

APPENDIX 2 | WHAT IS LOGISTIC REGRESSION?

Most people are familiar with the idea of expressing the chances of winning a bet in terms of its 'odds'. When the chance of winning is either very small (close to zero) or very large (close to one) then the odds of winning are approximately equal to the actual chance of winning.

Logistic Regression Model

In this report logistic regression models are used to describe how the odds that a person has good health vary with their other characteristics. The variables that measure those characteristics and appear in a logistic regression model are called 'predictor variables' because they are used to predict good health.

A logistic regression model includes an equation with the odds of good health on the left hand side and an expression involving the predictor variables on the right hand side. The model equation can be used to estimate the odds that a person with particular characteristics has good health. An 'odds ratio' can then be used to compare the health effects of two different sets of characteristics. An odds ratio is simply the odds of good health for one set of characteristics divided by the odds for a second set. If an odds ratio is less than 1.00 then the first set of characteristics is less likely to be associated with good health than the second set; if the odds ratio is greater than 1.00 then the first set of characteristics is more likely to be associated with good health.

A summary of the model equation is all that is usually presented. For example, the tables in Chapter 8 summarise the final model equations for the five indicators of good health considered in this report. They list the predictor variables that are in the final model equation and give the odds ratios associated with these variables. For each variable a 'base category' has been labelled, and the odds ratio for every other category of the variable is just the odds of good health for that category divided by the odds for the base category. For example, Table 8.1.1

summarizes the final model equation for excellent/very good general health. The local services score is in the final model equation, meaning that it has an (independent) effect on the odds of having excellent/very good general health that is not explained by the other study factors. As explained in Section 4.2.4 the score was grouped into three categories - 'Low', 'Middle' and 'High' – and in Table 8.1.1 this last category (reflecting the most positive view of local services) has been chosen as the base category. The odds ratio for a low local services score (compared to a high local services score) is given in Table 8.1.1 as odds ratio = 0.71. This means that, after adjusting for all the other variables in the final model equation, the odds that a respondent with a low local services score has excellent/very good general health is 0.71 times the odds for a respondent with a high local services score. Compared to a respondent with a high local services score, then, a respondent with a low local services score is approximately 30% less likely to have excellent/very good general health.

As we discussed in Section 4.3.1, these observed odds are not the true odds because they are based on a sample of people rather than the whole population. A summary of a model equation, then, also includes p-values which test statistical significance. These p-values are used to assess whether or not the observed odds ratio could simply be a chance deviation from a true value of 1.00 (corresponding to no underlying difference in the odds of good health). A small p-value suggests that the true odds ratio is statistically different from 1.00. For example; Table 8.1.1 gives the p-value associated with the odds ratio for a low local services score (compared to a high local services) as $p=0.0164$. This is generally taken to be a 'small' value (see Section 4.3.1): the odds a person with a low local services score has excellent/very good general health, is considered to be significantly lower than the odds for a person with a high local services score.

What's a 'Good' Model?

In most cases, a number of different model equations could be selected and a notion of a 'good' model is required.

For each respondent in the sample with good health, a model equation can be used to estimate the odds that a person with their characteristics has good health. Similarly, for each respondent in the sample who does not have good health, a model equation can be used to estimate the odds that a person with their characteristics does not have good health. Multiplying all the estimated odds of having (or not having) good health by one another, the 'likelihood' of observing the combined health experience of the whole sample can then be estimated.

A good' model equation is one that gives a high estimated likelihood of the health experience of the whole sample that was actually observed (it was recorded by the respondents when they completed the questionnaire). A 'good' equation is the one that, using the characteristics of the respondents, closely predicts their actual health experiences. It gives a good 'fit' with the observations.