Appendix H: Logistic Regression

In this appendix we shall be looking at a form of statistical modelling known as logistic regression. Like least squares regression it is used to test hypotheses and to explore data for associations between variables. However, whereas least squares regression is suitable for modelling data measured at interval or ratio level, logistic regression is used for modelling data where the dependent variable is dichotomous. By the end, you should:

- Understand how to construct and interpret a simple, logistic regression model
- Know how to check 'residuals' for evidence of linearity, homoscedacity and collinearity
- Be able to build a multiple regression model, including the use of 'dummy variables' as required

In Chapter 10 we saw that ordinary least squares regression requires a continuous dependent variable measured at the interval or ratio level. Independent variables too should be interval/ratio measures, but we saw that dichotomous categorical variables could be entered in a regression model as 'dummy variables'.

Suppose, though, that we have a series of interval/ratio independent variables, and we want to look at their influence on a dichotomous dependent variable. Can we simply use a dummy variable for our dependent variable?

The answer is that we cannot. Instead, we have to turn to a rather different modelling procedure called logistic regression.

WHY CAN'T WE USE A DICHOTOMOUS DEPENDENT VARIABLE IN ORDINARY LEAST SQUARES REGRESSION?

Calculating probabilities

Let us imagine that we wanted to use three independent variables to predict whether or not people smoke:

- Their self-reported levels of stress (Hypothesis 5 suggests that people who are more stressed may be more likely to smoke);

- Whether their parents – and particularly their mothers – smoked when they were growing up (we have rejected Hypothesis 1, which predicted that parental smoking would influence people's own smoking behaviour, but in Chapter 8 we found some tentative evidence that maternal smoking may be having an influence on their sons' later smoking behaviour);

- The proportion of their family and friends that currently smoke (the hypothesis of an association between respondent smoking behaviour and that of their peer group is hypothesis 2, which was supported by evidence analyzed in Chapter 8).

The variable in the Smoking Survey data set that indicates whether a respondent has smoked in the last week is var0003, a nominal level categorical variable. What we need to calculate is the chances of somebody ending up in one or other of these categories. What is the probability of being a smoker (coded 1) or a non-smoker (coded 0), and how much does this change when we take into account their stress level, their mother's smoking behaviour and the smoking patterns of their friends?

Given that 48% of our sample had smoked in the last week, and 52% had not, the predicted probability of being a smoker is obviously 0.48. This can be written as

$P(Y=1) = 0.48.$

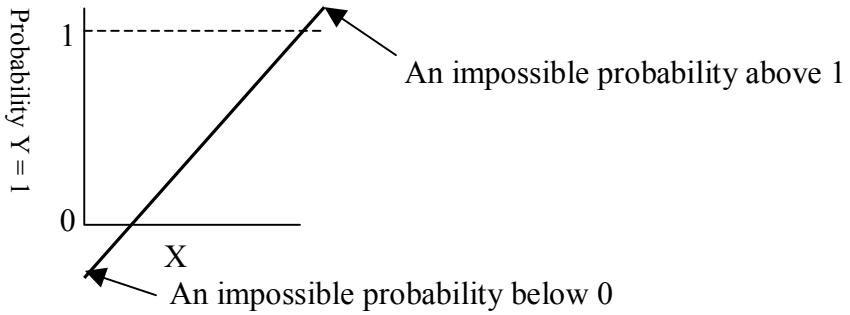
The predicted probability of not having smoked in the last week is simply 1 minus the probability of having smoked. In our example this is (1-0.48) = 0.52 and would be written as $P(Y=0) = 0.52$.

Creating impossible probabilities

Were we to set about examining how the probability of smoking varies according to values of the independent variable using ordinary least squares regression (OLS) we would hit problems. This is because when OLS is used to predict probabilities for a dichotomous dependent variable the regression equation can produce impossible probabilities either above 1 or below 0. This is the key reason why linear regression should not be used when the dependent variable is dichotomous.

