Individual Assignment 2: Logistic Regression Analysis

Alexandre

2024-11-18

1. Data Preparation.

(a). Loading data

```
# loading libraries
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
# loading  data
data <- read.csv("heart.csv")

# Display the first few rows of the selected data
head(data)
```

```
##   age sex cp trestbps chol fbs restecg thalach exang oldpeak slope ca thal
## 1  52   1  0      125  212   0       1     168     0     1.0     2  2    3
## 2  53   1  0      140  203   1       0     155     1     3.1     0  0    3
## 3  70   1  0      145  174   0       1     125     1     2.6     0  0    3
## 4  61   1  0      148  203   0       1     161     0     0.0     2  1    3
## 5  62   0  0      138  294   1       1     106     0     1.9     1  3    2
## 6  58   0  0      100  248   0       0     122     0     1.0     1  0    2
##   target
## 1      0
## 2      0
## 3      0
## 4      0
## 5      0
## 6      1
```

- Answer: The variables and their descriptions are:

1. age : Age of the individual in years.
2. sex : Gender of the individual (1 = male, 0 = female).
3. cp : Chest pain type (0-3).
4. trestbps : Resting blood pressure (in mm Hg).
5. chol : Serum cholesterol in mg/dl.

6. fbs : Fasting blood sugar > 120 mg/dl (1 = true, 0 = false).
7. restecg : Resting electrocardiographic results (0-2).
8. thalach : Maximum heart rate achieved.
9. exang : Exercise induced angina (1 = yes, 0 = no).
10. oldpeak : ST depression induced by exercise relative to rest.
11. slope : Slope of the peak exercise ST segment (0-2).
12. ca : Number of major vessels (0-3) colored by fluoroscopy.
13. thal : Thalassemia (1 = normal, 2 = fixed defect, 3 = reversible defect).
14. target : Diagnosis of heart disease (1 = presence, 0 = absence).

(b). Variable Selection:

```
# Select the relevant variables for analysis
selected_data <- data %>%
  select(target, age, sex, cp, chol)

# Display the first few rows of the selected data
head(selected_data)
```

```
##   target age sex cp chol
## 1      0  52   1  0  212
## 2      0  53   1  0  203
## 3      0  70   1  0  174
## 4      0  61   1  0  203
## 5      0  62   0  0  294
## 6      1  58   0  0  248
```

2. Data Cleaning

```
selected_data <- na.omit(selected_data) # Remove rows with missing values



# Check if cp exists: This form is used because we had an unknown error with the standard
if ("cp" %in% colnames(selected_data)) {
    # Convert cp to factor
    selected_data$cp <- as.factor(selected_data$cp)
} else {
    print("Variable 'cp' not found in the dataset.")
}
# Check if sex exists
if ("sex" %in% colnames(selected_data)) {
    # Convert sex to factor
    selected_data$sex <- as.factor(selected_data$sex)
} else {
    print("Variable 'sex' not found in the dataset.")
}



head(selected_data)
```

```
##   target age sex cp chol
## 1      0  52   1  0  212
## 2      0  53   1  0  203
```

```
## 3      0  70   1  0  174
## 4      0  61   1  0  203
## 5      0  62   0  0  294
## 6      1  58   0  0  248
```

```
#check the summary of the cleaned dataset
summary(selected_data)
```

```
##      target            age        sex     cp           chol
##  Min.   :0.0000   Min.   :29.00   0:312   0:497   Min.   :126
##  1st Qu.:0.0000   1st Qu.:48.00   1:713   1:167   1st Qu.:211
##  Median :1.0000   Median :56.00           2:284   Median :240
##  Mean   :0.5132   Mean   :54.43           3: 77   Mean   :246
##  3rd Qu.:1.0000   3rd Qu.:61.00                   3rd Qu.:275
##  Max.   :1.0000   Max.   :77.00                   Max.   :564
```

```
# Check for duplicated records
duplicates <- selected_data[duplicated(selected_data), ]
#print(duplicates)
# Display the duplicated records, if any
if (nrow(duplicates) > 0) {
    print("Duplicated records found:")
    #print(duplicates)
} else {
    print("No duplicated records found.")
}
```

```
## [1] "Duplicated records found:"
```

```
#summary(selected_data)
#head(selected_data)
```

```
# Deleting all the duplicating rows
df_heart <- selected_data[!duplicated(selected_data),]
```

```
# The dimensions of the cleaned dataset we will use for this analysis
dim(df_heart)
```

```
## [1] 302   5
```

Then, there are 302 rows and 05 columns in the cleaned dataset (df_heart).

