**Analyzing the Relevant Variables Using Logistic Regression in Order to Predict the Risk of Heart Disease**

Fatme Abboud, Gitasha Eland, Graham Eckel

Summary: Regression modelling techniques, in particular, logistic regression allow one to model independent categorical and/or numeric variables on a binary dependent variable. In this dataset, the presence of heart disease is a binary response. We propose the use of logistic regression modelling and associated statistical techniques to determine which attributes may help to predict heart disease. We apply our methods to determine the most significant covariates to include in a final model and show a reduced model, one that focuses on key features, has a better overall fit than a full model for this dataset.  

1. **Introduction**

According to the American Heart Association, the number one leading cause of death in the United States is heart disease (Benjamin Emelia J. et al., 2019). Heart disease (HD) accounted for approximately 75.5% of all cardiovascular disease (CVD) death in 2016 and had an estimated direct cost in 2014 to 2015 of $109.4 billion with an indirect cost of $218.7 billion (Benjamin Emelia J. et al., 2019). There are many risk factors associated with HD, with some of the main risk factors of HD being hypertension (prevalence: 46.0%), high cholesterol (38.2%), and high fasting blood sugar (prediabetes: 33.9%, diabetes: ~13.5%) (Benjamin Emelia J. et al., 2019). These risk factors also increase the mortality risk of CVD.

HD has a great impact on both the health of the population as well as the economy; thus, prevention and the ability to assess the risk of HD of individuals is extremely important in combating HD. We aim to build a model that exhibits the significance of various variables on heart disease. This is difficult as there are many contributing factors that can significantly impact cardiovascular health. In this paper, we analyze a dataset containing observations of 14 attributes across 303 individuals (*Heart Disease UCI*, n.d.). Many of these attributes have been found to contribute to or be indicative of heart disease. Analysis of this data was conducted in order to build an appropriate logistic regression model that may be used to predict an individual’s likelihood of developing heart disease. The significance of the various categorical and numeric variables in the dataset was explored in order to narrow down the potential model. Several models were built using various model selection techniques. Finally, various model comparisons were conducted on the built models in order to determine the most appropriate model for predicting an individual’s likelihood of developing heart disease.

1. **Cleaning of the Data**

The original dataset contained 303 observations, 2% of which were invalid.  An appropriate classification method for the missing data was not identified; thus, these rows were removed without significant impact on data integrity or model strength. Further, some data transformation was implemented to improve data usability. Categorical variables mistakenly classified as numeric datatype were updated to factor data type and dimensions were renamed and reordered for convenience.

1. **Exploratory Analysis of Numeric Variables**

The numeric variables present in the dataset included age, resting blood pressure, cholesterol, maximum heart rate, and ST depression (Table 1). In order to determine the variables that had the most significance, simple logistic regression modelling was done. Primarily, age showed a significant impact on the presence of heart disease in individuals. Resting blood pressure continued to show a significant effect on the presence of heart disease. Further, modelling the maximum heart rate on the presence of heart disease exhibited the most significance. However, contrary to previous research on cholesterol levels, regression of cholesterol on heart disease seemed to be insignificant. From this exploratory analysis, it was clear that each numeric variable, excluding cholesterol, had a significant effect on the presence of heart disease in individuals.

A boxplot was constructed to observe the presence of outliers and skewness in the numerical variables (Fig.1). We observed that chol, trestbps and oldpeak had a number of outlying data. In addition, age has some left skewness, indicating this is a relatively younger sample.

Next, a correlogram was made to explore the correlation of numerical covariates on each other (Fig.2). A strong negative correlation between age and thalach, as well as oldpeak and thalach was observed, while a strong positive correlation between age and trestbps was observed. Other notable, but less strong, positive correlations were seen for age and oldpeak, age and chol, and oldpeak and trestbps.

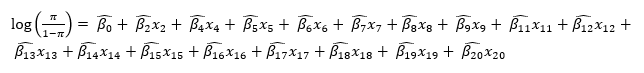
1. **Exploratory Analysis of Categorical and Binary Variables**

There are eight categorical and binary variables present in this dataset (Table 2). A Chi-squared test was conducted on each variable in order to assess its significance on the response variable. All variables, except fasting blood sugar, have extremely significant associations with the target response of heart disease (Table 2).  A Fisher’s exact test was performed on resting electrocardiogram results instead of a Chi-squared test as the expected values for that variable were less than 5 and thus violated the Chi-squared assumptions.

