12

From the log-rank test to the Cox proportional hazards model

12.1 Introduction

We will now address for censored time-to-event data what we outlined for complete data in Chapter 8: how to compare two groups. The problem is really the same, we can compare distributions by comparing the value of their e-CDFs at specific points, or by investigating percentiles, but instead of using the conventional e-CDF, we use the Kaplan–Meier estimator of the CDF with its associated variance. Because mean values usually are of minor interest in this context (we often have only partial knowledge about the CDF for large times) there is no t-test for this situation. There are, however, non-parametric tests, and as in our previous discussion on complete data we will focus on the Wilcoxon test. In doing this we will find that the Mantel–Haenszel test is buried inside most of these extensions; a variation of it, called the log-rank test, provides the building blocks. This test is actually the important test in this context, overshadowing the Wilcoxon test.

The most important models for time-to-event data are different from those for most other data. They are models for hazards, the two key examples being the AFT model and the proportional hazards model. The first of these is often analyzed within the framework of parametric models, using particular distributions, such as the Weibull distribution. This distribution is also useful for proportional hazards models, but in that context the use of parametric methods is completely overshadowed by non-parametric methods.

Working under the assumption of a homogeneous world, it is a direct extension to go from a two-group non-parametric test to a semi-parametric proportional hazards regression model. In the case of the log-rank test this is better known as the Cox proportional hazards model, or Cox model for short, one of the brightest stars in the firmament of biostatistics; nowadays it is so important in cancer research, that when you do not use it for the analysis of survival data, you may need to provide explicit excuses. The Cox model tries to explain the frailties in terms of specified covariates, and the log-rank test is simply the case where we only have the

Understanding Biostatistics, First Edition. Anders Källén.
 ISBN: 978-0-470-66636-4

group indicator variable available. The Cox model is a nonlinear model, and somewhat similar to the logistic regression model for binary data. Like it, there are consequences of omitting important and predictive covariates in the model, in that effects get diluted. We will end with an example of this, which also gives us an opportunity to compare a regression model to a stratified analysis in this setting.

12.2 Comparing hazards between two groups

In this section we will discuss some of the more immediate ways to compare two groups with respect to a time-to-event variable with censored data. In order to illustrate the different methods we will use the data described in the next example.

Example 12.1 In order to investigate whether a certain drug increases the risk of a particular cancer, an experiment was carried out on 150 female rats from 50 litters. One pup from each litter was chosen for drug treatment, together with two control animals. The rats were followed for the occurrence of a tumor for 2 years, after which they were sacrificed; the maximum observed time is therefore 104 weeks.

The overall result can be described as follows. Of the 50 drug-exposed rats, 21 died from the cancer, whereas of the 100 controls, 19 died from the cancer. The probability of cancer death in the drug-treated group is therefore estimated to be 0.42, whereas it is about half that for the controls, 0.19. Using these numbers only, we apply Fieller's method as described in Section 5.4.2 to obtain a risk ratio of 2.21 with 95% confidence interval (1.31, 3.98).

This analysis is an end-point analysis of the occurrence of the event cancer death in the presence of other causes of death. In fact, more rats died from other causes than from the particular cancer under study. This is illustrated in Figure 12.1, which shows the CIFs (see Section 11.4), both for all-cause mortality and for cancer deaths. The (right) end-point of the black curves corresponds to the result in Example 12.1.

To assess the effect of the drug, we want to analyze the cancer mortality in an environment free of competing causes of death. This is what we (try to) do with the Kaplan–Meier estimator,

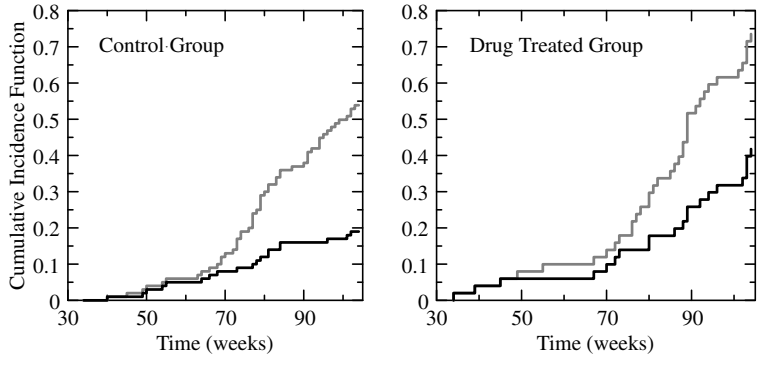

Figure 12.1 The cumulative incidence functions for the two groups. The gray curves show the all-cause mortality, and the black curves show the mortality for a particular cancer. The difference is the mortality for other causes.

