MATH3823 Assessed Practical

Faye Williams
ID - 201308646

July 23, 2023

Abstract

In this investigation, we find that blood pressure has an influence on how likely humans are to develop coronary heart disease (CHD). In particular, having a higher blood pressure puts individuals at greater risk of developing CHD. In addition to this, our serum cholesterol levels also have an impact, as people with a higher serum cholesterol level are more likely to develop CHD than those with a lower serum cholesterol level. We also find that this impact is at it's greatest when levels of serum cholesterol are above 219mg/100c. For those with a lower cholesterol, we are less likely to see such a large impact. Additionally, having a blood pressure of over 147mm of mercury will increase ones likelihood of developing CHD in comparison to those with a blood pressure below 147. However, the likelihood of CHD development increases even more for people with a blood pressure over 166mm of mercury. These two factors both influence the incidence of CHD, but there is no evidence to suggest there is an interaction between the two, meaning they are independent of each other.

1 Introduction

In this report we shall investigate the influence of the two covariates, blood pressure and cholesterol, on the likelihood of developing coronary heart disease (CHD). This investigation will be based on a data set of 1329 American men, each with their serum cholesterol level, in mg/100cc, and blood pressure, in mm of mercury, recorded for each man. We shall begin this investigation by looking at the box plots of incidence of CHD for these men in each of the four blood pressure groups and each of the four cholesterol level groups.

2 Box Plots

We have the box plots for each cholesterol group given as follows

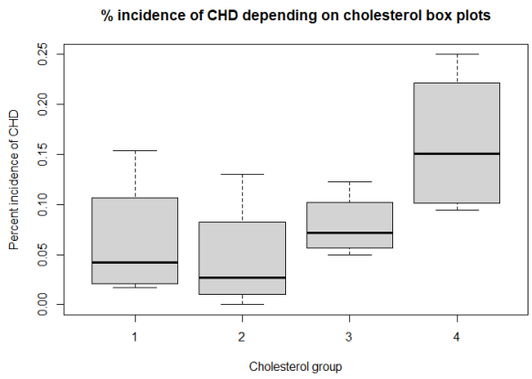

Figure 1: Box Plots for Cholesterol

I have included these two box plots as they allow us to see the spread and variation of data that cannot be expressed in the line graph, which we will see later on. I have also included both box plots as they each show the impact of separate factors affecting the incidence of Coronary Heart Disease (CHD) in our subjects, which are their cholesterol levels and blood pressure reading.

2.1 Figure 1

From figure 1, we see the average percentage incidence of our subjects having coronary heart disease (CHD) from each of the 4 groups for cholesterol. It is clear to see that, on average, the group with the highest chance of developing CHD is group 4, which denotes subjects with a serum cholesterol level of over 259mg/100c. Furthermore, the third group shows the second highest incidence of

