24
Testing hypotheses

The scientific imagination knows no bounds in the creation of theories and interesting models, but when should such elaboration end? The principle which is invoked to deal with this problem is *Occam's razor*. This principle holds that we should always adopt the simplest explanation consistent with the known facts. Only when the explanation becomes inconsistent are we justified in greater elaboration. Occam's razor has much in common with statistical tests of null hypotheses. Statisticians erect null hypotheses and seek positive evidence against them before accepting alternative explanations. This philosophical position should not be taken to imply that the absence of evidence against a null hypothesis establishes the null hypothesis as being true.

24.1 Tests involving a single parameter

An explanatory variable with two levels requires only one parameter to make a comparison between them. When the comparison is made using a rate ratio (or an odds ratio) the null value is 1.0, or zero on the log scale. The simplest way of testing for a zero null value is to use the Wald test, based on the profile log likelihood for the parameter being tested. This involves referring

$$\left(\frac{M-0}{S}\right)^2$$

to tables of the chi-squared distribution on one degree of freedom, where M is the most likely value of the log of the parameter and S is its standard deviation. These quantities are the ones listed in the computer output under estimate and standard deviation.

Exercise 24.1. Table 24.1 repeats the results of the regression analysis of the ischaemic heart disease data. Carry out the Wald test of the hypothesis of no effect of exposure on IHD incidence.

A log likelihood ratio test based on the profile likelihood for the exposure parameter can also be used to test the hypothesis in Exercise 24.1. The profile log likelihood ratio for a zero exposure effect is the difference between two log likelihoods: (a) the log likelihood when the exposure parameter is

Table 24.1. Program output for the ischaemic heart disease data

Parameter	Estimate	SD
Corner	−5.4180	0.4420
Exposure(1)	0.8697	0.3080
Age(1)	0.1290	0.4753
Age(2)	0.6920	0.4614

zero and the age parameters take their most likely values given that there is no exposure effect, and (b) the log likelihood evaluated when all parameters take their most likely values. The former is obtained by fitting a model which includes age but not exposure, and the latter is obtained by fitting a model which includes both age and exposure. The difference between these two log likelihoods gives the profile log likelihood ratio, and the test is carried out by referring minus twice this value to the chi-squared distribution with one degree of freedom. Some programs report the *deviance*, a quantity closely related to the log likelihood which we shall discuss in a later section of this chapter.

Exercise 24.2. The log likelihoods for the models

$$\begin{aligned} \log(\text{Rate}) &= \text{Corner} + \text{Age} + \text{Exposure} \\ \log(\text{Rate}) &= \text{Corner} + \text{Age} \end{aligned}$$

for the ischaemic heart disease data, are −247.027 and −251.176. How can you tell which likelihood was obtained for which model? Carry out the likelihood ratio test for a zero exposure effect and compare it with the Wald test calculated in the previous exercise.

The score test for a zero exposure effect is found from a quadratic approximation which has the same gradient and curvature as the profile log likelihood at the null value. Since the log likelihood ratio test is easy to obtain using a computer program the score test is rarely carried out, although some programs do offer this option.

24.2 Tests involving several parameters

When a variable has three levels two parameters are required to make comparisons between the levels. A test that just one of these parameters takes its null value is rarely of interest. The hypothesis that both take their null values is usually more relevant, because this corresponds to the variable having no effect on the response. We shall now consider the extension of the likelihood ratio test to cover this situation. A convenient example is provided by the problem of testing the effect of age in the analysis shown in Table 24.1, although this is a hypothesis of no scientific interest!

The same general principle as for one parameter is used: the log likelihood for the model

$$\text{Corner} + \text{Age} + \text{Exposure}$$

which includes the two age parameters, is subtracted from the log likelihood for the model

$$\text{Corner} + \text{Exposure},$$

in which the two age parameters are zero. This gives the log likelihood ratio for testing the hypothesis that both age parameters take their null values. Minus twice the log likelihood ratio is referred to the chi-squared distribution with *two* degrees of freedom, because two parameters have been set to their null values. In this case minus twice the log likelihood ratio is equal to 4.016, and the p-value is 0.134, showing that there is no significant effect of age on ischaemic heart disease in this study.

