yes

  partial paresis muscle stiffness Alopecia Obesity     class
0              No              Yes      Yes     Yes  Positive
1             Yes               No      Yes      No  Positive
2              No              Yes      Yes      No  Positive
3              No               No       No      No  Positive
4             Yes              Yes      Yes     Yes  Positive
```

Figure 3.10 – Binary logistic regression code

```
         Current function value: 0.899909
         Iterations 6
                          MNLogit Regression Results
==============================================================================
Dep. Variable:           prog_encoded   No. Observations:                  200
Model:                        MNLogit   Df Residuals:                      192
Method:                           MLE   Df Model:                           6
Date:                Fri, 23 May 2025   Pseudo R-squ.:                 0.1182
Time:                        03:08:38   Log-Likelihood:                -179.98
converged:                       True   LL-Null:                       -204.10
Covariance Type:            nonrobust   LLR p-value:                 1.063e-08
==============================================================================
prog_encoded=1       coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const              1.6894      1.227      1.377      0.169      -0.715       4.094
write             -0.0579      0.021     -2.706      0.007      -0.100      -0.016
ses_low            1.1628      0.514      2.261      0.024       0.155       2.171
ses_middle         0.6295      0.465      1.354      0.176      -0.282       1.541
------------------------------------------------------------------------------
prog_encoded=2       coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const              4.2355      1.205      3.516      0.000       1.874       6.597
write             -0.1136      0.022     -5.113      0.000      -0.157      -0.070
ses_low            0.9827      0.596      1.650      0.099      -0.185       2.150
ses_middle         1.2741      0.511      2.493      0.013       0.272       2.276
```

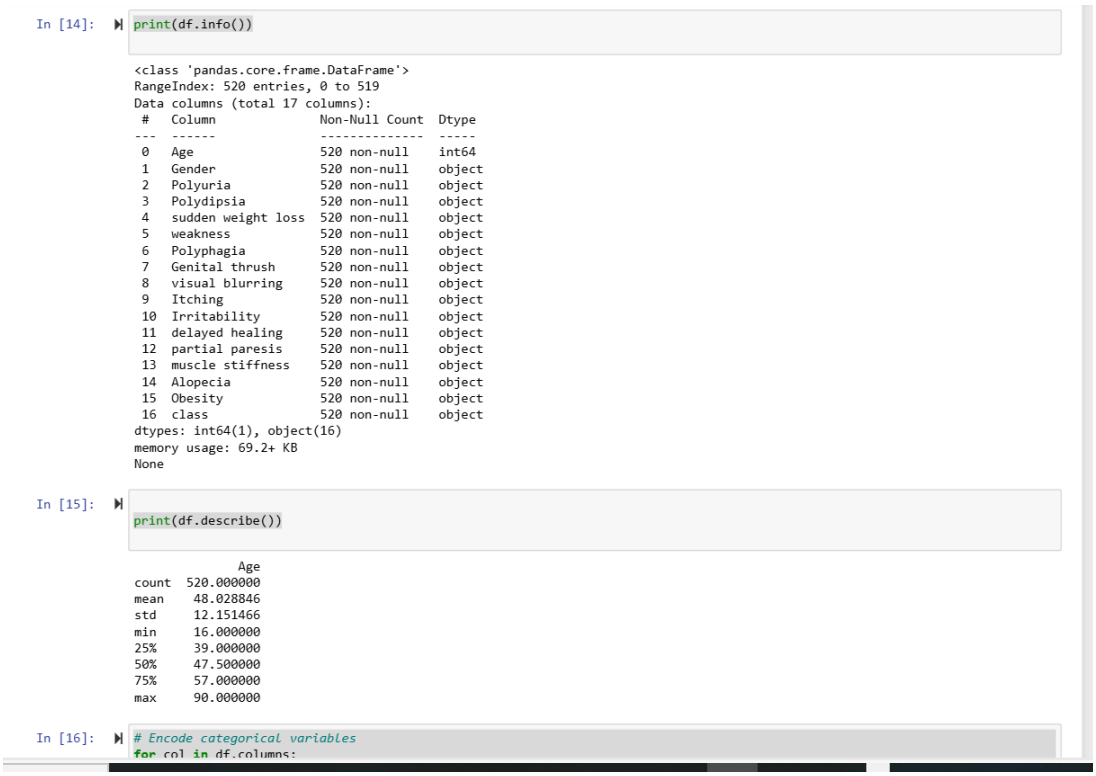
Figure 3.11 – Appendix: MLR summary

```
In [14]: print(df.info())

