15
Comparison of rates within strata

15.1 The proportional hazards model

Direct standardization is a very simple way of correcting for confounding but it does have some limitations. This chapter deals with the alternative and more generally useful approach of stratification. We shall again illustrate our argument using the study of the relationship between energy intake and IHD first introduced in Chapter 13 and further analysed in Chapter 14. There, in Table 14.1, we showed the data stratified by 10-year age bands and demonstrated that the low energy intake group is, on average, rather older. This might explain some, or all, of the increase in IHD incidence rate. The method of direct standardization predicts the marginal rates for energy intake groups with the same standard age distribution. This chapter explores the alternative approach which compares age-specific rates within strata. Table 15.1 extends Table 14.1 by calculating rate ratios within each age band. This demonstrates the main problem with this approach to confounding; holding age constant and making comparisons within age strata leads to variable and unreliable estimates, because the age-specific rates are based on so few data.

This problem is resolved is by combining the age-specific comparisons from the separate strata, but any such procedure carries with it a further modelling assumption, because combining the age-specific comparisons can only be legitimate if we believe that they all estimate the same underlying quantity. If we are prepared to believe that the rate ratio between exposure

Table 15.1. Rate ratios within age strata

	Exposed (< 2750 kcal)			Unexposed (≥ 2750 kcal)			Rate
Age	D	Y	Rate	D	Y	Rate	ratio
40–49	2	311.9	6.41	4	607.9	6.58	0.97
50–59	12	878.1	13.67	5	1272.1	3.93	3.48
60–69	14	667.5	20.97	8	888.9	9.00	2.33
Total	28	1857.5	15.07	17	2768.9	6.14	2.45

groups is constant across age-bands, the evidence from the three bands can be brought together to provide a single estimate of the (constant) age-specific rate ratio. Of course the model on which the estimate is based, like all models, is open to question and in later chapters we shall discuss ways in which we can test whether it holds. For the present, we shall be content to believe that the model holds in our example, and that the fluctuation of age-specific rate ratios in Table 15.1 is no more than we would expect given the small numbers of cases in each age band.

Our notation follows naturally from earlier chapters. The age bands are indexed by the superscript t and exposure groups are indexed by subscripts, so that λ_0^t and λ_1^t are the rate parameters in age band t for the unexposed and exposed subjects respectively. We shall write the rate ratio parameter as θ, so that the model of constant rate ratio may be written

$$\frac{\lambda_1^t}{\lambda_0^t} = \theta.$$

This is called the *proportional hazards* model. The parameter θ is called the rate ratio for exposure *controlled for* age, sometimes abbreviated to the *effect* of exposure controlled for age. In this chapter we discuss how θ can be estimated.

15.2 The likelihood for θ

When the rate ratio is constant across age bands, we can replace the rate parameters λ_1^t by $\theta\lambda_0^t$. In our example, this reparametrization replaces the original six rate parameters, which we assume to be constrained to obey the proportional hazards model, with four parameters which are free to take any positive value. One parameter, namely the rate ratio θ, is our prime interest, and the remaining three are regarded as nuisance parameters.

Since each age band serves as an independent study, it is a simple matter to write down the log likelihood for a stratified comparison. Constructing the log likelihood using the prospective argument, each age band contributes a term which depends upon θ and the appropriate λ_0^t. The total likelihood is obtained by adding these terms over age bands. For comparing rates between exposed and unexposed subjects, the parameters λ_0^t are nuisance parameters. As in Chapter 13, replacing these by their most likely value for given θ leads to a profile log likelihood for θ. With the caveat expressed at the end of section 13.3, this log likelihood can also be justified as a conditional likelihood based on the split of cases within each stratum.

