Generalised Linear Models, Assessed Practical

MSc in Statistical Science

P699

11/12/2019

Contents

1. Introduction

The data consists of records of coronary heart disease status for 4658 individuals. For each individual, heart disease status is recorded as well as some other biological measurements and social characteristics about the individual.

We will explore the data and examine the eventual relationships between the variables. Then we will build a model with explanatory variables and perform outlier analysis. Finally, we will interpret the model.

2. Exploratory data analysis

The outcome of the data is the heart disease status (numerical) which is coded as: 1 for coronary heart disease and 0 for healthy heart, 31% of the people in the dataset are suffering coronary heart disease. The other information we have about the individuals are:

1. Sex (categorical): Indicates whether the individual is a man or a woman
2. Age (numerical): Age of the individual
3. Education level (categorical): ranging from A= high school, to D= university degree
4. Cholesterol level (numerical): in mg/dL
5. Systolic blood pressure (numerical): in mm Hg
6. Body Mass Index (numerical): weight in kg / (height in m)^2

In the next subsections, we will examine the relationship between the incidence of heart disease and the available explanatory variables.

2.1 Effect of sex

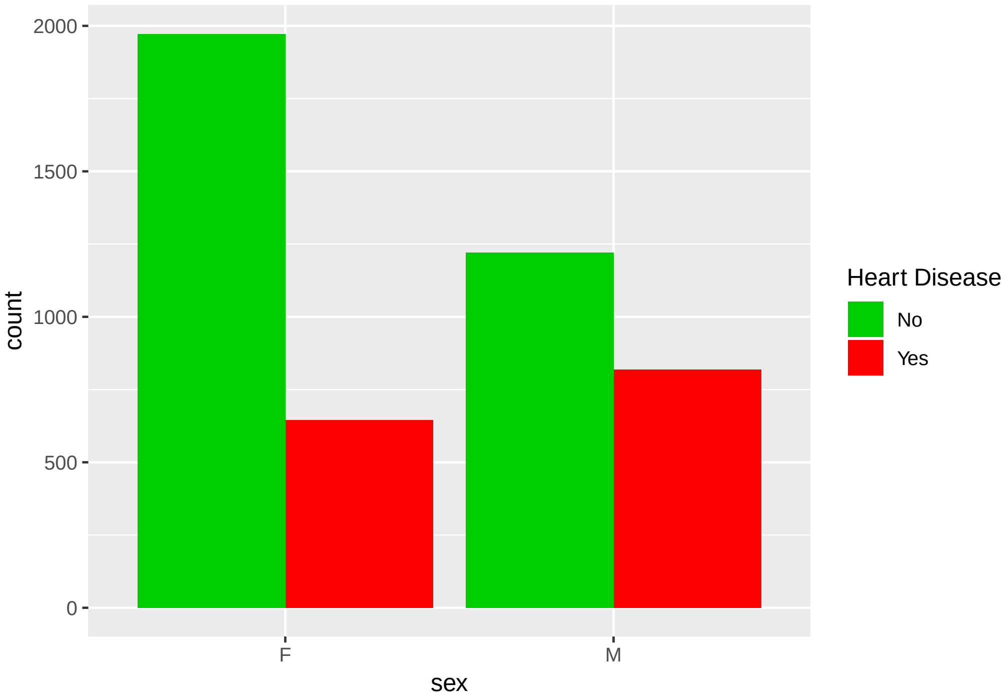

Figure 1: Heart disease incidence by sex

The data contains more records of women and Figure 1 seems to suggest that $\mathbb{P}(Disease|M) > \mathbb{P}(Disease|F)$: being a man increases the risk of coronary heart disease. Indeed, we find the following:

Table 1: Probability of Disease by sex

Sex	M	F
Proportion	43.8%	56.2%
P(Disease|sex)	0.401	0.246

This is not very surprising as we know that the frequence of many diseases depends on the sex. Nonetheless, we need to check that this observation is not biased by the distribution of age (i.e that the category of old people doesn't contain more men than women).