3. Exploratory Data Analysis

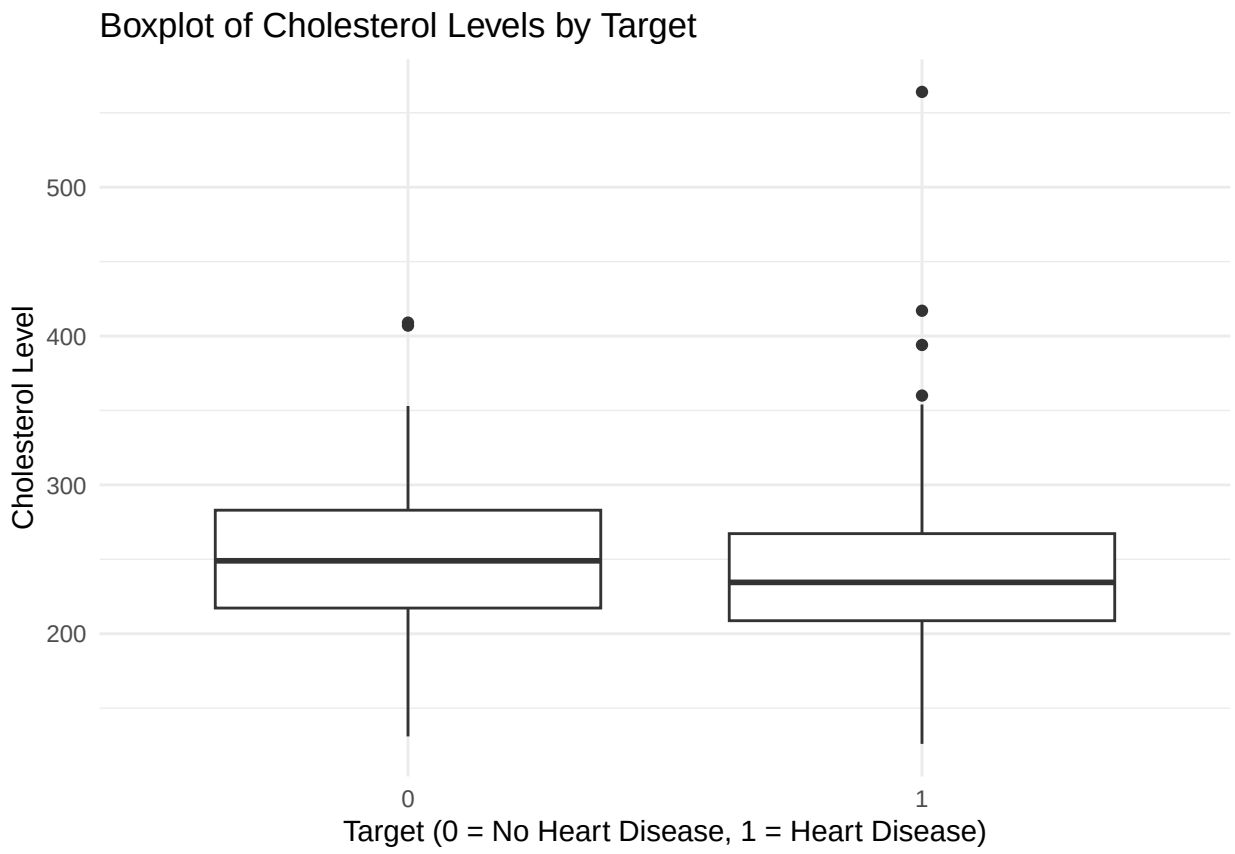
* Let's summarize the variables using the summary() function.

```
# Summarizing the variables
summary(df_heart)
```

```
##      target            age        sex     cp           chol
##  Min.   :0.000   Min.   :29.00   0: 96   0:143   Min.   :126.0
##  1st Qu.:0.000   1st Qu.:48.00   1:206   1: 50   1st Qu.:211.0
##  Median :1.000   Median :55.50           2: 86   Median :240.5
##  Mean   :0.543   Mean   :54.42           3: 23   Mean   :246.5
##  3rd Qu.:1.000   3rd Qu.:61.00                   3rd Qu.:274.8
##  Max.   :1.000   Max.   :77.00                   Max.   :564.0
```