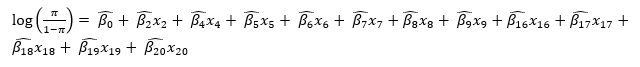
Risk analysis was conducted on the binary explanatory variables sex, fasting blood sugar, exercise-induced angina in order to further determine the significance of these variables. These results confirm the results of the Chi-squared test for these variables as well as provide the magnitude of the effect each variable has on heart disease (Table 3).

1. **Model Selection and Comparison**

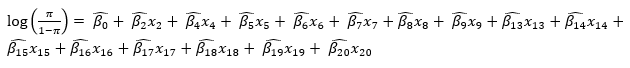
It was observed that the variables fbs and chol did not have a significant relationship with the dependent variable, target, and that chol did not have strong correlations with the other numerical variables. As a result, these two variables were not included in our initial model. The equation for our initial model is as follows:



After the initial model was created, backwards selection and stepwise selection were employed to determine the covariates in our final two reduced models. At every step of the backwards selection, a drop in deviance test was conducted between each successive reduced model and the previous model, beginning with the initial model. The backwards selection process ended when the drop in deviance test did not select for the next iteration. The equation for our backwards selection model is as follows:

**

For the stepwise selection, a minimum Akaike information criteria (AIC) approach was used when performing model comparisons. Beginning with covariate that produces the model with the smallest AIC, we continue to add or remove the next covariates in the same way. After each meaningful combination is explored, we select for the model that has the lowest AIC.   The equation for our stepwise model is as follows:

**

After performing the backwards selection and stepwise selection, a final model comparison was done using both minimum AIC and drop in deviance. The stepwise model was found to be necessary.

1. **Model Assumptions**

It was important to note that the model assumptions for logistic regression were met. The first assumption being that the dependent variable was to be binary, which agreed with our data. The response variable, target, indicated 1 for no heart disease and 0 for having heart disease. Second, the observations had to be independent of each other, which each variable was. The final assumption is that there is a linear relationship between log(1-) and the numeric variables, which was also found to be true (Fig. 3).

1. **Goodness of Fit**

To test our stepwise model for goodness of fit we used the residual deviance as a test statistic. Following this test, we calculated a pseudo R2. Our model has a residual deviance of 182.16 on 280 degrees of freedom (p=0.999) which provides strong evidence against rejecting the null hypothesis. The McFadden R2 provided a value of 0.554. Taking into consideration both of these results, we concluded that our model is an adequate fit.

1. **Discussion**

From our exploratory modelling, there were two variables, cholesterol and fasting blood sugar, that were not significant. However, from previous research, it is known that these variables increase the risk of heart disease (Benjamin Emelia J. et al., 2019). Fasting blood sugar levels above ~90mg/dL, tend to increase the risk for CVD for both men and women (Park et al., 2013); however, our analysis showed that fasting blood sugar is insignificant. This may be due to the imbalance between the individuals with fbs >120mg/dl and those with normal levels in our data. Approximately 85% of individuals in the dataset did not have fbs >120mg/dl, meaning they were neither diabetic nor pre-diabetic (*Diagnosis | ADA*, n.d.). This may have resulted in fbs to appear to be insignificant. Similarly, cholesterol is a factor that tends to increase the risk for CVD, if levels are too high. The suggested dosage of cholesterol is to be limited to <300mg/day, for those who are at risk of heart disease (McNamara, 2014). The lack of significance in our model could be due to the fact that the data did not specify the type of foods being eaten or the lifestyle of the patients. A study of the risk cholesterol has on heart disease suggested that it is not necessarily the level of cholesterol in the body, but the amount of saturated fatty acids and trans fats that contributes to the increased risk of CVD (Soliman, 2018). This could have caused biased results in the data and possibly is the reason why the impact of cholesterol levels on CVD was deemed insignificant.