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 520 entries, 0 to 519
Data columns (total 17 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   Age                 520 non-null    int64
 1   Gender              520 non-null    object
 2   Polyuria            520 non-null    object
 3   Polydipsia          520 non-null    object
 4   sudden weight loss  520 non-null    object
 5   weakness            520 non-null    object
 6   Polyphagia          520 non-null    object
 7   Genital thrush      520 non-null    object
 8   visual blurring     520 non-null    object
 9   Itching             520 non-null    object
 10  Irritability        520 non-null    object
 11  delayed healing     520 non-null    object
 12  partial paresis     520 non-null    object
 13  muscle stiffness    520 non-null    object
 14  Alopecia            520 non-null    object
 15  Obesity             520 non-null    object
 16  class               520 non-null    object
dtypes: int64(1), object(16)
memory usage: 69.2+ KB
None

In [15]: print(df.describe())

              Age
count  520.000000
mean    48.028846
std     12.151466
min     16.000000
25%     39.000000
50%     47.500000
75%     57.000000
max     90.000000

In [16]: # Encode categorical variables
         for col in df.columns:
```

Figure 3.12 – Binary logistic regression code

```
         Current function value: 0.899909
         Iterations 6
                          MNLogit Regression Results
==============================================================================
Dep. Variable:           prog_encoded   No. Observations:                  200
Model:                        MNLogit   Df Residuals:                      192
Method:                           MLE   Df Model:                           6
Date:                Fri, 23 May 2025   Pseudo R-squ.:                 0.1182
Time:                        03:08:38   Log-Likelihood:                -179.98
converged:                       True   LL-Null:                       -204.10
Covariance Type:            nonrobust   LLR p-value:                 1.063e-08
==============================================================================
prog_encoded=1       coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const              1.6894      1.227      1.377      0.169      -0.715       4.094
write             -0.0579      0.021     -2.706      0.007      -0.100      -0.016
ses_low            1.1628      0.514      2.261      0.024       0.155       2.171
ses_middle         0.6295      0.465      1.354      0.176      -0.282       1.541
------------------------------------------------------------------------------
prog_encoded=2       coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const              4.2355      1.205      3.516      0.000       1.874       6.597
write             -0.1136      0.022     -5.113      0.000      -0.157      -0.070
ses_low            0.9827      0.596      1.650      0.099      -0.185       2.150
ses_middle         1.2741      0.511      2.493      0.013       0.272       2.276
```

Figure 3.13 – Appendix: MLR summary

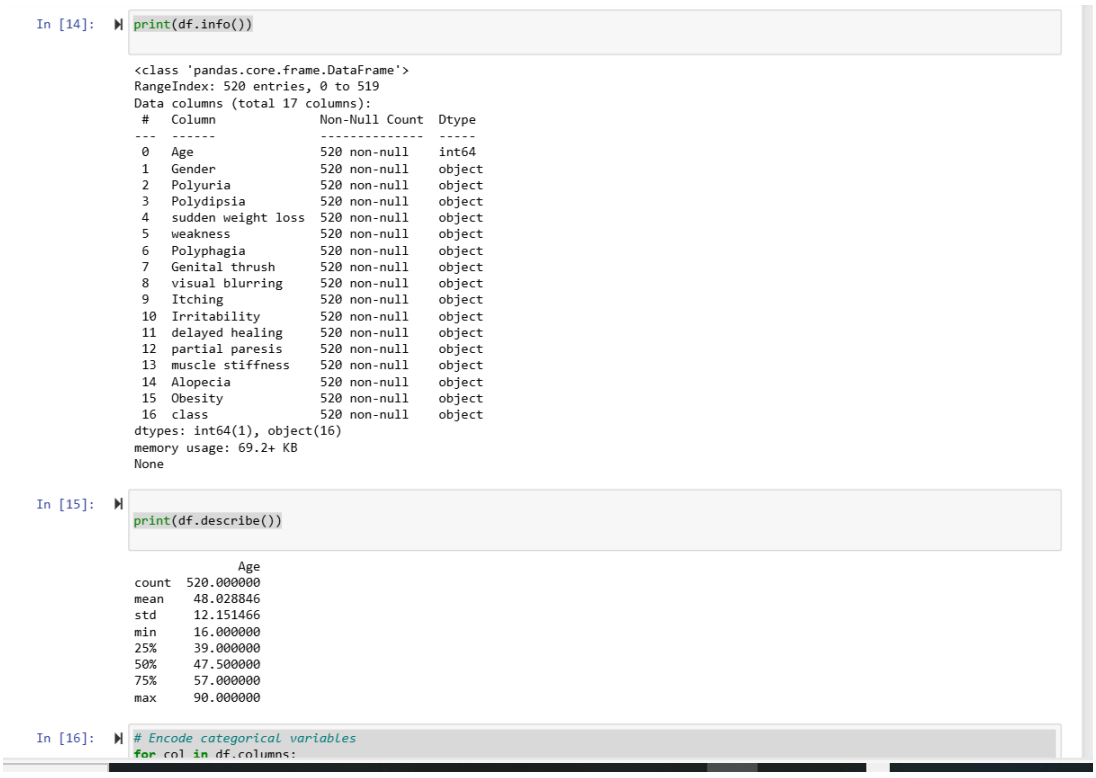