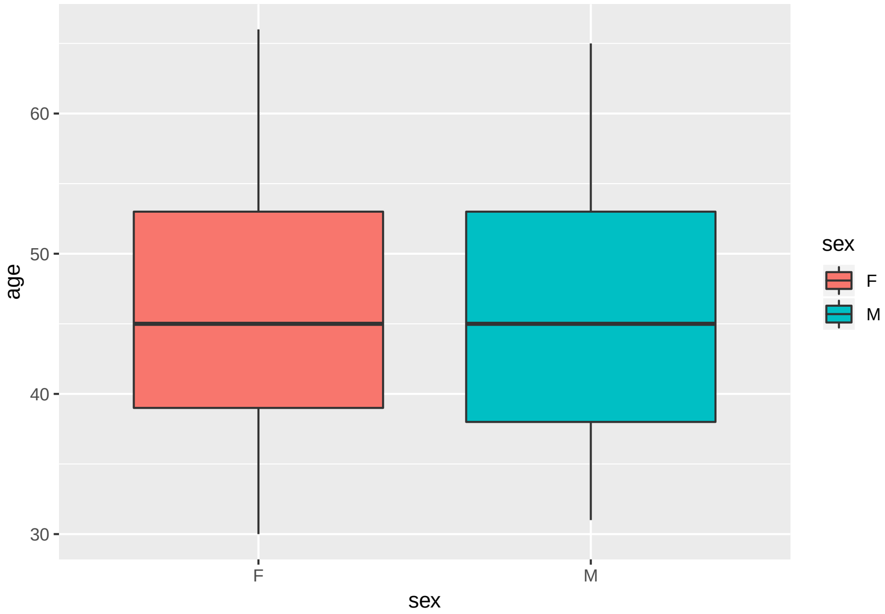

Figure 2: Age distribution by sex

This is obviously exluded because we see that the distribution of age doesn't differ by sex, we even have the opposite trend: men in the data tend to be younger so less prone to heart disease as we will see later. Similar plots with other continuous variables by sex show no significant difference in their distribution with respect to sex, so we may assume that sex is not very correlated to these variables, hence it will have a significant effect on the incidence of coronary heart disease.

2.2 Effect of Body Mass Index

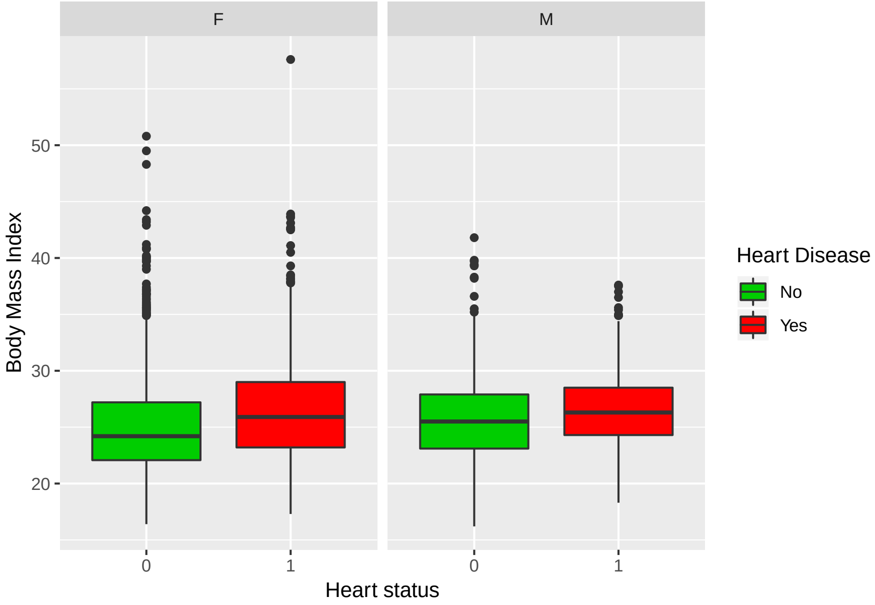

Figure 3: Boxplots of Body Mass index by disease status and sex

Figure 3 shows a higher risk of heart disease when the Body Mass Index is higher, once again, this is not surprising as we know that overweight is a risk for many heart diseases. Generally, the mean Body mass index of people suffering heart disease is 26.6 while it is 25.2 for healthy people.

