# MATH5824 Assessed Practical

Name: Dominic Lee
Student ID: 201691989
Date of submission: 31/03/2023

Throughout, BP (or bp), CHOL (or chol) and CHD stand for blood pressure, cholestorol and coronary heart disease respectively.

## 1 Summary

In conclusion, I've found that a simpler way to look at how the probability of having CHD depends on BP and CHOL, rather than splitting people into 16 individual categories as in the reference table given to us, is to assign a general "risk level" from 2 to 6 to each cell (see table 3). Note that in this model BP normal and BP elevated categories are combined, and so are the normal and medium risk CHOL categories. This way we aren't thinking of CHD risk through CHOL and BP separately, but rather through a general risk level based on both. Given a person and their risk level, we work out the linear predictor, in this case -4.69332+ $0.63089x_1$ where $x_1$ is the risk level. Then we calculate that the probability of them having CHD is $\frac{e^{-4.69332+0.63089x_1}}{1+e^{-4.69332+0.63089x_1}}$.

The interpretation of this model is that moving up levels in both BP and CHOL (other than moving up one level from normal which doesn't seem to increase the risk substantially) levels increases the "logit" of the probability of having CHD ($\log(\frac{p}{1-p})$) by roughly the same amount, so it's the model is more parsimonious, i.e. there are fewer parameters to estimate, when we combine the CHD and BP level first, and consider this as a quantitative explanatory variable rather than a qualititative one.

## 2 Building the model

The model for this situation should be binomial since each man either has CHD or doesn't. We could use the Poisson transformation of the multinomial model in a 3 way table with a fixed total of 1325, but since the response variable is binary there is no need, and the data is setup more conveniently to use a binomial model.

Because BP and CHOL levels are known to be linked to CHD, knowing which cell someone falls in should affect how likely they will be to have CHD, so it makes intuitive sense that each cell should reflect a different $p$ value in the underlying distribution. This is supported by table 1, showing the percentage of men in that cell that had CHD. So the model might look like: $Y_i \sim Bin(M_i, p_i)$ where $Y = (Y_i)_{1\leq i\leq 16}$ is the $y$ column of the table, and $M = (M_i)_{1\leq i\leq 16}$ is the $m$ column, and $p_i$ are parameters we will estimate.

| CHOL | BP normal | BP elevated | BP stage I | BP stage II |
|---|---|---|---|---|
| Normal | 2.542372881 | 3.278688525 | 5.769230769 | 16 |
| Medium risk | 3.370786517 | 2.941176471 | 2.43902439 | 12.5 |
| High risk | 7.142857143 | 5.479452055 | 9.589041096 | 14.58333333 |
| Dangerous | 10.66666667 | 9.821428571 | 19.64285714 | 27.90697674 |

Table 1: A table of the percentages of men tested that had CHD

This means that since $f(y_i) = \binom{m_i}{y_i}(\frac{p_i}{1-p_i})^{y_i}(1-p_i)_i^m = \exp\{y_i \text{logit} p_i + m_i \log(1-p_i) + \log(\binom{m_i}{y_i})\}$, we have $\phi = 1, \theta_i = \text{logit} p_i, b(\theta_i) = m_i \log(1+e^{\theta_i}), c(y,\phi) = \log\binom{m_i}{y_i}$

For the link function, we could choose any of the 4 commonly used: logit, probit, cloglog, cauchit functions. We will have to experiment to see which is best, however it's worth bearing in mind that using the canonical link function logit will give us

desirable properties such as being able to choose sufficient statistics through which the log-likelihood function depends on, potentially decreasing the model complexity.

To begin with, we will assume there are 16 different values of $p_i$ corresponding to the 16 different rows in the table in the chd.txt file, and use the logit link function. I fitted this model and the summary is shown below:

```
Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.85988  -0.15141  -0.00509   0.21420   0.76171

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.2868     0.3242 -10.137  < 2e-16 ***
bp2          -0.1121     0.2880  -0.389 0.697236
bp3           0.5165     0.3144   1.643 0.100460
bp4           1.1917     0.3109   3.833 0.000127 ***
chol2        -0.1522     0.4252  -0.358 0.720419
chol3         0.5424     0.3280   1.654 0.098208 .
chol4         1.2123     0.3243   3.738 0.000185 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 52.0260  on 15  degrees of freedom
Residual deviance:  2.8017  on  9  degrees of freedom
AIC: 71.415

Number of Fisher Scoring iterations: 4
```

I also plotted a graph of predicted probabilities of people having CHD in each category and the actual fractions that had it to check our model was giving reasonable answers, and as can be seen from figure 1, the model seemed to be doing well.

