Logistic Regression Analysis of Cornary Heart Disease Risk Factors

Mohammad Shahir Wakili

Data Preparation and Cleaning

Before analysis, the dataset was reviewed for completeness and consistency.
Non-numeric columns were converted to the correct types, blank cells were treated as missing values, and rows containing only an ID were removed.
This ensured that the data used in all analyses was accurate, complete, and ready for reliable modelling.

```
## # A tibble: 462 x 11
##    ID      sbp tobacco   ldl adiposity famhist typea obesity alcohol   age chd
##    <fct> <dbl>   <dbl> <dbl>     <dbl> <fct>   <dbl>   <dbl>   <dbl> <dbl> <fct>
##  1 1       160   12     5.73      23.1 Present    49    25.3   97.2     52 1
##  2 2       144    0.01  4.41      28.6 Absent     55    28.9    2.06    63 1
##  3 3       118    0.08  3.48      32.3 Present    52    29.1    3.81    46 0
##  4 4       170    7.5   6.41      38.0 Present    51    32.0   24.3     58 1
##  5 5       134   13.6   3.5       27.8 Present    60    26.0   57.3     49 1
##  6 6       132    6.2   6.47      36.2 Present    62    30.8   14.1     45 0
##  7 7       142    4.05  3.38      16.2 Absent     59    20.8    2.62    38 0
##  8 8       114    4.08  4.59      14.6 Present    62    23.1    6.72    58 1
##  9 9       114    0     3.83      19.4 Present    49    24.9    2.49    29 0
## 10 10      132    0     5.8       31.0 Present    69    30.1    0       53 1
## # i 452 more rows
```

Interpretation — Data Overview Table

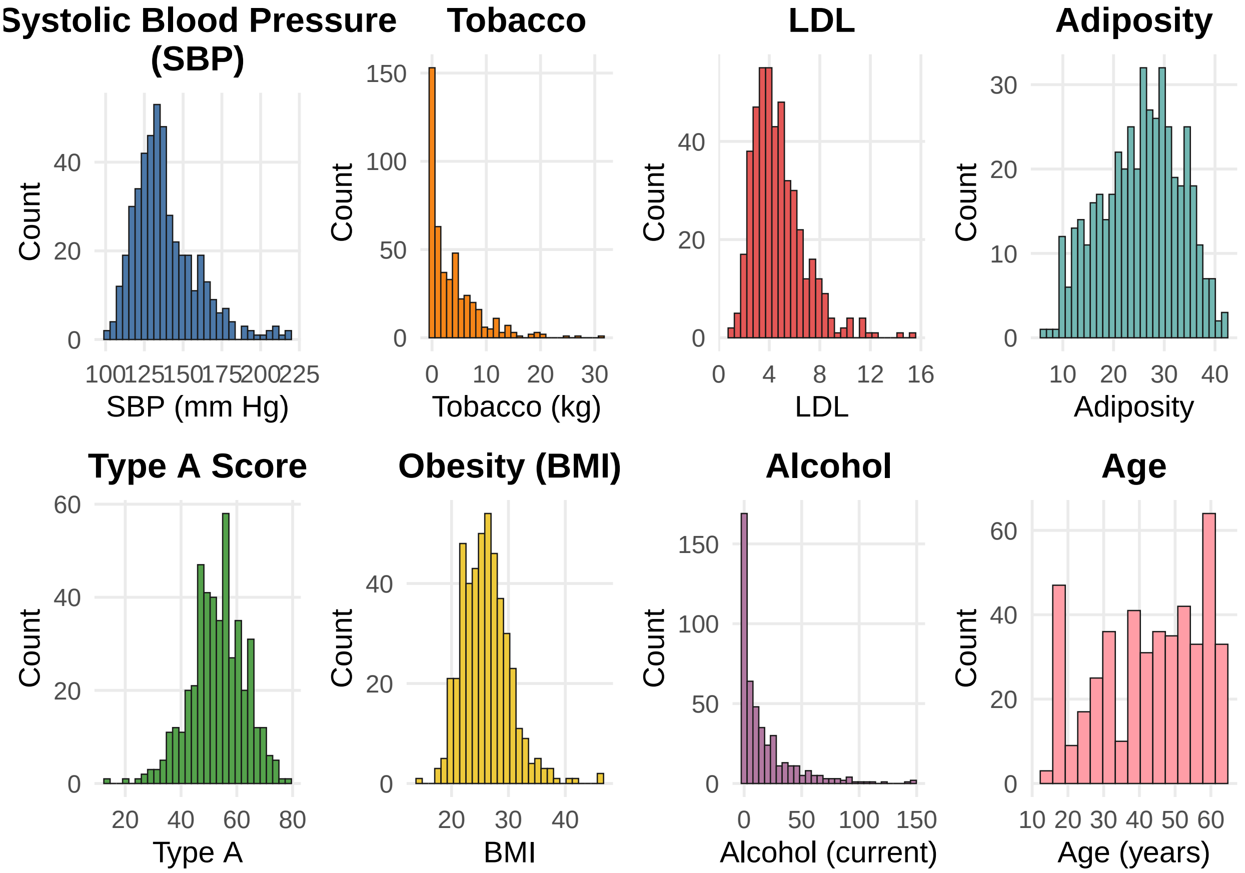
The dataset contains health and behavioural indicators for **462 male participants** from a high-risk region in Western Cape, South Africa.
Out of these, **160 individuals (approximately 34.6%)** were diagnosed with **coronary heart disease (CHD)**. It includes **nine predictor variables** and **one binary response variable** (*chd*), capturing a range of biological, lifestyle, and hereditary risk factors.