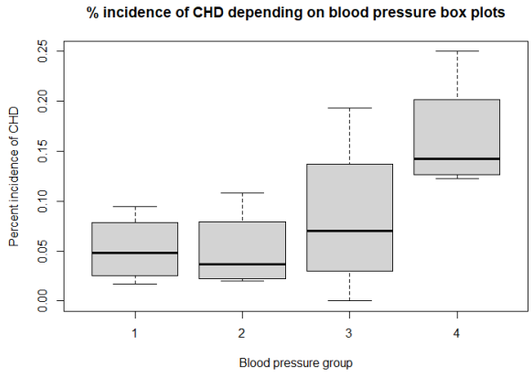

Figure 2: Box Plots for Blood Pressure

CHD, followed by the first and second groups, which show roughly the same percentage incidence on average. Therefore, from this we can conclude that the chance of developing CHD increases as ones serum cholesterol level increases. Not only this, but we also see a wider spread of data in group 4 than in the other groups, as shown by the range of the upper and lower quartile being around 12% and the full range from maximum to minimum value being roughly 15%. This suggests that although we can clearly see a higher percent incidence in this category, we cannot be sure of how great this difference is, the introduction of more participants in each of the different categories may help us understand this more.
We see that the box plots for cholesterol groups 1 and 2 are very similar in terms of their means, as well as the spread, as shown by the ranges both being around 13%. In addition to this, the interquartile range is shown to be similar, at approximately 8% for both groups. This suggests that the spread of data in these two groups and we may be able to combine these factors when making our model to give a better fit to the data, which we shall investigate in the future. Despite this, we see that the second group shows a slightly lower incidence of CHD occurring on average in comparison to the first, even though the general trend is that the incidence increases as cholesterol levels increase. This may be due to outliers, as we have no lower bound on the cholesterol levels, therefore someone with very low cholesterol levels may be more at risk of developing CHD. However we cannot be sure of this and would need more data specific to the cholesterol level reading when it is under 200mg/100c.

2.2 Figure 2.

Alternately, when comparing blood pressure groups we have the box plots above.Again, we see a very similar box plot for blood pressure groups 1 and 2 in figure 2, meaning we may also be able to combine these two factor levels when finding a suitable model for the data. We also again see an increase on average for incidence of CHD as the blood pressure levels increase. In particular, we have an average of around 15% in group 4, which decreases to 7.5% in group 3 and even lower to around 5% in groups 1 and 2. This suggests that the chance of developing CHD is greater when one's blood pressure is above average, strictly speaking this would be above 150 in our case.
However, we do see in the third blood pressure group that the minimum incidence of CHD is at 0%. Although this may suggest having a blood pressure within the range of 147 to 166 has no effect on the occurrence of CHD, the rest of the box plot suggests otherwise as we see the average percent incidence to be around 7.5%. Therefore, it is more likely that we can class the 0% value as an outlier, as we have sufficient evidence to suggest a blood pressure of above 147 will indeed increase ones likelihood of developing CHD.
Despite this box plot showing some similarities to figure 1, we see a much smaller interquartile range for groups 1, 2 and 4, in comparison to cholesterol groups 1, 2 and 4. In fact, the interquartile range in blood pressure groups 1 and 2 is around 5% and in group 4 around 7%. As this data is much more concentrated around the average incidence of CHD in each of the three box plots, we have more evidence to suggest that the average incidence of CHD in these groups is accurate, therefore strengthening the claim that blood pressure has an influence on CHD development, and perhaps a greater influence than cholesterol, as the ranges in figure 1 are generally larger. We shall investigate this interaction between the two factors further in the figures below.

2.3 Line Graphs

For further investigation, we can plot the percent incidence of CHD involving both the blood pressure group and cholesterol group of each subject. We do this by first taking a scatter plot of incidence against blood groups and connecting the points according to each cholesterol group, we have the graph below

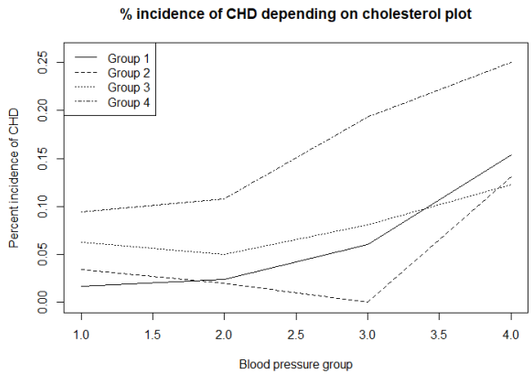

Figure 3: Plot of Blood Pressure against CHD incidence, with a key of Cholesterol groups

2.3.1 Figure 3

From figure 3, we can clearly see a correlation between the blood pressure group and incidence of CHD, with likelihood of developing CHD increasing as the blood pressure increases. This can be shown in all four lines, suggesting the influence of blood pressure on CHD incidence is regardless of the influence of cholesterol. The dot-dash line shows us the incidence of CHD for cholesterol level group 4, which we can clearly see is greater than the incidence of any of the other 3 groups for cholesterol, regardless of the blood pressure group. This suggests that having a serum cholesterol level of over 259mg/100c increases your chance of developing coronary heart disease, regardless of your blood pressure.