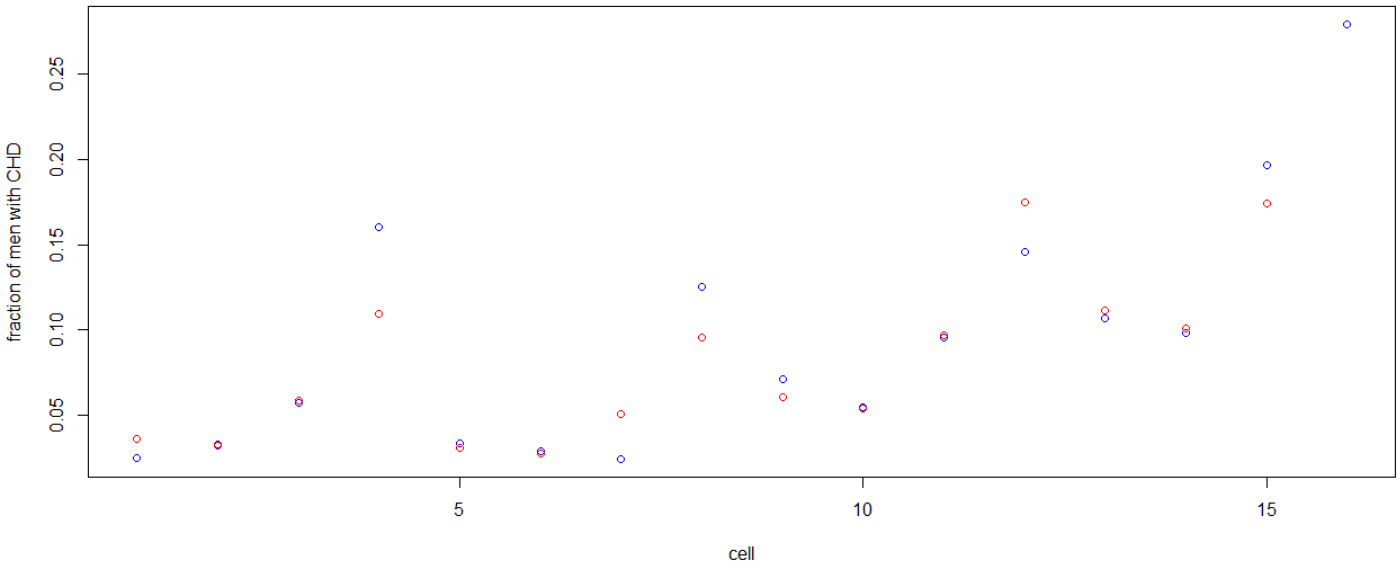


Figure 1: Scatterplot of fitted values of $p_i$ (red) and actual proportions of men with CHD (blue)

Note in this case $n = 16$, and since there are 7 D.O.F. (corresponding to the intercept and 3 additional for each factor) we can do a goodness of fit test with the $\chi^2_9$ test. The residual deviance is 2.8017 on 9 degrees of freedom, which is very small, being in the bottom 0.02835939 of the distribution. However, this model has a large number of parameters, and clearly we can remove some paramaters while still having a good fit.

To reduce the number of parameters in the model, we could combine factor levels which seem to have a similar effect, or change an ordered qualitative variable to a quantitative one if the output seems to depend linearly on its levels.

Looking at table 1, we can see that for 3/4 rows, the entry in the BP normal column is actually higher than the BP elevated column. The same is true for 3/4 columns and the Normal and Medium risk rows. This is contrary to what we know from the medical science we have, so it actually wouldn't make sense to make our model give a lower probability of having CHD for people with elevated BP than normal and the same for Medium risk and Normal which we can see is the case, since the bp2 and chol2 values are negative. Therefore I opted to make a change if our model isn't explainable it's likely to be based more off natural variance and is less likely to be useful for things such as predicting whether someone has CHD, so this suggests that for both, combining the 2 levels is a good idea.

Having done this, the new model output is:

```
Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.97750  -0.20493   0.00219   0.15582   0.86466

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.4128     0.2277 -14.989  < 2e-16 ***
bp2           0.5814     0.2686   2.164   0.0304 *
bp3           1.2550     0.2640   4.753 2.00e-06 ***
chol2         0.6010     0.2747   2.188   0.0287 *
chol3         1.2733     0.2705   4.707 2.52e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 52.0260  on 15  degrees of freedom
Residual deviance:  3.0845  on 11  degrees of freedom
AIC: 67.698

Number of Fisher Scoring iterations: 4
```

Now the residual deviance 3.0845 is in the bottom 0.01043866 of the $\chi^2_{11}$ distribution, which is an even lower number than before, meaning this is probably a good improvement, especially as it's simpler now.