2.3 Effect of the age

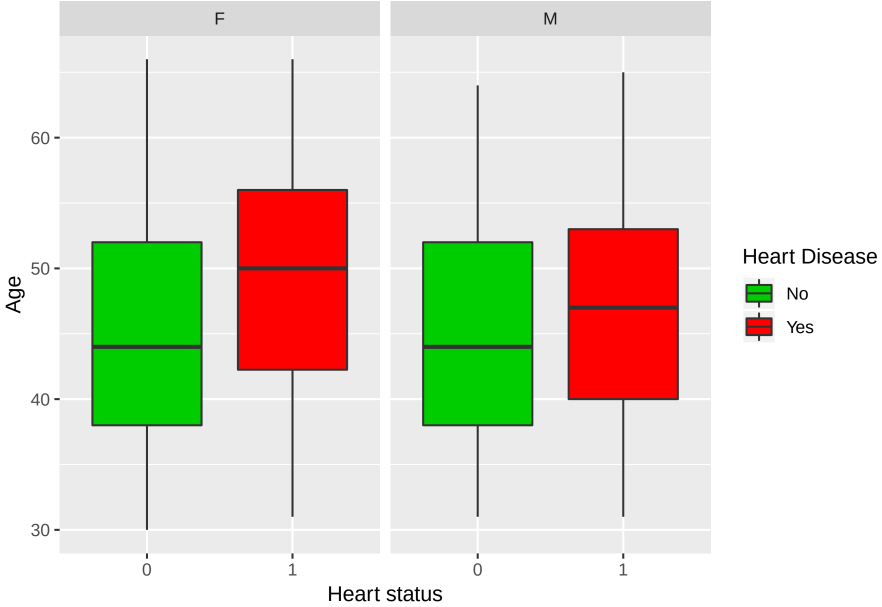

Figure 4: Boxplots of age by disease status and sex

As expected, older people are more affected by coronary heart disease. Generally, the average age of people suffering heart disease is 48 while it is 45 for healthy people.

2.4 Effect of Systolic blood pressure

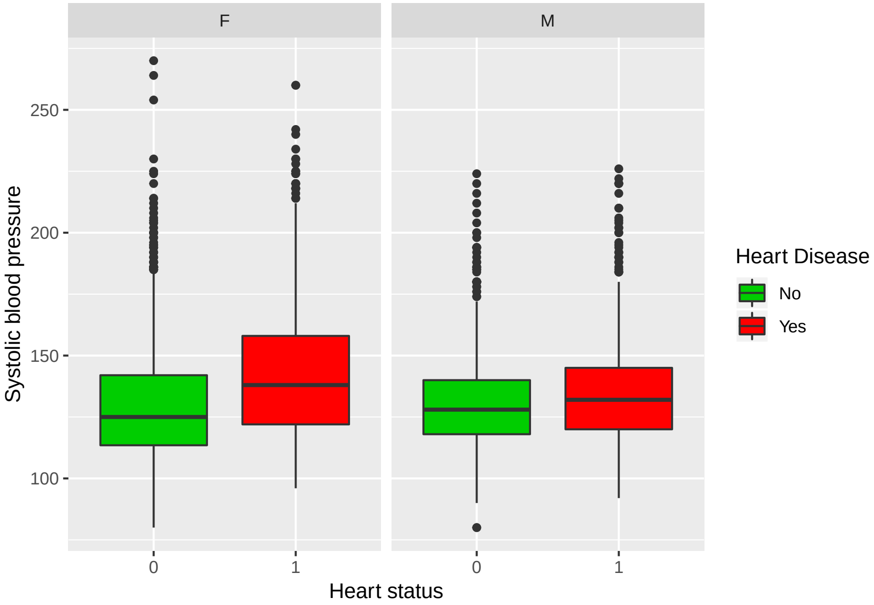

Figure 5: Boxplots of Systolic blood pressure by disease status and sex

The plot shows that a higher Systolic blood pressure increases the risk of coronary heart disease. We also notice that women's risk of heart disease seems more sensitive to Systolic blood pressure. Indeed:

Table 2: Average Systolic blood pressure by sex and disease status

Sex	Healthy	Suffering	Increase
M	130.19	135.12	3.7%
F	129.88	143.52	10.5%

This suggests a potential interaction term that we will investigate later.