The log likelihood ratio curve for $\log(\theta)$ in our illustrative example is shown in Figure 15.1. Using a computer, it is a simple matter to find the most likely value, M, and to use the curvature of the log likelihood ratio to compute a Gaussian approximation. In this case $M = 0.8697$

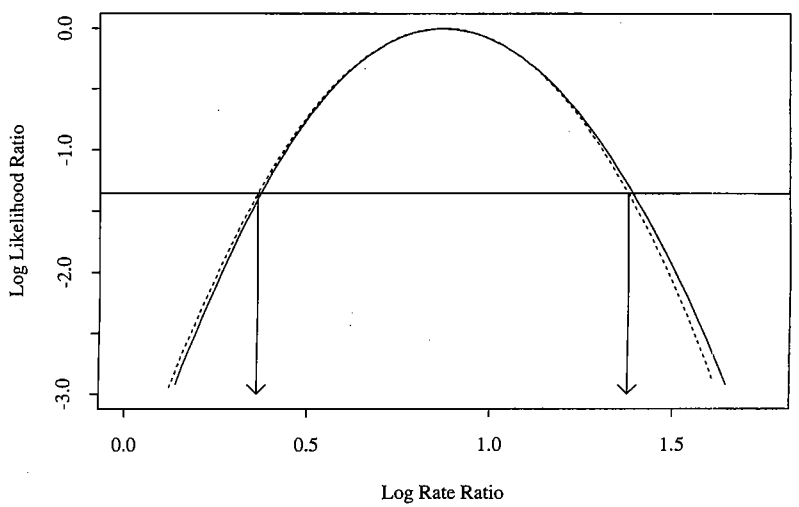

Fig. 15.1. Log likelihood ratio for the common rate ratio.

and $S = 0.3080$, and this approximation is shown as a broken line in the figure. The most likely value of the rate ratio is $\exp(0.8697) = 2.386$ and confidence intervals can be calculated using the error factor:

$$\exp(1.645 \times 0.3080) = 1.660.$$

The fact that the high energy-intake group is, on average, slightly younger than the low energy-intake group is the reason why the estimate of the rate ratio controlled for age is slightly smaller than the crude rate ratio (2.45). However, the difference is extremely small. This is not unusual; rather large differences between exposure groups in important variables are necessary for the effect of confounding to be appreciable.

Unfortunately it is not possible to calculate the values of M and S by hand using simple formulae. The computer programs which are used to carry out such computations are very flexible and allow more complicated models to be fitted. Accordingly discussion of these will be postponed until Part II and the remainder of this chapter will deal with methods which require only a hand calculator.

15.3 A nearly most likely value for θ

We saw in Chapter 13 that, in an unstratified analysis, both profile and conditional arguments led to the Bernoulli likelihood

$$D_1 \log(\Omega) - D \log(1 + \Omega),$$

where Ω, the odds for a case having been exposed, is $\theta Y_1/Y_0$. The gradient of the curve of log likelihood versus $\log(\theta)$ is

$$D_1 - D\frac{\Omega}{1+\Omega}$$

which, after substituting $\theta Y_1/Y_0$ for Ω and rearranging becomes

$$\frac{1}{Y_0+\theta Y_1}\left(D_1 Y_0 - \theta D_0 Y_1\right) = W\left(D_1 Y_0 - \theta D_0 Y_1\right),$$

where $W = 1/(Y_0+\theta Y_1)$. In a stratified analysis, the log likelihood is the sum of contributions of each stratum,

$$\sum\left[D_1^t \log(\Omega^t) - D^t \log(1+\Omega^t)\right]$$

and the gradient is similarly constructed by adding up gradient contributions:

$$\sum W^t\left(D_1^t Y_0^t - \theta D_0^t Y_1^t\right),$$

where $W^t = 1/(Y_0^t + \theta Y_1^t)$ are stratum *weights*.

The most likely value of θ occurs where the gradient is zero, that is, at

$$\theta = \frac{\sum W^t D_1^t Y_0^t}{\sum W^t D_0^t Y_1^t}.$$

Since calculation of the weights W^t involves θ, and this equation cannot be used directly to find the most likely value. However, it can be used *iteratively* as follows:

1. guess a value for θ, and use this to calculate initial weights;
2. using these, calculate a first estimate of θ;
3. using this new estimate, calculate more accurate weights.

The sequence of calculations may be repeated until there is no change in the estimate. Computer programs for maximum likelihood estimation use similar iterative methods of computation.