We could still reduce the number of parameters further, by turning a factor explanatory variable into a quantitave one, thereby only having one parameter for the whole explanatory variable. Since the old coefficients contributing to the predictor for a factor on level 1, 2, 3 were $\beta_1, \beta_2, \beta_3$ respectively, and now they will be $\beta, 2\beta, 3\beta$ respectively for some $\beta$, it would make sense to replace the factor where $\beta_3 - \beta_2$ is closest to $\beta_2 - \beta_1$ since in the new model these 2 quantities will be identical, namely $\beta$. In fact here, we can treat $\beta_1$ as the intercept for both factors. From looking at the output of the model2 summary, we can see that (chol3-chol2)-(chol2-intercept) = -3.3415, and (bp3-bp2)-(bp2-bp1) = -3.3206. So this would tell us that neither follow a particularly linear pattern, however bp is a better candidate to choose.

Having done this, the new model is:

```
Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.02256  -0.16146  -0.04685   0.16578   0.89199

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -4.0425     0.2964 -13.640  < 2e-16 ***
bp            0.6226     0.1296   4.804 1.56e-06 ***
chol2         0.6017     0.2747   2.191   0.0285 *
chol3         1.2732     0.2705   4.707 2.52e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 52.0260  on 15  degrees of freedom
Residual deviance:  3.1154  on 12  degrees of freedom
AIC: 65.728

Number of Fisher Scoring iterations: 4
```

Again, this is a very low residual deviance. So I tried to improve the model by doing the same for chol:

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.0023  -0.1816  -0.0361   0.1929   0.9160

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -4.6981     0.3644 -12.894  < 2e-16 ***
bp            0.6232     0.1295   4.811 1.50e-06 ***
chol          0.6392     0.1349   4.738 2.16e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 52.0260  on 15  degrees of freedom
Residual deviance:  3.1399  on 13  degrees of freedom
AIC: 63.753

Number of Fisher Scoring iterations: 4
```

Again a very low residual deviance. Now it's in the bottom 0.002618878 of the $\chi^2_{13}$ distribution. Also for the first time all the coefficients now have *** next to them implying a high significance which is a good sign. I decided to make a plot of the residuals (fig. 2). There were no obvious patterns I could see, meaning our model still looked ok!

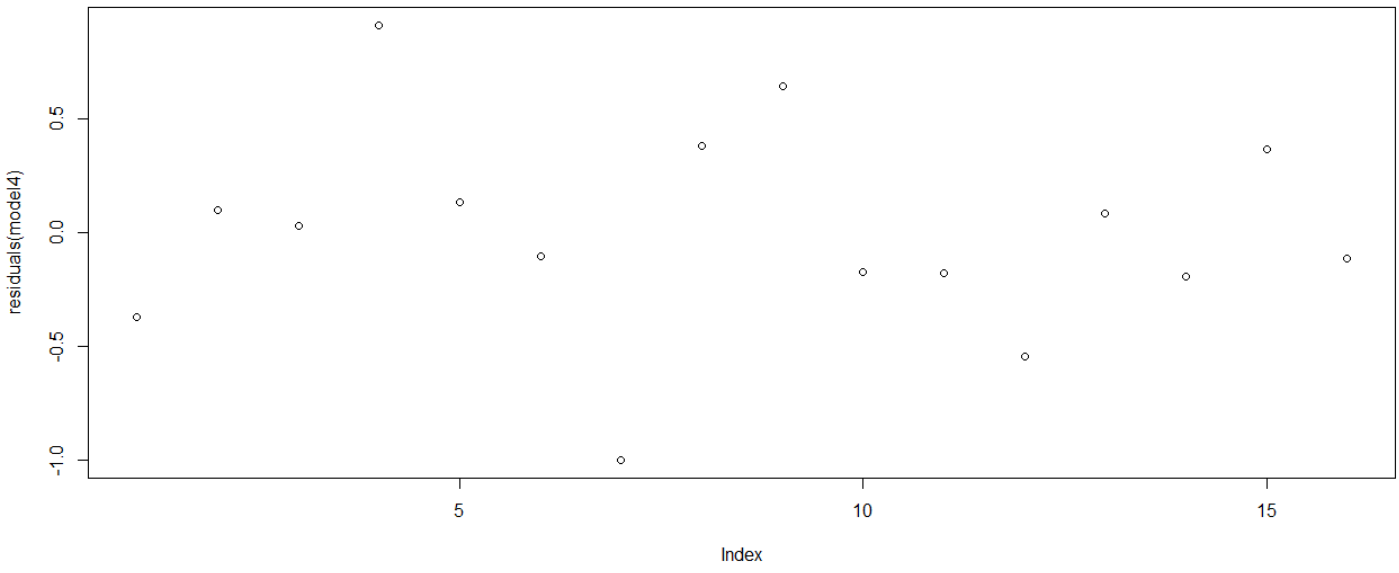


Figure 2: Scatterplot of residuals for model4

I also plotted another table of percentages with the new combined levels (table 2). Now it didn't make sense to me to combine any of the factor levels other than maybe CHOL1 and CHOL2, since there seemed to be a significant difference in percentages.

| CHOL | BP1 | BP2 | BP3 |
|---|---|---|---|
| CHOL1 | 3.0162413 | 4.3010753 | 14.2857143 |
| CHOL2 | 2.9166667 | 9.589041096 | 14.58333333 |
| CHOL3 | 10.1604278 | 19.64285714 | 27.90697674 |

Table 2: A table of the percentages of men tested that had CHD with combined factor levels

If at any point the models were nested, we could test the new models against the old one by using the fact that in 2 nested models with $r_1, r_2$ parameters resp.and $r_1 > r_2$ the deviances $D_1$ and $D_2$ satisfy $D_1 - D_2 \sim \chi^2_{r_1-r_2}$ asymptotically, however here they are not nested, we have just effectively applied conditions on the parameters from the original model (combining factor levels is equivalent to setting some parameters equal, and changing factors to quantitative variables is equivalent to setting some parameters equal to multiples of others.)

The next thing to do was check if the logit function could be significantly improved upon by another one. I decided to check the probit, cloglog and cauchit functions. The deviances were: 3.085, 3.237, 8.881 respectively, so only the probit function decreased the deviance. It only decreased it by 0.0549, which to me didn't justify losing the advantages that using the canonical link function gives, so I decided to stick with the logit function.

When I use the anova test, the output is:

```
 added sequentially (first to last)


     Df Deviance Resid. Df Resid. Dev
NULL                    15     52.026
bp    1   25.778        14     26.248
chol  1   23.108        13      3.140
```

This sharp rise from 3.140 to 26.248 and then again from 52.026 to me confirms that we do in fact need parameters corresponding to bp and chol in our model, along with the *** in the model summary above.

However I now noticed that the coefficients for bp and chol were very similar, so maybe I could actually combine the 2 explanatory variables into 1 by adding together the now numeric levels. When I did this the model output was:

```
Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.01802  -0.16693  -0.04974   0.18385   0.88413

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.69332    0.35930 -13.063  < 2e-16 ***
bpchol       0.63089    0.09015   6.998 2.59e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 52.0260  on 15  degrees of freedom
Residual deviance:  3.1468  on 14  degrees of freedom
AIC: 61.76

Number of Fisher Scoring iterations: 4
```

This model still looks fine, with the residual deviance barely any higher than before, and the plot of residuals still has no obvious patterns, so I decided it was an improvement. Now there is no way to reduce the number of paramters while keeping the dependence on bp and chol, since there are only 2 parameters.

Because of its simplicity and accuracy, I decided to stick with this model. So the final model has coefficients $\beta_1 = -4.69332, \beta_2 = 0.63089$, so the linear predictor is $-4.69332 + 0.63089x_1$, where $x_1$ is a number depending on the BP and CHOL level of the person (see table 3).

Using (5.4) from the notes the final density for $y$ is therefore: $Y \sim Bin(m, p)$ where $p = \frac{\exp(-4.69332+0.63089x_1)}{1+\exp(-4.69332+0.63089x_1)}$, $Y$ is the

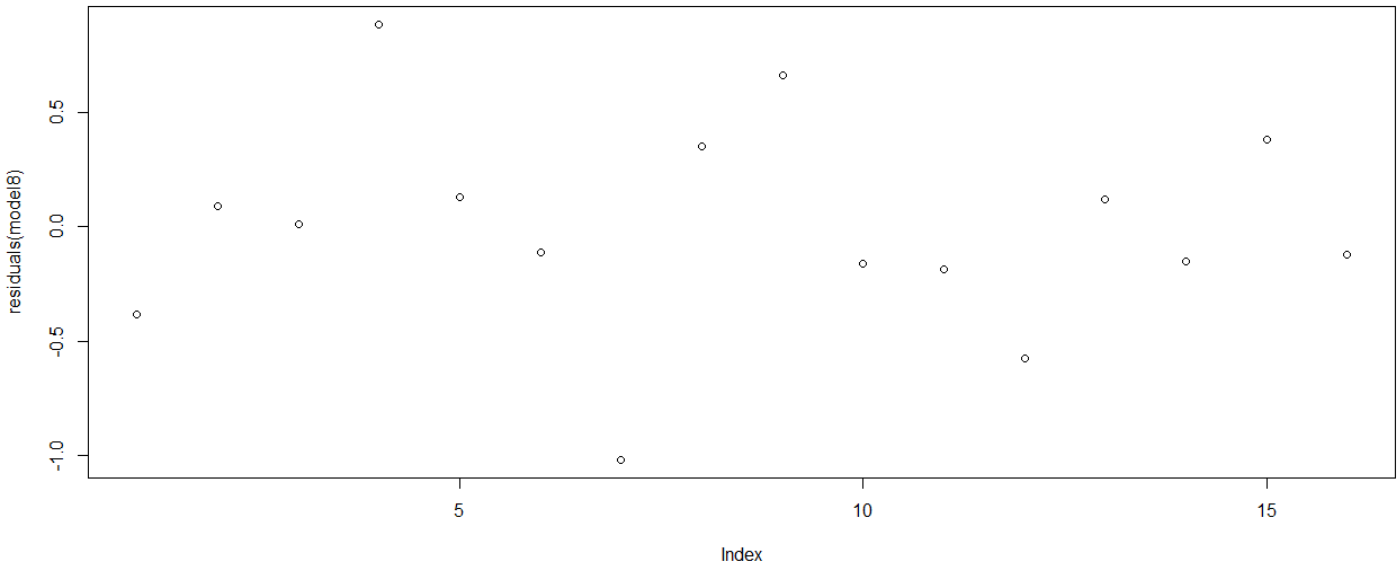


Figure 3: Scatterplot of residuals for model8

| CHOL | BP normal | BP elevated | BP stage I | BP stage II |
|---|---|---|---|---|
| Normal | 2 | 2 | 3 | 4 |
| Medium risk | 2 | 2 | 3 | 4 |