The graph in Figure H.1 illustrates the problem, for while the maximum and minimum possible values of the dependent variable (Y) are 1 and 0 respectively, the best fitting line predicts values of Y based on the dependent variable (X) above and below what is possible.

Figure H.1 A graph illustrating how impossible values of a dichotomous dependent variable can be predicted in linear regression



Suppose we ran a multiple regression with smoking/non-smoking as a dichotomous dependent variable, and we obtained the following regression coefficients (the coefficients are purely illustrative):

- Constant = 0.893
- Stress = 0.023
- Parents' smoking history = 0.018
- Friends/family smoke = -0.317.

Based on these figures, a person with a stress level of zero, where neither parent smoked and where no family or friends smoke, would have a predicted probability of having smoked in the last week (Y) of:

$$Y = 0.893 + (0 \times 0.023) + (1 \times 0.018) + (4 \times -0.259) = -0.125$$

But this is clearly an impossible probability, for probabilities can never fall below 0.

We therefore need a procedure that overcomes this problem. This is precisely what logistic regression achieves.

THE LOG ODDS TRANSFORMATION IN LOGISTIC REGRESSION

The idea of logistic regression is to find an equation similar to that used in linear regression to predict the probability of cases falling into one category as opposed to another (e.g. the probability of being a smoker compared to the probability of being a non-smoker), but to ensure that probabilities cannot be predicted below 0 or above 1.

To ensure that impossible probabilities cannot be predicted, logistic regression makes two transformations to the dependent variable (Y).

Transforming the probability into odds.

The first transformation is to substitute the probabilities for odds. The odds that the dependent variable will equal 1 (i.e. that a particular respondent will be a smoker) are:

Probability that Y = 1 divided by the probability that Y ≠ 0.

This can be written as:

$$\text{Odds } (Y=1) = \frac{P(Y=1)}{[1 - P(Y=1)]}$$

As we saw above, the probability that an individual smokes is 0.48 (which means that the probability that s/he does not smoke is 0.52). The odds of their being a smoker (Y = 1) are therefore (0.48/[1 – 0.48]) = 0.92.

Unlike probabilities, which cannot exceed 1, odds have no theoretical maximum value. This means that transforming the probability of Y= 1 by taking the odds stretches the highest possible value of Y to infinity. As Table H.1 shows, as the probability that Y = 1 moves towards 1, the odds become positive and increasingly large – stretching into infinity.

Table H.1: The relationship between high probabilities and the odds

Probability that Y = 1	Odds that Y = 1
0.9	9
0.99	99
0.999	999
1	∞ (infinity)

The fact that odds have no maximum value deals neatly with the problem of finding impossibly high values of the probability of Y = 1, for no matter how high the odds, they can never become impossible.

But this has only resolved part of our problem, for like probabilities, odds do have a theoretical *minimum* value of zero (a zero probability that Y = 1 translates into odds of 0/1 = 0). This therefore still leaves open the possibility that the regression equation will predict impossibly low values, below zero.

Transforming the odds by logging

To deal with this problem, the second transformation is to take the natural logarithm of the odds that Y = 1. This is referred to as *logit (Y)*. If we use 'ln' to stand for natural log, then the equation for logit (Y) is:

$$\ln\left[\frac{Y}{1-Y}\right]$$

Two types of logs

It is important to be aware of the fact that there are two types of logarithms that are used by mathematicians and statisticians. The most frequently used is known as the *common logarithm* or *base 10 logarithm*. This is the one we referred to in Chapter 7

when we transformed variables by logging them. It corresponds to the index of the number to base 10 (e.g. the common log of 100 is 2, because $10^2 = 100$).

In the remainder of this appendix, however, we will be using an alternative logarithm known as the *natural logarithm*. Unlike the common logarithm, which expresses numbers to base 10, the natural logarithm operates with a base which has a value of 2.72… (the number runs to an infinite number of decimal places). The natural log of 2.72 = 1, the natural log of $2.72^2 = 2$, and so on.

