Likelihood of Being Diagnosed with Heart Disease

Analysis of Categorical Data Course Project - Phase II

Daniel Evans

1st of November 2020

Contents

1. Introduction

Carrying on from phase one, this report aims to fit a logistic regression model to determine the likelihood of being diagnosed with heart disease based on a number of independent variables. A saturated logistic regression model was first fit to the data to determine which variables were significant in predicting heart disease. An analysis of variance was conducted to determine this. A reduced logistic regression model was then fit to the data using only the significant variables found. Residual analysis was then conducted to examine the assumption of homoscedasticity and to observe the accuracy of the model. A response analysis on the effects of blood pressure and diagnosis of heart disease showed that as blood pressure increased, the probability of being diagnose with heart disease also increased. The goodness of fit values indicated that both models fitted the data reasonably well. The likelihood of not being diagnosed with heart disease with a blood pressure 70 was found to be 0.77 with a 95% confidence Wald interval of $0.59 < \text{pi} < 0.88$. A hypothesis test of the model parameters was then conducted to check for significance. Odds ratios found that for every one unit increase in resting blood pressure, the odds of being diagnosed with heart disease increases by 0.93, for every ten unit increase in resting blood pressure, the odds of being diagnosed with heart disease increases by 1.18, and having 4 blood vessels coloured by fluoroscopy compared with 3 increases the odds of being diagnosed with heart diseases by 48 times.

2. Statistical Modelling

For the purpose of the logistic regression model, the categorical variables were converted to factors.

```
library(car)
```

```
## Warning: package 'car' was built under R version 3.6.2
```

```
## Loading required package: carData
```

```
heart <- read.csv("Project Group 62_data.csv")

heart$sex <- as.factor(heart$sex)
heart$cp <- as.factor(heart$cp)
heart$fbs <- as.factor(heart$fbs)
heart$restecg <- as.factor(heart$restecg)
heart$exang <- as.factor(heart$exang)
heart$slope <- as.factor(heart$slope)
heart$ca <- as.factor(heart$ca)
heart$thal <- as.factor(heart$thal)
heart$target <- as.factor(heart$target)
```

2.1 Model Fitting

Firstly, a saturated logistic model was fitted using all of the independent variables to determine which ones were relevant for predicting heart disease.

Full Model

```r
mod.fit <- glm(formula = target ~ ., family = binomial(link = logit),
               data = heart)
summary(mod.fit)
```

```
##
## Call:
## glm(formula = target ~ ., family = binomial(link = logit), data = heart)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.9459  -0.2738   0.1012   0.4515   3.1248
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.179045   3.705420   0.048 0.961461
## age          0.027819   0.025428   1.094 0.273938
## sex1        -1.862297   0.570844  -3.262 0.001105 **
## cp1          0.864708   0.578000   1.496 0.134645
## cp2          2.003186   0.529356   3.784 0.000154 ***
## cp3          2.417107   0.719242   3.361 0.000778 ***
## trestbps    -0.026162   0.011943  -2.191 0.028481 *
## chol        -0.004291   0.004245  -1.011 0.312053
## fbs1         0.445666   0.587977   0.758 0.448472
## restecg1     0.460582   0.399615   1.153 0.249089
## restecg2    -0.714204   2.768873  -0.258 0.796453
## thalach      0.020055   0.011859   1.691 0.090820 .
## exang1      -0.779111   0.451839  -1.724 0.084652 .
## oldpeak     -0.397174   0.242346  -1.639 0.101239
## slope1      -0.775084   0.880495  -0.880 0.378707
## slope2       0.689965   0.947657   0.728 0.466568
## ca1         -2.342301   0.527416  -4.441 8.95e-06 ***
## ca2         -3.483178   0.811640  -4.292 1.77e-05 ***
## ca3         -2.247144   0.937629  -2.397 0.016547 *
## ca4          1.267961   1.720014   0.737 0.461013
## thal1        2.637558   2.684285   0.983 0.325808
## thal2        2.367747   2.596159   0.912 0.361759
## thal3        0.915115   2.600380   0.352 0.724901
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 417.64  on 302  degrees of freedom
## Residual deviance: 179.63  on 280  degrees of freedom
## AIC: 225.63
##
## Number of Fisher Scoring iterations: 6
```

```r
Anova(mod.fit)
```

```
## Analysis of Deviance Table (Type II tests)
##
```

```
## Response: target
##          LR Chisq Df Pr(>Chisq)
## age         1.209  1  0.2715575
## sex        11.810  1  0.0005892 ***
## cp         21.419  3  8.617e-05 ***
## trestbps    5.043  1  0.0247270 *
## chol        0.991  1  0.3194764
## fbs         0.583  1  0.4452161
## restecg     1.458  2  0.4823037
## thalach     3.002  1  0.0831391 .
## exang       2.958  1  0.0854290 .
## oldpeak     2.809  1  0.0937236 .
## slope       9.402  2  0.0090856 **
## ca         40.038  4  4.250e-08 ***
## thal       13.703  3  0.0033389 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The analysis of variance of the saturated model indicates that sex, type of chest pain (cp), resting blood pressure (trestbps), slope of the peak during exercise (slope), number of major vessels coloured by fluoroscopy (ca), and whether or not someone had a heart defect (thal) are all good predictors of heart disease as their p values all fall below 0.05.

Reduced Model

The reduced model was then fitted using the relevant variables found from the analysis of variance of the saturated model.