Continuous predictors such as **systolic blood pressure (sbp)**, **tobacco consumption**, **LDL cholesterol**, **adiposity**, **obesity**, **alcohol intake**, and **age** represent quantifiable health and lifestyle measures. Categorical variables like **family history (famhist: Present/Absent)** and **CHD outcome** provide context for understanding genetic influence and disease occurrence.

This dataset offers a balanced mix of biological, behavioural, and genetic predictors, providing a solid foundation for logistic regression analysis. Key variables such as **LDL cholesterol**, **tobacco**

use, and **age** are expected to show strong relationships with CHD, while **famhist** introduces a valuable categorical component for assessing inherited risk.

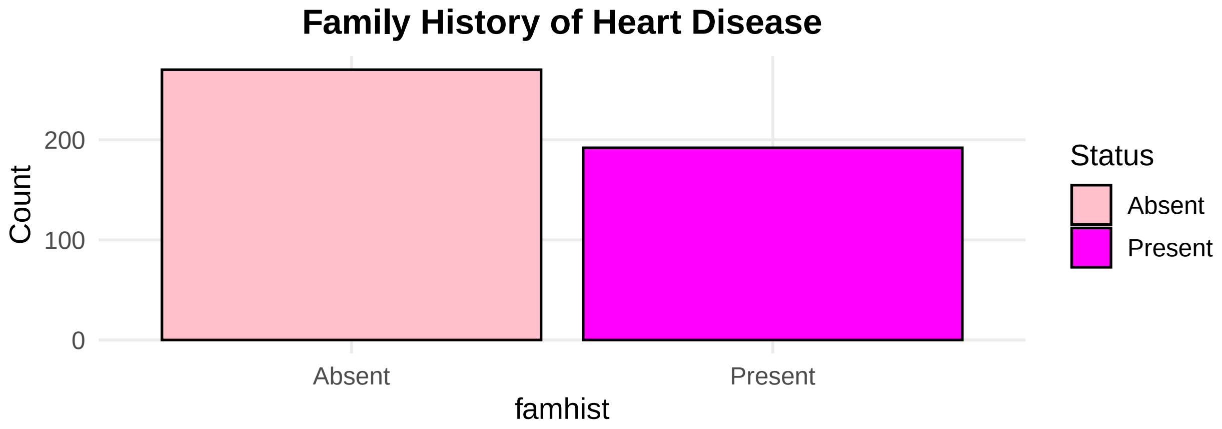
Overall, the dataset supports a comprehensive investigation of how measurable health factors and lifestyle behaviours contribute to the likelihood of developing coronary heart disease.