Furthermore, the dotted line for group 3 is almost completely above the lines for cholesterol groups 1 and 2, suggesting that individuals with cholesterol levels between 220 and 259 are more likely to develop coronary heart disease than individuals with a cholesterol level below 220mg/100c, no matter what their blood pressure is. This provides further evidence for our claim above.

We also see that this line has the greatest increase in percent incidence of CHD over the 4 different levels of blood pressure; in particular we see the incidence of CHD increase by over 15% from blood pressure group 1 to 4. This suggests that given a high level of cholesterol, the risk of developing CHD increases as ones blood pressure reading increases. This may suggest an interaction element between the cholesterol and blood pressure levels, where they influence each other in the development of CHD, we shall investigate this further in the model section to see if there is enough evidence to suggest an interaction between the two.

Moreover, we see a similar range in CHD incidence for subjects in group 1 for cholesterol, i.e. having a cholesterol level of less than 200mg/100c, over the 4 blood pressure groups. In particular, there is an almost 14% increase of incidence of CHD from blood pressure group 1 to group 4. This suggests that there may be a correlation between incidence of CHD and blood pressure levels when looking at subjects with low cholesterol levels, as well as subjects with high cholesterol levels as explained above.

We see a similar increase for subjects with cholesterol levels between 200 and 219, but this increase has a smaller range of roughly only 6%, suggesting that blood pressure has only a small influence on the likelihood of developing CHD for subjects with a healthy cholesterol level of 200-219mg/100cl.

Furthermore, we see that the right hand side of the graph has much higher plots for percentage incidence on average than the left hand side; in particular, for blood pressure in group 4 the percent incidences of CHD are all over 10%, no matter the cholesterol group, whereas in the first group for blood pressure they all fall below 10%. This suggests that blood pressure indeed has an impact on the likelihood of developing coronary heart disease, in fact the higher the individual's blood pressure, the more likely they are to develop CHD.

Lastly, we see that the 4th group for cholesterol clearly has the highest percent incidence, with groups 1 and 2 following behind, on average. This is what we would expect after the analysis of our box plots, and we shall investigate this claim further when generating a model for the data.

2.3.2 Figure 4

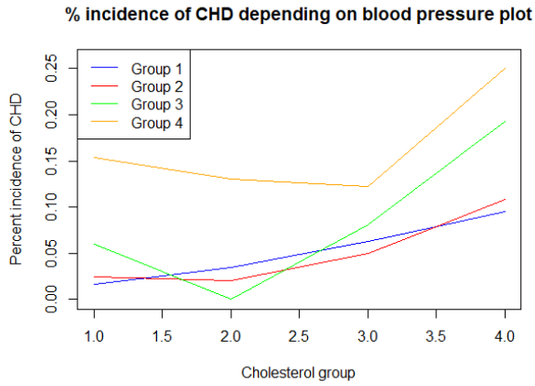

Figure 4: Plot of Cholesterol against CHD incidence, with a key of Blood Pressure groups

We can also use figure 4 for further analysis, this is a similar line graph but we instead have swapped the placement of the cholesterol levels and blood pressure groups around. Although this presents the same information, we may be able to see trends in the CHD incidence of the four blood pressure groups that were not as clear in figure 3.
As expected, we also see the positive correlation between cholesterol levels and incidence of CHD, as well as the positive correlation between blood pressure and CHD incidence. Not only this, but again we find the plots for blood pressure groups 1 and 2 to be very similar, therefore suggesting that a combination of these models may be needed to increase our residual degrees of freedom in our final model. Furthermore, we find the blood pressure group 3 to have varying levels of incidence of CHD, no matter the cholesterol level. We see a sharp drop to 0% in this instance for subjects with cholesterol between 220 and 259mg/100c, and as previously mentioned this may be considered an outlier, as we see a steady increase in CHD incidence for blood pressure group 3 if we were to exclude this reading.

