Chapter 10
Modelling Binary Outcomes

Logistic Regression

Gail M. Williams and Robert Ware

Abstract This chapter introduces regression, a powerful statistical technique applied to the problem of predicting health outcomes from data collected on a set of observed variables. We usually want to identify those variables that contribute to the outcome, either by increasing or decreasing risk, and to quantify these effects. A major task within this framework is to separate out those variables that are independently the most important, after controlling for other associated variables. We do this using a statistical model. We demonstrate the use of logistic regression, a particular form of regression when the health outcome of interest is binary; for example, dead/alive, recovered/not recovered.

The Generalized Linear Model (GLM)

Statistical models are mathematical representations of data, that is, mathematical formulae that relate an outcome to its predictors. An outcome may be a mean (e.g. blood pressure), a risk (e.g. probability of a complication after surgery), or some other measure. The predictors (or explanatory variables) may be quantitative or categorical variables, and may be causes of the outcome (as in smoking causes heart failure) or markers of an outcome (more aggressive treatment may be a marker for more severe disease, which is associated with a poor health outcome).

Generically, a fitted statistical model is represented by linear equations as shown in Fig. 10.1. 'Outcome' is the predicted value of the outcome for an individual who has a particular combination of values for predictors 1–3 etc. The coefficients are estimated from the data and are the quantities we are usually most interested in. The particular value of a predictor for an individual is multiplied by the corresponding coefficient to represent the contribution of that predictor to the outcome. So, in

G.M. Williams (✉) • R. Ware
School of Population Health, University of Queensland, Herston, QLD, Australia
e-mail: g.williams@sph.uq.edu.au

S.A.R. Doi and G.M. Williams (eds.), *Methods of Clinical Epidemiology*,
Springer Series on Epidemiology and Public Health,
DOI 10.1007/978-3-642-37131-8_10, © Springer-Verlag Berlin Heidelberg 2013

Fig. 10.1 A fitted GLM depicted mathematically

Predicted outcome = *constant coefficient*
+ *coefficient 1* × value of a predictor 1
+ *coefficient 2* × value of a predictor 2
+ *coefficient 3* × value of a predictor 3
+ ...

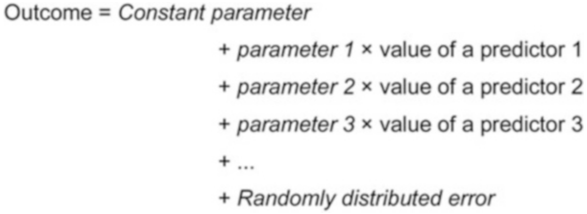
particular, if a coefficient for a predictor is estimated to be zero then that predictor makes no contribution to the outcome. The constant coefficient represents the predicted value of the outcome when the values of all of the predictors are zero. This may or may not be of interest or interpretable, because zero may not be in the observable range of the predictor.

So the model predicts values of an outcome from each person's set of values for predictors. This, of course, generally does not match that person's actual observed value. The difference between the observed value and the predicted value is called the residual, or sometimes the error. The term error does not imply a mistake but rather represents the value of a random variable measuring the effects on individual observed outcome values other than those due to the predictor variables included in the model. Adding more predictor variables to the model is expected to reduce the error. Mechanistically, the error or residual for a particular individual is the difference between the individual's observed and predicted values. An example is the difference between an individual's observed blood pressure and that predicted by a model that included age and body mass index.

The theory of model fitting and statistical inference from the model requires that we make an assumption about the distribution of the errors. In many cases, where we have a continuous outcome variable, the assumed distribution is a normal distribution. This is the classic regression model. A log-normal distribution might be used if a continuous variable is positively skewed. However, if we have a binary variable, we might assume a binomial distribution. Thus, the full theoretical specification of a model is represented by Fig. 10.2.

Fitting a Model

Fitting a model means finding the parameter estimates within the model equation that best fits with the observed data. So the parameters referred to in Fig. 10.2 are estimated from the data to give the coefficients referred to in Fig. 10.1. This may be done in different ways. One of the earliest methods proposed to do this was the Method of Least Squares, a general approach to combining observations, developed by the French mathematician Adrien Marie Legendre in 1805. Effectively, this identifies the parameter estimates that minimize the sum of squares of the errors as in Fig. 10.2. In this sense, we estimate the parameters by values that bring the predicted values as close as possible to the observed values. This works well with some probability distributions, but not with others. Currently, the statistically preferred technique is a process called maximum likelihood, or some variant of



Outcome = *Constant parameter*
+ *parameter 1* × value of a predictor 1
+ *parameter 2* × value of a predictor 2
+ *parameter 3* × value of a predictor 3
+ ...
+ *Randomly distributed error*

Fig. 10.2 The full general linear model depicted mathematically

this, which has the advantage of providing a more general framework covering different types of probability distributions. This method was pioneered by the influential English statistician and geneticist, Ronald Fisher, in 1912. The method selects the values of the parameters that would make our observed data more likely (under the chosen probability model) to have occurred than any other sets of values of the parameters. This approach has undergone considerable controversy, application and development, but now underlies modern statistical inference across a range of different situations.

Link Functions

The GLM generalizes linear regression by allowing the linear model to be related to the outcome variable via a link function and incorporating a choice of probability distributions which describes the variance of the outcomes. While this chapter focuses on using the logit link for modelling binary outcomes, it is not the only possible link function. The logit link (hence logistic regression) is linear in the log of the odds of the binary outcome and thus can be transformed to an odds ratio. However, if we want to model probabilities rather than odds, we need to use a log link rather than a logit link and then this can be transformed to a risk ratio. However, unlike the logistic regression model, a log-binomial model can produce predicted values which are negative or exceed one. Another concern is that it is not symmetric since the relative risks for the outcome occurring and the outcome not occurring are not the inverse of each other as with an odds ratio. Also, odds ratios and risk ratios diverge if the outcome is common. If the risk of the outcome occurring is greater than 50 %, it may be better to model the probability that the outcome does not occur to avoid producing predicted values which exceed one.