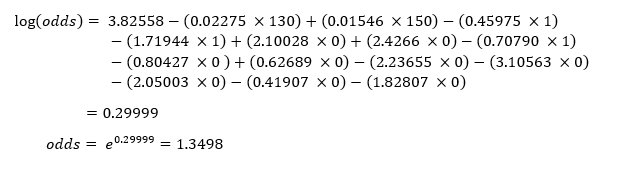
In order to discover the most appropriate model to interpret the data, multiple logistic regression models were built. It was intuitive to use logistic modelling as our response variable, target, is binary. First, a full model was conducted which included all the variables excluding cholesterol and fbs. The full model was then reduced using the backwards selection approach by dropping restecg, slope, age, exang, and thal. The drop in deviance tests performed between each backward selection indicated that certain variables were not significant enough in our model, which resulted in the final backwards model. Next, a stepwise selection was conducted by using the AIC to compare the two models. The final stepwise model excluded the variables age, chol, fbs and restecg, due to the insignificance they had on the variable target. Once the stepwise and backward selection models were devised, ANOVA was performed to determine the most appropriate model for the dataset. According to ANOVA, due to a p-value of 0.01028, it suggests that the stepwise selection model is the most appropriate.

In the course of exploring the data and developing our model we considered adding interaction terms. After determining the strength of correlation between the covariates and testing the linearity of the model, no specific interaction term appeared as an obvious choice. To be complete, we explored the population data from where our sample was derived. Given the disparity in geographic location – Budapest, Zurich, Basel and Cleveland, an obvious demographic-based interaction was not apparent. Despite our preliminary observations, interaction terms are an area that warrants further exploration.

The risk analysis conducted on the three binary variables sex, fbs, and exang provided great insight on the effects each variable has on the risk and odds of contracting heart disease. RD, RR, OR for the variable sex all indicated that sex has a positive association with the target response. The OR indicates that the odds of heart disease in males is approximately 3.7228 times the odds of heart disease in females. For fbs, RD, RR, and OR all indicate that the effects of fbs are neutral on heart disease. This indicates that there is no essential difference in the risk or the odds of heart disease in those who are healthy versus those who are diabetic or prediabetic. Finally, the RD, RR, and OR indicate that exang is positively associated with the target response. The OR of exang indicates that the odds of heart disease in those with exercise induced angina is 7.1094 times the odds of those without it.

1. **Conclusion**

The modelling conducted in this report is important in the discussion of what variables have the most significant effects on the presence of heart disease. By the stepwise model, it was demonstrated that the variables that were the most significant in the presence of CVD were the number of major vessels and chest pain type. This suggested that individuals that had one major heart vessel had a higher risk of heart disease than those with three. In addition, individuals that had chest pain type two, were more at risk of developing CVD than those who had chest pain type one. By using the final model, the odds that someone contracts heart disease can be calculated. For example, the odds of a male with a resting blood pressure of 130mm/Hg, a maximum heart rate of 150bpm, with an ST depression of 1, who experiences atypical angina as well as exercise induced angina, has a downsloping peak exercise ST segment, and whose the number of major vessels coloured by fluoroscopy is 0, with fixed coronary stenosis is as follows:



Further investigation into the insignificance of cholesterol and fbs on CVD may be conducted in order to determine why they were deemed insignificant. The speculation around these two variables led to questions that require reasoning into why two variables that are known to increase the risk of CVD, did not show significance in our models. The intuition of the population being biased due to age or demographics could be reasons surrounding this insignificance and warrant continued research.

**10. Data Availability Statement**

The data analyzed in this paper is openly available on Kaggle as well as in the Cleaveland database at <https://www.kaggle.com/ronitf/heart-disease-uci>.