3 Model Justification

3.1 Choice of distribution - Binomial

I chose to use the binomial model due to the nature of the dataset, we have recorded the people that developed CHD in each group as the successes, given by y, and the people in the group that did not develop CHD as the failures, which were given by m-y.

3.2 Choice of model explanatory variables

Firstly, we can exclude the two models which depend only on one of the factors, namely models $\frac{Y}{M}_C \sim C$ and $\frac{Y}{M}_B \sim B$, which assume an influence of cholesterol on CHD incidence only and an influence of blood pressure on CHD incidence only respectively.
We find that, when using the binomial distribution, the degrees of freedom are 12 in both cases, and the residual deviances are 26.805 for model $\frac{Y}{M}_C$ and 35.163 for $\frac{Y}{M}_B$ [Appendix figure 13]. The first model is shown to have a much lower deviance than the second, which makes it a better fit to the data in comparison to the second model. However despite this improvement, it is still a large value for deviance and as we will see, this becomes lower when we consider the model including both the influence of cholesterol levels and blood pressure.
For the model considering the blood pressure and cholesterol levels without interaction, given by $\frac{Y}{M}\tilde{C}+B$, we found the lowest deviance when using the 'probit' link function in the binomial model. In this case, we have a residual deviance of 7.5248 and 9 residual degrees of freedom in this case [Appendix figure 15]. To see if the simpler model, $\frac{Y}{M}\ C+B$, is a better fit to the data than the more complex model, $\frac{Y}{M}_{int}\tilde{C}+B+C:B$, which introduces the interaction between the blood pressure levels and cholesterol, we carry out the chi squared test. Sticking with the binomial distribution and the 'probit' link function, as this gave us the lowest residuals in both cases, we find that for the interaction model, the deviance is $4.49e^{-10}$.
Although we have a much lower deviance for the more advanced model, we also find this takes us to 0 degrees of freedom. This is called a saturated model and would not be suitable to use, not only that but the p value for the chi squared distribution with 9 degrees of freedom is 16.92, and as the deviance of the simpler model minus the more advanced model is less than this, we choose to reject the alternate model and stick with the original model, $\frac{Y}{M}\tilde{C}+B$.

3.3 Combination of factor levels

Next we considered whether it would be suitable to combine factor levels, with the aim of making the model smoother. When we don't combine any factor levels and stick to the 4 levels for cholesterol and 4 levels of blood pressure, we find the generalized linear model has a binomial distribution and a 'probit' link function. In this case, the residual deviance is given as 7.5248 with 9 residual degrees of freedom. We want to investigate whether the combination of the first two levels for blood pressure makes any difference, we shall find the generalized linear model in this circumstance. For this model, we have

Here we can carry out another hypothesis test to see whether the simpler model $\frac{Y}{M} \sim C + B_3$ is a better fit to the model than the alternative original model, $\frac{Y}{M} \sim C + B$, where in this case, we define B_3 as the covariate of blood pressure being split into 3 groups instead of 4, precisely we have 147, $147-166$ and 166. We find that the difference in residual deviance is 0.034756 and the change in degrees of freedom is 1. According to the chi squared test for 1 degree of freedom at the 5% significance level, we require the difference in deviance to be greater than 3.84. As our difference is less than this, there is not sufficient evidence to suggest the $\frac{Y}{M} \sim C + B$ model is a better fit to the data than the simplified $\frac{Y}{M} \sim C + B_3$ model, therefore we shall continue our investigation using the $\frac{Y}{M} \sim C + B_3$ model.

We can take this further by combining factors in the cholesterol group, and we find that the combination of the first 2 levels for cholesterol gives a more suitable fit to the model. In this case we have the model $\frac{Y}{M} \sim C_3 + B_3$, where C_3 denotes the 3 cholesterol groups as 219, $219-259$ and 259. Again, using a binomial distribution with the 'probit' link function, we have a residual deviance of 7.7296 and 11 residual degrees of freedom [Appendix figure 17]. To test if this even simpler model is a better fit to the data than the $\frac{Y}{M} \sim C + B_3$ model, we shall carry out another chi squared test. We find that, for the $\frac{Y}{M} \sim C + B_3$ model to be statistically significant we require the difference between our two deviances to be greater than the p value for chi squared with 2 degrees of freedom, which is 5.99. Clearly, the difference in deviance is smaller than this, therefore we do not have sufficient evidence to reject our simpler model and will choose the $\frac{Y}{M} \sim C_3 + B_3$ model as the best fit to the data.