This transformation stretches the lower values of the odds that Y = 1 so that the linear equation does not predict impossibly low values. As Table H.2 shows, as the odds decrease from 1 to 0, the logit value becomes negative and increasingly large (there is no upper maximum value). Therefore, as the odds increase from 1 to ∞, the logit value becomes negative and increasingly large.

Table H.2: The relationship between low odds and the logit

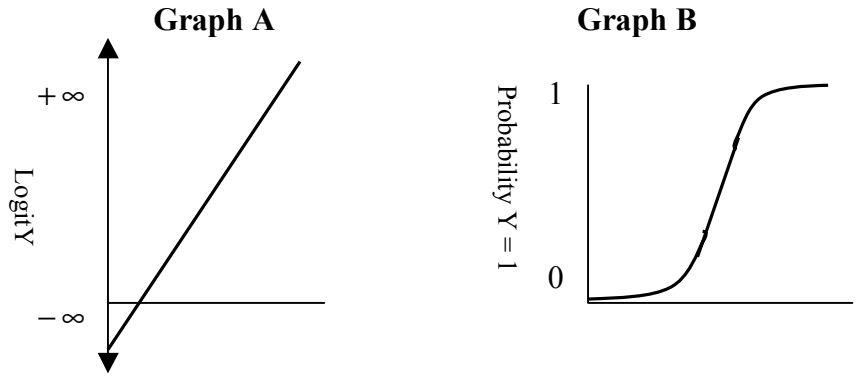
Odds that Y = 1	Log odds that Y = 1 or logit (Y)
0.1	-2.3
0.01	-4.6
0.001	-6.9
0	∞ (infinity)

Taken together, transforming the probability that Y = 1 into odds, and then taking the natural log of the odds, ensures that the transformed variable, logit (Y), has maximum and minimum values that can never be exceeded. The values of logit Y range from plus infinity to minus infinity. This means that when the best fitting line is calculated, it cannot produce impossible values of Y since the value of logit Y has no maximum or minimum values. This is illustrated in Figure H.2, Graph A.

The clever thing about all this is that, after the best fitting model has been estimated, we can convert the transformed variable back into probabilities, but the predicted probability will never produce impossible results below 0 or above 1. Figure H.2 illustrates this, for Graph B is the same model as that in Graph A except that the dependent variable has been converted from logit (Y) to the probability that Y=1. Because a probability of 0 or 1 in Graph A equates to the log of odds of minus infinity or plus infinity in Graph B, and since the regression line can never reach infinity,

when logit (Y) is converted back into probabilities, the best fitting line can never reach 1 or 0.

Figure H.2: Graphs showing a hypothetical best fitting line when the dependent variable is logit (Y) and when the dependent variable has been converted back into probabilities



The logistic regression equation

By transforming the dependent variable into the natural logarithm of the odds that the dependent variable = 1, we have ensured that the familiar linear regression equation can be retained:

$$\text{Logit}(Y) = a + b_1X_1 + b_2X_2 + b_3X_{3\ldots}$$

This produces a relationship similar to that in linear regression except that now each one-unit change in the independent variable(s) is associated with a change in the log odds of Y = 1 rather than a direct change in the values of the dependent variable. As in least squares regression, the equation can be used to predict values of logit (Y) for different specific values of the independent variables.

Converting back to probabilities

An obvious problem with trying to predict values of logit (Y), however, is that the logit coefficients in logistic regression do not have a meaning that is easily

interpreted. For this reason, we convert values of logit (Y) back to the more meaningful values of odds and probabilities.