Exercise 24.3. Does the fact that there is no significant effect of age on incidence in this study mean that there is no need to control for age when comparing exposure groups?

There is some temptation to scan the output for the model which includes both age and exposure and to try to interpret the separate tests of the two parameters for age, rather than making a joint test. Using the Wald test with the results in Table 24.1 shows that the data support both null values for age when tested separately, but it would be unwise to deduce from this that there is no effect of age. This is because both age effects are rather imprecisely estimated, due to the fact that only 6 heart attacks were observed in the first age band. When the corner is located where there is very little data it is common to see effects for both levels 1 and 2 which are small compared to their standard deviations, yet a highly significant effect from level 1 to level 2. The only safe way of testing the effect of age is to make a test of the joint hypothesis that both age effects take their null value. The Wald test can be generalized to do this (as can the score test), but the easiest test to use is the log likelihood ratio test.

24.3 Testing for interaction

The regression model used in the test for an exposure effect imposes the constraint that the effect of exposure is constant over age bands. Similarly for the test for age effects. An important question to ask is whether it is reasonable to impose these constraints, or whether the data better support different exposure effects in each age band, and different age effects in each exposure group. When the effects of exposure vary with age there is said to be *interaction* between exposure and age. Interaction between exposure and age automatically implies interaction between age and exposure and vice versa.

Table 24.2. Definition of interactions in terms of exposure

		Exposure	
		0	1
	0	5.0	15.0
Age	1	12.0	42.0
	2	30.0	135.0
	0	5.0	5.0×3.0
Age	1	12.0	12.0×3.5
	2	30.0	30.0×4.5
	0	5.0	5.0×3.0
Age	1	12.0	$12.0 \times 3.0 \times 1.167$
	2	30.0	$30.0 \times 3.0 \times 1.5$

To test for interaction it is necessary to choose new parameters in a way that allows for separate effects of exposure in the different age bands. This is done by choosing one parameter to measure the effect of exposure in the first age band and two to measure the extent to which the effects of exposure in the other two age bands differ from the effect in the first age band. The way this is done is best illustrated using numerical values for the parameters.

A set of illustrative values for the 6 rate parameters are shown at the top of Table 24.2. The rate ratios for exposure by levels of age are 3.0, 3.5, and 4.5, shown in the middle part of the table, so these rate parameters do not obey a multiplicative model. The extent of the departure from the multiplicative model can be measured by expressing 3.5 and 4.5 as ratios relative to 3.0, as shown in the third part of the table. These ratios, which take the values 1.167 and 1.5 in this case, are called *interaction* parameters.

Table 24.3 shows the same thing in terms of the rate ratios for age by levels of exposure. These rate ratios are 2.4 and 6.0 when exposure is at level 0 but 2.8 and 9.0 when exposure is at level 1. The extent to which these differ, measured as ratios relative to the rate ratios at level 0 of exposure, are again equal to 1.167 and 1.5. Thus the interaction parameters are symmetric in exposure and age.

Tables 24.2 and 24.3 are combined in Table 24.4. Using the terminology of regression models, the 6 original rate parameters are re-expressed in terms of the corner, the rate ratio for exposure when age is at level 0, the rate ratio for age when exposure is at level 0, and the two interaction parameters. This way of re- expressing the original rate parameters has not resulted in any reduction in the number of parameters; its sole purpose is to assess the extent of the departures from the multiplicative model. We

Table 24.3. Definition of interactions in terms of age

		Exposure	
		0	1
	0	5.0	15.0
Age	1	12.0	42.0
	2	30.0	135.0
	0	5.0	15.0
Age	1	5.0×2.4	15.0×2.8
	2	5.0×6.0	15.0×9.0
	0	5.0	15.0
Age	1	5.0×2.4	$15.0 \times 2.4 \times 1.167$
	2	5.0×6.0	$15.0 \times 6.0 \times 1.5$