2.5 Effect of Cholesterol level

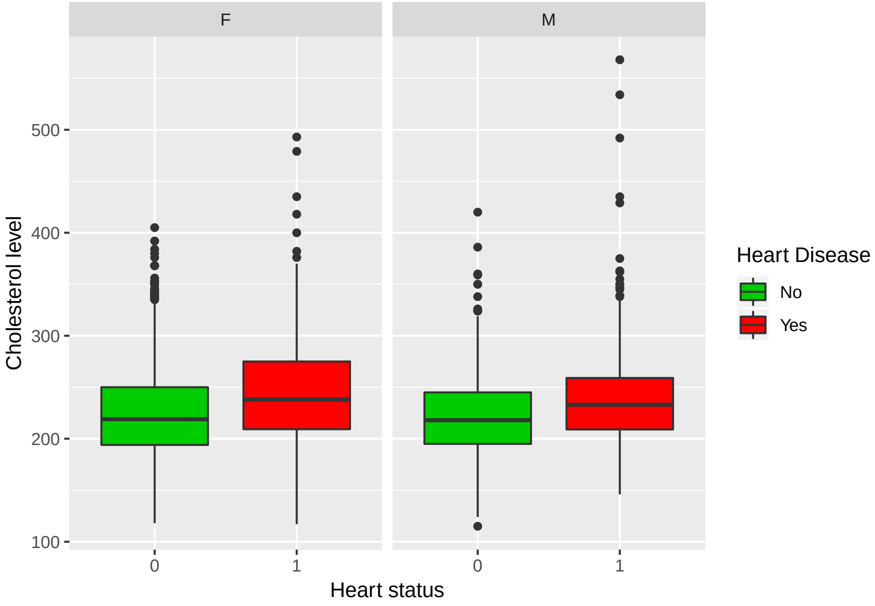

Figure 6: Boxplots of Cholesterol level by disease status and sex

As expected, a higher Cholesterol level increases the risk of coronary heart disease. Here again, we notice that the plots are different for each sex. This is less obvious than in the previous plots, but it may still suggest a potential interaction term as well.

2.6 Effect of Education level

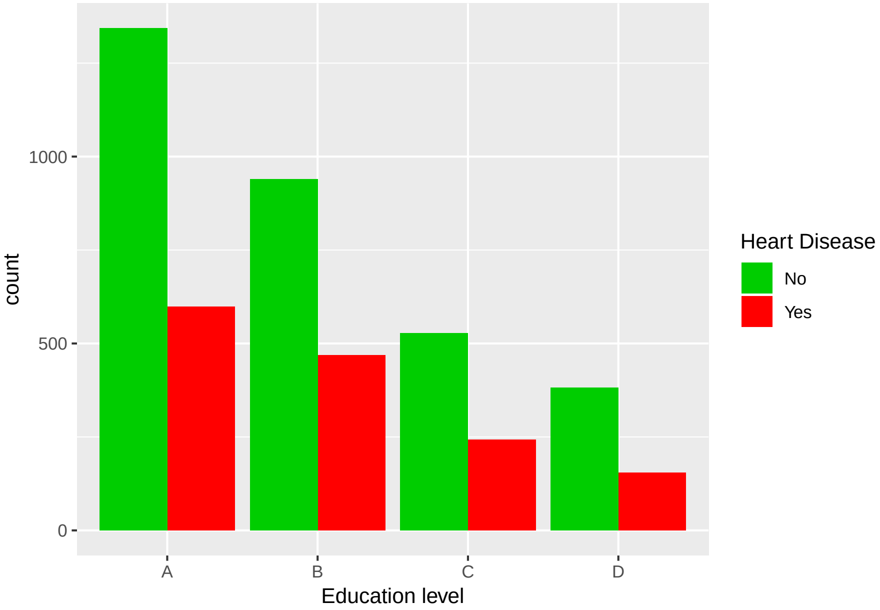

Figure 7: Barplot of heart disease incidence with respect to education level

Education level doesn't seem to have any effect on the outcome, indeed, the proportion of suffering people is approximately the same of each education level.

3. Data modeling

The outcome of interest in our dataset is the coronary heart disease status, it takes values 0 and 1, so we choose to model the outcome using a generalised linear model with binomial family (in our case a Bernoulli family).

3.1 Model selection

First, we include all the variables in our model without interaction term. It can be written as follows:

```r
glm1 = glm(hd ~ . , data=heartd, family=binomial)
```

All coefficients of this model are significant except for the Education level which confirms our finding in Figure 7. The Wald test tells us that none of the coefficients of education level is significant:

Table 3: P-value of the Wald test on the variable Education level

Coefficients	educB	educC	educD
P-value	0.19	0.88	0.28

So we choose to remove this variable from our model. When we fit again, all the variables are significant with p-values at most 2.2×10^{-5}.

3.2 Interactions

We can now investigate the potential interaction terms in the model. Based on the initial data analysis, some potential candidates are the interactions between sex and the Cholesterol level and between sex and the Systolic blood pressure. An other intuitively reasonable candidate would be the interaction between the age and the Body Mass Index. In fact, we can expect that being overweight leads to higher risk of heart disease for older people than for youngs. The following Figure illustrates this intuition:

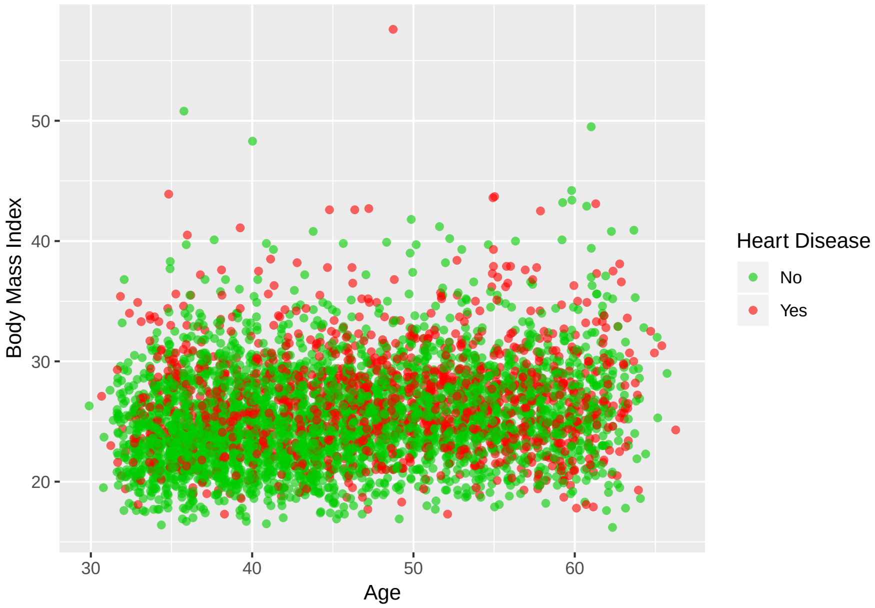

Figure 8: Heart disease incidence with respect to the age and the body mass index

In Figure 8, we can see that for a higher age, the effect of the Body Mass Index is lower and an individual would have a high risk of heart disease even with low Body Mass Index, which is not the case for younger people: The effect of Body Mass Index on heart disease incidence depends on the age.

As a result of the previous observations, we add those interaction terms and find that all the coefficients are now significant for the Wald test, though the terms Sex, Sex:Systolic blood pressure and Sex:Cholesterol Level are now only significant with level 10%. The significance of the variable Sex has been reduced because it is highly correlated with the interactions that include Sex. Nevertheless, the model now explains more variance and the Likelihood Ratio tests show that these interactions must be included as we can see in the following table:

Table 4: Analysis of deviance with interaction terms

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
NULL	NA	NA	4657	5801	NA
sex	1	126.9	4656	5674	1.982e-29
age	1	109.9	4655	5564	1.028e-25
chol	1	104	4654	5460	1.98e-24
sbp	1	70.25	4653	5390	5.223e-17
bmi	1	30.36	4652	5359	3.584e-08
age:bmi	1	5.485	4651	5354	0.01918
sex:sbp	1	2.896	4650	5351	0.08881
sex:chol	1	3.649	4649	5347	0.0561

After examining other interactions, we find that both Wald tests and Likelihood Ratio tests assert that they are not significant.

Now, let's check that our model minimises the AIC and that we don't need to drop any of the terms that we have considered so far:

Table 5: Stepwise AIC and deviance values

Model	Df	Deviance	AIC
Full		5347	5365
- sex:chol	1	5351	5367
- sex:sbp	1	5351	5367
- age:bmi	1	5353	5369

Table 5 confirms our previous findings: We can consider all the added terms.

3.3 Outlier Analysis

In this section, we will look for outliers. First let's point out that we can't rely on the plot of standardised deviance residuals or their QQ-plot. Indeed, we can't expect them to have a normal distribution because we are dealing with Bernoulli variables. Also, the residuals can only take two possible values for a given predicted response. Therefore, we will only use the plots of the leverage and cook's distance to have an approximate rule to find outliers.