**11. References**

Benjamin Emelia J., Muntner Paul, Alonso Alvaro, Bittencourt Marcio S., Callaway Clifton W., Carson April P., Chamberlain Alanna M., Chang Alexander R., Cheng Susan, Das Sandeep R., Delling Francesca N., Djousse Luc, Elkind Mitchell S.V., Ferguson Jane F., Fornage Myriam, Jordan Lori Chaffin, Khan Sadiya S., Kissela Brett M., Knutson Kristen L., … null null. (2019). Heart Disease and Stroke Statistics—2019 Update: A Report From the American Heart Association. *Circulation*, *139*(10), e56–e528.<https://doi.org/10.1161/CIR.0000000000000659>

*Diagnosis | ADA*. (n.d.). Retrieved April 7, 2020, from<https://www.diabetes.org/a1c/diagnosis>

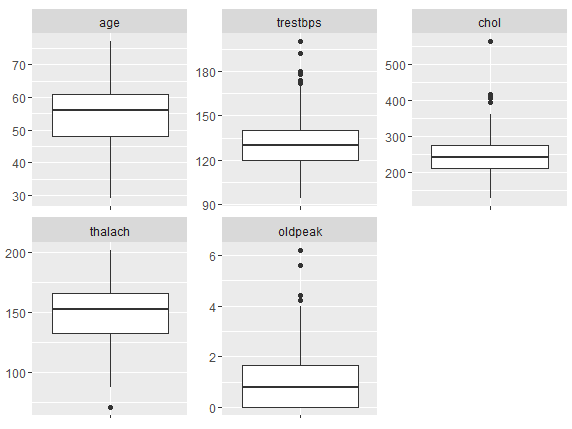
*Heart Disease UCI*. (n.d.). Retrieved April 5, 2020, from<https://kaggle.com/ronitf/heart-disease-uci>

McNamara, D. J. (2014). *Dietary cholesterol, heart disease risk and cognitive dissonance*. *73*(2), 161–166.<https://doi.org/10.1017/S0029665113003844>

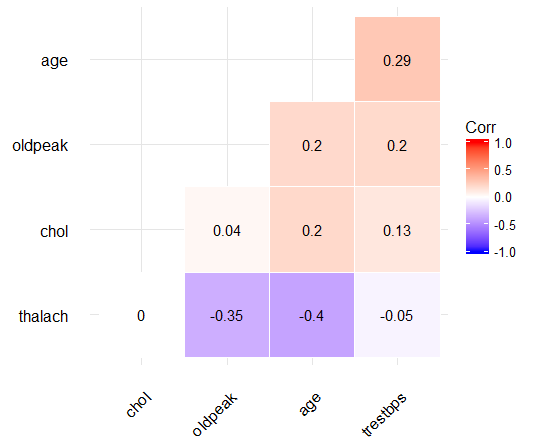
Park, C., Guallar, E., Linton, J. A., Lee, D.-C., Jang, Y., Son, D. K., Han, E.-J., Baek, S. J., Yun, Y. D., Jee, S. H., & Samet, J. M. (2013). Fasting glucose level and the risk of incident atherosclerotic cardiovascular diseases. *Diabetes Care*, *36*(7), 1988–1993. PubMed.<https://doi.org/10.2337/dc12-1577>

Soliman, G. A. (2018). Dietary Cholesterol and the Lack of Evidence in Cardiovascular Disease. *Nutrients*, *10*(6), 780. PubMed.<https://doi.org/10.3390/nu10060780>

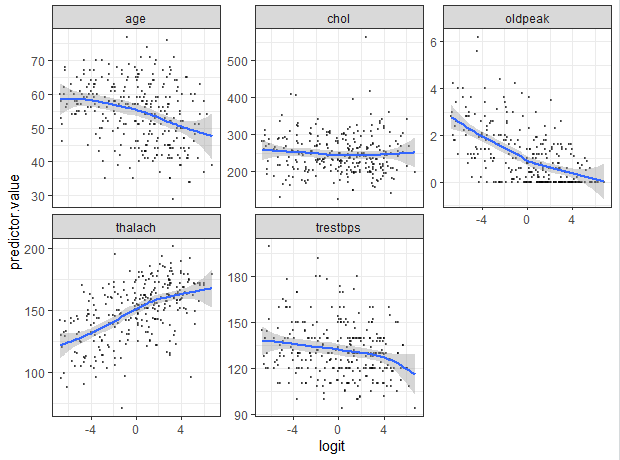
**12. Appendix**

****

**Figure 1:** Individual boxplots for the numeric variables age, trestbps, chol, thalch, and oldpeak. Each plot displays any visible outliers (if there are any).



**Figure 2:** Correlation matrix showing relationship between different variables. The matrix contains each numeric variable and summarizes the data associated with the correlation between each.