Table 24.4. Definition of interactions in terms of exposure and age

	Exposure	
Age	0	1
0	5.0	5.0×3.0
1	5.0×2.4	$5.0 \times 3.0 \times 2.4 \times 1.167$
2	5.0×6.0	$5.0 \times 3.0 \times 6.0 \times 1.5$

shall write the model with interaction in one or other of the forms

$$\begin{aligned} \text{Rate} &= \text{Corner} \times \text{Exposure} \times \text{Age} \times \text{Exposure·Age} \\ \log(\text{Rate}) &= \text{Corner} + \text{Exposure} + \text{Age} + \text{Exposure·Age}. \end{aligned}$$

To test for interaction it is necessary to fit the model with and without interaction parameters and to measure the log likelihood ratio for these two models. Minus twice this log likelihood ratio is then referred to tables of chi-squared on two degrees of freedom. The chi-squared has two degrees of freedom because the hypothesis being tested is that two interaction parameters take their null values. The instruction to include interaction parameters is done by including the term Age·Exposure in the model description. When this is done the output will include estimated values for the interaction parameters, but these are rarely of much use because they are chosen specifically to make the test for no interaction. If there is interaction then it will usually be best to report the effects of exposure separately for each age band. If there is no interaction then the effects of exposure and age should be obtained from the model without interaction parameters. Further details on how to report interactions are given in Chapter 26.

Table 24.5. Estimates of parameters in the model with interaction

Parameter	Estimate	SD
Corner	−5.0237	0.500
Exposure(1)	−0.0258	0.866
Age(1)	−0.5153	0.671
Age(2)	0.3132	0.612
Age(1)·Exposure(1)	1.2720	1.020
Age(2)·Exposure(1)	0.8719	0.973

Table 24.5 shows the output for the ischaemic heart disease data when fitting the model which includes the interaction between exposure and age. The interaction parameters are given names like Age(1)·Exposure(1) and Age(2)·Exposure(1). In general the number of interaction parameters between a variable on a levels and one on b levels is $(a-1)(b-1)$.

Exercise 24.4. Verify from Table 24.5 that the estimated corner parameter in the model with interaction is now the log of the observed rate for unexposed subjects in age band 0, and the estimated Exposure(1) parameter is now the observed rate ratio (exposed/unexposed) in age band 0. (The observed rates are in Table 22.6.)

24.4 Deviance

The log likelihood for a regression model, evaluated at the most likely values of the parameters, is a measure of *goodness-of-fit* of the model — the greater the log likelihood, the better the fit. Since the absolute value of the log likelihood is not itself of interest there is some advantage in always reporting a log likelihood ratio, compared to some other model. A convenient choice is the *saturated* which includes the maximum possible number of parameters. The output would then include the log likelihood ratio between the model being fitted and the saturated model. For use with tables of chi-squared it is slightly more convenient to report minus twice the log likelihood ratio, a quantity which is called the *deviance* for the model being fitted. Each deviance has degrees of freedom equal to the difference between the number of parameters in the model and the number in the saturated model.

The deviance is a measure of badness of fit; the larger the deviance the worse the fit. Two models are compared by comparing their deviances. The change in deviance is minus twice the log likelihood ratio for the two models because the log likelihood for the saturated model occurs in both deviances and cancels (see Fig. 24.1.) The degrees of freedom for this test are found by subtracting the degrees of freedom for the two deviances. For

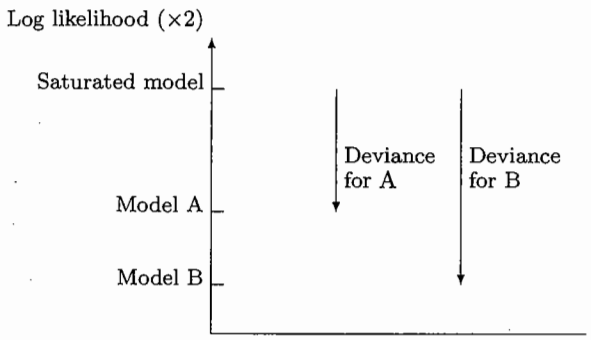

Fig. 24.1. Relationship between deviance and log likelihood

example, when fitting the models

$$\begin{aligned} \log(\text{Rate}) &= \text{Corner} + \text{Age} + \text{Exposure} \\ \log(\text{Rate}) &= \text{Corner} + \text{Exposure}, \end{aligned}$$

to the ischaemic heart disease data the corresponding values for the two deviances were 1.673 and 5.689. The difference between these is 4.016 which is the same as the result obtained earlier in the chapter for minus twice the log likelihood ratio.