```
sex <- heart$sex
chestPain <- heart$cp
bloodPressure <- heart$trestbps
slopeExercise <- heart$slope
numberBloodVessels <- heart$ca
heartDefect <- heart$thal
target <- heart$target

reducedModel <- glm(target ~ sex + chestPain + bloodPressure + slopeExercise
                    + numberBloodVessels + heartDefect,
                    family = binomial(link = logit))
summary(reducedModel)
```

```
##
## Call:
## glm(formula = target ~ sex + chestPain + bloodPressure + slopeExercise +
##     numberBloodVessels + heartDefect, family = binomial(link = logit))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.9398  -0.3840   0.1048   0.4500   3.0153
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)          2.32566    3.77789   0.616  0.53816
```

```
## sex1                -1.67331    0.50876  -3.289  0.00101 **
## chestPain1           1.54896    0.53579   2.891  0.00384 **
## chestPain2           2.40519    0.48356   4.974 6.56e-07 ***
## chestPain3           2.69310    0.69069   3.899 9.65e-05 ***
## bloodPressure       -0.02539    0.01058  -2.401  0.01636 *
## slopeExercise1      -0.46908    0.70569  -0.665  0.50624
## slopeExercise2       1.58584    0.72038   2.201  0.02771 *
## numberBloodVessels1 -2.27957    0.48512  -4.699 2.62e-06 ***
## numberBloodVessels2 -3.21428    0.70766  -4.542 5.57e-06 ***
## numberBloodVessels3 -2.33299    0.87390  -2.670  0.00759 **
## numberBloodVessels4  1.53817    1.58252   0.972  0.33106
## heartDefect1         2.44424    3.51654   0.695  0.48701
## heartDefect2         2.40656    3.44974   0.698  0.48542
## heartDefect3         0.81101    3.45251   0.235  0.81428
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 417.64  on 302  degrees of freedom
## Residual deviance: 194.96  on 288  degrees of freedom
## AIC: 224.96
##
## Number of Fisher Scoring iterations: 6
```

```
Anova(reducedModel)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: target
##                    LR Chisq Df Pr(>Chisq)
## sex                  11.822  1  0.0005854 ***
## chestPain            39.481  3  1.372e-08 ***
## bloodPressure         6.119  1  0.0133720 *
## slopeExercise        25.451  2  2.974e-06 ***
## numberBloodVessels   49.830  4  3.918e-10 ***
## heartDefect          17.439  3  0.0005739 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
coef(reducedModel)
```