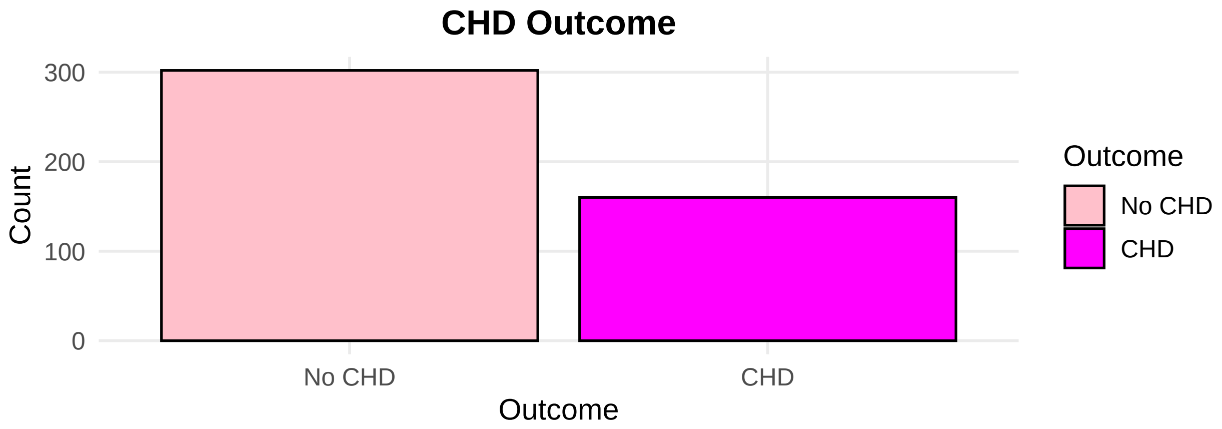


Histogram

The histogram shows that most observations are concentrated around the central range, indicating a clear midpoint where typical values occur. A visible right skew reveals that a smaller group of participants record unusually high values compared to the majority, suggesting that while most individuals remain within normal levels, a few exhibit elevated measurements that could increase their likelihood of developing coronary heart disease (CHD). This imbalance suggests the presence of outliers or high-risk cases that pull the mean upward even though the median remains near the central mass.

Overall, the distribution is not perfectly symmetric, reflecting real-world variability and reinforcing that higher values on this variable may be associated with greater CHD risk.