Exercise 24.5. How do you know which deviance was obtained for which model? How many degrees of freedom do the two deviances have?

When the data are entered as frequency records the saturated model has the same number of parameters as there are frequency records. In the case of the ischaemic heart disease data there are six records so the saturated model has 6 parameters. All models with six parameters are saturated and have the same log likelihood. The model which includes the interaction parameters between age and exposure has six parameters, and is saturated, so it follows that the deviance for the model

$$\log(\text{Rate}) = \text{Corner} + \text{Age} + \text{Exposure}$$

provides a test of no interaction between age and exposure. It may be referred directly to a chi-squared distribution with two degrees of freedom.

When the data are entered as individual records the saturated model has the same number of parameters as the number of individual records and the deviance measures minus twice the difference between the log likelihood for the fitted model and this saturated model. This is not a test of anything useful. There is no short cut for making a test of no interaction using individual records: it is necessary to obtain the deviances for the models

Table 24.6. Cases (controls) for oral cancer study

	Alcohol							
Tobacco	0		1		2		3	
0	10	(38)	7	(27)	4	(12)	5	(8)
1	11	(26)	16	(35)	18	(16)	21	(20)
2	13	(36)	50	(60)	60	(49)	125	(52)
3	9	(8)	16	(19)	27	(14)	91	(27)

Table 24.7. Case/control ratios for the oral cancer data

	Alcohol			
Tobacco	0	1	2	3
0	0.26	0.26	0.33	0.63
1	0.42	0.46	1.13	1.05
2	0.36	0.83	1.22	2.40
3	1.12	0.84	1.93	3.37

with and without the interaction parameters.

24.5 Models with two exposures

Because regression models treat all explanatory variables in the same way, models for studies with two exposures look very similar to models for studies with one exposure and one confounder. However, there are some differences in the way different hypotheses are interpreted.

Table 24.6 repeats the study of oral cancer introduced in Chapter 16, in which the numbers of cases and controls are tabulated by two exposures, alcohol consumption (on four levels) and tobacco consumption (also on four levels). For alcohol the levels are 0, 0.1–0.3, 0.4–1.5, and 1.6+ ounces per day (coded as 0, 1, 2, and 3). For tobacco the levels are 0, 1–19, 20–39, and 40+ cigarettes per day (also coded as 0, 1, 2, and 3). A summary table of case/control ratios by alcohol and tobacco is shown in Table 24.7. Because the frequencies in the table are small, there is a lot of random variation, but there is an overall tendency for the ratios to increase both from left to right along rows, and from top to bottom down columns. This indicates that *both* variables have an effect on cancer incidence; there is an effect of tobacco when alcohol intake is held constant, and vice versa.

An important question is whether the two exposures act independently of one another. In other words, are the effects of tobacco the same at all levels of alcohol, and are the effects of alcohol the same at all levels of tobacco? This question is answered by testing for no interaction between alcohol and tobacco, but it must be emphasized that the test depends on

how the effect parameters are defined. When they are defined as ratios the interaction parameters are also ratios and measure departures from a model in which the two exposures combine multiplicatively. By choosing to measure effects as ratios we have therefore chosen to interpret independent action as meaning that the two exposures act multiplicatively. In Chapter 28 we show how the effects can be defined as differences, in which case the interaction parameters are also differences and measure departures from a model in which the two exposures combine additively. In this case we have chosen to interpret independent action as meaning the two exposures act additively.

If there is a significant interaction then it will be necessary to report the effects of alcohol separately as odds ratios for each level of tobacco consumption, and the effects of tobacco separately as odds ratios for each level of alcohol. On the other hand, if there is no significant interaction then the two exposures may be assumed to act independently and we can estimate the effects of alcohol controlled for tobacco and the effects of tobacco controlled for alcohol. Note that even when the two exposures act independently it is still necessary to control each for the other. This is because people's drinking and smoking habits are not independent so ignoring one when studying the other could lead to biased estimates.

