Summary of Regression Project Written Report [1]

[1] Original final report was written by me and other three students

Data Description

Our dataset, titled "Heart Attack Analysis & Prediction Dataset", comes from Kaggle and was created by Rashik Rahman Pritom. The dataset has 303 rows and 14 columns that consist of 13 input variables and 1 output variable. The output variable is binary where 0 is that the patient did not have a heart attack and 1 is that they did. The input variables and their types are listed in the table below, with detailed definitions of the variables.

Variable Name	**Data Type**	**Description**
age	Integer (B)	Age of person
sex	Integer (B)	Gender of Person (1= male, 0 = female)
cp	Integer	Chest Pain Type (1: typical angina 2: atypical angina 3: non-anginal pain 4: asymptomatic)
trtbps	Integer	Resting Blood Pressure (mm HG)
chol	Integer	Cholesterol in mg/dl
fbs	Integer (B)	If fasting blood sugar > 120 mg/dl ((1 = true; 0 = false)
thall	Integer	Thal Rate (1-3)
restecg	Integer	Resting ECG (0: normal, 1: ST-T wave abnormality 2: definite left ventricular hypertrophy)
thalachh	Integer	Max Heart Rate
exng	Integer (B)	Exercise Induced Angina (1=yes, 0=no)
oldpeak	Decimal	ST depression induced by exercise relative to rest

slp	Integer	Slope of the ST segment (0,1,2)
caa	Integer	Number of Major Vessels (1-4)
output	Integer	Target Variable (0= less chance of heart attack , 1= more chance of heart attack)

Analysis

Training and Initial Model:

A. Dataset and Initial Splitting:
   - We first partitioned the data into training and testing subsets: 80% training and 20% testing

B. Exploratory Data Analysis (EDA):
   - Boxplots by response
   - Histograms of predictors (This part was done by another student)
   - Removing Outliers (This part was done by another student)
     - We removed outliers in chole and oldpeak predictors
   - Correlation Heatmap
     - No significant correlation among predictors was found

C. Training the Model:
   - First, trained the initial model with all predictors in the data set
   - Next, trained the bidirectional stepwise regression model for variable selection. We used the AIC criteria. The following are the model specification and the result output.

The logistic regression model is specified as follows:

$$\log\left(\frac{P(Y=1)}{P(Y=0)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_9 X_9$$

Where:

- Y: Binary response variable indicating heart attack (1 = heart attack, 0 = no heart attack)
- X_1: Sex of the patient (1 = male, 0 = female)
- X_2: Chest pain type (0 = typical angina, 1 = atypical angina, 2 = non-anginal pain, 3 = asymptomatic)
- X_3: Resting blood pressure (in mm Hg)
- X_4: Resting electrocardiographic results (0 = normal, 1 = ST-T wave abnormality, 2 = probable left ventricular hypertrophy)
- X_5: Maximum heart rate achieved
- X_6: Exercise-induced angina (1 = yes, 0 = no)
- X_7: Oldpeak (ST depression induced by exercise relative to rest)
- X_8: Number of major vessels colored by fluoroscopy (0-3)
- X_9: Thalassemia (1 = normal, 2 = fixed defect, 3 = reversible defect)
- β_0: Intercept
- $\beta_1, \beta_2, \ldots, \beta_9$: Coefficients for the predictors

```
> summary(stepwise_model)

Call:
glm(formula = output ~ sex + cp + trtbps + restecg + thalachh +
    exng + oldpeak + caa + thall, family = binomial, data = train)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.95042    2.25373   1.753 0.079630 .
sex         -1.32234    0.47819  -2.765 0.005687 **
cp           0.75964    0.19873   3.822 0.000132 ***
trtbps      -0.02357    0.01104  -2.135 0.032784 *
restecg      0.87577    0.39194   2.234 0.025454 *
thalachh     0.02256    0.01051   2.146 0.031870 *
exng        -0.99284    0.45020  -2.205 0.027431 *
oldpeak     -0.90147    0.22845  -3.946 7.94e-05 ***
caa         -0.60878    0.19648  -3.098 0.001946 **
thall       -1.16404    0.34440  -3.380 0.000725 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 323.52  on 234  degrees of freedom
Residual deviance: 168.02  on 225  degrees of freedom
AIC: 188.02

Number of Fisher Scoring iterations: 6
```

D. Performance Evaluation

a. VIF to check for multicollinearity

- VIF's were low suggesting no significant multicollinearity among predictors

```
> vif(stepwise_model)
     sex       cp   trtbps  restecg thalachh     exng  oldpeak      caa    thall
1.170818 1.268473 1.100897 1.086334 1.166022 1.145811 1.181221 1.041421 1.088348
```

b. Overall significance

- $H0$: $no\ significance$
- $H1$: $\sim H0$
- High test statistic that exceed the critical value resulted in low p-value, rejecting the null hypothesis

```
> # overall significance
> gstat = stepwise_model$null.deviance - deviance(stepwise_model)
> cbind(gstat, 1-pchisq(gstat,length(coef(stepwise_model))-1))
         gstat
[1,] 155.5082 0
```

c. Goodness of fit test

- $H0$: GOF
- $H1$: $\sim H0$

(a) Deviance

```
> # 1. test with deviance
> deviance_test <- c(deviance(stepwise_model), 1
+                    -pchisq(deviance(stepwise_model),df.residual(stepwise_model)))
> deviance_test
[1] 168.0163382   0.9982593
```