```
##         (Intercept)                sex1          chestPain1
##          2.32566108         -1.67331125           1.54896498
##          chestPain2          chestPain3       bloodPressure
##          2.40518911          2.69310209         -0.02539045
##      slopeExercise1      slopeExercise2 numberBloodVessels1
##         -0.46908272          1.58583828         -2.27956546
## numberBloodVessels2 numberBloodVessels3 numberBloodVessels4
##         -3.21427708         -2.33299256           1.53816817
##        heartDefect1        heartDefect2        heartDefect3
##          2.44423536          2.40655695          0.81101298
```

The summary() function found that certain levels of categorical variable were statistically significant and others not. This can be seen in the case of the number of blood vessels coloured by fluoroscopy which is significant for levels 1, 2, and 3 however 4 does not appear to be significant to the model. Contrary to this, the analysis of variance indicates that all of the independent variables in the model are significant.

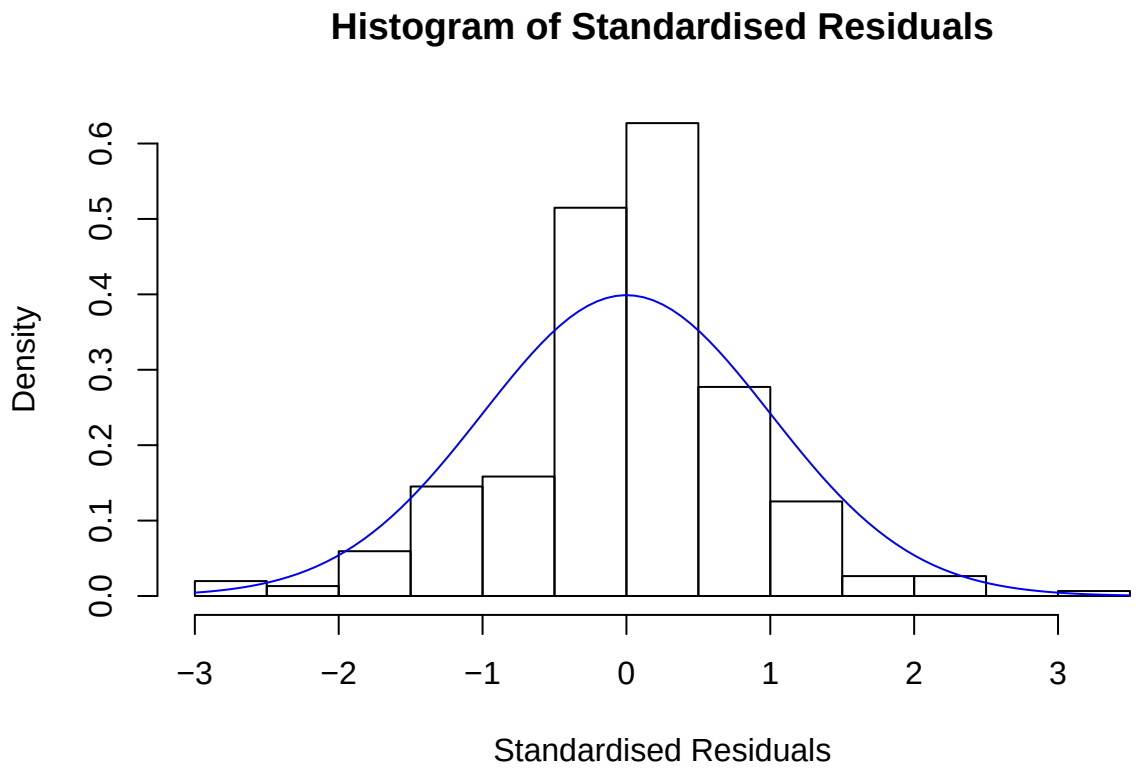
Final Model

Logit(Probability of Heart Disease) = 2.33 – 1.67 x Sex 1 + 1.55 x Chest Pain 1 + 2.41 x Chest Pain 2 + 2.39 x Chest Pain 3 – 0.03 x Blood Pressure – 0.47 x Slope Exercise 1 + 1.86 x Slope Exercise 2 – 2.78 x Blood Vessels 1 – 3.21 x Blood Vessels 2 – 2.33 x Blood Vessels 3 + 1.54 x Blood Vessels 4 + 2.44 x Heart Defect 1 + 2.41 x Heart Defect 2 + 0.81 x Heart Defect 3

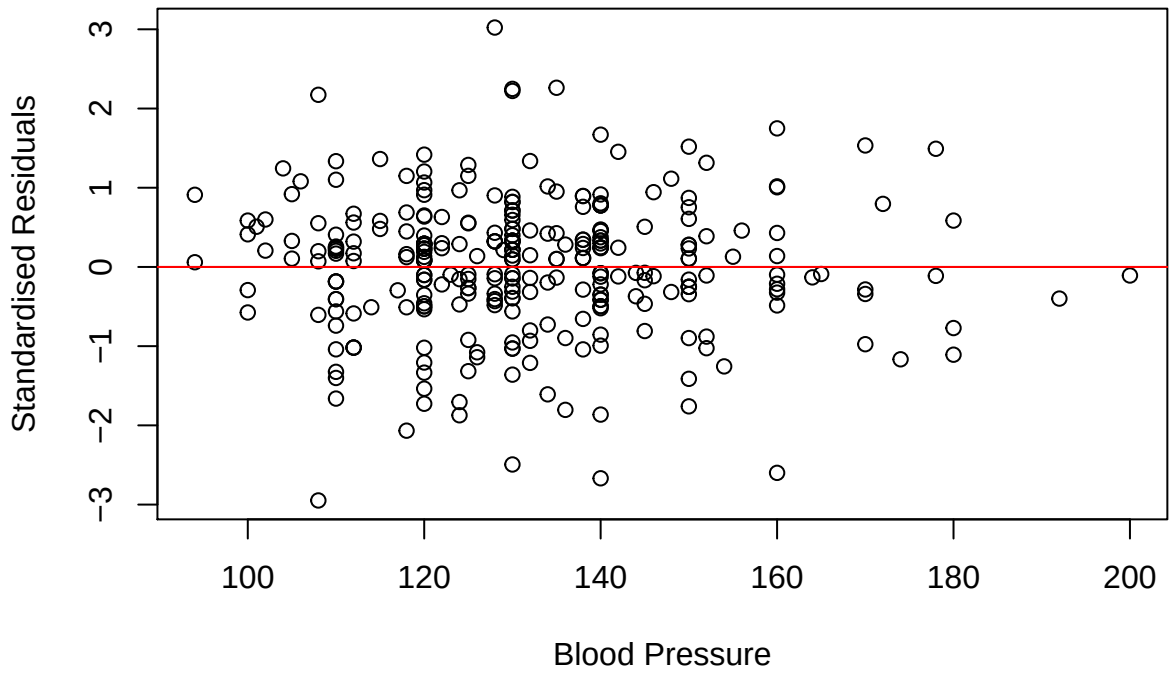
2.2 Residual Analysis