Models for Prediction Versus Establishing Causality

We can use models to establish causality or for prediction or a combination of the two. For causal models, we are usually interested in ensuring control of confounding, so we can assert that the exposure of interest (say smoking) is a likely cause of the outcome (heart failure); that is, that the association is not due to confounding by social class, diet, etc. In this situation, we usually need to examine closely the relationships between variables in the model. For prediction we try to produce an inclusive model that considers all relevant causes and/or markers of a particular outcome to enable us to predict the outcome in a particular individual. A predictive model thus focuses more on predictor–outcome associations, rather than being concerned with confounding per se.

Now that we have an understanding of a model and its components, we look at a type of model commonly used in clinical epidemiology – logistic regression.

A Preliminary Analysis

Data-set

The Worcester Heart Attack Study examined factors associated with survival after hospital admission for acute myocardial infarction (MI). Data were collected during 13 one-year periods beginning in 1975 and extending until 2001, on all MI patients admitted to hospitals in the Worcester, Massachusetts Standard Metropolitan Statistical Area. The 500 subjects in the data set are based on a 23 % random sample from the cohort in the years 1997, 1999 and 2001 yielding 500 subjects.

Of the 500 patients, 215 (43 %) died within their follow-up period. The median follow-up time was 3.4 years. All patients were followed up for at least 1 year and 138 (27.6 %) died within the first year following the MI. We are interested in examining the factors that predict death within the first year after the MI as the 500 subjects had complete follow-up to this time point.

Preliminary Results

When we examine the risk of death in the first year according to gender and age, we see a somewhat higher percentage of deaths in females than males, and that percentage of deaths increases markedly with age, from 7.2 % (95 % confidence limits (CL) 2.9, 11.6 %) to 49.4 % (41.7, 57.1 %) (Table 10.1). The 95 % confidence intervals are wider for smaller subgroups, but the age variation is substantial. Are these differences statistically significant? Because we are considering two categorical variables, evaluation of statistical significance uses the Pearson chi-square test, provided there are few small expected frequencies. This test examines the null hypothesis that the true risk of death is the same across all subgroups. Implicit in this assertion is an assumption that any observed differences in the estimated risk of death (e.g. 25.0 % vs. 31.5 % for males vs. females) are due to chance. The *P* value associated with the gender comparison is 0.111. Because the *P* value is not small enough (the usual criterion being <0.05), we do not reject our null hypothesis and we conclude the observed differences are not so large that they could not have occurred by chance. For age, however, $P < 0.0001$, and we conclude that observed differences are not consistent with chance variability.

Table 10.1 Percentage of deaths within the first year after an MI, by age group and gender, with 95 % confidence intervals (95 % CI) ($N = 500$)

Risk factor	N	Deaths	Deaths (%)	95 % CI	P value
Gender					
Male	300	75	25.0	20.1, 29.9	
Female	200	63	31.5	25.1, 37.9	0.111
Age group					
<60 years	138	10	7.2	2.9, 11.6	
60–69 years	86	12	13.9	6.6, 21.3	
70–79 years	114	36	31.6	23.0, 40.1	
>80 years	162	80	49.4	41.7, 57.1	<0.0001

If we are interested in identifying the significance of a trend for risk of death to systematically increase with age, we need to use a statistical test that takes into account how the age categories are ordered. There are various statistical tests and most are available in standard packages. They vary somewhat in their assumptions about the way in which the ordered categories are expressed, but they usually give similar answers, especially in large samples. One of the simplest forms assigns an ordinal score (1,2,3,...) to the categories and examines a linear regression of the prevalence on the score (as a predictor variable). For age groups, this test yields a P value < 0.0001.

We can go a step further and examine the relative risks (RRs), that is, the ratio of the percentage of deaths in a subgroup compared with that in a chosen reference group (Table 10.2). In anticipation of later analyses, the odds ratios (ORs) are also given in Table 10.2. Note that ORs are further away from 1 than are relative risks; for example, RR $= 6.81$ for the oldest age group compared with the youngest, with a corresponding OR of 12.49. This will always be the case, and the distance will increase as the risk of death increases. However, this does not change the formal statistical inference regarding this comparison. The P value for the difference between the percentages of deaths is <0.0001, based on a chi-square value of 63.0 (1 df), whether we choose to measure the age effect by an RR, OR, or, indeed a risk difference (49.4 % $-$ 7.2 % $=$ 42.2 %). Table 10.3 shows a similar analysis for selected characteristics of the MI.

We now wish to explore these relationships further to determine which factors, or combinations of them, are the most predictive of death within the first year. We know that the MI characteristics are associated and that they are also likely to be related to age group, itself a strong risk factor. We can explore this in several ways.

One approach is to carry out a stratified analysis: we stratify by a (suspected) confounding variable, and examine the effect of our exposure of interest within each stratum. Thus, to adjust the effect of congestive heart failure for age, we stratify by age groups. Before proceeding further, we collapse age into two categories (<70 years of age and $\geq$70 years) to increase the numbers in each category. Stratified analysis is shown in Table 10.4.

Recall that the RR associated with cardiogenic shock overall was 2.45 (95 % CI 1.73, 3.48) (Table 10.3). We see now that the risk of death is lower in younger

Table 10.2 Percentage of deaths within the first year after an MI, and RRs and ORs by age group and gender, with 95 % CIs ($N = 500$)

Risk factor	N	Deaths (%)	RR	95 % CI for RR	OR	95 % CI for OR
Gender						
Male	300	25.0	1		1	
Female	200	31.5	1.26	0.95, 1.67	1.38	0.93, 2.05
Age group						
<60 years	138	7.2	1		1	
60–69 years	86	13.9	1.92	0.87, 4.26	2.08	0.86, 5.04
70–79 years	114	31.6	4.36	2.26, 8.39	5.91	2.78, 12.57
>80 years	162	49.4	6.81	3.68, 12.63	12.49	6.12, 25.49

Table 10.3 Percentage of deaths within the first year after an MI, and RRs and ORs by MI characteristics, with 95 % CIs ($N = 500$)