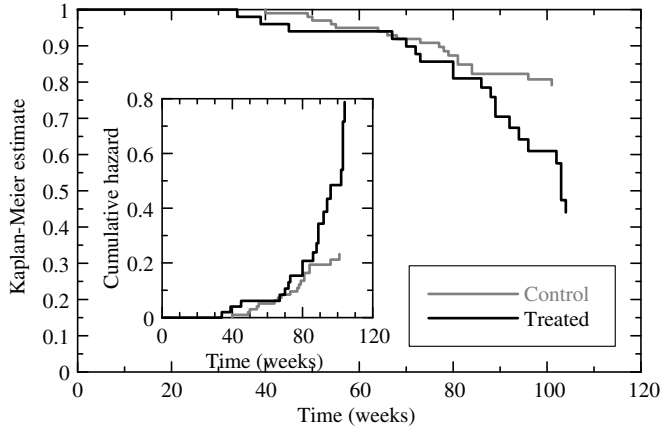

Figure 12.2 The larger graph shows the Kaplan–Meier estimates of the survival distributions for the drug-exposed rats (black) and their controls (gray). The inset graph shows the corresponding Nelson–Aalen estimates of the cumulative hazards.

which are shown in Figure 12.2 for the two groups. From this graph the probabilities of cancer death within 2 years are estimated to be 0.56 and 0.22 in the two groups, respectively. These numbers differ from, and are higher than, those in Example 12.1. They also give a different risk ratio, namely 2.49. With the same kind of analysis as in Example 12.1, but using Greenwood's variance estimate instead, the corresponding 95% confidence interval is (1.52, 4.45). The difference between this analysis and the previous one is that the new estimates address the mortality ratio when the only cause of death for the rats is the particular cancer, under the assumption that competing risks of death act independently of the risk of interest. The inset graph in Figure 12.2 displays the (Nelson–Aalen) estimates of the cumulative hazards. We see that nothing happens for about 30 weeks, whereafter the cumulative hazard increases sharply, most pronounced for the drug-treated group. From now on most of our discussion will be concerned with comparing two CDFs, which means censoring relevant competing risks. We will, however, make the occasional remark on the competing risk case as well.

With these analyses, both of which can be made more precise using profiling techniques that take all variability into account, we have small p-values associated with the hypothesis that the true relative risk is one.

An alternative way to compare $F(t)$ and $G(t)$ would be to compare some specified percentiles. To do this we can apply the methods described in Section 8.3, again modified so that they use the Greenwood variance estimate for the Kaplan–Meier estimators. All such comparisons of two survival functions are comparisons of a single aspect only, and we want to find methods which use the full Kaplan–Meier estimates when we compare groups. It is to this that we now turn.

12.3 Nonparametric tests for hazards

When we design tests to compare time-to-event data for two groups, we want these so constructed that they compare the hazards, not the CDFs. On a high level this is immaterial,

because of the relationship between the hazard and the CDF, but because of the nature of time-to-event data it is natural to model what holds true instantaneously. This also allows us to handle censored data smoothly.

There are a number of tests available for this situation, many of them from the 1960s and 1970s. The first to attain widespread use was Gehan's generalization of the Wilcoxon test, which was constructed by extending the Mann–Whitney score (defined on page 219) in such a way that it is set to zero when it is not known which of the two variables is the largest. At about the same time Mantel used a Mantel–Haenszel type of argument to propose a test that nowadays is known as the log-rank test. Further tests have been proposed by others, but most are variations on a theme.

The test construction process starts with equation (11.1), which shows how we build $F(t)$ from knowledge about the hazard $d\Lambda(t)$ (which we denote $d\Lambda_F(t)$ when we wish to emphasize its relation to the CDF $F(t)$) and the proportion at risk $F^c(t-)$. The basic idea for test construction is that, under a specific assumption about the relation between the two distributions, we can express $d\Lambda(t)$ in the CDFs $F(t)$ and $G(t)$, thereby providing a test statistic which should be close to zero if the model is correct.

To be more specific, we weight the differences $dF(t) - F^c(t-)d\Lambda(t)$ with a particular weight function $w(t)$, which means that we define

$$\Delta = \int_0^\infty w(t)(dF(t) - F^c(t-)d\Lambda(t)). \tag{12.1}$$

This is by definition zero when $\Lambda(t) = \Lambda_F(t)$, which is the important observation to bring forward. The choice of weight function $w(t)$ is subject to some constraints. First of all, we want to use weights constructed from the CDFs of the problem. Second, we want $w(t)$ to be estimated by predictable processes, so the function $w(t)$ should be continuous from the left. With these constraints an immediate choice would be to take

$$w(t) = a(\Psi^c(t-)) \tag{12.2}$$

for some function $a(u)$. This is not a necessary choice; we could use some other function of $F^c(t-)$ and $G^c(t-)$. The particular choice $a(u) = u^\rho$ defines, varying $\rho \in [0, 1]$, the Fleming–Harrington family (of tests). The two border cases $\rho = 0$ and $\rho = 1$ in this family are of particular interest. As can be seen in Box 12.1, the choice $\rho = 1$ gives us the Wilcoxon test, while the choice $\rho = 0$ is what defines the fundamental log-rank test.