In practice, the estimate obtained is not very sensitive to changes in the values of the weights — rather large changes make only a relatively small difference to the estimate. Additionally, it may be argued that it is only really important to achieve the closest approximation to the log likelihood when estimating rate ratios which are fairly close to 1. These considerations suggest using the weights corresponding to the choice $\theta = 1$, and to go no further with the calculations. These weights are the reciprocal of the person-years observations in each age band:

$$W^t = \frac{1}{Y_0^t + Y_1^t} = \frac{1}{Y^t}.$$

Use of these weights leads to the *Mantel–Haenszel* estimate of the rate ratio*,

$$\frac{\sum D_1^t Y_0^t/Y^t}{\sum D_0^t Y_1^t/Y^t}.$$

In this expression, each age band makes contributions of

$$Q^t = \frac{D_1^t Y_0^t}{Y^t}, \qquad R^t = \frac{D_0^t Y_1^t}{Y^t}$$

to the top (numerator) and bottom (denominator) of the estimate respectively. The estimate of the rate ratio for age band t is Q^t/R^t and the combined estimate of the constant rate ratio is Q/R, where $Q = \sum Q^t$ and $R = \sum R^t$.

Exercise 15.1. Calculate Q^t and R^t for each of the three age bands in Table 15.1, and hence calculate the Mantel–Haenszel estimate of the rate ratio. Compare this with the most likely value.

15.4 Calculating p-values and confidence intervals

Approximate p-values are most easily calculated using the score test. Since the log likelihood for θ for the age-stratified comparison is the sum of contributions from each age band, it follows that its gradient, and hence the *score*, is the sum of scores for each stratum. Similarly, the curvature is the sum of the curvatures of the separate contribution of each stratum so that the overall score variance is the sum of score variances for each stratum. That is,

$$U = \sum U^t, \qquad V = \sum V^t.$$

Thus to carry out the test we first calculate scores and score variances for each stratum separately and then sum these over strata to obtain the total score and score variance. We then compare $(U)^2/V$ with the chi-squared distribution in the usual way. The contribution of stratum t to the score and score variance are of the same form as given at the end of section 13.2, namely

$$U^t = D_1^t - D^t\pi_\varnothing^t, \qquad V^t = D^t\pi_\varnothing^t(1-\pi_\varnothing^t),$$

where $\pi_\varnothing^t = Y_1^t/Y^t$, the ratio of exposed to total person years.

Exercise 15.2. For our example, what is the p-value for the null hypothesis that, after controlling for age, the rate ratio is 1.

*In fact Mantel and Haenszel did not propose this method but an extremely similar one for case-control studies. We shall discuss this in Chapter 18.

As before, the value of U may be interpreted as the difference between the number of cases who had been exposed and the number expected under the null hypothesis, taking into account the age structures of exposed and unexposed groups.

The calculation of the score variance, V, also allows us to calculate an approximate confidence interval around the Mantel–Haenszel estimate. A Gaussian approximation on the $\log(\theta)$ scale, with

$$S = \sqrt{\frac{V}{QR}}.$$

can be used to calculate an error factor and the approximate confidence interval in the usual way.[†]

Exercise 15.3. Calculate the standard deviation, S, of the log Mantel-Haenzsel estimate for the energy intake data. Use this to calculate a 90% confidence interval for the rate ratio adjusted for age.

These results are very close to those obtained using a computer program to find the Gaussian approximation to the log likelihood curve. The computer method is better in the sense that, as the quantity of data increases, the approximate interval of support approachs the correct likelihood-based interval, while the Mantel–Haenszel interval remains *slightly* wider no matter how much data we collect. The discrepancy is rarely important.

★ 15.5 The log-rank test

Our example in this chapter has involved stratification by a time scale, age, into three rather broad bands. In clinical follow-up studies time is measured from diagnosis or start of treatment and the incidence of events may vary rapidly, requiring the choice of narrow bands. This, together with the fact that choice of bands may introduce an arbitrary element into the analysis, has led to the popularity of a version of the test in which time is stratified infinitely finely into clicks, with no click containing any more than one event. This test is called the *log rank*[‡] or *Mantel–Cox* test.