Risk factor	N	Deaths (%)	RR	95 % CI for RR	OR	95 % CI for OR
Cardiogenic shock						
Absent	478	25.9	1		1	
Present	22	63.6	2.45	1.73, 3.48	5.00	2.05, 12.20
Congestive heart failure						
Absent	345	17.4	1		1	
Present	155	50.3	2.89	2.19, 3.82	4.81	3.16, 7.32
MI type						
Non-Q wave	347	30.6	1		1	
Q wave	153	20.9	0.68	0.48, 0.97	0.60	0.38, 0.94
History of cardiovascular disease						
Absent	125	24.0	1		1	
Present	375	28.8	1.20	0.85, 1.70	1.28	0.80, 2.04
Atrial fibrillation						
Absent	422	26.1	1		1	
Present	78	35.9	1.38	0.98, 1.93	1.59	0.95, 2.65
Complete heart block						
Absent	489	27.2	1		1	
Present	11	45.4	1.67	0.86, 3.24	2.23	0.67, 7.43
Previous MI						
Absent	329	25.2	1		1	
Present	171	32.2	1.27	0.96, 1.70	1.41	0.94, 2.11

patients (9.8 % vs. 42.0 %). However, within these groups (i.e. controlling for patient age, at least up to a point) we also see that the risk of death increases with the presence of cardiogenic shock, although the RRs have decreased, because of the confounding of the overall effect with age; older patients are more likely to have cardiogenic shock. However, the CIs for these RRs are now wider, reflecting the fact that we are now dealing with subgroups of the data, rather than the entire sample (Table 10.4).

Table 10.4 Percentage of deaths within the first year after an MI, and RRs and ORs by presence of cardiogenic shock and age group, with 95 % CIs

Age	Cardiogenic shock	*N*	Deaths (%)	RR	95 % CI for RR	OR	95 % CI for OR
<70 years	Absent	218	9.2	1		1	
	Present	6	33.3	3.63	1.09, 12.14	4.95	0.85, 28.73
	Total	224	9.8				
≥70 years or more	Absent	260	40.0	1		1	
	Present	16	75.0	1.88	1.36, 2.58	4.50	1.41, 14.33
		276	42.0				

Table 10.5 Percentage of deaths within the first year after an MI, and RRs by presence of cardiogenic shock, with 95 % CIs, unadjusted RR and adjusted by the Mantel–Haenszel method (RR_A) for the effect of age

Cardiogenic shock	*N*	Deaths (%)	RR	95 % CI for RR	RR_A	95 % CI for RR_A
Absent	478	25.9	1	1	1	
Present	22	63.6	2.45	1.73, 3.48	2.02	1.48, 2.76

Using the Mantel–Haenszel technique, we can then pool the stratum-specific RRs, with weightings that reflect stratum size to obtain adjusted RRs. This provides us with the best overall estimate (provided the stratum-specific RRs are consistent) and gives us greater precision, that is, narrow confidence intervals (Table 10.5).

We now see clearly that the effect of adjustment for age appears to have been to decrease the RR associated with cardiogenic shock, since we have adjusted for the fact that patients with cardiogenic shock are also older and age carries its own separate risk.

The Mantel–Haenszel approach to adjustment is an effective method of adjusting for confounders, and is a useful way of identifying confounders one variable at a time. However, it is obvious that this will become tedious when we have multiple confounders to take into account; we would have to construct all the strata related to all combinations of confounder categories, and then perform an analysis on each (some strata would be small, with wide confidence intervals for within-stratum effect estimates) and then pool these estimates. Regression modelling provides us with an effective approach, but, as we will see, involves some additional assumptions.

Logistic Regression

As explained earlier, a regression model consists of two major components: (a) a probability model, which specifies a theoretical distribution (our choice of this is based partly on empirical observations and partly on our theory about the underlying processes that generated the observations) and (b) specification of relevant predictors based on the research questions or hypotheses we want to examine. We now require a probability model for an outcome variable that takes only two values, such as disease/no disease, dead/alive, etc. A further assumption we make is that our observations are independent in the sense that one person dying within the first year after MI does not affect the probability that another person dies in the first year (this may not be true, e.g. if we had included two MI episodes in the study for the same patient). With this assumption, the number of deaths in the first year out of a sample of size *N* would be expected to follow a binomial distribution. Apart from *N*, this distribution depends on a parameter *p*, which is the probability of an event (death in first year). We can estimate this overall by our proportion of deaths, 27.6 % or $p = 0.276$.

However, as we have seen, the risk of death varies according to age and the characteristics of the MI itself. Thus, our *p* parameter is allowed to take various values, according to various predictors; indeed this is what we want to model. Our outcome variable is the proportion of events of interest (death), out of a given number of possible outcomes, when the probability of a single event is *p* (which may depend on the predictors of interest).

If we simply model the probability of an event as a function of predictors, it is possible to obtain predicted values that do not lie between 0 and 1. We could, for example, predict a prevalence of −0.05 or −5 % or 1.06 or 106 %. This is a very undesirable feature of a theoretical model.

Several different approaches have been tried to overcome this problem, by transforming the outcome probability to a quantity that must lie between 0 and 1. Currently, the most widely used transformation is the logit transformation, first proposed by Joseph Berkson in 1944. It is effectively a log-odds transformation. If *p* is the probability of the event of interest (say disease), the logit of *p* is given by

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) = \log(\text{odds of disease})$$

where log is the logarithm function, to base e.

We can see that this transformation accommodates the constraints on modelling a proportion. If we invert the transformation, we can see that the probability of the event, *p*, is

$$p = \frac{\exp(\text{logit}(p))}{1+\exp(\text{logit}(p))}$$

where exp is the exponential or antilog function. This is always greater than zero, because the exponential function cannot take negative values. The denominator is larger than the numerator, so it can never be greater than 1. So proportions must lie between 0 and 1.

The main reason for the popularity of this transformation, however, is the consequent interpretation of the regression coefficients when it is used. Putting this transformation together with a model based on age group (reverting to four age groups), we have the logistic regression model as follows:

$$\begin{aligned}&\text{logit(Probability of death)} = \log(\text{odds of death})\\&\quad = a + b_1 \times (60 - 69 \text{ years})\\&\qquad + b_2 \times (70 - 79 \text{ years})\\&\qquad + b_3 \times (80 \text{ years and over})\end{aligned}$$

where the notation following the coefficients means: if the statement inside the brackets is true, the value inside the brackets takes the value 1, otherwise it takes the value 0. These are sometimes referred to as indicator variables. This is a compact way of indicating that the coefficients b_1, b_2, b_3 are associated with the categories 60–69 years to 70 years or older, in order, and that the omitted category, <60 years, is the reference category. The above model fits the framework give in Fig. 10.1, where the predicted outcome is the logit(Probability of death), the coefficients are a, b_1, b_2, b_3, and the values of the predictors are given by the indicator variables for each age group.