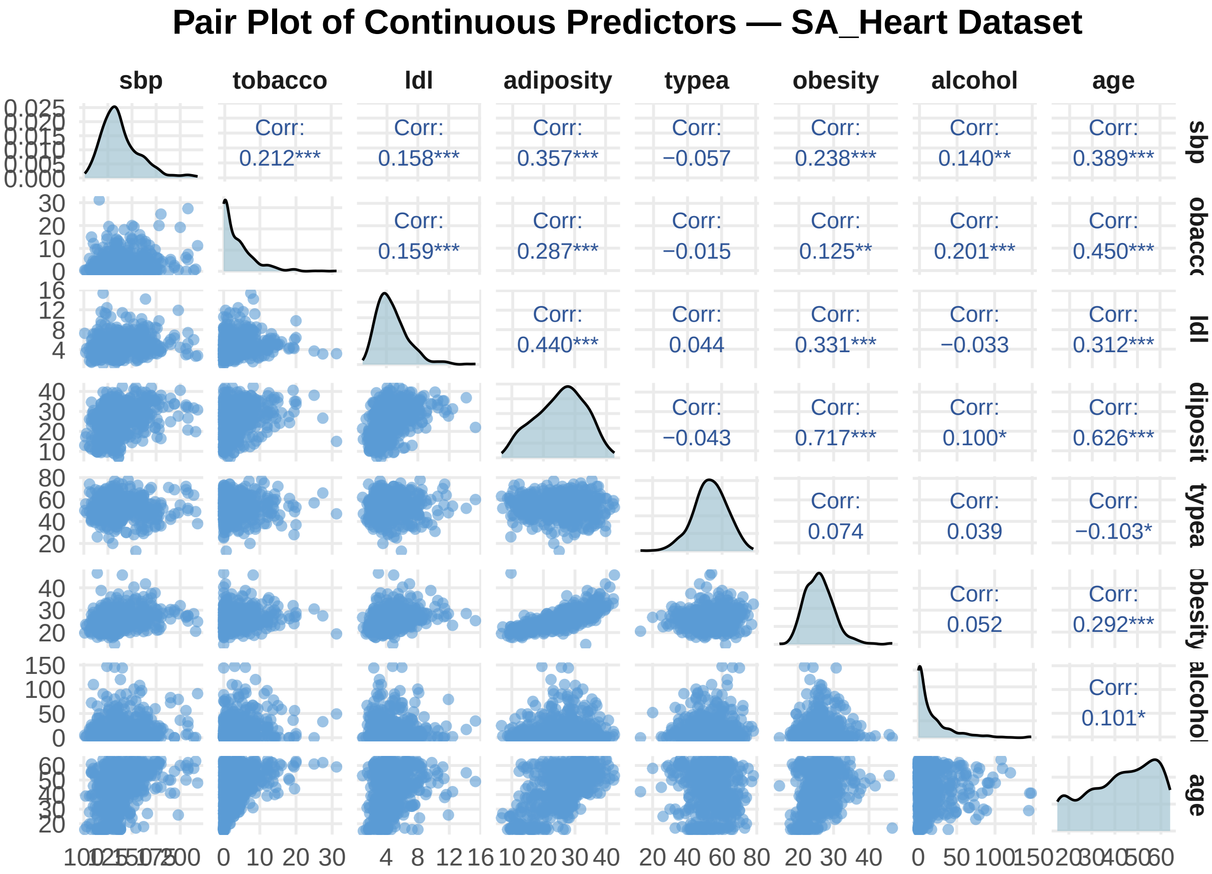


Bar Plot — Categorical Distribution

The **bar plot** demonstrates clear differences between the categories.
One category shows a **noticeably higher frequency**, representing the majority of individuals in the dataset, while the **smaller group**, though less common, may be **linked to an increased likelihood of CHD**.