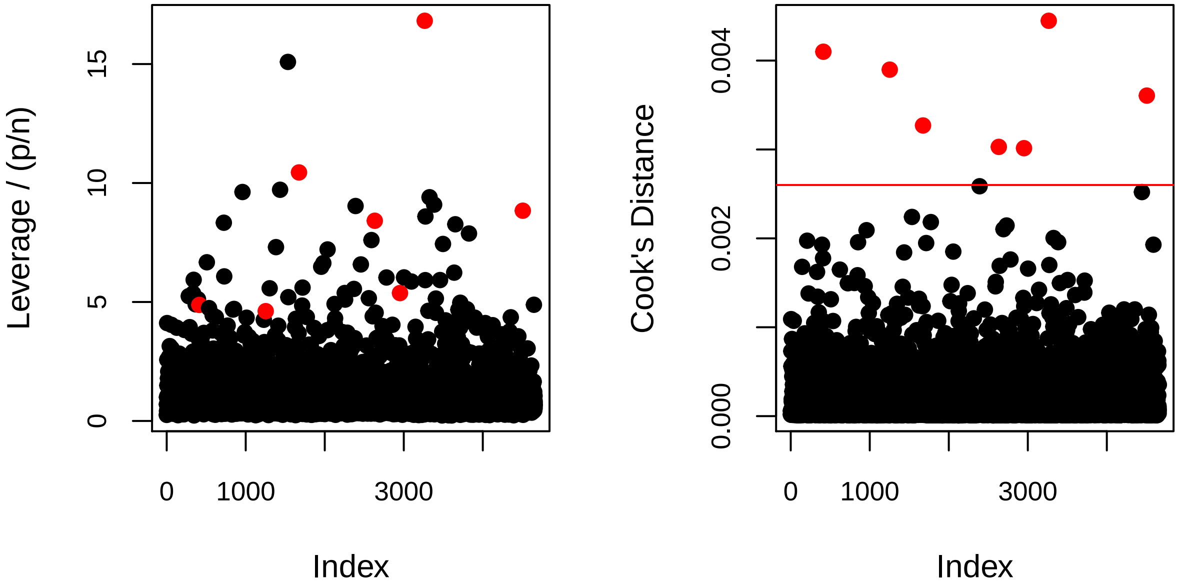

Figure 9: Leverage and influence of observations

Figure 9 shows the observations considered as outliers in red. Instead of using the theoretical treshold $\frac{8}{n-2p} \approx 0.00173$, we decide to use a relaxed treshold = 0.0026 to determine outliers in order to avoid removing too many observations and because there are observations with much higher cook's distance. This gives us 7 outliers as shown in the plot at the right.

We see in the left plot of Figure 9 that many observations have high leverage but not all of them have a high influence on the parameters of the model. We decide to remove the outliers and refit the model. We perform the same analysis once again and find that we have 2 observations with cook's distance slightly higher than the treshold, for this reason we decide to ignore them.

4. Model interpretation

After removing the outliers and refitting the model, the Residual Deviance drops from 5347 to 5330 which is a good improvement given that we only removed 7 observations: the drop per observation is 2.43 whereas Residual Deviance divided by the number of observations is 1.14.

We can notice that compared to the degrees of freedom 4642, the Residual Deviance 5330 can seem quite high, but given that we are modelling a Bernoulli variable, the Residual Deviance doesn't have $\mathcal{X}^2$ distribution.

Table 6: Parameters of the final model

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-9.102	1.225	-7.43	1.082e-13
sexM	0.8934	0.521	1.715	0.08642
age	0.07788	0.02513	3.1	0.001936
chol	0.005118	0.001062	4.821	1.426e-06
sbp	0.01404	0.002049	6.854	7.198e-12
bmi	0.1557	0.04606	3.381	0.0007228
age:bmi	-0.002294	0.0009464	-2.424	0.01536
sexM:sbp	-0.006553	0.003058	-2.143	0.03209
sexM:chol	0.003335	0.001578	2.114	0.03453

Table 6 shows that all the parameters are significant to the level 5% except for the variable sex that has a p-value 0.086.

Before interpreting the parameters, we should point out that a higher parameter doesn't necessarily mean a higher effect in this model. Indeed, we must take into account the standard deviation of the variables in question before interpreting. To develop this observation, we first summarise the standard deviations of the continuous variables of our data in Table 7:

Table 7: Standard deviation of variables

Variables	age	chol	sbp	bmi
Standard Deviation	8.49	44.45	22.49	4.053

Now, we see from Table 6 that an increase of 1 in age implies an increase of 0.07788 in the link function, whereas an increase of 1 in Cholesterol level implies an increase of only 0.005118 in the link function. But if we look at the standard deviations of the variables, an increase of 1 in age is roughly as likely as an increase of $\sigma(chol)/\sigma(age) \approx 5$ in the Cholesterol level which corresponds to an increase of ≈ 0.027 in the link function. The same analysis tells us that if we rescale the variables by dividing on their standard deviation, we would find that the effect of Cholesterol level and Systolic blood pressure are approximately the same.

This analysis tells us that there are variables that change more easily than others, which inflates their effect on the outcome.

We define the odds of coronary heart disease as $\frac{\mathbb{P}(disease)}{\mathbb{P}(healthy)}$, knowing that $\hat{\pi} = \frac{e^{\hat{\eta}}}{1+e^{\hat{\eta}}}$, we have that $O\hat{d}ds = \exp(\hat{\eta}) = \exp(x^T\hat{\beta})$. From Table 6, we can compute the multiplicative effects of the variables on the odds of coronary heart disease:

Table 8: Effect of an increase of 1 on the odds of heart disease

	sexM	age	cholesterol	blood pressure	body mass index
Effect	2.443	1.081	1.005	1.014	1.169

From Table 8:

- Being a man multiplies the odds of heart disease by 2.44
- For an average weight person (body mass index ≈ 25), an increase of 1 in the age multiplies the odds of heart disease by 1.0207 (we can't neglect the interaction effect in this case). Indeed, 1.081 would be the multiplicative effect on a person with weight 0 which does not make sense
- For a woman, an increase of 1 in the Cholesterol level multiplies the odds of heart disease by 1.005
- For a woman, an increase of 1 in Systolic blood pressure multiplies the odds of heart disease by 1.014
- For a 40 years old person, an increase of 1 in the Body Mass Index multiplies the odds of heart disease by 1.0665 (we can't neglect the interaction effect in this case). Indeed, 1.169 would be the multiplicative effect on a person with age 0, but it wouldn't be reasonable to extrapolate the model to this value

Table 6 also confirms our initial intuition on the interaction Age:Body Mass Index (Figure 8): For younger people, the Odds of heart disease increase more with the body mass index, which means that by losing a fixed amount of weight, young people reduce the risk of heart disease more efficiently than an old person

To assess the effect of the interaction terms Sex:Systolic blood pressure and Sex:Cholesterol level. We compute the multiplicative effects of those continuous variables for men. We find that:

- An increase of 1 in Cholesterol level multiplies the odds of heart disease by 1.005 for a woman and by 1.0084 for a man: the effect of Cholesterol level growth is more severe for men
- An increase of 1 in Systolic blood pressure multiplies the odds of heart disease by 1.014 for a woman and by 1.0074 for a man: the effect of Systolic blood pressure growth is more severe for women

5. Conclusion

We modeled the probability of having coronary heart disease using generalised linear models and including Sex, Age, Cholesterol level, Systolic blood pressure and Body Mass Index as well as the interactions Age:Body Mass Index, Sex:Systolic blood pressure and Sex:Cholesterol level.

The initial units of the variables seem to have been chosen so that we obtain similar standard deviations, but it may be interesting to do some preprocessing to the data and rescale the variables in order to have the same standard deviation for all continuous variables and obtain clearer effects in the final model.

Finally, although our parameter estimates are significant, we can still improve the accuracy of our model using further specific information such as smoking, drinking, exercising habits as well as diabetes signs.