To further clarify the interpretation of the coefficients, and the role of the reference category, consider a patient who is less than 60 years of age. This patient's predictive model is as follows:

$$\log(\text{odds}(\text{death})), \quad \text{if patient} < 60 \text{ years} = a$$

A patient who is 60–69 years of age has the following predictive model:

$$\log(\text{odds}(\text{death})) \text{ if patient } 60 - 69 \text{ years} = a + b_1$$

Subtracting these last two expressions (the first from the second), we see that

$$\begin{aligned}&\log(\text{odds}(\text{death})) \text{ if patient } 60 - 69 \text{ years}\\&- \log(\text{odds}(\text{death})) \text{ if patient} < 60 \text{ years}\\&= b_1\end{aligned}$$

Using the fact that $(\log A - \log B) = \log(A/B)$, we see that

$$\begin{aligned}b_1 &= \log\left(\frac{\text{odds(death) if patient } 60 - 69 \text{ years}}{\text{odds(death) if patient} < 60 \text{ years}}\right)\\&= \log\left(\begin{array}{l}\text{odds ratio of death for patient aged } 60 - 69 \text{ years,}\\ \text{compared with patient aged} < 60 \text{ years}\end{array}\right)\end{aligned}$$

So the regression coefficients are directly interpretable as log(ORs) and we can then obtain the actual OR by exponentiation or antilogs of the parameter estimates.

Table 10.6 Parameter estimates from logistic regression of death in the first year, with age group as a predictor

Parameter		Value	Parameter estimate	OR	95 % CL for OR	*P* value
Intercept	a		−2.55			
Age at MI	Reference: Age < 60 years					
	b_1	60–69 years	0.73	2.08	0.86, 5.04	0.106
	b_2	70–79 years	1.78	5.91	2.78, 12.57	<0.0001
	b_3	≥80 years	2.52	12.49	6.12, 25.49	<0.0001

Fitting the logistic regression model is done using maximum likelihood estimation of the model parameters (a, b_1, b_2, b_3 in the age group model), as has been described previously. It is not important to understand the details of this process, but it is important to understand that the process does not always work, in the sense that a solution may not be found, often due to sparseness of data or unusual distributions. Depending on the software you use, you may receive a warning that convergence has not been attained, or you may simply observe results that look meaningless, such as extremely large standard errors of estimates. You should always scrutinize parameter estimates and their standard errors (or CLs) to look for values that differ greatly from your single variable or preliminary analyses.

Again (Table 10.6) we see that the trend for ORs from the logistic regression, as for RRs (Table 10.2), increases as age increases. However, we see that the ORs from the logistic regression are exactly the same as those in Table 10.2. This is because they are mathematically equivalent; this equivalence does not hold as we include more variables in the analysis. The parameters b_1, b_2, b_3 thus represent the outcome (death) log ORs for each group, compared with the reference group, which is the group omitted from the parameter list in the model. The parameter a is usually not of interest; it represents the log(odds) of the event, within the reference category. In this case, the reference category is the youngest age group, and the odds of death for this group is $10/(138 - 10) = 0.078 = e^{(-2.55)}$.

We can also calculate what our model predicts for the probability of death for each age group by substituting for the parameters a, b_1, b_2, b_3.

$$\begin{aligned}
\text{Probability of death} &= \frac{\exp(\text{logit}(p))}{1+\exp(\text{logit}(p))}\\
&= \frac{\exp(-2.55\ )}{1+\exp(-2.55)} = 0.072 \quad \text{if patient} < 60 \text{ years}\\
&= \frac{\exp(-2.55+0.73)}{1+\exp(-2.55+0.73)} = 0.139 \quad \text{if patient } 60-69 \text{ years}\\
&= \frac{\exp(-2.55+0.1.78\ )}{1+\exp(-2.55+1.78)} = 0.316 \quad \text{if patient } 70-79 \text{ years}\\
&= \frac{\exp(-2.55+2.52\ )}{1+\exp(-2.55+2.52)} = 0.494 \quad \text{if patient} \geq 80 \text{ years}
\end{aligned}$$

We see that the univariable model replicates the observed proportions, which is what we would expect.

Multivariable Logistic Regression

Categorical Predictors

Although univariate logistic regression gives the same results as a simple cross-tabulation, the major advantage of embarking on a logistic regression approach obviously comes from the ability to include additional variables, either as confounders, or as risk factors or predictors in their own right. Later, we also deal with interactions, but for now we examine a logistic regression model that includes age group as a possible confounder to the cardiogenic shock effect. This may be written out exactly as we have done previously, by adding additional terms and regression coefficients to the right-hand side of the model equation:

$$\begin{aligned}
&\text{logit}(\text{Probability death}) \\
&\quad = \log(\text{odds}(\text{death})) \\
&\quad = a + b_1 \times (60 - 69 \text{ years}) \\
&\qquad + b_2 \times (70 - 79 \text{ years}) \\
&\qquad + b_3 \times (\geq 80 \text{ years}) \\
&\qquad + c \times (\text{cardiogenic shock present})
\end{aligned}$$

The maximum likelihood estimates are given in Table 10.7.