After more investigation, combining other factor levels, we find that this model is the overall best fit for the data, as the comparison with even simpler models showed this model to be the best fit at significant a significance level of 5%.

3.4 Link function

Further, I have used the 'probit' link function over the 'logit' or 'cloglog' link functions for my model as it gave me the lowest residual deviance out of the three, meaning the 'probit' function gave the best fit to the data. In particular, when comparing the link functions for the model $\frac{Y}{M} \sim C_3 + B_3$ with the binomial distribution, the 'probit' link function gave a deviance of 7.7296, meanwhile the 'logit' and 'cloglog' functions gave 8.2971 and 8.5526 respectively.

4 Final model

Overall, the best fit model to the data is given by the model $\frac{Y}{M} \sim C_3 + B_3$ with a binomial distribution and a 'probit' link function, where C_3 is the 3 cholesterol levels and B_3 is the 3 blood pressure levels as detailed below,

	Group 1, g1	Group 2, g2	Group 3, g3
Cholesterol	219	219-259	259
Blood Pressure	147	147-166	166

The final generalised linear model is given as follows

$$\frac{Y_i}{M_i} = -1.96 + 0.64B_g3 + 0.31C_g2 + 0.71C_g3 \tag{1}$$

where the residual deviance is given as 7.7296 and we have 11 residual degrees of freedom.

4.0.1 Residual Plots

We have the residual plots for our generalised linear model as follows,

As we can see they all express a similar result. That is, a plot with a somewhat negative correlation between the residuals and fitted values, concentrating at the lower end of the fitted values and proportionately higher end of the residual values.

Despite this general trend, we also see some outliers that are clear in all three plots, specifically on the far right hand side of the graph where we have a plot with a high fitted value and a varying

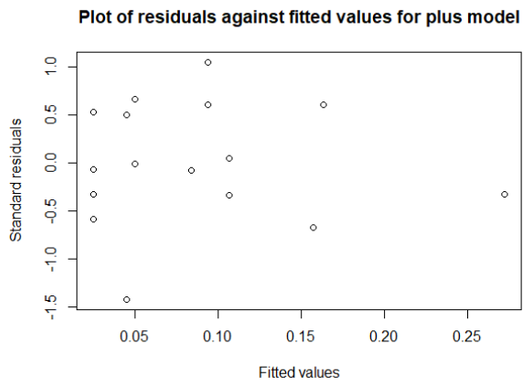

Figure 5: Pearson Residuals

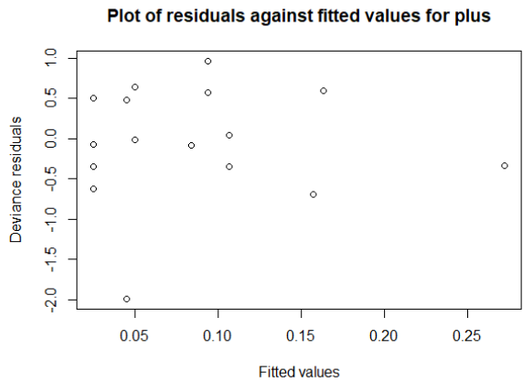

Figure 6: Deviance Residuals

residual value, depending on the figure we are looking at. There is also another outlier at the lower end of the fitted values scale, as well as the residuals scale. In all three figures this plot is shown to be low on both scales, which would suggest a positive correlation on it's own. However, since most of the values are concentrated in the top left corner, we are more likely to suggest a negative correlation between our two variables in this case.

Therefore, we can conclude that in our generalised linear model, the higher the residual deviance, the lower fitted values we expect to see.