****

**Figure 3:** Scatter plots between each numeric explanatory variable and the logit of the response. This was done to test the linearity assumption of the model. The following scatter plots seem to show that all the numeric explanatory variables are fairly linearly associated with the target variable. This indicates that there is no evidence to suggest that the linearity assumption has been violated

**Table 1:** Summary table of numeric explanatory variables. For each variable, the associated name in the dataset, coefficient in models, mean, minimum, maximum values and p-value is listed. The p-value is indicating the significance of the relationship between the explanatory variable and the response. This was calculated by running a simple linear regression model for each variable regressed on the target variable.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Variable** | **Name in Dataset** | **Coefficients** | **Mean** | **Min** | **Max** | **P-value** |
| Age | age | x1, 1 | 54.52 | 29.0 | 77.0 | 0.000147 |
| Resting Blood Pressure (mmHg) | trestbps | x2, 2 | 131.16 | 94.0 | 200.0 | 0.01152 |
| Serum Cholesterol Levels (mg/dl) | chol | x3, 3 | 247.20 | 126.0 | 564.0 | 0.191 |
| Maximum Heart Rate Achieved | thalach | x4, 4 | 149.60 | 71.0 | 202.0 | 1.90 10-11 |
| ST depression induced by exercise relative to rest | oldpeak | x5, 5 | 1.059 | 0.00 | 6.20 | 2.31 10-11 |

**Table 2:** Summary of categorical and binary explanatory variables. For each variable, the associated name in the dataset, levels, coefficients in models for each level (if binary, the coefficient is for 1), and p-value is listed. Levels indicates the number and the associated description of each level of the categorical or binary variable. The p-value is indicating the significance of the relationship between the explanatory variable and the response. This was calculated using a Chi-square test for all variables except fbs which required the use of the Fisher’s exact test.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable** | **Name in Dataset** | **Levels** | **Coefficients** | **P-value** |
| Sex | sex | 0- female  1- male | x6, 6 | 1.719 10-6 |
| Chest Pain Type | cp | 0- typical angina  1- atypical angina   2- non-anginal pain  3- asymptomatic | 1- x7, 7  2- x8, 8  3- x9, 9 | 2.2 10-16 |
| Fasting Blood Sugar (>120mg/dl) | fbs | 0- false  1- true | x10, 10 | 1 |
| Resting Electrocardiogram Results | restecg | 0- probable or definite left ventricular hypertrophy  1- normal  2- ST-T wave abnormality | 1- x11, 11  2- x12, 12 | 0.005866 |
| Exercise Induced Angina | exang | 0- no  1- yes | x13, 13 | 6.517 10-13 |
| Slope of the Peak Exercise ST Segment | slope | 0- downsloping  1- flat  2- upsloping | 1- x14, 14  2- x15, 15 | 2.116 10-10 |
| Number of Major Vessels Coloured by Fluoroscopy | ca | 0-3 | 1- x16, 16  2- x17, 17  3- x18, 18 | 7.996 10-16 |
| Nuclear Stress Test (thallium tracer testing for coronary stenosis) | thal | 1- fixed defect  2- normal  3- reversible defect | 2- x19, 19  3- x20, 20 | 2.2 10-16 |

**Table 3:** Summary of risk analysis on the binary explanatory variables. The risk difference, risk ratio, and odds ratio of the variables sex, fasting blood sugar, and exercise induced angina and their associated 95% confidence intervals (C.I.) are summarized.

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **Risk Difference** | **Risk Ratio** | **Odds Ratio** |
| Sex | RD = 0.3046  95% C.I = (0.1935, 0.4167) | RR = 2.2056  95% C.I. = (1.5278, 3.1842) | OR= 3.7228  95% C.I. = (2.1694, 6.3886) |
| Fasting Blood Sugar (>120mg/dl) | RD = 0.006  95% C.I = (-1.546, 0.1678) | RR = 1.10144  95% C.I. = (0.7167,1.4358) | OR= 1.0270  95% C.I. = (0.5371, 1.9638) |
| Exercise Induced Angina | RD = 0.4513  95% C.I = (0.3450, 0.5577) | RR = 2.4486  95% C.I. = (1.9369, 3.0956) | OR= 7.1094  95% C.I. = (4.0780, 12.3943) |