The coefficients and ORs for age group have now changed because of the inclusion of an additional variable, cardiogenic shock. They are now the estimated effects, after adjusting (controlling) for the effect of cardiogenic shock. Reciprocally, the effects of cardiogenic shock have been adjusted for age group. To see this, consider a patient who is 60–69 years of age and does not have cardiogenic shock. This patient's predictive model is as follows.

$$\begin{aligned}
&\log(\text{odds}(\text{death})), \quad \text{if patient } 60 - 69 \text{ years does not have cardiogenic shock} \\
&\quad = a + b_1
\end{aligned}$$

A patient who is 60–69 years of age and has cardiogenic shock has the following predictive model:

$$\begin{aligned}
&\log(\text{odds}(\text{death})) \text{ if patient } 60 - 69 \text{ years has cardiogenic shock} \\
&\quad = a + b_1 + c
\end{aligned}$$

Table 10.7 Parameter estimates and 95 % CIs from logistic regression of death, with age and presence of cardiogenic shock

Parameter		Value	Parameter estimate	OR	95 % CL for OR	*P* value
Intercept	a		−2.59			
Age group	Reference: <60 years					
	b_1	60–69	0.66	1.94	0.79, 4.75	0.147
	b_2	70–79	1.73	5.63	2.63, 12.04	<0.0001
	b_3	≥ 80	2.48	11.98	5.85, 24.55	<0.0001
Cardiogenic shock	Reference: Absent					
	c	Present	1.49	4.46	1.68, 11.82	0.0027

Subtracting these last two expressions (the first from the second), we see that

$$\begin{aligned}&\log(\text{odds}(\text{death}))\text{ if patient }60-69\text{ years has cardiogenic shock}\\&-\log(\text{odds}(\text{death}))\text{ if patient }60-69\text{ years does not have cardiogenic shock}\\&\quad=c\end{aligned}$$

Using the fact that ($\log A\ -\ \log B) = \log(A/B)$, we see that

$$\begin{aligned}b_1&=\log\left(\frac{\text{odds}(\text{death})\text{ if patient }60-69\text{ years has cardiogenic shock}}{\text{odds}(\text{death})\text{ if patient }60-69\text{ years does not have cardiogenic shock}}\right)\\&=\log\left(\begin{array}{l}\text{OR of death associated with having}\\\text{cardiogenic shock in patient aged }60-69\text{ years}\end{array}\right)\end{aligned}$$

So we have controlled for age by virtue of holding it constant at 60–69 years. It is easy to see that had we held age constant at some other age group, 70–79 years say, then the same result would have been obtained for the age-adjusted effect of cardiogenic shock. This is an assumption that we make: the effects of variables are constant across values of other variables in the model. This assumption can be relaxed at the cost of making the model more complex; see later section on effect modification.

Returning to the results, we now see similar effects to those we saw with the Mantel–Haenszel analysis for the association between death and cardiogenic shock for age: the effect decreases. We can also see that age is a significant predictor of death. Although these results are consistent with the effects we saw in the Mantel–Haenszel process, they are not the same, largely because ORs are not the same as RRs (except when the outcome rate is very low), but also partly because the method of adjustment by logistic regression is mathematically different from the Mantel–Haenszel approach.

Regression modelling using maximum likelihood fitting also produces likelihood ratio tests, which examine the significance of variables overall. These tests each compare two models: a model that excludes the variable of interest, and one that includes it. The chi-square statistic is a measure of the difference between the

Table 10.8 Likelihood ratio tests for logistic regression of death, with patient age group and presence of cardiogenic shock as predictors

Source	df	Chi-square	*P* value
Patient age group	3	77.70	<0.0001
Cardiogenic shock	1	9.61	0.0019

models and thus can be assessed for statistical significance. These are shown in Table 10.8, and confirm the significance of each of the risk factors independently of the other.

Continuous Predictors

In the above analysis we have grouped age into categories. However, risk increases with increasing age and so it may make sense to treat age as a continuous variable. A simple logistic regression model relating death in the first year to age at MI is then as follows:

$$\begin{aligned}&\text{logit(Probability of death)}\\&\quad=\text{log(odds of death)}\\&\quad=a+b\times\text{Age at MI (years)}\end{aligned}$$

Again we see the meaning of the regression coefficients by considering particular values, say a patient who is 65 years at the MI episode.

$$\begin{aligned}&\text{logit(Probability of death)}_{65}\\&\quad=\text{log(odds of death)}\\&\quad=a+b\times 65\ \text{(years)}\end{aligned}$$

Compare this with a patient who is 64 years at the MI episode.

$$\begin{aligned}&\text{logit(Probability of death)}_{64}\\&\quad=\text{log(odds of death)}\\&\quad=a+b\times 64\ \text{(years)}\end{aligned}$$

Subtracting these, we have

$$\begin{aligned}&\text{log(odds of death if patient is 65 years)}\\&-\text{log(odds of death if patient is 64 years)}\\&\quad=b\end{aligned}$$

Using the fact that $\log A - \log B = \log(A/B)$, we see that

Table 10.9 Parameter estimates and 95 % CLs from logistic regression of death, with age in years as a continuous predictor

Parameter		Value	Parameter estimate	OR	95 % CL for OR	*P* value
Intercept	*a*		−6.86			
Age (years)	*b*	per year	0.080	1.08	1.06, 1.11	<0.0001

$$\begin{aligned} b &= \log\left(\frac{(\text{odds of death if patient is 65 years})}{(\text{odds of death if patient is 64 years})}\right) \\ &= \log(\text{odds ratio for death for a 1-year increase in age at MI}) \end{aligned}$$

Again we see that the regression coefficient is interpretable as a log(OR). Here, however, we do not have a fixed reference group: the OR refers to a fixed increase of 1 unit of the predictor variable. It follows that we cannot interpret the coefficient for a continuous variable unless we know the units in which it is measured. To then get the actual OR we need to exponentiate or antilog the coefficient. Fitting the model for age in years yields Table 10.9.

The OR associated with age is 1.09 or an increase in odds of death by around 9 %. This seems very modest until we remember that this represents the increase associated with only 1 year of age. The predicted increase in risk for an increase of 10 years of age (similar to the age groups we used earlier) can be calculated as follows:

$$\begin{gathered} \text{Increase in } \log(\text{odds death}) \text{ for 1 year of age} = 0.084 \\ \text{Increase in } \log(\text{odds death}) \text{ for 10 years of age} = 0.084 \times 10 = 0.84 \\ \text{Increase in } (\text{odds death}) \text{ for 10 years of age} = \mathrm{e}^{0.84} = 2.32 \end{gathered}$$

Thus, a decade increase in age at MI increases the odds of death in the first year by 2.32-fold.

