Solutions to the exercises

26.1 The multiplicative effect of work is the ratio of the prevalence odds for non-agricultural workers to the prevalence odds for agricultural workers.

26.2 The degrees of freedom for the deviances are

$$\begin{aligned} 20-(1+4+1+1) &= 13 \\ 20-(1+4+1+1+4) &= 9 \\ 20-(1+4+1+1+1) &= 12 \end{aligned}$$

The change of deviance with inclusion of the Work.Age interaction is 1.84 with 4 degrees of freedom, and for the Work.Sex interaction it is 0.41 with 1 degree of freedom. Neither is significant.

26.3 The change in log odds over the age range of 35 to 75 is approximately +4. The gradient is therefore approximately +1 per 10 year age band.

26.4 The Wald test for interaction between the linear effect of age and work is

$$\left(\frac{0.053}{0.188}\right)^2 = 0.079,$$

which is not significant.

26.5 There would be two parameters for this interaction term.

27
Choice and interpretation of models

Previous chapters have illustrated the use of regression models using simple bodies of data containing relatively few variables. More commonly, we are faced with large data files containing many variables. Sometimes derived variables such as Quetelet's weight-for-height index are included in the model in addition to or in place of the original variables. In such situations it can be difficult to know where to begin, and all too easy to lose one's way. This chapter offers some guidance towards the sensible use of regression methods.

27.1 Variable selection strategies

A lot has been written about the process of finding the 'best' regression model in problems involving many variables. Much of this activity has been concerned with the search for an optimal strategy, and the relative merits of different approaches have been hotly debated. Many computer programs implement one or more of these strategies in an automatic model selection option called *stepwise regression*. These programs usually work by a combination of the *step-up* strategy (examining the effect of inclusion of variables not yet in the model) and the *step-down* strategy (examining the effect of of removing variables currently in the model). With the recent increased speed and reduced cost of computers, some programs now offer an exhaustive search of *all subsets* from a list of possible explanatory variables.

In assessing the value of such procedures it is important to note that regression models have two very different uses in epidemiology. Historically they were first used to derive *risk scores* designed to classify subjects into graded categories with respect to risk of developing disease. Later, when attention turned to interpretation of the parameter estimates and the close relationship between regression and stratification methods became apparent, regression models became important tools for analyses whose aim was the advancement of scientific knowledge. For convenience we refer to these two uses as *prediction* and *explanation*, respectively.

When the aim is prediction, the best model is the one which best predicts the fate of a future subject. This is a well defined task and automatic strategies to find the model which is best in this sense are potentially use-

ful. However, when used for explanation the best model will depend on the scientific questions being asked, and automatic selection strategies have no place.

An important tool for assessing how well a model predicts the fate of a future subject is *cross-validation* — a technique in which each subject in turn is removed from the dataset and the actual outcome for that subject is compared with the predicted outcome using the model based on the remaining observations. The deviance for a model will always decrease with the introduction of more parameters, but prediction of future observations is not always improved. There comes a point at which increasing the complexity of the model to gain a slightly better fit to the observed data will reduce the accuracy of its predictions. Cross-validation measures the predictive properties of the model directly and therefore reflects the adverse consequences of fitting too many parameters.

Cross-validation is potentially expensive in computer time, but simple approximate criteria have been developed which allow the assessment of whether any step up or down in an automatic model selection procedure would be expected to improve prediction. The best known is *Akaike's information criterion*, namely

$$\text{(Reduction in deviance)} - 2 \times \text{(Increase in number of parameters)}.$$

If this is positive the increased complexity would be expected to improve prediction and if negative, to degrade prediction.

27.2 Explanatory variables and natural experiments

This book has been entirely concerned with the use of models whose aim is explanation. In such analyses there is a clear distinction between the roles of exposures and confounders but this distinction is lost when using regression models — both become explanatory variables. Ignoring the distinctions between different types of explanatory variable is appropriate when using regression models for prediction, since all variables have the same role, but in a scientific analysis of data different explanatory variables may play quite different roles.

The distinction between exposure and confounder, as described in this book, relies heavily on the idea of experiments of nature. An exposure is something which we can intervene to change while a confounder is a variable which we would have held constant had we designed the experiment rather than leaving it to nature. It is helpful to think of regression analysis as simulating an experiment, in the same way. For example, the effects of A in the model

$$\log(\text{Rate}) = \text{Corner} + \text{A} + \text{B} + \text{C}$$

are the effects of changing the level of A in a simulated experiment in

which B and C are held constant. Similarly, the effects of B are the effects of changing the level of B in a simulated experiment in which the levels of A and C are held constant. Thus regression analysis does not simulate a single experiment but many. This flexibility of the regression approach is undoubtedly useful, but in practice it can also become its most serious weakness. To extend our analogy, the data analyst is in a position like that of an experimental scientist who has the capability to plan and carry out many experiments within a single day. Not surprisingly a cool head is required! Before embarking on a regression analysis it is essential to spend an hour or so, preferably away from the computer, to list the main scientific questions and to think how these can be answered by fitting a series of models. Analyses which follow such thought are always simpler and more incisive than those which are born of uncritical use of the computer or worse, of a stepwise regression program.