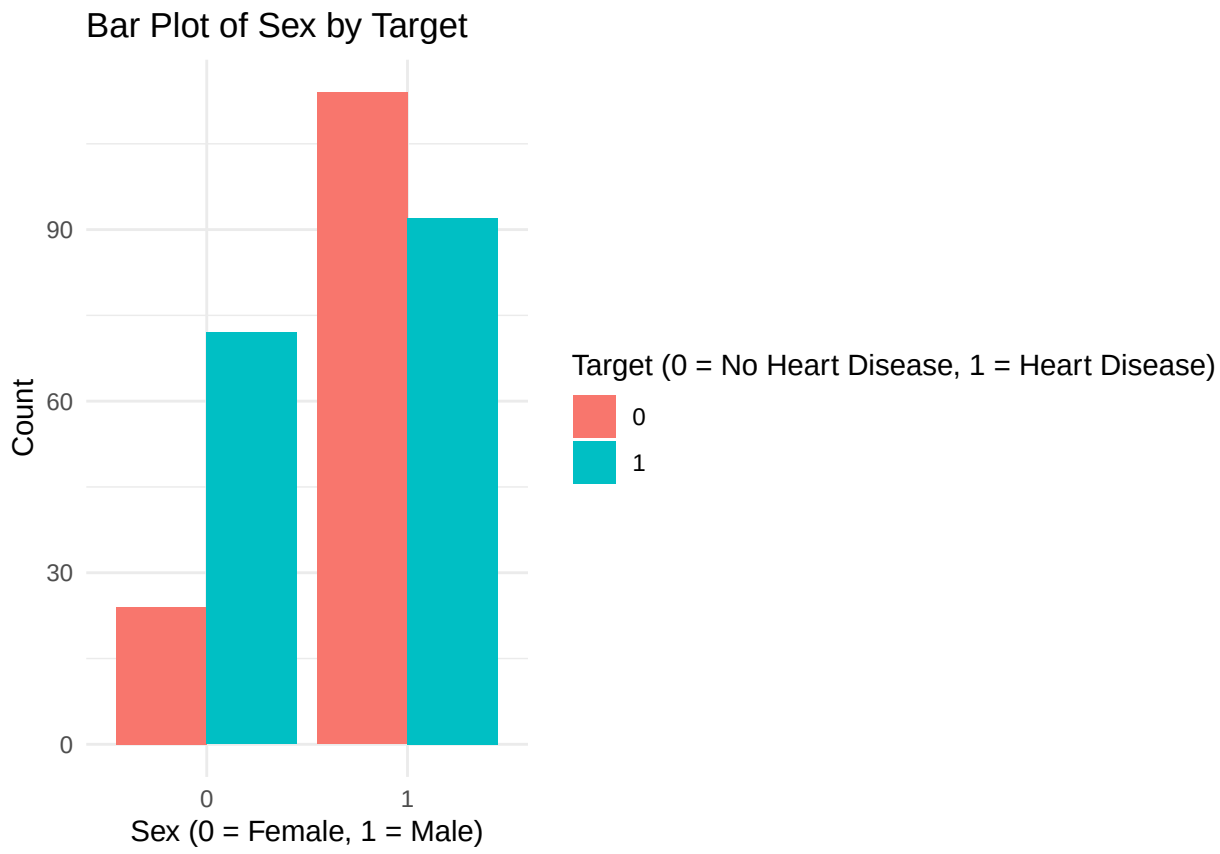
* Create boxplots of chol across levels of target.

```r
# Create boxplot of cholesterol levels across target levels
ggplot(df_heart, aes(x = factor(target), y = chol)) +
  geom_boxplot() +
  labs(title = "Boxplot of Cholesterol Levels by Target",
       x = "Target (0 = No Heart Disease, 1 = Heart Disease)",
       y = "Cholesterol Level") +
  theme_minimal()
```