We need to be extremely careful in interpreting ORs as RRs. It is well known that ORs approximate RRs when the risk of the outcome is small. Small usually means less than about 15 %. The OR is further from 1 than is the RR, as we can see from Tables 10.2 and 10.3. Thus, if the OR is uncritically interpreted as an approximate RR, it will consistently overestimate the strength of the association.

Let us now examine the predictions from our model. Our fitted model (Table 10.9) is

$$\begin{aligned} &\text{logit}(\text{Probability of death}) \\ &\quad = \log(\text{odds of death}) \\ &\quad = -6.86 + 0.084 \times \text{Age (years)} \end{aligned}$$

When we do the algebra to express the probability of death in terms of age at MI we get

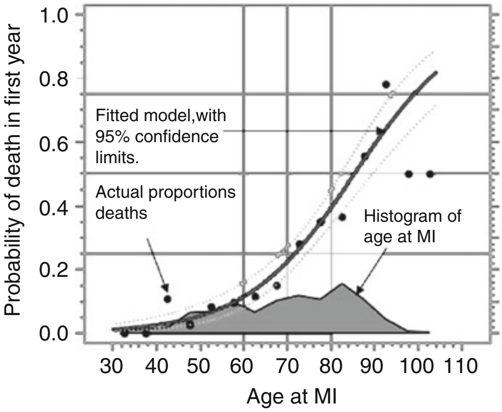


Fig. 10.3 Logistic regression model giving the probability of death in the first year after MI, as predicted by age at MI. The figure shows the classic S-shaped logistic curve; the probability of the outcome increases with the predictor, slowly at first, then increasingly so, and then flattening out. It also shows 95 % CLs for the predicted proportions with the outcome, the risk of the outcome. The dots show the observed risk of death within a centred 5-year age group

$$\text{Probability of death} = \frac{\exp(-6.86 + 0.084 \times \text{Age})}{1 + \exp(-6.86 + 0.084 \times \text{Age})}$$

While this may look a little complex, it is relatively easy to calculate given any particular value of age. Most statistical software programs that fit regression models can calculate these values for all values of predictors that occur in the sample used for fitting the model. We see these in Fig. 10.3 for the current example.

The shaded area at the bottom of the graph in Fig. 10.3 shows the distribution of age, with the three vertical lines showing cut-offs at 60, 70 and 80 years. The RR, comparing two values of the predictor is simply the ratio of the heights of the curve at those outcomes. These values can be read from the graph or calculated from the formula given above. Table 10.10 shows these values, as well as the calculated ORs and RRs, comparing each increase in risk (whether measured by the odds or the proportion of deaths) associated with 1 year increase in age.

Table 10.10 confirms that the OR is constant; this is not surprising because this is a condition of the model. It also confirms that when the predicted probability of death is small (less than 15 %), the RR is very close to the value of the OR. However, as age increases and the predicted risk of death correspondingly increases, the RR diminishes, although it is always >1.

Figure 10.3 is also revealing in terms of the strength of the association between age and death. We see that if a patient is 80 years old or more at the MI, he or she has at least a 50 % chance of dying in the first year after the MI. A patient in the ninth decade of life has an 80 % chance of death in the first year after MI.

Using age as a continuous variable implies that we are fitting a linear effect (on the logit scale) for age; that is, the OR is constant. We may be interested in testing

Table 10.10 Logistic regression model of death with age as a predictor: predicted probabilities, ORs and RRs for each year of age, compared to year below

Age at MI (years)	Predicted probability of death in first year	OR (death) comparing age with age − 1	RR (death) comparing age with age − 1
55	0.08014	1.08	1.08
56	0.08627	1.08	1.08
57	0.09282	1.08	1.08
58	0.09981	1.08	1.08
59	0.10727	1.08	1.07
60	0.11522	1.08	1.07
61	0.12367	1.08	1.07
62	0.13265	1.08	1.07
63	0.14218	1.08	1.07
64	0.15227	1.08	1.07
65	0.16294	1.08	1.07
66	0.17421	1.08	1.07
67	0.18608	1.08	1.07
68	0.19856	1.08	1.07
69	0.21167	1.08	1.07
70	0.22539	1.08	1.06
71	0.23974	1.08	1.06
72	0.25470	1.08	1.06
73	0.27026	1.08	1.06
74	0.28641	1.08	1.06
75	0.30312	1.08	1.06
76	0.32036	1.08	1.06
77	0.33812	1.08	1.06
78	0.35634	1.08	1.05
79	0.37498	1.08	1.05
80	0.39401	1.08	1.05
81	0.41336	1.08	1.05
82	0.43298	1.08	1.05
83	0.45282	1.08	1.05
84	0.47280	1.08	1.04

whether this is a reasonable fit to the data. We can do this by including a square or quadratic term in the model. It is usually helpful to centre continuous variables before including them in polynomial or interaction terms. Centring means subtracting a central value (mean or median) from each value. When we do this we obtain Table 10.11.

We see that the quadratic term is clearly non-significant, indicating the linearity assumption is supported.

Table 10.11 Logistic regression model of death with age as a predictor and a quadratic term

Parameter		Value	Parameter estimate	OR	95 % CL for OR	P value
Intercept	a		−1.26			
Age at MI						
Age − 70	b	per year	0.08	1.08	1.06, 1.10	<0.0001
$(\text{Age} - 70)^2$	c	per $(\text{year})^2$	0.0002	1.00	1.00, 1.00	0.772

Combining Categorical and Continuous Predictors

We can combine categorical and continuous predictors in a model provided we keep in mind the appropriate interpretation of the regression coefficients. We now add the effect age as a continuous variable to a model incorporating cardiogenic shock and gender (both categorical variables), as follows:

$$\begin{aligned}
&\text{logit}(\text{Probability of death}) \\
&\quad = \log(\text{odds of death}) \\
&\quad = a + b \times \text{Age at MI (years)} \\
&\qquad + c \times (\text{Patient is male}) \\
&\qquad + d \times (\text{Cardiogenic shock is present})
\end{aligned}$$

The maximum likelihood estimates of the model parameters are now given in Table 10.12.