Derivation of this test from that of the previous section is straightforward. The first thing to notice is that clicks which contain no event (i.e. with $D^t = 0$) make no contribution either to the score, U, or the score variance, V. We therefore need only consider those clicks in which we observe the occurrence of an event in one of the groups ($D^t = 1$). These are are

[†]This approximation is not widely known, but it would not appropriate to justify it here. It suffices to say that it is adequate for all our purposes.

[‡]This nomenclature may seem rather obscure, since the calculation of the test requires neither logarithms or ranks! It arises from an alternative derivation.

Table 15.2. Survival times in two groups of patients

Group	Time (days)
Test treatment ($N = 20$)	86, 99*, 119*, 123*, 139*, 161*, 185*, 212*, 231, 253*, 262*, 281*, 303*, 355*, 360*, 380*, 392, 467*, 499*, 514*
Control ($N = 20$)	73, 91, 102*, 120*, 135, 160*, 194, 202*, 209*, 220*, 252, 270*, 296, 330*, 347*, 375*, 390*, 414, 475*, 485*

known as *informative* time points.[§] Since each click is very short, we need not consider variation in the time spent by different subjects in the band, and the null probability that a failure was exposed becomes

$$\pi^t_{\oslash} = \frac{N^t_1}{N^t} = \frac{\text{Number of } \textit{exposed} \text{ subjects in study at time } t}{\textit{Total} \text{ number of subjects in study at time } t}.$$

Each failure makes a contribution to the score of the difference between the observed number of events in the exposed group, which is either 0 or 1, and the expected number, which is simply $\pi^t_{\oslash}$. The score variance is obtained by adding the contributions

$$V^t = \pi^t_{\oslash}(1 - \pi^t_{\oslash}).$$

Exercise 15.4. Table 15.2 shows times between entry to a clinical trial and relapse for patients receiving two methods of therapy. (The data are only illustrative — a real trial with so much censoring would need to be much larger than this!) The times marked with an asterisk represent times at which observation ceased without occurrence of relapse. Construct a table showing the times of occurrence of relapses, the number of patients in each group under study at each of these times, and the corresponding observed and expected relapses in the test group. Use this table to carry out the score test.

15.6 Comparison with reference rates: the SMR

An important special case concerns the comparison of age-specific rates in a study cohort, λ^t, with those in a *reference population*, which we shall denote by λ^t_{R}. We have discussed this informally in Chapter 6. A more formal treatment follows as a simple case of the methods discussed above.

The proportional hazards model holds that the ratio of age-specific rates in the study cohort to the reference rates is constant across age bands,

$$\frac{\lambda^t}{\lambda^t_{\mathrm{R}}} = \theta.$$

[§]Since clicks have no duration, we assume that no more than one event occurs at any time point.

If we observe D^t failures in Y^t person years of observation in each age band of the cohort, the log likelihood contribution is

$$D^t \log(\lambda^t) - \lambda^t Y^t$$

and making the substitution $\lambda^t = \theta\lambda_R^t$ this becomes

$$D^t \log(\theta) + D^t \log(\lambda_R^t) - \theta\lambda_R^t Y^t.$$

Since the reference rates λ_R^t are calculated from very large populations, they are effectively known constants, and the above log likelihood depends only on one unknown parameter, θ. The second term in the log likelihood does not depend on θ and can be ignored, and the third term may be simplified after noting that $\lambda_R^t Y^t$ is the expected number of failures obtained by multiplying the age-specific reference rate by the corresponding person-years of observation of the study cohort (see Chapter 6). Denoting this by E^t, the log likelihood contribution of one age band becomes

$$D^t \log(\theta) - \theta E^t$$

and summation over age bands leads to the total log likelihood

$$D \log(\theta) - \theta E,$$

where D, E are the total observed and expected numbers of failures. This is a Poisson log likelihood, but the rate ratio parameter θ replaces the rate parameter λ, and the expected number of failures E replaces the person-years Y. Thus estimating θ in this case is just the same as estimating a rate. The most likely value is the ratio of observed to expected cases, D/E, and in epidemiology this is called the standardized mortality ratio, or *SMR*. A 90% confidence interval can be calculated using the error factor

$$\exp\left(1.645\sqrt{\frac{1}{D}}\right).$$

An approximate p-value for the null hypothesis $\theta = 1$ can be carried out using the score and score variance

$$U = D - E, \qquad V = E.$$

Comparison of rates with reference rates in this way is known in epidemiology as *indirect standardization.*