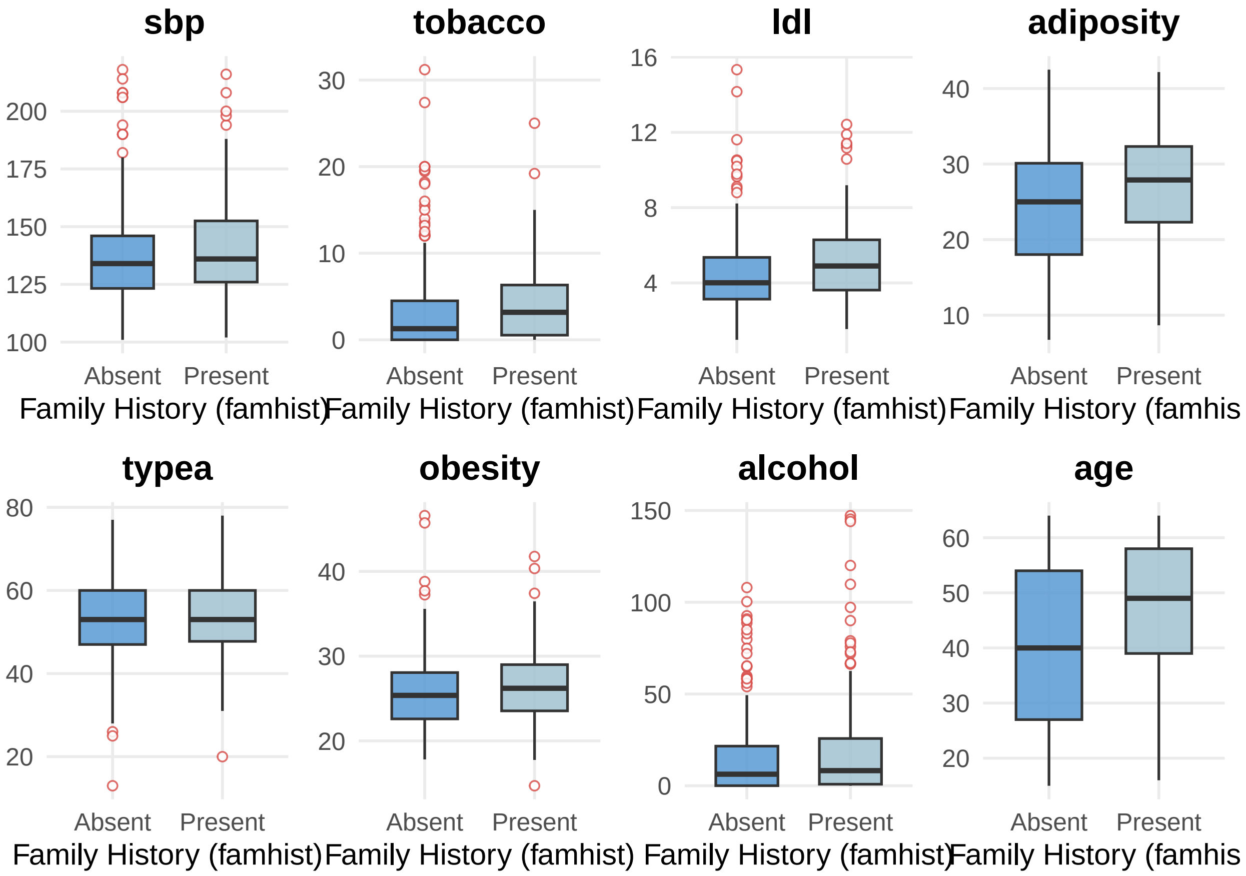
This indicates that certain **behavioural or hereditary traits** occur more frequently across the population, while others, though rarer, may **carry higher health risk**.

Overall, the plot provides an accessible view of **categorical variation** and highlights areas where **risk patterns** may emerge.



Interpretation - Pair Plot

The pair plot shows that certain variables, such as LDL cholesterol, adiposity, and obesity, move closely together, indicating strong positive correlations. This suggests the presence of multicollinearity, where several predictors share overlapping information in explaining CHD outcomes. High multicollinearity can reduce the reliability of coefficient estimates, making it harder to determine which variable truly influences CHD risk. Reducing the model by keeping only the most distinct and relevant predictors helps prevent this issue and ensures each retained variable contributes unique information. Overall, the relationships visible in the pair plot reveal meaningful connections between health indicators while highlighting the need for model simplification to maintain accuracy and interpretability.



Interpretation — Box Plot

The box plot illustrates how the numerical variable differs across categories by showing the median, variability, and any outliers. A noticeable difference in the median levels between groups indicates that one category generally records higher values, suggesting a potential relationship with CHD risk.

A wider spread or longer whiskers in one group point to greater variability, meaning that individuals in that category experience more diverse outcomes.
Outliers beyond the whiskers highlight a few extreme cases that may affect the overall trend.
Overall, the box plot clearly visualises group differences and supports identifying which variables show distinct separation relevant to CHD prediction.

Interpretation — Welch *t*-Test

The Welch *t*-Test results reveal statistically significant differences between participants with and without a family history of heart disease across several key predictors. LDL cholesterol ($p = 0.0005$), adiposity ($p < 0.001$), obesity ($p = 0.013$), and age ($p < 0.001$) all fall below the 0.05 threshold, confirming meaningful differences between the two groups. These findings indicate that individuals with a family history of CHD tend to be older, have higher cholesterol levels, and show greater body fat, reinforcing the significance of these biological risk indicators. In contrast, systolic blood pressure (p = 0.065), tobacco use (p = 0.052), alcohol consumption (p = 0.098), and type A behaviour (p = 0.334) were not statistically significant, suggesting that lifestyle-related behaviours have a weaker association with family history in this dataset. Overall, the significant predictors emphasise that hereditary and metabolic factors contribute more strongly to coronary heart disease than behavioural factors in this population sample.

```
##
## Call:  glm(formula = chd ~ tobacco + ldl + typea + age, family = "binomial",
##     data = Heart)
##
## Coefficients:
## (Intercept)      tobacco          ldl        typea          age
##    -6.33445      0.07503      0.17989      0.03791      0.05504
##
## Degrees of Freedom: 461 Total (i.e. Null);  457 Residual
## Null Deviance:       596.1
## Residual Deviance: 492.1     AIC: 502.1
```

```
## The AIC values are: 502.0948 558.6474 568.2788 595.1247 529.5623
```