```
reducedModel.stdres = rstandard(reducedModel)

hist(reducedModel.stdres, freq = FALSE,
     main = "Histogram of Standardised Residuals",
     xlab = "Standardised Residuals")
curve(dnorm(x), add=TRUE, col = "blue")
```



```r
plot(bloodPressure, reducedModel.stdres,
     ylab = "Standardised Residuals",
     xlab = "Blood Pressure")
abline(h = c(0,0), col = "red")
```



The histogram of standardised residuals is normally distributed with most of the values falling close to 0 which indicates the model is working well. There are a few values which have errors greater than 3 standard deviations which is something to note.

The assumption of homoscedasticity of residuals for blood pressure appears to be true as the variance appears to be constant. Most of the errors fall close to zero however there are a few which fall greater than 3 standard deviations.

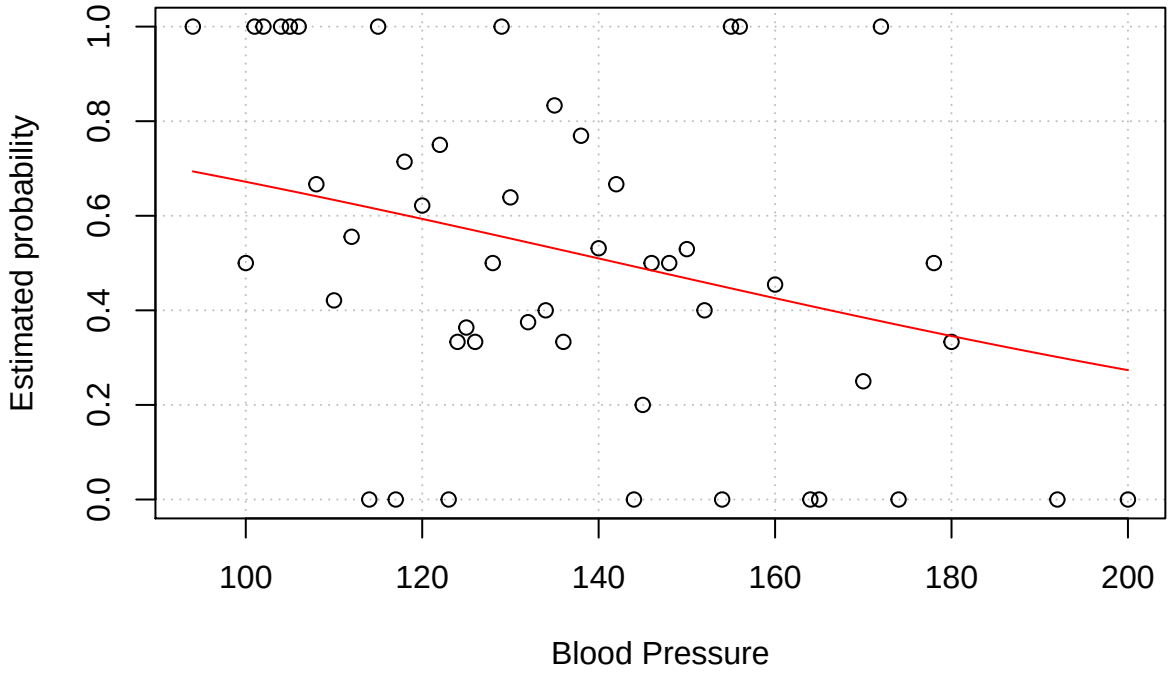
2.3 Response Analysis

```r
heart <- read.csv("Project Group 62_data.csv")

w <-aggregate(formula = target ~ trestbps, data = heart,
              FUN = sum)
n<-aggregate(formula = target ~ trestbps, data = heart,
               FUN = length)

trestbps.fit <- glm(formula = target ~ trestbps, data = heart,
               family = binomial(link = logit))
```

```
plot(x = w$trestbps, y = w$target/n$target,
     xlab="Blood Pressure",
     ylab="Estimated probability", panel.first =
       grid(col = "gray", lty = "dotted"))
curve(expr = predict(object = trestbps.fit, newdata =
    data.frame(trestbps = x), type = "response"), n = 303, col =
        "red", add = TRUE)
```



I was intrigued at analysing the effect of blood pressure on the probability of being diagnosed with heart disease. One would expect that the higher the blood pressure, the greater chance of being diagnosed with heart disease. To examine the effects of Blood Pressure alone, a model was fit with that independent variable only. The above response analysis indicates the original hypothesis to be true. One must remember that in this dataset, a target value of 0 indicates diagnosis of heart disease so as this plot shows, as the blood pressure increases, values tend toward 0. A logistic regression cure was fitted however it appears more as a straight line as there instances where patients with low blood pressure are diagnosed with heart disease and vice versa.

2.4 Goodness of Fit