It will rarely be necessary to include a large number of variables in the analysis, because only a few exposures are of genuine scientific interest in any one study, and there are usually very few variables of sufficient *a priori* importance for their potential confounding effect to be controlled for. Most scientists are aware of the dangers of analyses which search a long list of potentially relevant exposures. These are known as *data dredging* or *blind fishing* and carry a considerable danger of false positive findings. Such analyses are as likely to impede scientific progress as to advance it. There are similar dangers if a long list of potential confounders is searched, either with a view to explaining the observed relationship between disease and exposure or to enhancing it — findings will inevitably be biased. Confounders should be chosen *a priori* and not on the basis of statistical significance. In particular, variables which have been used in the design, such as matching variables, must be included in the analysis.

Recently there has been some dispute between 'modellers', who support the use of regression models, and 'stratifiers' who argue for a return to the methods described in Part I of this book. Logically this dispute is based on a false distinction — there is no real difference between the methods. In practice the difference lies in the inflexibility of the older methods which thereby imposes a certain discipline on the analyst. Firstly, since stratification methods treat exposures and confounders differently, any change in the role of a variable requires a new set of computations. This forces us to keep in touch with the underlying scientific questions. Secondly, since strata must be defined by cross classification, relatively few confounders can be dealt with and we are forced to control only for confounders of *a priori* importance. These restraints can be helpful in keeping a data analysis on the right tracks but once the need for such discipline is recognized, there are significant advantages to the regression modelling approach.

EXAMPLE: DIETARY FAT AND TOTAL ENERGY INTAKE

The analogy between regression models and imaginary experiments is very useful in making decisions about whether to include a variable in a regression model or not. An interesting illustration arises in nutritional epidemiology when considering the relationship between total energy intake and the incidence of coronary heart disease. This relationship was first detected because relationships were observed between intake and disease risk for a large number of nutrients — the more that was eaten, the lower the risk. A relationship with total energy intake, possibly reflecting energy expenditure, was considered the most likely explanation.

However, once this relationship is recognized, how should the relationship between risk and other aspects of the diet, notably fat intake, be analysed? One way is to measure *nutrient density*, which is the ratio of daily intake of fat to the total energy intake. This approach is open to the criticism that such nutrient densities are not usually independent of total energy intake — subjects with high energy intakes typically have a different pattern of nutrient densities from subjects with low energy intakes.

If energy intake is to be regarded as a confounder, then it should be controlled for, either by stratification or with a regression model. In the latter case we fit a model such as

$$\log(\text{Rate}) = \text{Corner} + \text{Fat} + \text{Energy}$$

and interpret the parameters representing the effect of fat in terms of an experiment in which fat intake is varied but the total energy content of the diet is held constant. Of course, such an experiment would require other constituents of the diet such as carbohydrate to vary in order to maintain the total energy intake and this must be born in mind when interpreting parameters.

Exercise 27.1. How would you interpret the effect of fat in the model

$$\log(\text{Rate}) = \text{Corner} + \text{Fat} + \text{Carbohydrate} + \text{Energy}?$$

Other authors have approached the problem of allowing for total energy expenditure by dividing total calories between calories from fat and calories from other sources, and fitting the model

$$\log(\text{Rate}) = \text{Corner} + \text{Fat-calories} + \text{Other-calories}.$$

The parameters representing the effect of fat intake must now be interpreted in terms of an experiment in which fat intake is varied while intake of other calories is held constant. In this experiment a reduction of fat intake would result in a reduction of total energy intake. Such an experiment would be difficult to interpret, even if it could be carried out.

Finally we should point out that a real public health intervention to reduce dietary fat intakes would be unlikely to mimic either of the above imaginary experiments. When dietary fat intake is reduced in free–living subjects, some of the energy intake is made up from other sources, but typically there is a net reduction in energy intake. This demonstrates that the use of models to predict the effect of intervention usually requires considerable extra knowledge. In particular, we need to have some understanding of the mechanism by which change will be effected.