The test for no interaction is carried out by comparing the fit of the multiplicative model

$$\log(\text{Odds}) = \text{Corner} + \text{Alcohol} + \text{Tobacco},$$

with that of the model which includes the interaction parameters,

$$\log(\text{Odds}) = \text{Corner} + \text{Alcohol} + \text{Tobacco} + \text{Alcohol} \cdot \text{Tobacco}.$$

Since the second of these models is saturated the test can be based directly on the deviance for the multiplicative model. Provided the data support the hypothesis of no interaction it is then possible to test for an effect of alcohol, controlled for tobacco, by comparing the models

$$\begin{aligned} \log(\text{Odds}) &= \text{Corner} + \text{Alcohol} + \text{Tobacco} \\ \log(\text{Odds}) &= \text{Corner} + \text{Tobacco}. \end{aligned}$$

Similarly the test for an effect of tobacco is made by comparing the models

$$\begin{aligned} \log(\text{Odds}) &= \text{Corner} + \text{Alcohol} + \text{Tobacco} \\ \log(\text{Odds}) &= \text{Corner} + \text{Alcohol}. \end{aligned}$$

In each of these tests the smaller of the two models being compared is obtained from the larger by setting some parameters to zero. The smaller

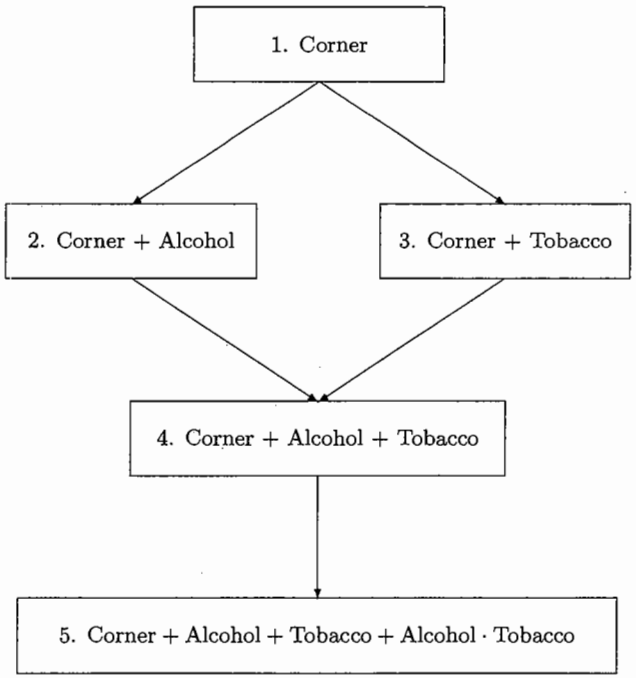

Fig. 24.2. Nesting of models.

model is then said to be *nested* in the larger model. Comparisons between models where neither is nested in the other are not allowed since they do not correspond to a hypothesis in which some parameter values are set equal to zero. Fig. 24.2 shows the five possible models which could be fitted to the alcohol and tobacco data. The arrows indicate nesting so any two models joined by an arrow correspond to a hypothesis which can be tested. For example, a comparison of models 4 and 5 is a test of no interaction, and a comparison of models 4 and 2 is a test of no effect of tobacco (controlling for alcohol). In model 1 both alcohol and tobacco parameters are set to zero so it is nested in all of the other models.

Exercise 24.6. For the models set out in Fig. 24.2, the deviances are (1) 132.561, (2) 37.951, (3) 61.880, and (4) 6.689. What are the degrees of freedom associated with each of these deviances? Carry out the four tests corresponding to the arrows in the figure. What is the interpretation of these tests?