```
reducedModel.rdev <- reducedModel$deviance
reduced.dfr <- reducedModel$df.residual
reducedModel.fit <- reducedModel.rdev/reduced.dfr

fullModel.rdev <- mod.fit$deviance
```

```
fullModel.dfr <- mod.fit$df.residual
fullModel.fit <- fullModel.rdev/fullModel.dfr

Models <- c("Reduced", "Full")
Res.dev <- c(reducedModel.rdev, fullModel.rdev)
df <- c(reduced.dfr, fullModel.dfr)
fit <- c(reducedModel.fit, fullModel.fit)
good.fit.data <- data.frame(Models, Res.dev, df, fit, stringsAsFactors=FALSE)
good.fit.data
```

```
##    Models  Res.dev  df       fit
## 1 Reduced 194.9598 288 0.6769437
## 2    Full 179.6307 280 0.6415383
```

Both the full and reduced models appear to be relatively good fits for the data with goodness of fit values greater than 0.6. The reduced model performs slightly better than the saturated model ($0.68 > 0.64$)

2.5 Confidence Intervals

```
# Predicting likelihood of not having heart dieases with a blood pressure of 70
predict.data<-data.frame(trestbps = 70)
predict(object = trestbps.fit, newdata = predict.data, type =
            "link")
```

```
##        1
## 1.224302
```

```
predict(object = trestbps.fit, newdata = predict.data, type =
            "response")
```

```
##         1
## 0.7728197
```

```
alpha<-0.05
linear.pred<-predict(object = trestbps.fit, newdata =
                       predict.data, type = "link", se = TRUE)
pi.hat<-exp(linear.pred$fit) / (1 + exp(linear.pred$fit))
CI.lin.pred<-linear.pred$fit + qnorm(p = c(alpha/2, 1-alpha/2))*linear.pred$se
CI.pi<-exp(CI.lin.pred)/(1+exp(CI.lin.pred))
data.frame(predict.data, pi.hat, lower = CI.pi[1],
                 upper = CI.pi[2])
```

```
##   trestbps    pi.hat     lower     upper
## 1       70 0.7728197 0.5912401 0.8888955
```

I decided to predict the likelihood of not being diagnosed with heart disease with a blood pressure of 70. The logistic regression model found that with a blood pressure of 70, the probability of not begin diagnosed with heart disease was 0.77. the 95% Wald confidence interval for this prediction was $0.59 < \text{pi} < 0.88$. Thus, the probability of not being diagnosed with heart disease with a blood pressure as low as 70 is quite high.

2.6 Hypothesis Tests for Regression Parameters

```
summary(reducedModel)
```

```
##
## Call:
## glm(formula = target ~ sex + chestPain + bloodPressure + slopeExercise +
##     numberBloodVessels + heartDefect, family = binomial(link = logit))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.9398  -0.3840   0.1048   0.4500   3.0153
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)           2.32566    3.77789   0.616  0.53816
## sex1                 -1.67331    0.50876  -3.289  0.00101 **
## chestPain1            1.54896    0.53579   2.891  0.00384 **
## chestPain2            2.40519    0.48356   4.974 6.56e-07 ***
## chestPain3            2.69310    0.69069   3.899 9.65e-05 ***
## bloodPressure        -0.02539    0.01058  -2.401  0.01636 *
## slopeExercise1       -0.46908    0.70569  -0.665  0.50624
## slopeExercise2        1.58584    0.72038   2.201  0.02771 *
## numberBloodVessels1 -2.27957    0.48512  -4.699 2.62e-06 ***
## numberBloodVessels2 -3.21428    0.70766  -4.542 5.57e-06 ***
## numberBloodVessels3 -2.33299    0.87390  -2.670  0.00759 **
## numberBloodVessels4  1.53817    1.58252   0.972  0.33106
## heartDefect1          2.44424    3.51654   0.695  0.48701
## heartDefect2          2.40656    3.44974   0.698  0.48542
## heartDefect3          0.81101    3.45251   0.235  0.81428
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 417.64  on 302  degrees of freedom
## Residual deviance: 194.96  on 288  degrees of freedom
## AIC: 224.96
##
## Number of Fisher Scoring iterations: 6
```