27.3 Endogenous and exogenous explanatory variables

The 'effects' of an explanatory variable are defined in terms of differences in log rate (or log odds) between groups of subjects with different levels of the variable. Thus the effect of cigarette smoking is defined by contrasting rates in smokers and non-smokers, and the effect of serum cholesterol concentration (classified as high or low) is defined as the difference in log rate between subjects with high cholestrol concentration and subjects with low cholesterol concentration. This language encourages people to interpret 'effects' as the change in rates to be expected as a result of intervention to change the level, but this is a big step. How are the subjects to alter their level? For a variable like serum cholesterol there is no direct way to alter its level and any intervention would have to be indirect, for example by change of diet or by cholesterol lowering drugs. However, there is no guarantee that such mechanisms will bear any relationship to the mechanism which led the the study subjects to have different levels in the first place. The effect of indirectly changing the levels of serum cholesterol in a group of subjects may be completely different from that estimated by comparing groups of subjects who just happen to have different levels of cholesterol.

The same problem arises in an even more acute form when studying the effects of two or more interrelated variables, such as blood pressure and obesity in relation to the incidence of coronary disease. The effect of blood pressure controlled for obesity might now be interpreted as the expected effect of changing blood pressure while keeping obesity constant. However, is it be possible to intervene to change blood pressure while keeping obesity constant? While this could be achieved, for example by using drug treatment, this method of intervention would bear little relation to the mechanism that led subjects to their current levels in the first place, and it might have different effects. Intervention aimed at life style changes are more likely to duplicate these conditions, but might be expected to change both blood pressure and obesity simultaneously. In this case the estimated effects of blood pressure controlled for obesity, or obesity controlled for blood pressure could be poor predictors of the effect of the intervention.

The position is much clearer when considering environmental exposures, such as radiation dose, occupational exposure to toxic chemicals, and even

cigarette smoking. In such cases, it is entirely reasonable to imagine an experiment in which exposure of groups is directly varied without any consequent change in other variables, and the parameters of regression models are easier to interpret.

Variables such as cholesterol concentration, blood pressure, and obesity are called *endogenous*. The word endogenous means 'growing from within'. Variables such as smoking, diet and occupation are called *exogenous*. The distinction between endogenous and exogenous variables is borrowed from the behavioural sciences and, although the distinction is not hard and fast, is useful in drawing attention to the different assumptions which it is necessary to make for the two kinds of variable when interpreting the parameters of regression models as expected effects following intervention.

27.4 Interpretation of interaction

An underlying theme of this chapter is that while distinctions between different types of explanatory variable are not relevant to the mechanical process of estimating the parameters of a regression model, they are essential to the strategy adopted in the analysis and to the interpretation of results. This is particularly true when dealing with interaction. The word describes a purely mathematical concept in regression models. Its relationship to the scientific language of epidemiology requires further consideration of the nature of the variables involved.

We shall first consider interaction between two confounders. There seems to be no word to describe this in epidemiology, almost certainly because the phenomenon is of no scientific interest. Whether we include such terms in a model or not is a purely technical matter of trading the number of parameters against freedom from assumptions. Usually if there are two strong confounders such as age and sex, the gain in efficiency from assuming no interaction between them is extremely modest and it will usually be safer to include an interaction term regardless of its significance. However, if we are worried about the aggregate effect of five or six weak confounders, then omission of interaction terms is unlikely to have a major effect on estimates of parameters of interest.

Interaction between a confounder and an exposure of interest is known in epidemiology as *effect modification* and is clearly of considerable scientific importance, since the *consistency* of an effect in diverse study groups would usually be considered relevant to labelling a relationship as 'causal', in the sense of predicting the effect of future interventions. The ease with which we can test for such interaction in the framework of regression models represents a clear advance over earlier stratification methods in which the absence of such interaction is a hidden assumption.

Finally, the question of interaction between two exposures of interest is usually of considerable importance, both for the scientific interpretation of

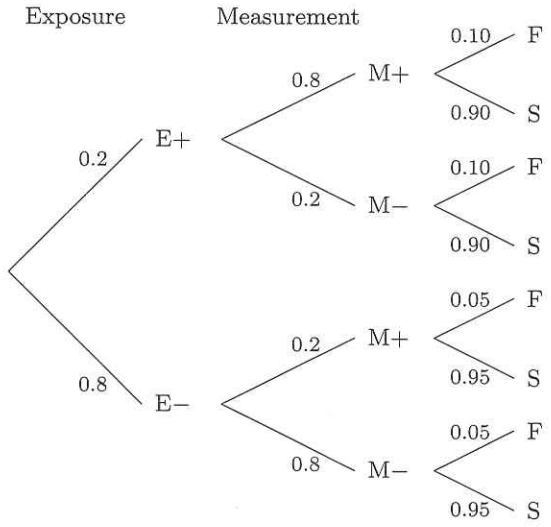

Fig. 27.1. Misclassification of exposure.

an analysis and for its implications for preventive intervention. We shall deal with this in more detail in Chapter 28.