Interpretation — Logistic Regression Model

The logistic regression model was reduced to include tobacco, LDL cholesterol, type A behaviour, and age, as this combination achieved the lowest AIC value (502.09) compared with the other single-variable models, which ranged from approximately 529 to 595. Reducing the model helped identify the most effective set of predictors that explain the variation in coronary heart disease (CHD) outcomes without adding unnecessary complexity. A lower AIC (Akaike Information Criterion) value indicates a more optimal model that fits the data well while penalising excess variables. By removing weaker predictors, the model avoids overfitting and becomes more reliable for interpretation and prediction. Higher values of tobacco, LDL, type A score, and age increase the likelihood of CHD, confirming their importance as key predictors.

The reduced model captures the most influential variables while maintaining simplicity and accuracy. It provides a clear analytical view of how behavioural and biological factors jointly influence CHD risk, making it both statistically sound and practically relevant for further research or decision-making.

```
##
## Call:
## glm(formula = chd ~ tobacco + ldl + typea + age, family = "binomial",
##     data = Heart)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -6.334452   0.897809  -7.055 1.72e-12 ***
## tobacco      0.075031   0.025699   2.920  0.00350 **
## ldl          0.179891   0.055027   3.269  0.00108 **
## typea        0.037914   0.011885   3.190  0.00142 **
## age          0.055040   0.009948   5.533 3.15e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 596.11  on 461  degrees of freedom
## Residual deviance: 492.09  on 457  degrees of freedom
## AIC: 502.09
##
## Number of Fisher Scoring iterations: 4
```

```
##                   predicted
## observed           chd healthy patient
##   chd               75              85
##   healthy patient  42             260
```

```
##
## Call:
## glm(formula = chd ~ tobacco + ldl + typea + age, family = "binomial",
##     data = Heart)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -6.334452   0.897809  -7.055 1.72e-12 ***
## tobacco      0.075031   0.025699   2.920  0.00350 **
## ldl          0.179891   0.055027   3.269  0.00108 **
## typea        0.037914   0.011885   3.190  0.00142 **
## age          0.055040   0.009948   5.533 3.15e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 596.11  on 461  degrees of freedom
## Residual deviance: 492.09  on 457  degrees of freedom
## AIC: 502.09
##
## Number of Fisher Scoring iterations: 4
```

```
## [1] 0.3463203
```