4.1 Conclusion

Overall, this model suggests that the influence of a high blood pressure level and high cholesterol levels are likely to increase the risk of developing CHD. We find that having a blood pressure of over 147mm of mercury will increase ones likelihood of developing CHD in comparison to those with a blood pressure below this, as shown by the 0.64 coefficient in the model. Similarly, we see that the percent incidence of CHD is at its greatest for those with a serum cholesterol level above 219mg/100c, compared to those below, at the rate of 0.31 as given in the model. However, the likelihood of CHD development increases even more for people with a serum cholesterol level of over 259mg/100c, as we have a coefficient of 0.71 in this case on the model. These two factors both influence the incidence of CHD, but there is no evidence to suggest there is an interaction between the two, meaning they are independent of each other.

Despite these conclusions we have drawn being valid based on the fit of the model being accurate, we have some faults in our data that may alter the reliance we can put on our conclusions. In particular, the study was only carried out on US men, therefore we cannot be certain that we will see the same trend in incidence of CHD for women or people outside of the US. For instance, we may see a lower incidence of CHD for women with high blood pressure or cholesterol levels, as on average women are expected to live longer than men, meaning these factors may have less of an impact on the health of women.

A Appendices

R CODE AND R OUTPUT

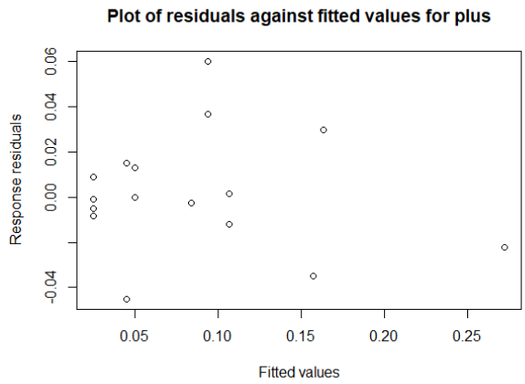

Figure 7: Response Residuals

```
chd$chol = as.factor(chd$chol)

figure1 = plot(chd$chol, chd$net, ylab= "Percent incidence of CHD",
               xlab="Cholesterol group",
               main="% incidence of CHD depending on cholesterol box plots")
```

Figure 8: Code for figure 1

```
chd$bp = as.factor(chd$bp)

figure3 = plot(chd$bp, chd$net, ylab= "Percent incidence of CHD",
               xlab="Blood pressure group",
               main="% incidence of CHD depending on blood pressure box plots")
```

Figure 9: Code for figure 2

```
library(dplyr)
chd$chol = as.numeric(chd$chol)
chd$bp = as.numeric(chd$bp)

g1 <- chd %>%
  filter(chol == "1")
g2 <- chd %>%
  filter(chol == "2")
g3 <- chd %>%
  filter(chol == "3")
g4 <- chd %>%
  filter(chol == "4")

figure_g1 = plot(g1$bp, g1$net, ylab= "Percent incidence of CHD",
                 xlab="Blood pressure group",
                 main="% incidence of CHD depending on cholesterol plot",
                 type="l", ylim=c(0,0.26))
show(figure_g1)
lines(g2$bp, g2$net, lty="dashed")
lines(g3$bp, g3$net, lty="dotted")
lines(g4$bp, g4$net, lty="dotdash")
labels <- c("Group 1", "Group 2", "Group 3", "Group 4")
legend("topleft", legend = labels, pch =NULL, lty=c(1,2,3,4) )
```

Figure 10: Code for figure 3

```
chd$bp = as.factor(chd$bp)
chd$chol = as.numeric(chd$chol)

b1 <- chd %>%
  filter(bp == "1")
b2 <- chd %>%
  filter(bp == "2")
b3 <- chd %>%
  filter(bp == "3")
b4 <- chd %>%
  filter(bp == "4")

figure_bp1 = plot(b1$chol, b1$net, ylab= "Percent incidence of CHD",
                  xlab="Cholesterol group",
                  main="% incidence of CHD depending on blood pressure plot",
                  ylim=c(0,0.26), type="l",
                  col = "blue")
show(figure_bp1)
lines(b2$chol, b2$net, col = "red")
lines(b3$chol, b3$net, col = "green")
lines(b4$chol, b4$net, col = "orange")
labels <- c("Group 1", "Group 2", "Group 3", "Group 4")
legend("topleft", legend = labels, col =c("blue", "red", "green", "orange"),
       lty=c(1,1,1,1))
```

