# The Framingham Heart Disease Data

Arden Miller

## Executive Summary

The analysis of the Framingham data clearly shows that the probability of developing CHD is higher for males than for females. It is also higher for people in the $50-62$ age group than those in the $30-49$ age group, and increases as the level of serum cholesterol increases. The lowest estimated probability, 0.01, occurs for females with low cholesterol who are in the 30-49 age group, whereas the highest estimated probability, 0.25, occurs for males with high cholesterol who are from 50-62.

## The Framingham Data

Framingham is an industrial town located approximately 30 km from Boston. In 1948 a study was begun with the aim of identifying factors that are related to the occurrence of coronary heart disease (CHD). At the start of the study, a large proportion of the town's inhabitants were examined for the presence of CHD. Measurements were made on a number of potential risk factors. The individuals who were found to be free of CHD at that time were followed up for twelve years and those who developed CHD during that period were identified. The following dataset was extracted from that data and relates the proportions developing CHD to the initial serum cholesterol level (mg per 100 ml) of individuals, age, and sex.

| | | Serum Cholesterol Level | | | |
|---|---|---|---|---|---|
| Sex | Age | $< 190$ | $190-219$ | $220-249$ | $\geq 250$ |
| Male | $30-49$ | 13/340 | 18/408 | 40/421 | 57/362 |
| | $50-62$ | 13/123 | 33/176 | 35/174 | 49/183 |
| Female | $30-49$ | 6/542 | 5/552 | 10/412 | 18/357 |
| | $50-62$ | 9/58 | 12/135 | 21/218 | 48/395 |

| | | Serum Cholesterol Level | | | |
|---|---|---|---|---|---|
| Sex | Age | $< 190$ | $190 - 219$ | $220 - 249$ | $\geq 250$ |
| Male | $30 - 49$ | 0.04 | 0.04 | 0.09 | 0.16 |
| | | (0.02, 0.07) | (0.02, 0.07) | (0.07, 0.13) | (0.12, 0.21) |
| | $50 - 62$ | 0.14 | 0.18 | 0.20 | 0.25 |
| | | (0.09, 0.23) | (0.13, 0.25) | (0.14, 0.26) | (0.19, 0.32) |
| Female | $30 - 49$ | 0.010 | 0.011 | 0.03 | 0.05 |
| | | (0.006, 0.019) | (0.006, 0.020) | (0.02, 0.04) | (0.03, 0.07) |
| | $50 - 62$ | 0.07 | 0.09 | 0.10 | 0.13 |
| | | (0.04, 0.13) | (0.06, 0.14) | (0.07, 0.14) | (0.10, 0.17) |

Table 1: Estimated probabilities of coronary heart disease (CHD).

# Analysis of the Data

This data was analysed using logistic regression to investigate how gender, age and serum cholesterol level affect the probability of developing CHD. The estimated probabilities for the different combinations of sex, cholesterol level, and age group are given in Table 1. An interval of feasible values for the true probability is given in each case. The estimated probabilities range from 0.01 for females with low cholesterol who are in the 30-49 age group, to 0.25 for males with high cholesterol who are from 50-62. In general the probability of CHD is higher for males than for females, is higher for people in the $50 - 62$ age group than those in the $30 - 49$ age group, and increases as the level of serum cholesterol increases.

The size of effect that gender has on the probability of developing CHD depends on the age group. For individuals in the $30 - 49$ age group, the estimated probability for males is from 3 to 4 times the probability for females for each level of serum cholesterol. For the $50 - 62$ age group the probabilities for males are approximately 2 times those for females.

The size of effect that serum cholesterol level has on the probability of developing CHD also depends on age group. For individuals in the $30 - 49$ age group, the estimated probability for the highest serum cholesterol level ($\geq 250$) is from 4 to 5 times that for the lowest serum cholesterol level ($< 190$). Whereas for the $50 - 62$ age group, the estimated probability for the highest serum cholesterol level is approximately 2 times that for the lowest serum cholesterol level.

There are 2 unusual observations in the data. The observed proportion for males in the $< 190$ level of serum cholesterol and the 50-62 age group is much lower than the model predicts and the observed proportion for females in the $< 190$ level of serum cholesterol and the 50-62 age group is much higher than predicted. These observations should be checked to make sure they have been correctly recorded. These points only have a large impact for the estimated probabilities for males and females in the $< 190$ level of serum cholesterol and the 50-62 age groups. The other estimated probabilities stay much the same if either or both of these observations are removed.