To obtain the odds that Y = 1 we 'unlog' logit (Y). This is done by taking the anti-log (or *exponent* written as 'e'). The equation for this is:

$$\text{Odds}(Y=1) = e^{a + bX}$$

To fully solve the equation so that we can return to the probability that Y = 1, we must also reverse the calculation which turned the probability into odds:

The probability that $Y=1 = e^{a + bX}$ divided by $1+ e^{a + bX}$

An example

To see how this works, let us calculate the probability that people smoke, from their reported stress level. Do not worry about how we obtained the coefficients for the moment – simply look at how the linear equation works, and how we can move between logit(Y), the odds and probabilities.

A respondent who reports a very low level of stress (for example, a score of 1) will have an estimated probability of being a smoker of:

$$\text{Logit}(Y) = a + bX$$

With a constant (a) of –0,8987 and a coefficient for stress of 0.1638, this translates as:

$$\text{Logit}(Y) = -0.8987 + (1 \times 0.1638) = -0.735.$$

Therefore the odds of smoking will be:

$$\text{Odds}(\text{smoker}=1) = e^{-0.735} = 0.48$$

This can be interpreted as saying that respondents reporting very low levels of stress are about half as likely to smoke as not smoke. The probability that they smoke will be:

Probability of smoking = odds of smoking ÷(1 + odds of smoking) = 0.33 or 33%.

Therefore, the probability of smoking if the reported level of stress is very low, is just 33 per cent. On the other hand, if the respondent reported a very high level of stress (for example, a stress score of 10), he or she will have an estimated probability of smoking of:

Logit (smoker) = -0.8987 + (10 x 0.1638) = 0.738.
Odds(smoker=1) = $e^{0.738}$ = 2.09

This indicates that the odds of being a smoker are just over twice as high as those of not being a smoker if the respondent reported a very high level of stress. From this we can calculate the probability:

Probability of smoking = odds of smoking ÷(1 + odds of smoking) = 0.68 or 68%.

Our findings therefore suggest that a respondent with a high level of stress has a 68 per cent probability of being a smoker, as compared with a 33 per cent probability if they report a low level of stress.

Of course, it would be dangerous to assume from these results that stress causes a greater likelihood of smoking. It may be, for example, that the relationship between high levels of stress and taking up smoking is explained by some other variable. This is why, just as with linear regression, it is important to include all relevant independent variables in the analysis.

In logistic regression the principle behind including multiple independent variables in a model is exactly the same as in linear regression. Independent variables can be

measured at the ratio, interval, ordinal or nominal level (the latter two should be transformed into dummy variables) and the equation that is produced is linear in form:

$$\text{Logit}(Y) = a + b_1X_1 + b_2X_2 + b_3X_{3\ldots}$$

ESTIMATING THE COEFFICIENTS AND INTERPRETING A LOGISTIC REGRESSION MODEL

So far we have examined the logistic equation without worrying about how the coefficients are calculated.

In ordinary least squares regression, the coefficients (a and b) are calculated through a formula that attempts to minimise the sum of squared distances of the data points to the regression line. However, this method cannot be applied in logistic regression, precisely because we have ensured that the dependent variable may have infinitely large or small values. This means that a best fitting line can never be drawn, for such a line would go on forever, and the computer could never complete the plotting of all the possible values of Y against all the values of X.

Therefore, instead of ordinary least squares, an algorithm is used (an algorithm is a mathematical process designed to solve a problem). This is known as *maximum likelihood estimation*, and it is used to estimate the logit coefficients.

Rather than calculating coefficients that minimize the sum of squared errors (as in linear regression), maximum likelihood estimation in logistic regression calculates coefficients that maximize the *log likelihood*. The log likelihood reflects how likely it is (the odds) that the observed values of the dependent variable may be predicted from the observed values of the independent variables.

The algorithm starts with an initial 'guess' of what the logit coefficients could be, and it then determines the direction and size of change which is needed to improve the likelihood that the observed values of the dependent variable can successfully be predicted from the observed values of the independent variables (i.e. the log likelihood). The initial guess is usually poor (as indicated by the residuals), so a more

informed estimate is made, and this process is repeated, refining the estimates each time, until the log likelihood can no longer be improved. The estimated coefficients are then placed into the equation.