(b) Person residual

```
> # 2. test with pearson residual
> pearres <- residuals(stepwise_model, type = "pearson")
> pearson_tvalue <- sum(pearres^2)
> pearson_test <- c(pearson_tvalue, 1 - pchisq(pearson_tvalue,
+                                              df.residual(stepwise_model)))
> pearson_test
[1] 225.2909483   0.4819979
```

- The result of both tests suggest that it is hard to reject the null hypothesis, thereby goodness-of-fit is attained in our model

d. K-fold cross validation (k = 10)

- No significance variance among individual fold's accuracies were observed
- KCV accuracy (mean of fold accuracies) = around 83%
- Train accuracy = around 86 %
- No significant overfitting observed

```
> # KCV (k = 10)
> control <- trainControl(method = "cv", number = 10)
>
> cv_model <- train(
+   as.factor(output) ~ ., data = train,
+   method = "glm", family = "binomial",
+   trControl = control, metric = "Accuracy"
+ )
>
> print(cv_model)
Generalized Linear Model

235 samples
 13 predictor
  2 classes: '0', '1'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 211, 212, 212, 212, 212, 211, ...
Resampling results:

  Accuracy   Kappa
  0.8304348  0.6571049

>
> #KCV result
> print(cv_model$resample) # Accuracy for each fold
    Accuracy     Kappa Resample
1  0.6250000 0.2702703   Fold01
2  0.8260870 0.6461538   Fold02
3  0.9565217 0.9105058   Fold03
4  0.8260870 0.6377953   Fold04
5  0.7826087 0.5593870   Fold05
6  0.8333333 0.6643357   Fold06
7  0.9130435 0.8230769   Fold07
8  0.7916667 0.5714286   Fold08
9  0.9583333 0.9166667   Fold09
10 0.7916667 0.5714286   Fold10
```

```
> # train accuracy
> train_preds <- predict(cv_model, train)  # train set prediction
> train_accuracy <- mean(train_preds == train$output)  # Accuracy
> print(train_accuracy)
[1] 0.8680851
```

E. Test

a. ROC

```
> # ROC
> probs <- predict(cv_model, test, type = "prob")[, 2]
> roc_curve <- roc(test$output, probs)
Setting levels: control = 0, case = 1
Setting direction: controls < cases
> plot(roc_curve)
> auc(roc_curve)
Area under the curve: 0.9286
```

b. Confusion Matrix

```
> # prediction values using test dataset
> predictions <- predict(stepwise_model, newdata = test, type = "response")
>
> # turn the output into 0 and 1 (threshold = 50%)
> predicted_class <- ifelse(predictions > 0.5, 1, 0)
>
> # Confusion Matrix
> confusionMatrix(factor(predicted_class), factor(test$output))
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 20  2
         1  8 30

               Accuracy : 0.8333
                 95% CI : (0.7148, 0.9171)
    No Information Rate : 0.5333
    P-Value [Acc > NIR] : 1.056e-06

                  Kappa : 0.6606

 Mcnemar's Test P-Value : 0.1138

            Sensitivity : 0.7143
            Specificity : 0.9375
         Pos Pred Value : 0.9091
         Neg Pred Value : 0.7895
             Prevalence : 0.4667
         Detection Rate : 0.3333
   Detection Prevalence : 0.3667
      Balanced Accuracy : 0.8259

       'Positive' Class : 0
```

Summary

Our final logistic regression model equation is as follows:

$$\begin{aligned}\text{logit}(\text{output}) = {} & 3.95042 - 1.32234(\text{sex}) \\ & + 0.75964(\text{cp}) - 0.02357(\text{trtbps}) \\ & + 0.87577(\text{restecg}) + 0.02256(\text{thalachh}) \\ & - 0.99284(\text{exng}) - 0.90147(\text{oldpeak}) \\ & - 0.60878(\text{caa}) - 1.16404(\text{thall})\end{aligned}$$

The most significant factors found that determine the odds of having a heart attack were found to be `restecg`, `cp`, `sex`, and `thall`. Some of the notable insignificant factors found, factors that were removed during the bidirectional stepwise regression, were `age`, `fbs`, and `chol`. **Age and cholesterol were interesting because medical studies indicate that people who are at an older age and/or have high blood cholesterol are at a higher chance of getting a heart attack. So some of our model results suggest findings that deviate from standard medical findings regarding risk factors for heart attacks.** This is also true for factors that were included in our model, primarily sex. The beta for sex was found to be -1.32234 which means the odds ratio is 0.26651. This indicates that the odds of heart disease are 26.65% higher for females than males, but men are known to have a higher chance of having a heart attack than women. These differing results are likely due to the small sample size of the data, so more data should be collected.

Further work that could be done with the model is to investigate the reasons for large differences between groups that do not support known medical data, and to collect more data to see if the model continues to support this theory, or aligns more with medical findings. Additional work that could be done is to break up certain variables like age into groups and use them as dummy variables in the model to see if that would make them become significant. The probit function could also be used instead of the logit to see if that improves the model. Lastly, additional potential factors could be looked into not included in the dataset that possibly raise the risk of having a heart attack.