Exercise 15.5. In the follow-up study of ankylosing spondylitis patients discussed in Chapter 6, the observed number of deaths from leukaemia was 31 while the expected number calculated from reference rates was 6.47. Calculate the 90% confidence interval for the common ratio of cohort age-specific rates to reference rates. Also calculate an approximate p-value for the null hypothesis $\theta = 1$.

Exercise 15.6. The calculation of the expected number of deaths in the ankylosing spondilitis study was based on person-years classified by both age and calendar period (see Chapter 6). What further modelling assumption is formally necessary to justify the analysis carried out in the previous exercise?

15.7 Comparing standardized rates

We showed in Chapter 14 that standardized rates estimate the marginal rates when the age distributions are corrected to a common standard. These are weighted sums of age-specific rates. In the case of three age bands, the marginal rate is

$$W^1\lambda^1 + W^2\lambda^2 + W^3\lambda^3$$

where (W^1, W^2, W^3) are the relative frequencies of the three age bands in the standard distribution, and the ratio of two marginal rates, corrected to the same age distribution, is

$$\frac{W^1\lambda_1^1 + W^2\lambda_1^2 + W^3\lambda_1^3}{W^1\lambda_0^1 + W^2\lambda_0^2 + W^3\lambda_0^3}.$$

When the proportional hazards model holds, every term in the numerator of this expression is θ times the corresponding term in the denominator, and it follows that the ratio of marginal rates will also be θ — the relationship between marginal rates is the same as that between the conditional (age-specific) rates. Thus, the ratio of standardized rates can be used as an estimate of θ. However it may not be a very good estimate if the standard age distribution gives high weight to age bands with few failures.

Note that the equivalence demonstrated above between the conditional and marginal comparisons does not hold for *all* stratification models. For example, if the ratio of the age-specific *odds* of failure for exposed and unexposed subjects is a constant, θ, for all ages then the ratio of marginal odds is not equal to θ, even when there is no confounding and the age distributions are identical. Thus we cannot always rely on the method of direct standardization if we are interested in comparisons within strata. In Chapter 18 we shall encounter an important example of this.

15.8 Comparison of SMRs

Although the ratio of standardized rates can be used as an alternative estimate of θ, there has been some controversy as to whether the ratio of two SMRs can also be used in this way.

An understanding of the formal model which lies behind indirect standardization clarifies this argument. Calculation of an SMR for an exposed cohort, using reference rates λ_R^t implies the model

$$\lambda_1^t = \theta_1 \lambda_R^t,$$

where θ_1 is the constant ratio of rates in this cohort to reference rates. Similarly, calculation of an SMR for an unexposed cohort implies the model

$$\lambda_0^t = \theta_0 \lambda_R^t.$$

A direct consequence of these two models is that the ratio of rates for the two cohorts is also constant across age. This can be demonstrated by simply dividing the two equations, when λ_R^t cancels leaving

$$\frac{\lambda_1^t}{\lambda_0^t} = \frac{\theta_1}{\theta_0} = \theta.$$

Thus if the age-specific rates for both exposed and unexposed cohorts are proportional to the reference rates, the comparison of SMRs is legitimate. Since the likelihoods for θ_1 and θ_0 are Poisson in form, with expected numbers of failures E_1 and E_0 replacing person-years observation Y_1 and Y_0, the likelihood for their ratio, θ, is the same as for the rate ratio in Chapter 13.