**R Code:**

library(DataExplorer)

library(ggplot2)

library(ggcorrplot)

library(dplyr)

library(MASS)

library(gridExtra)

library(reshape2)

library(fmsb)

library(DescTools)

library(tidyverse)

library(broom)

library(magrittr)

library(car)

#data file from Kaggle stored in project directory

dir = "E:\\Google Drive\\...\\Group Project\\"

file1 = "heart.csv"

dfHeartDiseaseOriginal = read.table(file=paste(dir,file1, sep=""), header=TRUE, sep=',')

####

####DATA CLEANING

####

#Remove rows with NaN's

which(dfHeartDiseaseOriginal$ca==4)

which(dfHeartDiseaseOriginal$thal==0)

dfHDFull = dfHeartDiseaseOriginal[-c(49,93,159,164,165,252,282), ]

#Cleaning headers

colnames(dfHDFull)[1] = "age"

#Changing categorical variables from numerics to factors

dfHDFull = mutate(dfHDFull,

                  sex = as.factor(sex),

                  cp = as.factor(cp),

                  fbs = as.factor(fbs),

                  restecg = as.factor(restecg),

                  exang = as.factor(exang),

                  slope = as.factor(slope),

                  ca = as.factor(ca),

                  thal = as.factor(thal),

                  target = as.factor(target))

#reorder columns for easier exploratory analysis

dfHDFull = dplyr::select(dfHDFull, target, sex, fbs, exang, cp, restecg, slope, ca, thal, everything())

####

####EXPLORATORY ANALYSIS OF NUMERICAL VARIABLES

####

#check for missing values

PlotMissing(dfHDFull)

summary(dfHDFull)

#correlation matrix for independent numerical variables

dfHDFull.cor = cor(dfHDFull[10:14])

ggcorrplot(dfHDFull.cor, hc.order = TRUE, outline.color = "white", type = "lower", lab = T)

#boxplot of numerical variables to explore outliers

ggplot(melt(dfHDFull[,10:14]), aes(variable, value)) +

  geom\_boxplot() +

  facet\_wrap(~variable, scales = "free") +

  theme(axis.text.x= element\_blank()) +

  xlab("") + ylab("")

#Checking correlation of numerical covariates on categorical dependent

heart.model1 = glm(target~age, family = binomial, data = dfHDFull)

summary(heart.model1)

heart.model2 = glm(target~chol, family = binomial, data = dfHDFull)

summary(heart.model2)

heart.model3 = glm(target~trestbps, family = binomial, data = dfHDFull)

summary(heart.model3)

heart.model4 = glm(target~thalach, family = binomial, data = dfHDFull)

summary(heart.model4)

heart.model5 = glm(target~oldpeak, family = binomial, data = dfHDFull)

summary(heart.model5)

####

####EXPLORATORY ANALYSIS OF CATEGORICAL VARIABLES

####

#checking bias of categorical dependent

ggplot(dfHDFull, aes(x=dfHDFull$target, fill=dfHDFull$target)) +

  geom\_bar() +

  xlab("Heart Disease") +

  ylab("Count") +

  ggtitle("Heart Disease Occurances") +

  scale\_fill\_discrete(name = "Heart Disease", labels = c("No","Yes")) +

  theme(axis.text.x= element\_blank())

#Checking correlation of categorical covariates on categorical dependent

chisq.test(dfHDFull$target, dfHDFull$sex)

chisq.test(dfHDFull$target, dfHDFull$cp)

chisq.test(dfHDFull$target, dfHDFull$fbs)

chisq.test(dfHDFull$target, dfHDFull$restecg)

chisq.test(dfHDFull$target, dfHDFull$exang)

chisq.test(dfHDFull$target, dfHDFull$slope)

chisq.test(dfHDFull$target, dfHDFull$ca)

chisq.test(dfHDFull$target, dfHDFull$thal)

#fisher's test for small values

table(dfHDFull$target, dfHDFull$restecg)

fisher.test(dfHDFull$target, dfHDFull$restecg)

#exploring risk difference, risk ratio and odds ratio for binary covarites

table(dfHDFull$target, dfHDFull$sex)

table(dfHDFull$target, dfHDFull$fbs)

table(dfHDFull$target, dfHDFull$exang)