```
In [14]: print(df.info())

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 520 entries, 0 to 519
Data columns (total 17 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   Age                 520 non-null    int64
 1   Gender              520 non-null    object
 2   Polyuria            520 non-null    object
 3   Polydipsia          520 non-null    object
 4   sudden weight loss  520 non-null    object
 5   weakness            520 non-null    object
 6   Polyphagia          520 non-null    object
 7   Genital thrush      520 non-null    object
 8   visual blurring     520 non-null    object
 9   Itching             520 non-null    object
 10  Irritability        520 non-null    object
 11  delayed healing     520 non-null    object
 12  partial paresis     520 non-null    object
 13  muscle stiffness    520 non-null    object
 14  Alopecia            520 non-null    object
 15  Obesity             520 non-null    object
 16  class               520 non-null    object
dtypes: int64(1), object(16)
memory usage: 69.2+ KB
None

In [15]: print(df.describe())

              Age
count  520.000000
mean    48.028846
std     12.151466
min     16.000000
25%     39.000000
50%     47.500000
75%     57.000000
max     90.000000

In [16]: # Encode categorical variables
         for col in df.columns:
```

Figure 3.14 – Binary logistic regression code

```
         Current function value: 0.899909
         Iterations 6
                          MNLogit Regression Results
==============================================================================
Dep. Variable:           prog_encoded   No. Observations:                  200
Model:                        MNLogit   Df Residuals:                      192
Method:                           MLE   Df Model:                           6
Date:                Fri, 23 May 2025   Pseudo R-squ.:                 0.1182
Time:                        03:08:38   Log-Likelihood:                -179.98
converged:                       True   LL-Null:                       -204.10
Covariance Type:            nonrobust   LLR p-value:                 1.063e-08
==============================================================================
prog_encoded=1       coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const              1.6894      1.227      1.377      0.169      -0.715       4.094
write             -0.0579      0.021     -2.706      0.007      -0.100      -0.016
ses_low            1.1628      0.514      2.261      0.024       0.155       2.171
ses_middle         0.6295      0.465      1.354      0.176      -0.282       1.541
------------------------------------------------------------------------------
prog_encoded=2       coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const              4.2355      1.205      3.516      0.000       1.874       6.597
write             -0.1136      0.022     -5.113      0.000      -0.157      -0.070
ses_low            0.9827      0.596      1.650      0.099      -0.185       2.150
ses_middle         1.2741      0.511      2.493      0.013       0.272       2.276
```

Figure 3.15 – Appendix: MLR summary



```
In [14]: print(df.info())

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 520 entries, 0 to 519
Data columns (total 17 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   Age                 520 non-null    int64
 1   Gender              520 non-null    object
 2   Polyuria            520 non-null    object
 3   Polydipsia          520 non-null    object
 4   sudden weight loss  520 non-null    object
 5   weakness            520 non-null    object
 6   Polyphagia          520 non-null    object
 7   Genital thrush      520 non-null    object
 8   visual blurring     520 non-null    object
 9   Itching             520 non-null    object
 10  Irritability        520 non-null    object
 11  delayed healing     520 non-null    object
 12  partial paresis     520 non-null    object
 13  muscle stiffness    520 non-null    object
 14  Alopecia            520 non-null    object
 15  Obesity             520 non-null    object
 16  class               520 non-null    object
dtypes: int64(1), object(16)
memory usage: 69.2+ KB
None

In [15]: print(df.describe())

              Age
count  520.000000
mean    48.028846
std     12.151466
min     16.000000
25%     39.000000
50%     47.500000
75%     57.000000
max     90.000000