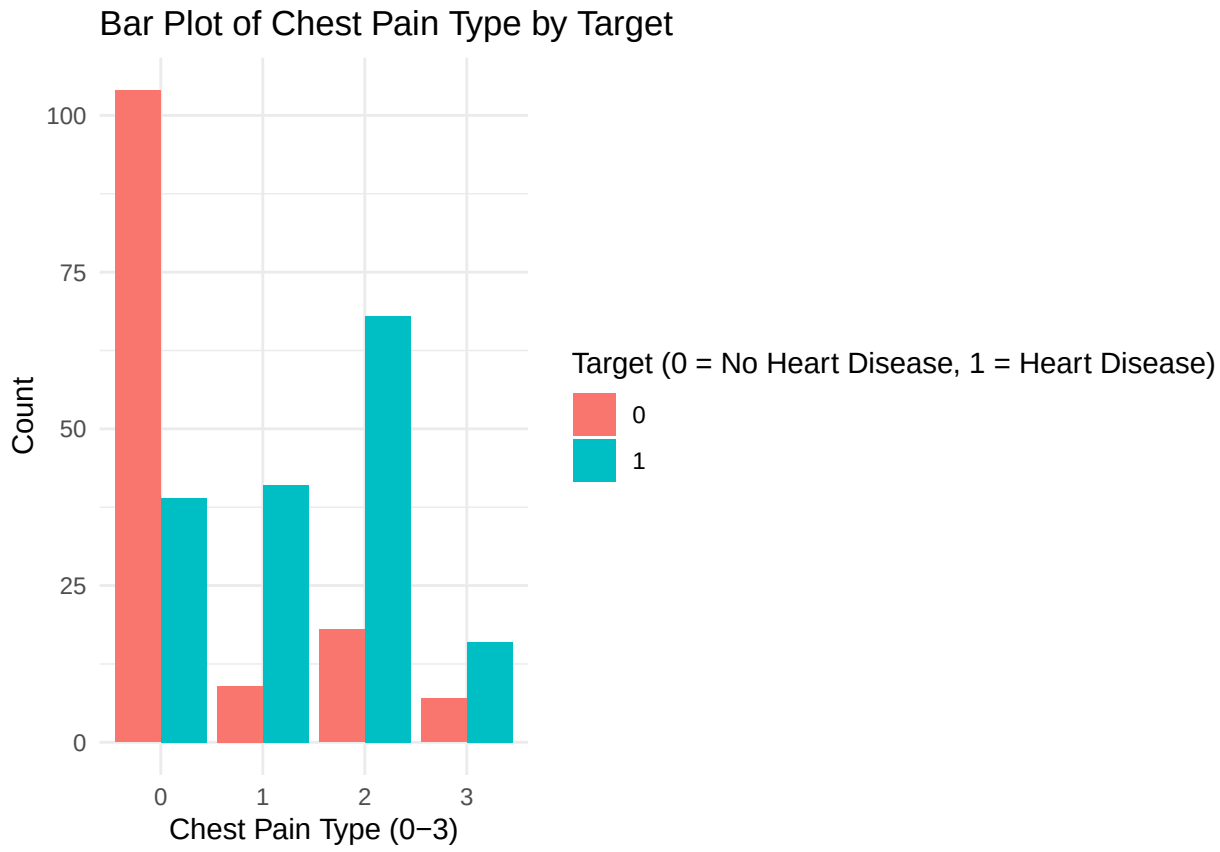
- Bar plot for sex with respect to target

```r
# Bar plot for sex with respect to target
ggplot(df_heart, aes(x = factor(sex), fill = factor(target))) +
  geom_bar(position = "dodge") +
  labs(title = "Bar Plot of Sex by Target",
       x = "Sex (0 = Female, 1 = Male)",
       y = "Count",
       fill = "Target (0 = No Heart Disease, 1 = Heart Disease)") +
  theme_minimal()
```



* Bar plot for cp with respect to target

```
# Bar plot for cp with respect to target
ggplot(df_heart, aes(x = factor(cp), fill = factor(target))) +
  geom_bar(position = "dodge") +
  labs(title = "Bar Plot of Chest Pain Type by Target",
       x = "Chest Pain Type (0-3)",
       y = "Count",
       fill = "Target (0 = No Heart Disease, 1 = Heart Disease)") +
  theme_minimal()
```



- What patterns do we observe in the data?

1. Cholesterol level and heart diseases: The boxplots of cholesterol across levels of target shows that the cholesterol level between the two statuses seems very close.
2. For this population, in the case of females, there are more diseased than non-diseased, whereas in the case of males, there are more non-diseased than diseased.