| High risk | 3 | 3 | 4 | 5 |
| Dangerous | 4 | 4 | 5 | 6 |

Table 3: A table of the value of $x_1$ for a person corresponding to the cell of the table they fall in

number of men who have CHD out of a sample of $m$ men all of who fall in a cell of the table with the same $x_1$ value, and therefore the link function is the canoncial logit function, $\text{logit}(p) = \log(\frac{p}{1-p}) = -4.69332 + 0.63089x_1$.

As a final check, this means the probability of the people at the highest risk having CHD is 0.2874133695, and the people at lowest risk have a 0.03132417034 chance of having CHD, which seem to be sensible probabilities!

# 3 Appendix

Here is all the R code I used in order:

```
data = read.csv("C:\\Users\\xboxl\\OneDrive\\GLAM practical\\chd.csv")
library(forcats)

data$chol = as.factor(data$chol)
data$bp = as.factor(data$bp)
data[, 5] = data[, 2] - data[, 1]
names(data)[5] = "m-y"

model1 = glm(cbind(y,m-y) ~ bp + chol, data = data, family = binomial)
summary(model1)

data[, 6] = data[, 1] / data[, 2]
names(data)[6] = "fraction"

plot(x, data$fraction, col = "blue", ylab = "fraction of men with CHD", xlab = "cell")
points(x, fitted.values(model), col = "red")
```

```
pchisq(2.8017,df=9)

data$bp = fct_collapse(data$bp, "1"= c(1,2), "2" = c(3), "3" = c(4))
data$chol = fct_collapse(data$chol, "1"= c(1,2), "2" = c(3), "3" = c(4))

model2 = glm(cbind(y,m-y) ~ bp + chol, data = data, family = binomial)
summary(model2)

data$bp = as.numeric(data$bp)
model3 = glm(cbind(y,m-y) ~ bp + chol, data = data, family = binomial)
summary(model3)

data$chol = as.numeric(data$chol)
model4 = glm(cbind(y,m-y) ~ bp + chol, data = data, family = binomial)
summary(model4)

plot(residuals(model4))

model5 = glm(cbind(y,m-y) ~ bp + chol, data = data, family = binomial(probit))
model5
model6 = glm(cbind(y,m-y) ~ bp + chol, data = data, family = binomial(cloglog))
model6
model7 = glm(cbind(y,m-y) ~ bp + chol, data = data, family = binomial(cauchit))
model7

anova(model4)

data$bpchol = data$bp+data$chol
model8 = glm(cbind(y,m-y) ~ bpchol, data = data, family = binomial)
summary(model8)
```