27.5 Errors of measurement of explanatory variables

In the models discussed in this book it is assumed that explanatory variables are correctly measured. This assumption is often unjustified in practice, but epidemiologists have generally been prepared to ignore measurement errors. Some have believed that to do so is justifiable providing there is no relationship between errors of measurement of exposure and disease outcome, that is if there is no *differential misclassification*. This is now known to be false.

To illustrate the effect of ignoring measurement error we consider the hypothetical situation illustrated in Fig. 27.1, in which exposure E is measured imperfectly by measurement M. As a result of this misclassification there is a probability of 0.2 that an exposed subject is misclassified as unexposed, and a probability of 0.2 that an unexposed subject is misclassified as exposed. The probability of failure depends only on true exposure, taking the value 0.1 for exposed subjects and 0.05 for unexposed subjects. An epidemiological study observes only the marginal relationship between measured exposure and failure.

Exercise 27.2. Calculate probabilities for each of the eight tips of the tree in Fig. 27.1. By collapsing over exposure categories, calculate the probabilities for each of the four possible combination of measured exposure and disease (failure) status. Hence derive the probability tree expressing the probability of failure

Table 27.1. Diastolic blood pressure (DBP) and rate ratios for stroke

Baseline DBP	Rate ratio	Mean DBP at baseline	Mean DBP after 2 years
≤ 69	0.276	63.6	72.7
70–79	0.395	73.8	77.0
80–89	0.595	83.6	83.0
90–99	1.000	93.5	91.2
100–109	1.904	103.4	99.2
≥ 110	3.875	116.4	107.3

conditional upon measured exposure.

It is clear from this exercise that the effect of exposure is decreased by the measurement error: whereas the risk ratio for true exposure is 2, the risk ratio for measured exposure is only 1.42. It is worth noting that 20% misclassification would be regarded as acceptable in many branches of epidemiology.

Similar considerations apply when exposure takes on more than two levels. The observed dose-response relationship between measured exposure and disease outcome is less steep than the underlying relationship with true exposure, under any realistic assumptions about the dose-response relationship. This is illustrated by the data of Table 27.1 which concern the relationship between diastolic blood pressure and subsequent incidence of stroke.* These data are taken from a re-analysis of seven cohort studies, and the first two columns of the table summarize the relationship between diastolic blood pressure at a single initial visit (the 'baseline' measurement) and subsequent incidence. Note that in the rate ratios the fourth category is taken as reference. These were obtained by fitting the model

$$\log(\text{Rate}) = \text{Corner} + \text{Study} + \text{DBP}$$

where study is a categorical variable with one level for each study, so that confounding of the relationship due to differences between the study cohorts is eliminated. The third column shows the mean of the baseline diastolic pressures for each of the five categories. The log rate ratios are plotted against the mean baseline values in Fig. 27.2 (solid line). This line represents the apparent dose-response relationship between a single measurement of diastolic blood pressure and the incidence of stroke. It is approximately log-linear, so that essentially the same relationship would have been obtained by fitting the model

$$\log(\text{Rate}) = \text{Corner} + \text{Study} + [\text{DBP}],$$

*From Macmahon, S. *et al.* (1990), *The Lancet*, **335**, 765–774.

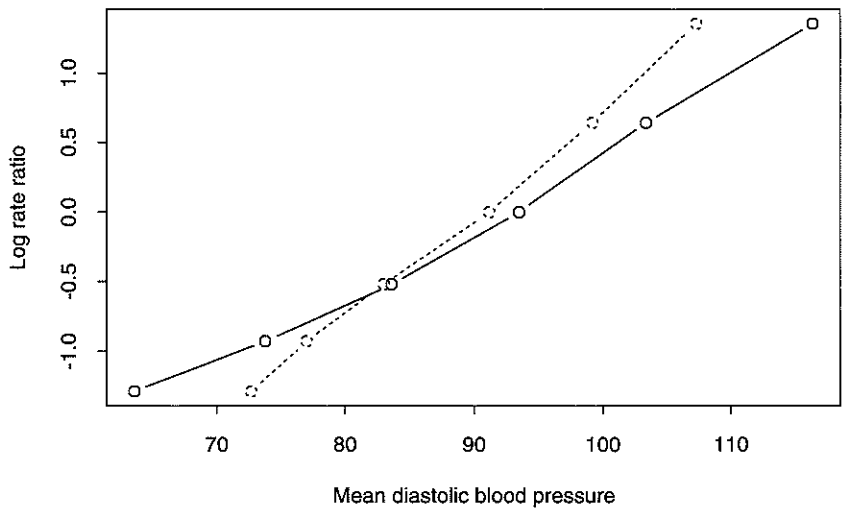

Fig. 27.2. Apparent and true dose-response relationships.