3.The proportion of non-diseased hearts with 0 pain is very high compared to diseased hearts with no pain (which means that the majority of people with 0 pain are non-diseased) and in contrast, the majority of those with at least one pain are diseased.

4. Fitting a Logistic Regression Model

* Fitting the model to predict target using Age, Sex, cp and chol

```
# Logistic regression model
model <- glm(target ~ age + sex + cp + chol,
data = df_heart, family = "binomial")

summary(model)
```

```
##
## Call:
## glm(formula = target ~ age + sex + cp + chol, family = "binomial",
##     data = df_heart)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
```

```
## -2.3720  -0.7044   0.2467   0.7322   2.2976
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  4.956253   1.275701   3.885 0.000102 ***
## age         -0.063368   0.017892  -3.542 0.000398 ***
## sex1        -1.891184   0.361858  -5.226 1.73e-07 ***
## cp1          2.538803   0.448243   5.664 1.48e-08 ***
## cp2          2.360416   0.356639   6.619 3.63e-11 ***
## cp3          2.237305   0.535818   4.175 2.97e-05 ***
## chol        -0.004796   0.002862  -1.676 0.093777 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 416.42  on 301  degrees of freedom
## Residual deviance: 289.87  on 295  degrees of freedom
## AIC: 303.87
##
## Number of Fisher Scoring iterations: 5
```

Interpretation of coefficients:

1. Intercept (4.956253) represents the log-odds of having heart disease when all predictors are zero. In our case, it is not directly interpretable since age and cholesterol cannot be zero.
2. Age (-0.063368) indicates that for each additional year of age, the log-odds of having heart disease decrease by approximately 0.0634. This suggests that older peoples are less likely to have heart disease in our model.
3. Sex (sex1 = -1.891184) for sex1 (where 1 indicates male) shows that males are less likely to have heart disease compared to females.
4. Chest Pain Types: All these coefficients(cp1, cp2, cp3) suggest that higher chest pain types are strongly associated with an increased likelihood of heart disease compared to those without chest pain (cp = 0)

5. Let's convert the coefficients into odds ratios using exp(coef()).

```r
# Compute odds ratios
exp(coef(model))
```

```
## (Intercept)         age        sex1         cp1         cp2         cp3
## 142.0605407   0.9385984   0.1508931  12.6645085  10.5953589   9.3680525
##        chol
##   0.9952157
```

* Interpretations:

* Sex1 = 0.1508931: This means that males are about 85% less likely to have heart disease compared to females

* cp1 = 12.6645085: poeple with chest pain type 1 have odds of heart disease that are approximately 12.66 times higher compared to those without chest pain (cp = 0).

* cp2 = 10.5953589: Similarly, individuals with chest pain type 2 have odds of heart disease that are about 10.60 times higher than those without chest pain.

* cp3 = 9.3680525: For chest pain type 3, the odds of heart disease are approximately 9.37 times higher compared to those without chest pain.

* Chol = 0.9952157: indicates that for each unit increase in cholesterol level, the odds of having heart disease decrease by about 0.48% (since 1-0.9952157 = 0.00481).

6. Model Comparison

1. Fit a reduced model excluding the chol variable.

```
# Reduce model
reduced_model <- glm(target ~ age + sex + cp,
data = df_heart, family = "binomial")
```

2. Perform a likelihood ratio test between the full and reduced models using anova().

```
# Likelihood ratio test
anova(reduced_model, model, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: target ~ age + sex + cp
## Model 2: target ~ age + sex + cp + chol
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       296     292.65
## 2       295     289.87  1   2.7835  0.09524 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Interpretation:
- The p-value (P = 0.09524 > 0.05) indicates that adding chol does not significantly improve the model.

7. Model Predictions and Performance

1. Predict probabilities of heart disease for all individuals.

```
# Predict probabilities
df_heart$pred_prob <- predict(model, type = "response")
```

2. Convert these probabilities into binary predictions using a threshold of 0.5.

```
# Convert probabilities to binary predictions
df_heart$predicted <- ifelse(df_heart$pred_prob > 0.5, 1, 0)
```

3. Create a confusion matrix to evaluate the model's performance.