Figure 11: Code for figure 4

```
chol.glm = glm(matrix(c(chd$y,chd$failures), ncol=2, byrow=F) ~
                 chd$chol, binomial("cloglog"))
deviance(chol.glm)
df.residual(chol.glm)

bp.glm = glm(matrix(c(chd$y,chd$failures), ncol=2, byrow=F) ~
                chd$bp, binomial("cloglog"))
deviance(bp.glm)
df.residual(bp.glm)
```

Figure 12: Code for Chol and BP models only

```
> deviance(chol.glm)
[1] 26.80498
> df.residual(chol.glm)
[1] 12
> bp.glm = glm(matrix(c(chd$y,chd$failures), ncol=2, byrow=F) ~
+                 chd$bp, binomial("cloglog"))
> deviance(bp.glm)
[1] 35.16305
> df.residual(bp.glm)
[1] 12
```

Figure 13: Deviance and degrees of freedom of Chol and BP only

```
probit.glm = glm(matrix(c(chd$y,chd$failures), ncol=2, byrow=F) ~
                   chd$bp + chd$chol, binomial("probit"))
deviance(probit.glm)
df.residual(probit.glm)
```

Figure 14: Probit model code

```
> deviance(probit.glm)
[1] 7.524783
> df.residual(probit.glm)
[1] 9
```

Figure 15: Probit model deviance and residuals

```
chd$bp.in.3 <- 1*(bp==1) + 1*(bp==2) + 2*(bp==3) + 3*(bp==4)
chd$bp.in.3 <- as.factor(chd$bp.in.3)
chd$chol.in.3 <- 1*(chol==1) + 1*(chol==2) + 2*(chol==3) + 3*(chol==4)
chd$chol.in.3 <- as.factor(chd$chol.in.3)

final.glm = glm(matrix(c(chd$y,chd$failures), ncol=2, byrow=F) ~
                  chd$bp.in.3 + chd$chol.in.3, binomial(link="probit"))
summary(final.glm)
```

Figure 16: GLM model code

```
Call:
glm(formula = matrix(c(chd$y, chd$failures), ncol = 2, byrow = F) ~
    chd$bp.in.3 + chd$chol.in.3, family = binomial(link = "probit"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9953  -0.3431  -0.0404   0.5216   0.9713

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)     -1.9563     0.1100 -17.782  < 2e-16 ***
chd$bp.in.32     0.2634     0.1404   1.876   0.0606 .
chd$bp.in.33     0.6383     0.1452   4.396 1.10e-05 ***
chd$chol.in.32   0.3125     0.1363   2.293   0.0218 *
chd$chol.in.33   0.7118     0.1381   5.153 2.56e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 58.7262  on 15  degrees of freedom
Residual deviance:  7.7296  on 11  degrees of freedom
AIC: 68.833

Number of Fisher Scoring iterations: 4
```

Figure 17: GLM model output

```
fv.final = fitted.values(final.glm)
res.final.1 = residuals(final.glm, type="pearson")
res.final.2 = residuals(final.glm, type="deviance")
res.final.3 = residuals(final.glm, type="response")

plot(x=fv.final, y=res.final.1, xlab="Fitted values",
     ylab="Standard residuals",
     main="Plot of residuals against fitted values for plus model")

plot(x=fv.final, y=res.final.2, xlab="Fitted values",
     ylab="Deviance residuals",
     main="Plot of residuals against fitted values for plus")
plot(x=fv.final, y=res.final.3, xlab="Fitted values",
     ylab="Response residuals",
     main="Plot of residuals against fitted values for plus")
```

Figure 18: Code for residual plots