# Statistical Appendix

I used the step.glm function to select a suitable model. It identified the model that contained the 3 main effects plus the age:serum and sex:age interactions. The analysis of deviance table for this model is:

```
Analysis of Deviance Table
Binomial model
Response: num/tot
Terms added sequentially (first to last)
          Df Deviance Resid. Df Resid. Dev    Pr(Chi)
     NULL                    15   301.2147
      sex  1  79.7103         14   221.5044 0.00000000
      age  1 133.1238         13    88.3806 0.00000000
    serum  3  61.7984         10    26.5822 0.00000000
age:serum  3  13.3384          7    13.2438 0.00395912
  sex:age  1   5.6590          6     7.5847 0.01736560
```

A goodness of fit test for this model has a P-value of 0.27, $\Pr(\chi^2_6 \geq 7.5847$, so this model provides an adequate description of this data.

The fitted probabilities in Table 1 were calculated using this model. The intervals were produced by first creating 95% confidence intervals for logit $\pi$ and then using the logistic transformation to convert these to intervals for $\pi$.

Diagnostic plots for the fitted model are given in Figure 1. These indicate two unusual points (5 and 13). These points both correspond to the lowest level of serum cholesterol and the 50-62 age group: observation 5 corresponds to males and observation 13 to females. The deviance residuals indicate that the observed proportion for 13 is higher that expected while that for point 5 is lower than expected. In fact, this is the only combination of age group and serum cholesterol level for which the observed proportion for females is higher than that for males. These points should be investigated to make sure that they are the correct.

The output from influence.measures indicates that dropping either of these points will have a big impact on the fitted coefficients for both of the interactions. In both cases, the interactions will become less important if the point is dropped. A comparison of fitted values indicates that dropping one or both of these observations has very little effect on the estimated probabilities for the other cells.