Interpreting a logistic regression model

The interpretation of a logistic regression is relatively straightforward and has many analogies to linear regression.

For our example, we have taken as our dependent variable var00003 (whether the respondent smoked in the last week), and for our independent variables we shall include self-reported stress (var00029), the proportion of family and friends who smoke (var00016) and whether the respondent's mother smoked when the respondent was a child (the derived variable, *mumsmoke,* which is dichotomous and is treated as a dummy variable in exactly the same way as in ordinary least squares regression.

In previous chapters we have seen that all of these independent variables appear to be associated with smoking behaviour. Using logistic regression, we now have the opportunity to assess the combined power of these variables to predict the probability of smoking, as well as assessing their relative impact on the probability (and odds) that somebody will be a smoker.

Examining model fit

The first thing we want to know is whether the relationship between the independent variables and the dependent variable is statistically significant.

In linear regression, we assess this with an F test. The F statistic is arrived at by expressing the mean square of the regression and the mean square of the residual as a ratio. A high ratio indicates that the degree of fit of the model is unlikely to have occurred by chance.

If you are unsure about how the F statistic is calculated, you should return to chapter 10 to familiarise yourself with it, for you will need to understand how the F statistic is calculated to be able to understand this section.

In logistic regression, we use the log likelihood multiplied by –2 to assess the model fit. We have already seen that the log likelihood measures the odds that the value of the dependent variable can be correctly predicted from the observed values of the independent variables. It is analogous to the sum of squares residual in linear regression, for it tells us what variance in the dependent variable has been left unexplained after all the independent variables have been entered into the model. By multiplying this statistic by –2, the distribution comes to approximate the Chi Square distribution.

In our model, the –2 log likelihood is 341.192 (see Table H.3). To find how well the model fits, we compare this –2 log likelihood with all of the independent variables entered with the *initial* –2 log likelihood when none of the independent variables have been entered (this is analogous to the total sum of squares in linear regression). If we subtract the –2 log likelihood from the initial –2 log likelihood we obtain the model chi square statistic. This is analogous to the F statistic in linear regression and indicates the increase in the amount of variance explained by adding the independent variables.

Table H.3 Model fit statistics for the logistic model

```
Dependent Variable..   VAR00003   Smoked in last week?
VAR00029  How stressed (scale 0 to 10)
VAR00016  Family/friends smoke?
MUMSMOKE

-2 Log Likelihood   425.32849


-2 Log Likelihood      341.192
Goodness of Fit        312.607
Cox & Snell - R^2         .240
Nagelkerke - R^2          .320

                    Chi-Square    df Significance

 Model                  84.137     3        .0000
```

In our example, the initial –2 log likelihood is 425.328. Therefore, the model chi square is (425.33 – 341.19) = 84.137. If the model chi square is significant at the 0.05 level (as it is in our example), we can conclude that the model provides a significantly better prediction of the probability that the value of the dependent variable = 1 than any prediction that could be made without including the independent variables.

Measuring strength of association

Our model is significantly better than a model assuming no relationship between the dependent and independent variables. But to what *extent* do the independent variables correctly predict the probability of the independent variable equalling 1?

In linear regression we use R^2 to indicate the proportion of variation in the model explained by the independent variables. This provides us with a summary statistic of the strength of the model and is calculated by dividing the sum of squares of the regression by the total sum of squares. To obtain the equivalent measure in logistic regression we divide the model chi square by the initial –2 log likelihood statistic. This is referred to as R_L squared (written R_L^2). For our example in our model:

$$R_L^2 = \frac{84.137}{425.32849} = 0.198$$

The interpretation of this is that the inclusion of the independent variables reduces the 'badness of fit' of the initial model (i.e. the model without any independent variables) by nearly 20 per cent.