The inclusion of age as a continuous variable and gender has reduced the effect of cardiogenic shock as a predictor of death, but only slightly. Although females have a higher odds of death than males, this was not significant, and it is likely that the adjustment to the effect of cardiogenic shock was largely due to the strong effect of age, which appears unaffected by adjusting for gender and cardiogenic shock.

Likelihood ratios tests also show the overall significance of effects (Table 10.13), and confirm the predominance of the age and cardiogenic shock effects.

Effect Modification

The models considered so far assume that the effects of predictors are additive on a logit scale; there is only one parameter for the effect of cardiogenic shock, for example, and its effects are assumed to be the same over all age groups. If we wish to allow for effects to vary across values of another variable we need to incorporate an interaction term, which allows for effect modification.

To see how this works, consider the effect of congestive heart failure stratified by age group. Again, for simplicity we divide age into two groups: <70 years and ≥70 years. The stratified analysis is given in Table 10.14.

Table 10.12 Logistic regression model of death, with patient age (as a continuous variable) and gender, and presence of cardiogenic shock

Parameter	Value	Parameter estimate	OR	95 % CL for OR	*P* value
Intercept	*a*	−7.15			
Age					
	b Years	0.08	1.09	1.06, 1.11	<0.0001
Gender	Reference: Males				
	c Females	0.19	1.21	0.77, 1.90	0.404
Cardiogenic shock	Reference: Absent				
	d Present	1.46	4.29	1.62, 11.33	0.003

Table 10.13 Likelihood ratio tests for logistic regression model of death with patient age (as a continuous variable) and gender, and presence of cardiogenic shock

Source	df	Chi-square	*P* value
Age (years)	1	85.2	<0.0001
Gender	1	0.70	0.403
Cardiogenic shocks	1	9.13	0.0025

Table 10.14 Percentage of deaths within first year after an MI, and RRs and ORs by presence of congestive heart failure and age group, with 95 % confidence intervals

Age	Congestive heart failure	*N*	Deaths (%)	RR	95 % CI for RR	OR	95 % CI for OR
<70 years	Absent	181	5.0	1		1	
	Present	43	30.2	6.08	2.78, 13.30	8.28	2.94, 23.75
	Total	224	9.8				
≥70 years	Absent	164	31.1	1		1	
	Present	112	58.0	1.87	1.41, 2.46	3.06	1.80, 5.21
		276	42.0				

We see that the effect of congestive heart failure is much greater in those who are <70 years of age. Note again that the ORs are further away from 1 than are the RRs. The logistic regression model incorporating age group and congestive heart failure is:

$$\begin{aligned}&\text{logit(Probability of death)} = \log(\text{odds of death})\\&\quad = a + b \times (\text{Age } \geq 70 \text{ years}) + c \times (\text{Congestive heart failure is present})\end{aligned}$$

If we fit this logistic regression (first without allowing for an interaction), we get the results in Table 10.15. The OR for the association of age and death is 5.52 and 95 % CI (3.29, 9.24). The OR for the association of congestive heart failure and death is 3.85 (2.47, 6.01). We see that the logistic regression estimate for congestive heart failure falls between the two age stratum-specific estimates in Table 10.15. Thus, the model averages in some way over the stratum-specific estimates, as it has only one parameter.

Table 10.15 Parameter estimates, ORs and 95 % CIs from logistic regression of death, with age and presence of congestive heart failure

Parameter	Value		Parameter estimate	OR	95 % CL for OR	*P* value
Intercept	*a*		−2.60			
Age group	Reference: <70 years					
	b	≥70 years	1.71	5.52	3.29, 9.24	<0.0001
Congestive heart failure	Reference: Absent					
	c	Present	1.35	3.85	2.47, 6.01	<0.0001

The next step is to estimate effects of congestive heart failure within age groups. This is achieved in logistic regression by including additional terms in the predictor part of the model. These additional parameters allow an increment to the congestive heart failure effect for the older age group compared with the younger age group, These parameters are denoted by the *d* parameter in the following formula:

$$\begin{aligned}&\text{logit}(\text{Probability of death})\\&\quad=\log(\text{odds of death})\\&\quad=a+b\times(\text{Age }\geq 70\text{ years})\\&\qquad+c\times(\text{Congestive heart failure is present})\\&\qquad+d\times(\text{Age }\geq 70\text{ years})\times(\text{Congestive heart failure is present})\end{aligned}$$

After maximum likelihood fitting of the interaction model we have the results in Table 10.16.

Table 10.16 shows that the interaction parameter *d* falls just short of significance, although as it is very close, we may still be interested in reporting the result. We need to take care in interpreting the above parameter estimates. The antilog of the *c* parameter for age group (e^c) is the OR for those with congestive heart failure compared with those without, within the reference category for age (patients <70 years). It does not represent the overall effect of congestive heart failure (indeed we have assumed there is no overall effect, because it is modified by age). To get the estimated OR for congestive heart failure for those 70 years of age, we add the parameters *c* and *d* together and then antilog to obtain 3.06. Equivalently we can multiply the OR associated with the reference category for age (8.28) by the OR calculated for the interaction parameter (0.37). We usually present model output involving an interaction as in Table 10.17. This table shows the separate ORs for each age group explicitly (which Table 10.16 does not), and the results of the test for interaction. Notice that no overall effects are given for variables involved in the interaction.

As a final example, Table 10.18 displays a model combining cardiogenic shock, age group and congestive heart failure, incorporating the effect modification of congestive heart failure by age group. To demonstrate the parameterization of the model, the model equation is given below.