where [DBP] is measured per mm Hg. However, this line is a poor representation of the true relationship between blood pressure and the incidence of stroke. Blood pressure is subject to both short-term fluctuations and to measurement errors, neither of which will be reflected in the risk of stroke which is determined by the longer-term average level of blood pressure. The final column of Table 27.1 shows the mean blood pressure taken two years later in representative samples taken from each of the five groups. These figures provide a better estimate of long-term average blood pressure in the six groups as the short-term fluctuations and measurement errors are washed out. Plotting the rate ratios for stroke against these new values for mean diastolic blood pressure provides a truer estimate of the relationship between stroke incidence and the long- term average level of diastolic pressure. This plot is shown in Fig. 27.2 as a broken line and clearly represents a stronger relationship than the apparent relationship based on a single baseline measurement. This finding is true in general. When an explanatory variable suffers from measurement error or within subject variability the linear effects of this variable will be closer to zero than when there is no error or variability. This is known as *regression dilution*

This second example demonstrates both the attenuation of relationships owing to exposure measurement error and one of the methods which has been suggested for correcting for it. An alternative approach is to formally adopt probability models such as that illustrated in Fig. 27.1 and to estimate the conditional probabilities for every branch of the tree. Validation

substudies are required in order to estimate the misclassification probabilities. A difficulty with this approach is that when there are several levels of exposure, the number of parameters in the model can become very large.

In summary, when exposures are subject to measurement error, the apparent exposure effects will be less pronounced than the true underlying relationships. When confounders are measured inaccurately, the consequences are even more serious. Since the relationship between disease and confounder is not correctly estimated in these circumstances, it follows that the analysis will not properly control for confounding. If both exposure and confounder are measured inaccurately, there exists the possibility that the two sets of errors may be interrelated, so that the apparent relationship between exposure and confounder may be quite different from that between the underlying variables. In these circumstances models for relationships between measured exposure, measured confounder, and response have no interpretation in terms of an imaginary experimental intervention and may be scientifically meaningless. Such might well be the position in our example involving dietary fat and total energy intake. Measured intakes of total energy and of each specific nutrient are usually derived from the same dietary records, taken over a period of several days. Not only are such measurements very imperfect measures of long-term intake, but it is reasonable to believe that errors in the measured fat intake will be closely related to errors in measured energy intake, since the former is an important contributor to the latter. Regression models which include total energy as well specific nutrients may, therefore, not be interpretable in practice.

Solutions to the exercises

27.1 The parameter(s) measure the effect of changes in fat intake while holding both total energy intake and carbohydrate intake constant. To reduce fat intake while holding both total energy and carbohydrate intake constant would be very difficult for an individual to do and would require large changes in other components of the total energy intake, such as protein.

27.2 From top to bottom the probabilities are 0.016, 0.144, 0.004, 0.036, 0.008, 0.152, 0.032, and 0.608. The remaining calculations are shown in Fig. 27.3. The probability of failure conditional upon having been measured as exposed is 0.075, while the failure probability conditional upon having been measured as unexposed is 0.053.

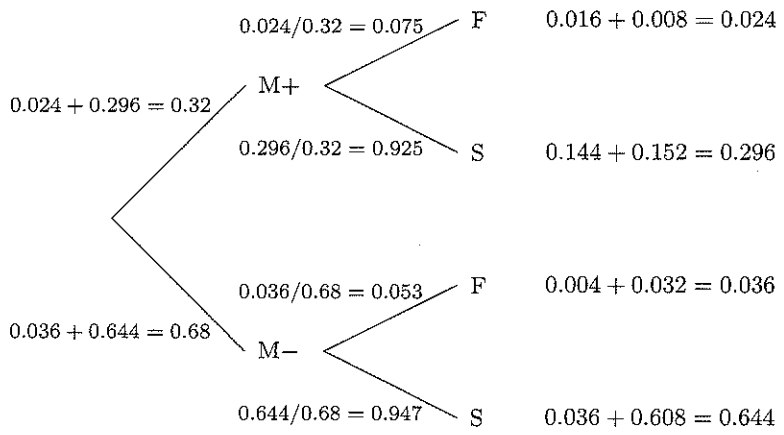

Fig. 27.3. Failure probabilities conditional upon measured exposure.