One problem with R_L^2, however, is that it often underestimates the proportion of variation explained by the independent variables. An alternative that tries to correct for this is Cox and Snell's R^2. However, this statistic also has a flaw, for its theoretical maximum cannot reach 1. Nagelkerke's R^2 corrects for this. In our example, its value is 0.32 which suggests a moderately strong association between the independent variables and the probability of the dependent variable equalling 1.

Predicted classifications

The output for the logistic regression also provides information on how well the model is able to predict the values of the dependent variable. In most cases we are more interested in predicting probabilities and odds for the dependent variable. However, we can also examine the ability of the model to correctly allocate cases to each of the categories of the dependent variable.

Table H.4: Logistic regression classification table

```
Classification Table Hor VAR00003
The Cut Value is .50
                    Predicted
                   no      yes     Percent Correct
                     n   ⇔    y
Observed          ⇧⇩⇩⇩⇩⇩⇩⇩⇧⇩⇩⇩⇩⇩⇩⇩⇧
   no       n     ⇔  121   ⇔   37   ⇔   76.58%
                  ⇧⇩⇩⇩⇩⇩⇩⇩⇧⇩⇩⇩⇩⇩⇩⇩⇧
   yes      y     ⇔   44   ⇔  105   ⇔   70.47%
                  ⇧⇩⇩⇩⇩⇩⇩⇩⇧⇩⇩⇩⇩⇩⇩⇩⇧
                            Overall  73.62%
```

The classification table (Table H.4) classifies all of those cases where the expected probability of the dependent variable equalling 1 is above 0.5 as having an expected

value of 1. Those cases where the expected probability of the dependent variable equalling 1 is less than 0.5 are classified as having an expected value of 0.

In our example, if we sum each of the columns in the classification table we can see that the model predicts (44 + 105) = 142 cases in our sample to be smokers and (121 + 37) = 165 cases to be non-smokers.

These predicted values are cross-classified with the actual (observed) values to find what proportion of cases has been correctly predicted by the model. If the model had made perfect predictions, then the proportion of cases correctly classified would be 100 per cent and the predicted value of the dependent variable would be exactly the same as for the observed value for every case. In our example, the model seems better able to predict non-smokers than smokers (77 per cent accuracy compared with 70 per cent accuracy) and the overall predictive accuracy is 74 per cent. This can be interpreted as indicating that knowing the values of the independent variables enables us to correctly predict who will smoke and who will not smoke nearly three times out of four.

Interpreting the logit coefficients

Table H.5 gives information about the variables in the model, their logit coefficients, the significance of these coefficients and the odds ratio for each independent variable.

The B coefficients are not terribly useful in helping us understand the effects of the variables since they predict the value of logit (Y) – a rather meaningless value. Therefore we will probability want to convert these coefficients to the more meaningful statistics such as probabilities or odds.

Table H.5: Logit coefficient and odds ratio output in logistic regression

---------- Variables in the Equation ----------

Variable	B	S.E.	Wald	df	Sig	R	Exp(B)
VAR00029	.1247	.0534	5.4522	1	.0195	.0901	1.1328
VAR00016	-1.0319	.1412	53.4120	1	.0000	-.3477	.3563
MUMSMOKE	.1677	.2706	.3841	1	.5354	.0000	1.1826
Constant	1.8821	.5094	13.6490	1	.0002		

If we wanted to calculate the probability that somebody smokes who is highly stressed (a score of 10 for var00029), most of whose friends smoke (a score of 1 for var00016) and whose mother smoked when he or she was growing up, we would get:

$$\text{Logit}(Y) = a + b_1X_1 + b_2X_2 + b_3X_3$$

$$\text{Logit}(Y) = 1.8821 + (0.1247 \times 10) + (1 \times -1.0319) + (1 \times 0.1677) = 2.2649$$

$$\text{Odds} = e^{2.2649} = 9.63$$

Therefore, the probability of such an individual being a smoker = 9.63/(1+9.63) = 0.91 or 91 per cent.