#sex

riskdifference(112, 24, 201, 95, CRC = FALSE, conf.level = 0.95)

riskratio(112, 24, 201, 95, conf.level = 0.95)

oddsratio(112, 24, 89, 71, conf.level = 0.95)

#fbs

riskdifference(20, 116, 43, 253, CRC = FALSE, conf.level = 0.95)

riskratio(20, 116, 43, 253, conf.level = 0.95)

oddsratio(20, 116, 23, 137, conf.level = 0.95)

#exang

riskdifference(74, 62, 97, 199, CRC = FALSE, conf.level = 0.95)

riskratio(74, 62, 97, 199, conf.level = 0.95)

oddsratio(74, 62, 23, 137, conf.level = 0.95)

####

#### MODEL SELECTION & COMPARISON

####

#Full model

dfHDFull.modelFull = glm(target~., family=binomial, data=dfHDFull)

#Through exploratory analysis, we determined that fbs and chol were not significant

#They are dropped from the dataset in use

drops = c("fbs", "chol")

dfHDRed = dfHDFull[ , !(names(dfHDFull) %in% drops)]

#The full model of the reduced data

dfHDRed.modelFull = glm(target~., family=binomial, data=dfHDRed)

#########STEPWISE SELECTION############

#using min AIC and going forwards and backwards

dfHDFull.modelStep = dfHDFull.modelFull %>% stepAIC(direction = "both", trace = FALSE)

dfHDRed.modelStep = dfHDRed.modelFull %>% stepAIC(direction = "both", trace = FALSE)

#the min AIC stepwise model selection produces the same results

#for the full dataset as the reduced dataset

#########BACKWARD SELECTION############

#starting with the full model on reduced data, then dropping in order: restecg, slop, age, exang, thal

heartred = glm(target~ sex+exang+cp+restecg+slope+ca+thal+age+trestbps+thalach+oldpeak, family = binomial, data = dfHDRed)

heart1 = glm(target~ sex+exang+cp+slope+ca+thal+age+trestbps+thalach+oldpeak, family = binomial, data = dfHDRed)

heart2 = glm(target~ sex+exang+cp+ca+thal+age+trestbps+thalach+oldpeak, family = binomial, data = dfHDRed)

heart3 = glm(target~ sex+exang+cp+ca+thal+trestbps+thalach+oldpeak, family = binomial, data = dfHDRed)

heart4 = glm(target~ sex+cp+ca+thal+trestbps+thalach+oldpeak, family = binomial, data = dfHDRed)

heart5 = glm(target~ sex+cp+ca+trestbps+thalach+oldpeak, family = binomial, data = dfHDRed)

####

#### MODEL COMPARISON

####

#testing each model in backwards selection using drop in deviance

anova(heart3, heart2, test = "Chisq")

anova(heart4, heart3, test = "Chisq")

anova(heart5, heart4, test = "Chisq")

#testing which is better: backwards or stepwise using drop in deviance

anova(heart4, dfHDRed.modelStep, test = "Chisq")

#model summaries, initial full model, backwards model, stepwise (final) model

summary(dfHDRed.modelFull)

summary(heart4)

summary(dfHDRed.modelStep)

#Goodness of Fit

pchisq(179.78, 295, lower.tail = F)

pchisq(194.51, 283, lower.tail = F)

pchisq(182.16, 280, lower.tail = F)

####

#### TEST MODEL ASSUMPTIONS

####

#check linearity

probabilities = predict(dfHDFull.modelStep, type = "response")

dfHDFull.lin = dfHDFull %>%

  dplyr::select\_if(is.numeric)

predictors = colnames(dfHDFull.lin)

dfHDFull.lin = dfHDFull.lin %>%

  mutate(logit = log(probabilities/(1-probabilities))) %>%

  gather(key = "predictors", value = "predictor.value", -logit)

ggplot(dfHDFull.lin, aes(logit, predictor.value)) +

  geom\_point(size = 0.5, alpha = 0.5) +

  geom\_smooth(method = "loess") +

  theme\_bw() +

  facet\_wrap(~predictors, scales = "free\_y")