24.6 Goodness-of-fit tests

A question which is often asked is whether a model provides an adequate fit to the data. Because the absolute value of the log likelihood has no meaning this question can only be answered by comparing the model with other more complicated models and asking whether the extra complication is justified. The saturated model represents the most complicated model which could be used and the deviance automatically provides a comparison of the model currently being fitted with the saturated model. For this reason the deviance for a model is often put forward as a test of goodness of fit (really badness-of-fit) of the model. There are several cautions which need to be borne in mind when interpreting the deviance in this way.

1. Comparisons with the saturated model are meaningless when the data are entered as individual records.

2. Comparisons with the saturated model which are on many degrees of freedom will lack power to discriminate; in this case it will be better to make comparisons with models which are less complicated than the saturated model.

3. The deviance is only approximately distributed as chi-squared and this approximation gets worse as the degrees of freedom increase.

24.7 Collinearity

In a study in which tobacco and alcohol consumption were very highly associated it would be very difficult to make an estimate of the effects of alcohol controlled for tobacco (or of the effects of tobacco controlled for alcohol). This is because controlling for tobacco involves fixing the level of tobacco consumption and then estimating the effects of alcohol from subjects whose tobacco consumption is at this level. If alcohol and tobacco are highly associated then nearly all subjects at a fixed tobacco level will have the same level of alcohol consumption and it will therefore be difficult to estimate the effects of alcohol. In extreme cases fixing the level of tobacco might fix the level of alcohol completely, in which case it would be impossible to estimate the effects of alcohol. In such a case the two variables are said to be *collinear*. This situation is not uncommon, particularly when working with derived variables.

Solutions to the exercises

24.1 In the Wald test $(0.8697/0.3080)^2 = 7.97$ is referred to the chi-squared distribution with one degree of freedom, giving a p-value of 0.005.

24.2 The larger likelihood, -247.027, corresponds to the first model because this has more parameters than the second. The log likelihood ratio for the two models is $-251.176 - (-247.027) = -4.149$. Minus twice this is 8.298 which is quite close to the Wald chi-squared value obtained in the

previous exercise. Referring 8.30 to the chi-squared distribution with one degree of freedom gives $p = 0.004$.

24.3 No. When taking account of confounding variables it is best to play safe and to control for them regardless of whether their effects are significant or not. Very little is lost by doing this.

24.4 The Corner, Exposure(1), Age(1) and Age(2) parameters are

$$\begin{aligned} \log(6.580/1000) &= -5.0237 \\ \log(6.412/6.580) &= -0.0258 \\ \log(3.931/6.580) &= -0.5153 \\ \log(9.00/6.58) &= 0.3132. \end{aligned}$$

24.5 The smaller deviance corresponds to the larger model since this will be a better fit. The degrees of freedom are 2 and 4 respectively.

24.6 The number of parameters in models 1 to 5 are 1, 4, 4, 7, and 16, respectively. The number of parameters in the saturated model is 16, so the degrees of freedom for the deviances are $16-1=15$, $16-4=12$, $16-4=12$, $16-7=9$, and $16-16=0$ respectively. Note that model 5 has 16 parameters so it is saturated. The table below shows the comparisons of models in terms of the change in deviance.

Comparison	Change in deviance	Change in df
(1) vs (2)	$132.56 - 37.95 = 94.61$	$15 - 12 = 3$
(1) vs (3)	$132.56 - 61.88 = 70.68$	$15 - 12 = 3$
(2) vs (4)	$37.95 - 6.69 = 31.26$	$12 - 9 = 3$
(3) vs (4)	$61.88 - 6.69 = 55.19$	$12 - 9 = 3$
(4) vs (5)	$6.69 - 0 = 6.69$	$9 - 0 = 9$

The last of these comparisons shows that there is no significant interaction. This means that the next two comparisons (working up from the bottom) make sense. The change in deviance from model 3 to model 4 shows that there is a significant effect of alcohol after controlling for tobacco; similarly the change in deviance from model 2 to model 4 shows that there is a significant effect of tobacco after controlling for alcohol. All of the models can be compared with model 1, but these comparisons have little interest. For example, a comparison of model 1 with model 2 is a test of the alcohol effects (ignoring tobacco) while a comparison of model 1 with model 4 is a joint test of the alcohol effects (controlling for tobacco) *and* the tobacco effects (controlling for alcohol).