Table 10.16 Parameter estimates, ORs and 95 % CIs from logistic regression of death, with age and presence of congestive heart failure (CHF) and interaction effects

Parameter		Value	Parameter estimate	OR	95 % CL for OR	*P* value
Intercept	*a*		−2.95			
Age group	Reference: <70 years					
	b	≥70 years	2.15	8.63	4.09, 18.21	<0.0001
CHF	Reference: Absent					
	c	Present	2.11	8.28	3.25, 21.08	<0.0001
Age × CHF	*d*	Reference: Age < 70 years or CHF absent				
		Age < 70 years and CHF present	−0.99	0.37	0.13, 1.07	0.066

Table 10.17 Logistic regression model of death within first year after an MI with age group and presence of congestive heart failure as predictors, allowing for effect modification

Parameter		Value	Parameter estimate	OR	95 % CL for OR	*P* value
Age group						
<70 years						
Congestive heart failure		Reference category: absent				
	c	Present	2.11	8.28	3.25, 21.08	<0.0001
≥ 70 years						
Congestive heart failure		Reference category: absent				
	c + *d*	Present	1.12	3.06	1.86, 5.05	<0.0001
Age × Congestive heart failure		Reference category: <70 years. No congestive heart failure				
≥70 years, No congestive heart failure	*d*		2.15	8.63	4.09, 18.21	<0.0001

$$
\begin{aligned}
&\text{logit}(\text{Probability of death}) \\
&\quad = \log(\text{odds of death}) \\
&\quad = a + b \times (\text{Cardiogenic shock is present}) \\
&\qquad + c \times (\text{Age } \geq 70 \text{ years}) \\
&\qquad + d \times (\text{Congestive heart failure is present}) \\
&\qquad + e \times (\text{Age } \geq 70 \text{ years}) \times (\text{Congestive heart failure is present})
\end{aligned}
$$

The way in which the effect of cardiogenic shock is presented has not changed, because it is not involved in an interaction. However, its value has reduced somewhat from its previous value (Table 10.7). This is because of the additional adjustment for congestive heart failure. In the presence of an interaction in the model, other coefficients will be adjusted for all combinations of the interacting variable (equivalent in this case to stratifying by age group and congestive heart failure simultaneously (four groups) and examining the cardiogenic shock effect within each).

Table 10.18 Logistic regression model of death within first year after an MI with cardiogenic shock, age group and presence of congestive heart failure as predictors, allowing for effect modification

Parameter		Value	Parameter estimate	OR	95 % CL for OR	*P* value
Cardiogenic shock		Reference category: absent				
	b	Present	1.27	3.57	1.26, 10.10	0.016
Age group						
<70 years						
Congestive heart failure		Reference category: absent				
	d	Present	2.10	8.13	3.18, 20.81	<0.0001
≥ 70 years						
Congestive heart failure		Reference category: absent				
	$d + e$	Present	1.04	2.83	1.70, 4.70	<0.0001
Age × Congestive heart failure		Reference category: <70 years, No congestive heart failure				
≥ 70 years, No congestive heart failure	e		2.17	8.75	4.13, 18.53	<0.0001

Likelihood ratio tests are available for each of the terms in our model. For the model in Table 10.18 these are given in Table 10.19.

P values for likelihood ratio tests in Table 10.19 are slightly different from those for parameter estimates given in Table 10.18; for example, the *P* value for the interaction term is $P = 0.066$ in Table 10.18 and 0.051 in Table 10.19. This is because these are estimated in different ways. The likelihood ratio tests are based on the likelihood function for the interaction model compared with the non-interaction model, whereas the *P* values for individual parameters are based on Wald statistics, which relate to the parameter estimates themselves and their standard errors. The likelihood ratio test is generally preferred for various statistical reasons, but both usually give similar answers. It is important to remember that calculation of each of these and indeed many *P* values is an approximate process that relies on large enough sample sizes and is based on assumptions that are sometimes slightly different.

Extensions and Variations of Logistic Regression

Case–Control Studies

Case–control studies address questions of associations between risk factors, commonly called exposures, and health outcomes. Typically a series of cases is first defined. These are persons experiencing the event of interest, for example, successful recovery from an illness. A series of controls is then chosen, according to criteria such that a selected control would have become a case, had he or she had the

Table 10.19 Likelihood ratio tests for logistic regression model of obesity, at the 21-year follow-up, with maternal smoking and child's exercise at age 14 years as predictors, with interaction effects

Source	df	Chi-square	*P* value
Cardiogenic shock	1	6.21	0.0127
Age group	1	44.12	<0.0001
Congestive heart failure	1	19.04	<0.0001
Age × Congestive heart failure	1	3.81	0.051

particular health outcome of interest. An example might be a series of patients experiencing a nosocomial infection during a hospital stay, with controls being chosen from other in-patients who did not experience an infection. In such a case, the variable indicating caseness (case/control) is used as the outcome variable and potential risk factors are included in the logistic regression model in the usual way. If the controls are matched in some way to the cases (e.g. by age, type of ward, admission diagnosis) then a technique called conditional logistic regression is needed to take the matching into account.

Multinomial and Ordinal Regression

The logistic regression model can be extended to the situation when the outcome variable has more than two categories (multinomial regression) and when these categories fall into a natural order (ordinal regression). These models are very similar to the logistic regression but allow the incorporation of additional hypotheses concerning these additional categories of outcome. In many instances it is possible to address the questions dealt with by these more complex models, by using a series of simpler logistic regressions.

Conclusion

Logistic regression is a very general model that can be used to analyse the determinants or predictors of a binary outcome arising in a process in which events are independent. Because of the nature of the logit transformation, the model gives rise to regression coefficients that are interpretable as log(ORs), which allows a useful interpretation, after exponentiation.

As with other regression models, multivariate models can be built up by including additional predictor variables, such that effects are mutually adjusted.

Logistic regression may be applied to continuous variables, or a mix of continuous and categorical variables. Detailed examination of relationships with continuous variables may be valuable in detecting curvilinear effects.

Caution must be exercised in interpreting ORs as RRs. When the outcome becomes more common (at least 15 %), this interpretation may be misleading.

Bibliography

Hosmer DW, Lemeshow S (2000) Applied logistic regression. Wiley, New York

Kirkwood BR, Sterne JAC (2003) Essential medical statistics, 2nd edn. Blackwell Science, Malden, Part C

Kleinbaum DG, Kupper LK, Muller KE (1988) Applied regression analysis and other multivariable methods, 2nd edn. PWS-Kent, Boston, Chapters 21–25

Tonne C, Schwartz J, Mittleman M, Melly S, Suh H, Goldberg R (2005) Long-term survival after acute myocardial infarction is lower in more deprived neighbourhoods. Circulation 111:3063–3070

Vittinghoff E, Glidden DV, Shiboksi SC, McCulloch CE (2005) Regression methods in biostatistics: linear, logistic, survival, and repeated measures models. Springer, New York, Chapter 6.1, 6.2