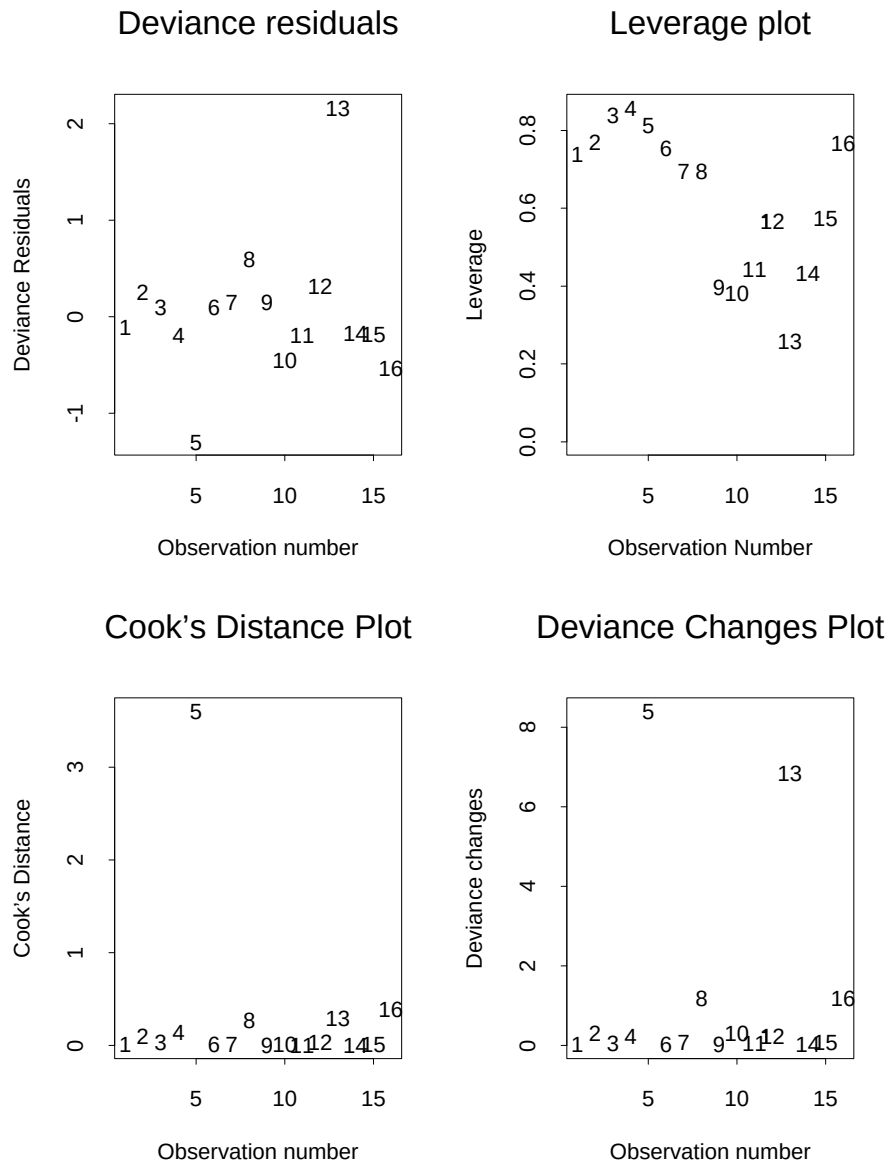


Figure 1: Diagnostic Plots for the Fitted Model

# Discussion for Odds

The discussion could have been done in terms of estimated odds instead of estimated probabilities. In this case, a table of the estimated odds would be

| | | Serum Cholesterol Level | | | |
|---|---|---|---|---|---|
| Sex | Age | $< 190$ | $190 - 219$ | $220 - 249$ | $\geq 250$ |
| Male | $30 - 49$ | 0.04 | 0.04 | 0.10 | 0.19 |
| | $50 - 62$ | 0.17 | 0.23 | 0.24 | 0.33 |
| Female | $30 - 49$ | 0.010 | 0.011 | 0.03 | 0.05 |
| | $50 - 62$ | 0.08 | 0.10 | 0.11 | 0.15 |

The odds of CHD is higher for males than for females, is higher for people in the $50 - 62$ age group than those in the $30 - 49$ age group, and increases as the level of serum cholesterol increases.

The effect of gender on the odds of developing CHD depends on the age group. For individuals in the $30 - 49$ age group, the estimated probability for males is close to 4 times the odds for females for each level of serum cholesterol. For the $50 - 62$ age group the odds for males are just over 2 times those for females.

The effect of serum cholesterol level on the odds of developing CHD also depends on age group. For individuals in the $30 - 49$ age group, the estimated probability for the highest serum cholesterol level ($\geq 250$) is approximately 5 times that for the lowest serum cholesterol level ($< 190$). Whereas for the $50 - 62$ age group, the estimated odds for the highest serum cholesterol level is approximately 2 times that for the lowest serum cholesterol level. This is true for both genders.

# 475.330 Assignment 4: Marking Guide

This assignment asks the students to use logistic regression to analyse the Framingham data. The relationship between the probability of CHD and three explanatory variables is explored. The assignment is worth a total of 20 marks.

The assignment should consist of two parts: a report that is understandable by a non-statistics major and a statistical appendix that explains their analysis.

### The Report

The report should contain the following:

1. An executive summary. This should be a short paragraph that summarises their report. In this case it should summarise their findings about the way the 3 explanatory variables impact the probability of CHD.
2. A brief description of the data. A plot is not necessary.
3. They should clearly explain how the probability of CHD or the odds of CHD is related to the explanatory variables. They should discuss how each of the explanatory variables affects $\pi$. They should also discuss the effects of the age:sex and age:serum interactions. It is not necessary to present a fitted model but they can if they like.
4. They should present the values of $\pi$ that are produced by their model (or the estimated odds) for different combinations of the explanatory variables.
5. They need to identify points 5 and 13 as being unusual and discuss their impact on the estimated probabilities.

### Statistical Appendix

The statistical appendix should contain:

1. A justification of the model they selected. I think the only suitable model is the one identified in the model answers.
2. They need to include diagnostic plots and a discussion of the impact of points 5 and 13.

**Allotment of marks**

- Give 5 marks for presentation. The writing should be clear and concise. The report and the statistical appendix should include all the parts listed above. In the report, look for explanations that would be understandable to non-statistics students. Graphs should have informative captions. The statistical appendix should give a explanation of their analysis.
- Give 3 marks for identifying a suitable model and justifying this model in the statistical appendix.
- Give 7 marks for explaining how the 3 factors affect the occurrence of CHD. They do not need to give the model in the report but they do need to explain what it indicates about the explanatory variables. Make sure that they identify the consequences of the 2 active interactions.
- Give 2 marks for identifying points 5 and 13 as being influential.
- Give 3 marks for discussing the influential points. They should identify 13 as representing a large observed proportion than expected and 5 as a smaller observed proportion than expected. They should also discuss the impact of these points on the fitted model.

**Note:**

- Include short comments indicating why a student has lost marks.