```
# Confusion matrix
table(Predicted = df_heart$predicted, Actual = df_heart$target)
```

```
##          Actual
## Predicted   0   1
##         0 101  31
##         1  37 133
```

4. Accuracy of the model

```
# Calculate accuracy
mean(df_heart$predicted == df_heart$target)
```

```
## [1] 0.7748344
```

- Answer:
- The model correctly classify 77.48344 of the observations.
- The are 37 False Positives and 31 False Negatives

8. ROC Curve and AUC

1. Generate an ROC curve and calculate the AUC for the model using the pROC package.

- ROC Curves

```
# ROC curve
library(pROC)
```

```
## Type 'citation("pROC")' for a citation.
```

```
##
## Attaching package: 'pROC'
```

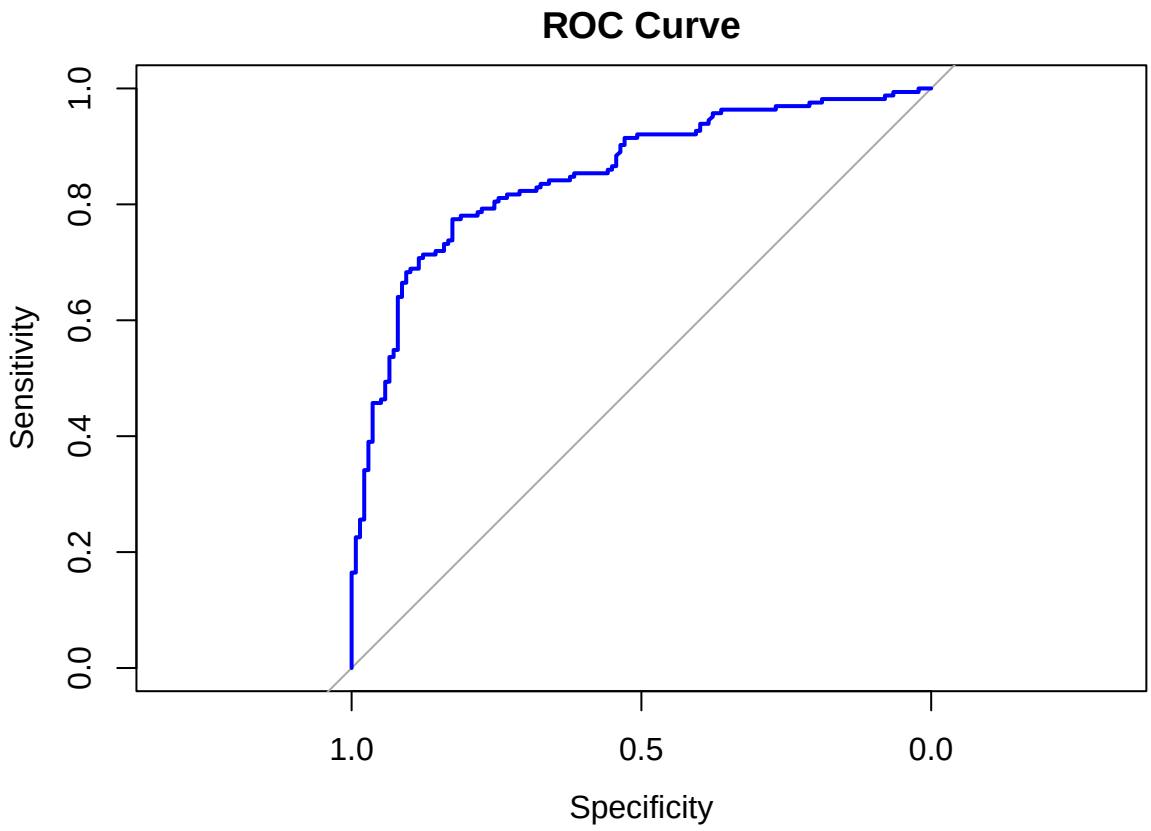
```
## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

```
roc_obj <- roc(df_heart$target, df_heart$pred_prob)
```

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```

```
plot(roc_obj, main = "ROC Curve", col = "blue", lwd = 2)
```



- Calculate AUC

```r
# Calculate AUC
auc(roc_obj)
```

```
## Area under the curve: 0.8517
```

2. Question: What does the AUC tell you about the model's performance? Is the model good at distinguishing between individuals with and without heart disease?

Answer: The AUC = 0.8517 indicates that our logistic regression model has strong predictive performance in distinguishing between individuals with and without heart disease