R code

```r
#Libraries
library(ggplot2)
library(gridExtra)
library(ggExtra)
library(knitr)
library(MASS)
library(leaps)
library(pander)


#Loading the data
heartd <- read.csv("heartd.csv")
attach(heartd)
```

```r
#### Data Exploration ####


#2.1 Effect of Sex

ggplot(heartd, aes(x = sex, fill = as.factor(hd))) +
  geom_bar(stat = "count", position="dodge") + scale_colour_brewer(palette = "Set1") +
  scale_fill_manual(name="Heart Disease", labels=c("No", "Yes"),
                    values = c("0" = 3, "1" = 2))

disease_sex = data.frame("Sex" = c("Proportion", "P(Disease|sex)"),
                         "M" = c("43.8%", 0.401), "F" = c("56.2%", 0.246))

kable(disease_sex, caption = "Probability of Disease by sex")

ggplot(heartd, aes(x = sex, fill = sex))  + geom_boxplot(aes(y = age))

#2.2 Effect of Body Mass Index

ggplot(heartd, aes(x = as.factor(hd), fill = as.factor(hd)))  +
  geom_boxplot(aes(y = bmi)) + facet_grid(cols = vars(sex)) +
  scale_colour_brewer(palette = "Set1") +
  labs( x = "Heart status", y = "Body Mass Index") +
  scale_fill_manual(name="Heart Disease", labels=c("No", "Yes"),
                    values = c("0" = 3, "1" = 2))

#2.3 Effect of the age

ggplot(heartd, aes(x = as.factor(hd), fill = as.factor(hd)))  +
  geom_boxplot(aes(y = age)) + facet_grid(cols = vars(sex)) +
  scale_colour_brewer(palette = "Set1") + labs( x = "Heart status", y = "Age") +
  scale_fill_manual(name="Heart Disease", labels=c("No", "Yes"),
                    values = c("0" = 3, "1" = 2))

#2.4 Effect of Systolic blood pressure

ggplot(heartd, aes(x = as.factor(hd), fill = as.factor(hd)))  +
  geom_boxplot(aes(y = sbp)) + facet_grid(cols = vars(sex)) +
  scale_colour_brewer(palette = "Set1") +
  labs( x = "Heart status", y = "Systolic blood pressure") +
  scale_fill_manual(name="Heart Disease", labels=c("No", "Yes"),
                    values = c("0" = 3, "1" = 2))

sbp_sex = data.frame("Sex" = c("M", "F"), "Healthy" = c(130.19, 129.88),
                     "Suffering" = c(135.12, 143.52), "Increase" = c("3.7%", "10.5%"))

kable(sbp_sex, caption = "Average Systolic blood pressure by sex and disease status")

#2.5 Effect of Cholesterol level

ggplot(heartd, aes(x = as.factor(hd), fill = as.factor(hd))) +
  geom_boxplot(aes(y = chol)) + facet_grid(cols = vars(sex)) +
  scale_colour_brewer(palette = "Set1") +
  labs( x = "Heart status", y = "Cholesterol level") +
  scale_fill_manual(name="Heart Disease", labels=c("No", "Yes"),
                    values = c("0" = 3, "1" = 2))


#2.6 Effect of Education level

ggplot(heartd, aes(x = educ, fill = as.factor(hd))) +
  geom_bar(stat = "count", position="dodge") + scale_colour_brewer(palette = "Set1") +
  labs( fill = "Heart status", x = "Education level") +
  scale_fill_manual(name="Heart Disease", labels=c("No", "Yes"),
                    values = c("0" = 3, "1" = 2))


#### Data Modeling ####


#3.1 Model selection

glm1 = glm(hd ~ . , data=heartd, family=binomial)

education = data.frame("Coefficients" = c("P-value"), "educB" = c(0.19),
                       "educC" = c(0.88), "educD" = c(0.28))

kable(education, caption = "P-value of the Wald test on the variable Education level")

#3.2 Interactions

ggplot(heartd, aes(x = age, y = bmi, color = as.factor(hd))) +
  geom_jitter(height = 0, width = 0.4, alpha = .6)  +
  labs( color = "Heart status", x = "Age", y = "Body Mass Index") +
  scale_colour_manual(name="Heart Disease", labels=c("No", "Yes"),
                      values = c("0" = 3, "1" = 2))

glm2 = glm(hd ~ . - educ + bmi*age +sex*(sbp+ chol), data=heartd, family=binomial)

summary2 = summary(glm2)

anova2 = anova(glm2, test = "Chisq")

pander(an, caption = "Analysis of deviance with interaction terms")

AIC = step(glm2)

aic_tab = data.frame("Model" = c("Full", "- sex:chol", "- sex:sbp", "- age:bmi"),
                     "Df" = c("", 1, 1, 1), "Deviance" = c(5347, 5351, 5351, 5353),
                     "AIC" = c(5365, 5367, 5367, 5369))

kable(aic_tab, caption = "Stepwise AIC and deviance values")

#3.3 Outlier Analysis

p <- 9
```

```
n <- 4658
treshold = 8/(n-2*p)
soft_treshold = 0.0026

outliers = cooks.distance(glm2) >= soft_treshold

num_outliers = sum(outliers)

mar <- par("mar")
mar[c(2, 3)] <- 3.8
par(mfrow = c(1, 2), mar=mar)

plot(influence(glm2)$hat/(p/n), pch=19, col=1+outliers, ylab='Leverage / (p/n)',
     yaxt="n", xaxt="n")
axis(1,cex.axis=.8)
axis(2,cex.axis=.8)

plot(cooks.distance(glm2), pch=19, col=1+outliers, ylab="Cook's Distance",
     yaxt="n", xaxt="n")
axis(1,cex.axis=.8)
axis(2,cex.axis=.8)
abline(a=soft_treshold, b=0, col = "red")

#Removing outliers and fitting again

heartnew = heartd[-which(outliers),]

glm3 = glm(hd ~ . - educ + bmi*age +sex*(sbp+ chol), data=heartnew, family=binomial)

par(mfrow = c(1, 2), mar=mar)

plot(influence(glm3)$hat/(p/n), pch=19, ylab='Leverage / (p/n)',yaxt="n")
axis(2,cex.axis=1)

plot(cooks.distance(glm3), pch=19, ylab="Cook's Distance",yaxt="n")
axis(2,cex.axis=1)
abline(a=soft_treshold, b=0, col = "red")


#### Model interpretation ####


#Final coefficients

summary3 = summary(glm3)

pander(summary3$coefficients, caption = "Parameters of the final model")

#Standard deviations of the variables

sd_vars = apply(heartnew, 2, sd)
sd_vars = as.data.frame(t(as.matrix( tab[!is.na(sd_vars)][-1] )))
```

```
pander(cbind(data.frame("Variables" = c("Standard Deviation")), sd_vars),
       caption = "Standard deviation of variables")

#Multiplicative effect

options(digits = 5)

exp_coeff = exp(summary3$coefficients[,1])

vars = c("sexM" , "age" , "cholesterol" , "blood pressure" , "body mass index" )

values = c(2.4434 , 1.0810 , 1.0051 , 1.0141 , 1.1685 )

exp_effect = as.data.frame(t(data.frame("Effect"=values)), header = TRUE)

colnames(exp_effect) = vars

pander(exp_effect, caption = "Effect of an increase of 1 on the odds of heart disease")
```