This method, however, relies on the assumption that both sets of age-specific rates are proportional to the reference rates. If they are proportional to each other, but not to the reference rates, then the ratio of SMRs will not correctly estimate the rate ratio θ. Because of this additional assumption concerning reference rates, estimation of θ by the ratio of SMRs is not usually to be recommended.

Solutions to the exercises

15.1 The calculations are as follows:

Age	Q^t	R^t
40–49	$2 \times 607.9/919.8 = 1.32$	$4 \times 311.9/919.8 = 1.36$
50–59	$12 \times 1272.1/2150.2 = 7.10$	$5 \times 878.1/2150.2 = 2.04$
60–69	$14 \times 888.9/1556.4 = 8.00$	$8 \times 667.5/1556.4 = 3.43$
Total	16.42	6.83

The Mantel–Haenszel estimate is $16.42/6.83 = 2.40$ while the most likely value is 2.39.

15.2 The score is:

$$U = \left(2 - 6\frac{311.9}{919.8}\right) + \left(12 - 17\frac{878.1}{2150.2}\right) + \left(14 - 22\frac{667.5}{1556.4}\right)$$

$$= 28 - 18.41$$
$$= 9.59$$

and the score variance is

$$V = 6 \times \frac{311.9 \times 607.9}{(919.8)^2} + 17 \times \frac{878.1 \times 1\,272.1}{(2\,150.2)^2} + 22 \times \frac{667.5 \times 888.9}{(1\,556.4)^2}$$
$$= 1.34 + 4.11 + 5.39$$
$$= 10.84.$$

The chi-squared value (1 degree of freedom) is $(9.59)^2/10.84 = 8.48$ and $p < 0.005$.

15.3 The standard deviation for the approximation is

$$S = \sqrt{\frac{V}{QR}} = \sqrt{\frac{10.84}{16.42 \times 6.83}} = 0.311.$$

The error factor for the 90% confidence interval is $\exp(1.645 \times 0.311) = 1.67$, and recalling that the Mantel–Haenszel estimate was 2.40, the confidence limits are $2.40/1.67 = 1.44$ (lower limit) and $2.40 \times 1.67 = 4.01$ (upper limit).

15.4 The times at which events occurred, the numbers of patients under observation, and the observed and expected relapses in the test group are shown below.

t	N_1^t	N_0^t	N^t	D_1^t	E_1^t
73	20	20	40	0	$20/40 = 0.50$
86	20	19	39	1	$20/39 = 0.51$
91	19	19	38	0	$19/38 = 0.50$
135	16	16	32	0	$16/32 = 0.50$
194	13	14	27	0	$13/27 = 0.48$
231	12	10	22	1	$12/22 = 0.55$
252	11	10	21	0	$11/21 = 0.52$
296	8	8	16	0	$8/16 = 0.50$
392	4	3	7	1	$4/7 = 0.57$
414	3	3	6	0	$3/6 = 0.50$

The overall score is

$$U = 3 - (.50 + .51 + .50 + \cdots + .57 + .59) = -2.13$$

and the score variance is

$$V = (.50 \times .50) + (.51 \times .49) + \cdots + (.50 \times .50) = 2.49.$$

The score test is $(U)^2/V = 1.82$ and $p > 0.10$. This test is the score test for $\theta = 1$ in the proportional hazards model which holds that the ratio of the relapse rates of the two treatments is constant (at θ) regardless of time since entry into the trial.

15.5 The most likely value of θ is the SMR,

$$\frac{31}{6.47} = 4.791.$$

The error factor is

$$\exp\left(1.645\sqrt{\frac{1}{31}}\right) = 1.344,$$

so that the 90% confidence interval is from $4.791/1.344 = 3.56$ to $4.791 \times 1.344 = 6.44$.
The score test is

$$\frac{(31 - 6.47)^2}{6.47} = 93.00$$

and $p < 0.001$.

15.6 Follow-up was stratified by both age and calendar period when calculating the expected number of deaths. The model which underlies the above analysis therefore assumes that the ratio of rates in the ankylosing spondilitis cohort to those in the reference population is constant for all ages and for all calendar periods.