Logistic Regression Analysis

Ronak Fathi

January 24, 2025

Packages and Data Setup

```
# Install required packages (if not already installed)
# install.packages('tidymodels')
# install.packages('glmnet')

library(tidyverse)
library(gtsummary)
library(ggplot2)
library(ggpubr)
library(GGally)
library(readr)
library(tidymodels)
library(labelled)
```

```
# Set working directory and load data
setwd("C:/Users/0&1/OneDrive/Documents/R-Youtube")
data <- read.csv("CHDdata.csv")
```

Exploratory Data Analysis

```
# Basic structure
dim(data)
```

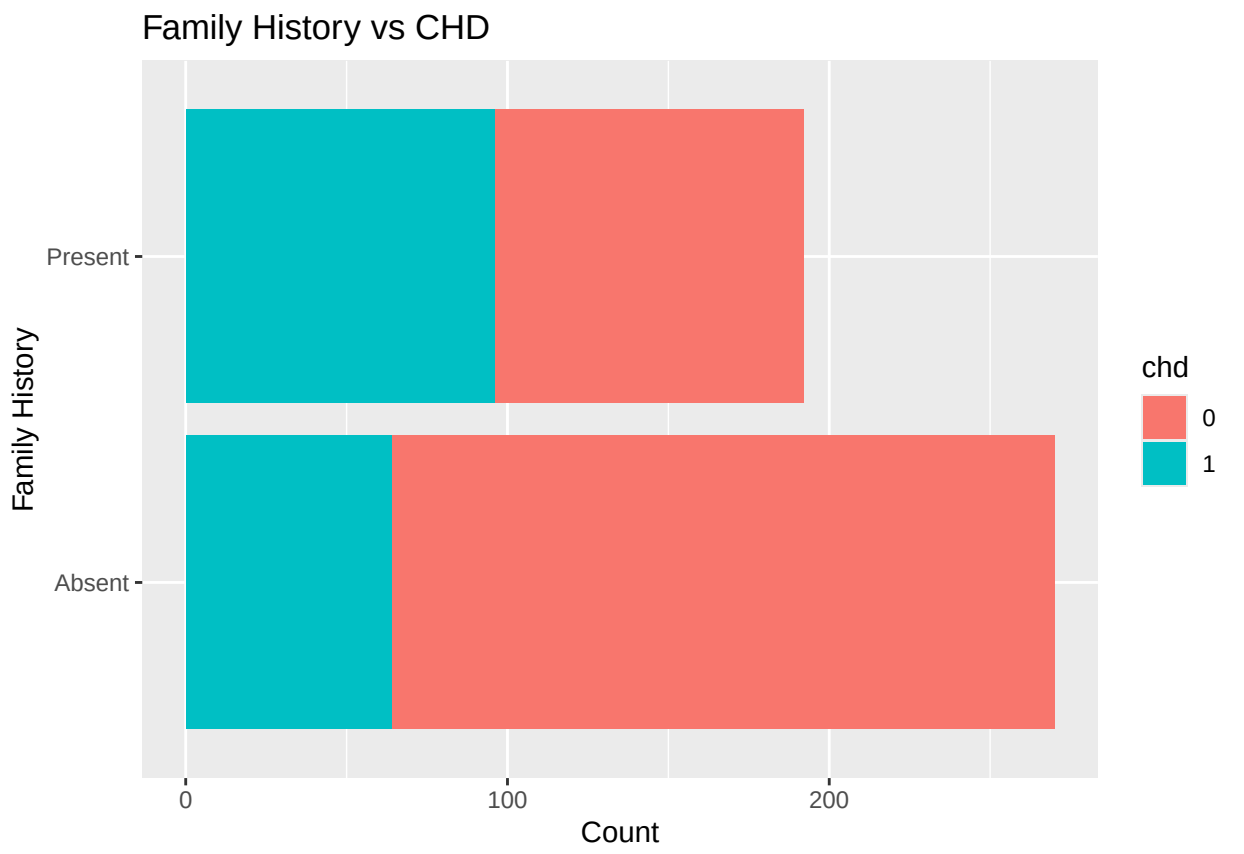
```
## [1] 462  10
```

```
head(data)
```

```
##   sbp tobacco  ldl adiposity famhist typea obesity alcohol age chd
## 1 160   12.00 5.73     23.11 Present    49   25.30   97.20  52   1
## 2 144    0.01 4.41     28.61  Absent    55   28.87    2.06  63   1
## 3 118    0.08 3.48     32.28 Present    52   29.14    3.81  46   0
## 4 170    7.50 6.41     38.03 Present    51   31.99   24.26  58   1
## 5 134   13.60 3.50     27.78 Present    60   25.99   57.34  49   1
## 6 132    6.20 6.47     36.21 Present    62   30.77   14.14  45   0
```

```
# Convert target variable to factor
data$chd <- as.factor(data$chd)

# Plot family history vs CHD
ggplot(data, aes(famhist, fill = chd)) +
  geom_bar() +
  coord_flip() +
  labs(title = "Family History vs CHD", x = "Family History", y = "Count")
```



Train-Test Split

```
set.seed(421)
split <- initial_split(data, prop = 0.8, strata = chd)
train <- training(split)
test <- testing(split)
```

Logistic Regression Model

```
model <- logistic_reg(mixture = 0, penalty = 0) %>%
  set_engine("glmnet") %>%
  set_mode("classification") %>%
  fit(chd ~ ., data = train)
```

```
tidy(model)
```

```
## # A tibble: 10 x 3
##    term           estimate penalty
##    <chr>             <dbl>   <dbl>
##  1 (Intercept)    -6.28          0
##  2 sbp             0.00883       0
##  3 tobacco         0.0692        0
##  4 ldl             0.148         0
##  5 adiposity       0.0260        0
```

```
##  6 famhistPresent  0.868         0
##  7 typea           0.0347        0
##  8 obesity        -0.0467        0
##  9 alcohol        -0.00111       0
## 10 age             0.0374        0
```

Model Predictions

```r
# Class predictions
pred_class <- predict(model, new_data = test, type = "class")

# Class probabilities
pred_proba <- predict(model, new_data = test, type = "prob")

# Combine predictions
results <- test %>%
  select(chd) %>%
  bind_cols(pred_class, pred_proba)
```

Model Evaluation

```r
accuracy(results, truth = chd, estimate = .pred_class)
```

```
## # A tibble: 1 x 3
##   .metric  .estimator .estimate
##   <chr>    <chr>          <dbl>
## 1 accuracy binary         0.720
```

Summary

- This analysis trains a logistic regression classifier to predict coronary heart disease (chd) using multiple features.
- Data was split into training and testing sets using stratified sampling.
- Model accuracy on the test set was calculated using the yardstick::accuracy()metric.