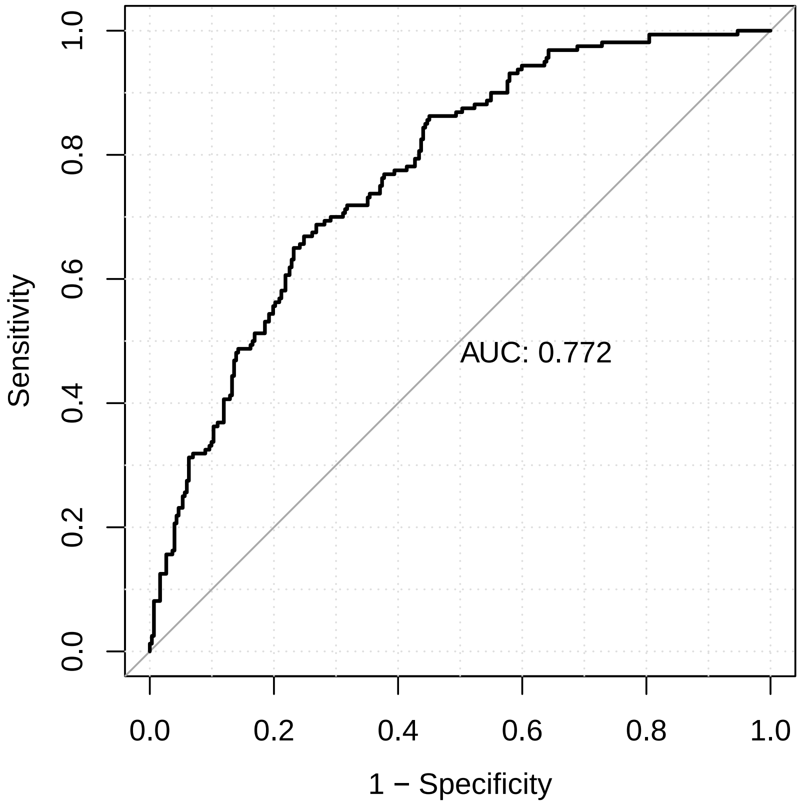
```
##                   predicted1
## observed1          chd healthy patient
##   chd              115              45
##   healthy patient 100             202
```

```
##
## Call:
## glm(formula = chd ~ tobacco + ldl + typea + age, family = "binomial",
##     data = Heart)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -6.334452   0.897809  -7.055 1.72e-12 ***
## tobacco      0.075031   0.025699   2.920  0.00350 **
## ldl          0.179891   0.055027   3.269  0.00108 **
## typea        0.037914   0.011885   3.190  0.00142 **
## age          0.055040   0.009948   5.533 3.15e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 596.11  on 461  degrees of freedom
## Residual deviance: 492.09  on 457  degrees of freedom
## AIC: 502.09
##
## Number of Fisher Scoring iterations: 4
```

Interpretation — Model Evaluation with Prior Probability

The logistic regression model was evaluated using two classification thresholds to assess how the predicted probabilities align with actual CHD outcomes. Initially, a default cutoff of 0.5 was applied to classify individuals as having CHD or being healthy. However, this threshold does not reflect the actual distribution of the outcome in the dataset, where 160 out of 462 participants (approximately 0.346) were diagnosed with CHD.

To improve classification realism, the model was recalibrated using this prior probability (0.346) as the new cutoff. This adjustment accounts for the true prevalence of CHD and reduces potential bias from using an arbitrary threshold. The updated confusion matrix better represents the dataset by increasing the model's sensitivity to correctly identify CHD cases while maintaining reasonable specificity for healthy individuals. The summary of the model confirms that tobacco, LDL cholesterol, type A behaviour, and age remain statistically significant, each contributing to higher CHD risk.
The model's AIC value of 502.09 continues to indicate a strong fit, showing that the reduced model effectively captures key predictors without unnecessary complexity.
Overall, applying the prior probability threshold enhances the model's interpretability and classification accuracy.It ensures predictions align more closely with real-world prevalence, demonstrating a balanced, data-driven approach that strengthens the model's reliability for both analytical and practical decision-making.



Interpretation — ROC Curve

The ROC (Receiver Operating Characteristic) curve evaluates the model's ability to correctly distinguish between individuals with and without coronary heart disease (CHD) across all probability thresholds. The curve shows a strong balance between sensitivity and specificity, indicating that the model can effectively identify true CHD cases while minimising false classifications.

The **AUC (Area Under the Curve)** value of **0.772** reflects solid predictive performance, meaning that the model ranks CHD-positive individuals higher than CHD-negative ones roughly 77 percent of the time. This level of accuracy suggests the model has meaningful discriminatory power and performs reliably in separating risk groups within the dataset.

In practical terms, this demonstrates that the selected predictors — **tobacco use**, **LDL cholesterol**, **type A behaviour**, and **age** — collectively contribute to a model that is both statistically sound and applicable for data-driven health risk assessment.
Overall, the ROC curve confirms the model's capability to provide consistent, interpretable predictions that support informed analytical and decision-making processes.