```
# Likelihood Ratio Test (LRT)
Anova(reducedModel)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: target
##                    LR Chisq Df Pr(>Chisq)
## sex                  11.822  1  0.0005854 ***
## chestPain            39.481  3  1.372e-08 ***
## bloodPressure         6.119  1  0.0133720 *
## slopeExercise        25.451  2  2.974e-06 ***
```

```
## numberBloodVessels   49.830  4  3.918e-10 ***
## heartDefect          17.439  3  0.0005739 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The hypothesis test indicates that there is sufficient evidence to suggest that sex, type of chest pain (cp), resting blood pressure (trestbps), slope of the peak during exercise (slope), number of major vessels coloured by fluoroscopy (ca), and whether or not someone had a heart defect (thal) all have an effect on the diagnosis of heart disease as their respective p values are all below the alpha significance level of 0.05. The remaining variables in the data were deemed to be insignificant from the hypothesis test.

2.7 Sensitivity Analysis

```
# Every 1mm/HG increase in resting blood pressure
exp(trestbps.fit$coefficients[2])
```

```
##  trestbps
## 0.9832135
```

```
# Every 10mm/HG increase in resting blood pressure
1/exp(10*trestbps.fit$coefficients[2])
```

```
## trestbps
## 1.184463
```

```
# Wald confidence interval for odds ratio
beta.ci<-confint.default(object = trestbps.fit, parm = "trestbps", level = 0.95)
rev(1/exp(beta.ci*10)) #Invert OR C.I. with c=10
```

```
## [1] 1.036635 1.353371
```

```
# Having 4 blood vessels colour by flouroscopy vs 3
exp(reducedModel$coefficients[12] - reducedModel$coefficients[11])
```

```
## numberBloodVessels4
##            47.99807
```

The above results can be interpreted as follows:

- For every one unit increase in resting blood pressure, the odds of being diagnosed with heart disease increases by 0.93.
- For every ten unit increase in resting blood pressure, the odds of being diagnosed with heart disease increases by 1.18
- Having 4 blood vessels coloured by fluoroscopy compared with 3 increases the odds of being diagnosed with heart diseases by 48 times.

3. Critique and Limitations

This analysis assumed linearity between the dependent variable and the independent variables. It also assumed that there was no multicollinearity between the independent variables. Further analysis should conduct a covariance matrix between the independent variables to check for multicollinearity. A limitation of this study is the amount of data. With only 303 instances, we cannot be sure of the results determined. A greater number of instances might provide further evidence to what we have concluded in this analysis. A strength of this study is the reliability of the data. The data was collected at a medical centre with reliable instruments providing reliability and confidence in the data that was collect.

4. Summary and Conclusions

This analysis showed that there is a relationship between sex, type of chest pain (cp), resting blood pressure (trestbps), slope of the peak during exercise (slope), number of major vessels coloured by fluoroscopy (ca), and whether or not someone had a heart defect (thal) on the diagnosis of heart disease. Phase one of the project prepared the data for analysis and found that gender, age, and maximum heart rate during exercise could be good predictors for Heart Disease in the logistic regression model.

In relation to the original goal, we were able to determine the probability of a patient being diagnosed or not diagnosed with heart disease based on certain attribute information. We were also able to analyse the effects of varying levels of significant predictor variables and how these effect the odds of being diagnosed with heart disease. This is what the analysis was originally set out to achieve. By understanding what contributes to heart disease and how changes to personal health increase the odds of being diagnosed with heart disease, we can better educate and treat people against Australia's leading cause of death.