In [16]: # Encode categorical variables
         for col in df.columns:
```

Figure 3.16 – Binary logistic regression code

references,directories

Bibliography

[1] Alan Agresti. *Foundations of Linear and Generalized Linear Models*. Hoboken, NJ: Wiley, 2015.

[2] Alan Agresti. *An Introduction to Categorical Data Analysis*. 3rd. Hoboken, NJ: Wiley, 2019.

[3] Marzia Ahmed. *Maternal Health Risk*. UCI Machine Learning Repository. DOI: https://doi.org/10.24 2020.

[4] Jay L. Devore et al. *Simple Linear Regression model (PDF)*. https://www.colorado.edu/amath/sites/default/files/attached-files/ch12_0.pdf. [Online; accessed July 2025]. 2025.

[5] Walter T Ambrosius. *Methods in Molecular Biology,Topics on biostatistics*. Humana Press Inc, Totowa,New Jersey, 2007.

[6] J. Bruin. *newtest: command to compute new test @ONLINE*. Feb. 2011. URL: https://stats.oarc.ucla.edu/stata/ado/analysis/.

[7] Y H Chan. "Biostatistics 201: Linear Regression Analysis". In: *Singapore Med J* 45(2):55 (2004).

[8] Annette J. Dobson and Adrian G. Barnett. *An Introduction to Generalized Linear Models*. 4th. CRC Press, 2018. ISBN: 9781138741515.

[9] Norman R. Draper and Harry Smith. *Applied Regression Analysis*. 3rd. Wiley, 1998.

[10] *Early Stage Diabetes Risk Prediction*. DOI: https://doi.org/10.24432/C5VG8H. 2020.

[11] J. J. Faraway. *Extending the Linear Model with R*. CRC Press, 2006.

[12] Jenine K Harris. "Primer on binary logistic regression". In: *Fam Med Com Health* 9:e001290. doi:10.1136/ fmch-2021-001290 (2021).

[13] Jenine K Harris. "Statistics with R: solving problems using real-world data." In: *SAGE Publications* (2020).

[14] David W. Hosmer, Stanley Lemeshow, and Rodney X. Sturdivant. *Applied Logistic Regression*. 3rd. New York: Wiley, 2013.

[15] Gareth James et al. *An Introduction to Statistical Learning*. 2nd. New York: Springer, 2021.

[16] USA Joseph M.Hilbe Jet Propulsion Laboratory California Institute of Technology and USA Arizona State University. *Practical Guide to Logistic Regression*. CRC Press Taylor Francis Group, 2015.

[17] Michael H. Kutner et al. *Applied Linear Statistical Models*. 5th. McGraw-Hill/Irwin, 2005.

[18] Peter McCullagh and John A. Nelder. *Generalized Linear Models*. 2nd. Chapman and Hall/CRC, 1989.

[19] Douglas C. Montgomery, Elizabeth A. Peck, and G. Geoffrey Vining. *Introduction to Linear Regression Analysis*. 5th. Wiley, 2012.

[20] Todd G. Nick and Kathleen M. Campbell. "Logistic Regression". In: *Methods in Molecular Biology* 404 (2007), pp. 273–301. DOI: 10.1007/978-1-59745-530-5_14.

[21] Shalabh. "Regression Analysis". In: *IIT Kanpur* (), chapter 2, Simple Linear Regression.

[22] Wikipedia contributors. *Linear regression — Wikipedia, The Free Encyclopedia*. https://en.wikipedia.org/wiki/Linear_regression. [Online; accessed July 2025]. 2025.

[23] Wikipedia contributors. *Proofs involving ordinary least squares — Wikipedia, The Free Encyclopedia*. https://en.wikipedia.org/wiki/Proofs_involving_ordinary_least_squares. [Online; accessed July 2025]. 2025.

[24] Wikipedia contributors. *Simple linear regression — Wikipedia, The Free Encyclopedia*. https://en.wikipedia.org/wiki/Simple_linear_regression. [Online; accessed July 2025]. 2025.

Abstract

This thesis explores the application of logistic regression models—binary, multinomial, and ordinal—for analyzing categorical outcomes in real-world scenarios. It begins with foundational concepts of linear regression and transitions into logistic regression, emphasizing maximum likelihood estimation (MLE) for parameter estimation. Key statistical tools like odds ratios, log-odds, and confidence intervals are interpreted to reveal predictor-outcome relationships. Practical applications include predicting diabetes diagnosis, classifying student program choices, and assessing maternal mortality risks, evaluated using confusion matrix metrics and diagnostic tests. The study highlights logistic regression's efficiency as both a predictive and explanatory tool, offering actionable insights for decision-making in healthcare and education.

Keywords: Binary Logistic Regression, Multinomial Logistic Regression, Ordinal Logistic Regression, Maximum Likelihood Estimation, Odds Ratios, Confusion Matrix.