Remember that the logit coefficients are unstandardized, which means that we cannot directly compare the relative strength of effect of the independent variables in predicting logit (Y). The standardized coefficients are not provided, and this is a serious shortcoming if we are interested in comparing the relative effects that the independent variables have on logit (Y).

The simplest way of dealing with this is to standardize all the variables *before* running the regression. Had we done this we would have obtained the following standardized logit coefficients:

Stress (var00029)	zb= 0.320
Friends/family smoke (var00016)	zb= -1.107
MUMSMOKE	zb= 0.083

The standardized logit coefficients indicate that the proportion of friends and family who smoke has by far the most impact on the dependent variable. We can say that for every standard deviation increase in the proportion of friends and family that smoke, the predicted value of logit(Y) decreases by –1.107 standard deviations. In comparison, the stress variable appears to have about only one third of the influence on the predicted value of logit(Y) since a standard deviation increase in level of stress only results in 0.320 of a standard deviation increase in logit(Y).

Statistical significance of independent variable effects

Returning to Table H.5, the statistical significance of the effect of each variable is derived from the *Wald statistic* which is the ratio of the unstandardized logit coefficient to its standard error. To obtain the Wald statistic for each variable, the logit coefficient is divided by the standard error of the logit coefficient and the result is squared. For example, for the stress variable (var00029) the Wald statistic = $(0.1247/0.0534)^2 = 5.4$

The Wald statistic has the same distribution as the chi square statistic and, once the degrees of freedom have been calculated, we can obtain the probability. In this example, a chi square value of 5.4 with 1 df = p 0.0195.

Interpreting the odds ratios

The right-hand side of the output under the column labelled 'Exp(B)' in Table H.5 contains the odds ratios for each variable. As the label suggests, the odds ratios are simply calculated by taking the exponential (the antilog) of the logit coefficient. For example, the odds ratio for the stress variable is $e^{0.1247} = 1.13$.

Let us work through the odds ratios for each independent variable to try to understand how to interpret them:

- *Self-reported stress*: The stress variable has an odds ratio of 1.13. This can be interpreted as saying that for each unit increase in self-reported level of stress, the odds of being a smoker increase by 13 per cent. Looking at this another way, an individual with an extremely high level of stress (a score of 10) would have an odds of 1 + (10 x 0.13) = 2.13 of being a smoker as compared with somebody with a low level of stress (a stress score of 1). This means that, controlling for the other independent variables, a highly stressed person has more than twice the odds of being a smoker than a person who has a low level of self-reported stress.

- *Proportion of friends and family that smoke*: For var00016 the interpretation of the odds is made more complex by the fact that an increase in the level of the variable represents a decrease in the proportion of friends and family that smoke. Therefore, a one-unit increase in this variable (a reported decrease in the proportion of friends and family who smoke) results in a decrease in the odds of smoking of 64 per cent (or nearly two-thirds). We obtain 64 per cent because the odds of being a smoker are increased by a factor of .3563, which is .64 less than 1.

- *Mother's smoking history*: The final independent variable is a dichotomous variable that records whether the respondent's mother smoked when the respondent was growing up. The odds of somebody smoking if his or her mother smoked (a value of 1) are increased by 18 per cent as compared to somebody whose mother did not smoke. We should, however, treat this odds ratio with caution because the coefficient for this variable is not significant ($p = > 0.05$) and it would not be wise to imply that this effect would be found in the population from which the sample was drawn.

A note on logistic regression assumptions

Many of the assumptions that are made in linear regression also hold in logistic regression. Logistic regression assumes linearity in the dependent variable, logit (Y), a low degree of collinearity and the absence of large residuals (i.e. outliers). Evidence for any of these should be examined carefully. The most common symptoms that one

or more of the assumptions is being violated is that the standard errors for one or more the independent variables will be large.