In this section our discussion will address the null hypothesis of no group difference. In other words, we assume that $G(t) = F(t)$, which implies that the hazard for the first group is the same as the hazard for the combined sample: $d\Lambda(t) = d\Lambda_\Psi(t)$. Under this assumption we wish to find an estimator of the parameter Δ in the presence of right-censored data. The obvious choice is to estimate $d\Lambda(t)$ with the Nelson–Aalen estimator of the combined sample, which gives us the stochastic variable

$$\hat{\Delta} = \int_0^\infty \hat{w}(t) \left(dN_n(t) - Y_n(t) \frac{dN_+(t)}{Y_+(t)} \right). \tag{12.3}$$

Here the single subscript n refers to data from the first group (with n subjects) and the subscript $+$ means that we sum over both groups. We have ignored a proportionality factor. The expected value of $\hat{\Delta}$ is zero under the null hypothesis of no group difference, so we can use this test statistic to test the null hypothesis. For this purpose we need to derive an

Box 12.1 The limits of the Fleming–Harrington family

The parameter Δ defined by equation (12.1) is of particular interest in the following two cases.

Case $\rho = 1$**.** If we take $a(u) = u$ and change the order of integration in a double integral, we find that

$$\begin{aligned}\Delta = \int_0^\infty \Psi^c(t-)dF(t) - \int_t^\infty dF(s)d\Psi(t) &= \int_0^\infty (\Psi^c(t-) - \int_0^t d\Psi(s))dF(t) \\ &= \int_0^\infty (1 - (\Psi(t-) + \Psi(t)))dF(t) \\ &= 2\left(\frac{1}{2} - \int_0^\infty \frac{G(t-) + G(t)}{2} dF(t)\right).\end{aligned}$$

The condition $\Delta = 0$ is therefore equivalent to equation 8.6, which defines the Wilcoxon test.

Case $\rho = 0$**.** If we take $a(u) = 1$ we find that

$$\Delta = 1 - \int_0^\infty F^c(t-)d\Lambda_\Psi(t) = \int_0^\infty (1 - \Lambda_\Psi(t))dF(t),$$

which leads to the log-rank test. The name is justified because for continuous distributions the right-hand side can be written as $1 + \int_0^\infty \ln(\Psi^c(t))dF(t)$.

Like the Wilcoxon test, the log-rank test defines a rank test for complete data; it can be written as

$$\int_0^\infty \left(1 - \int_{-\infty}^t \frac{d\Psi_{mn}(s)}{\Psi^c_{mn}(s-)}\right) dF_n(t) = 1 - \frac{1}{n}\sum_{i=1}^n a(R_{mn}(x_i)),$$

where (assuming no ties) we have

$$a(k) = \sum_{i=1}^k \frac{1}{n + m + 1 - i}.$$

These scores are the expected value of the kth-order statistic in a sample of size $n + m$ from a Exp(1) distribution and were originally introduced by Leonard Savage in order to test the null hypothesis of equal distributions against the alternative that $G(x) \geq F(x)$ with strict inequality at at least one point; he proved that it was the best test for the one-parameter model (Lehmann alternative) $G(x) = F(x)^\theta$, $\theta > 1$. Note that if there are no ties (and no censoring) then the Nelson–Aalen estimator for the cumulative hazard is $\Lambda_{nm}(t_k) = a(k)$.

estimate of its variance, and then appeal to the CLT. The choice of $\hat{w}(t)$ is not unique even if we have decided on the weight function, because we can estimate $\Psi^c(t-)$ in different ways. One choice is to use the Kaplan–Meier estimate for $\Psi(t)$, lagged one time step to ensure

predictability. Alternatively, we can estimate it by $Y_+(t)/(n+m)$. These different tests have slightly different interpretations. If we use the Kaplan–Meier estimate we take weights from an environment that is free of other risks, whereas if we use the second version we take weights that depend on the competitive risks present when the data were collected. For the Wilcoxon test, if we choose the Kaplan–Meier estimator when we estimate $\Psi^c(t-)$, we get either the Prentice version of the Wilcoxon test for censored data, or a variant due to Peto and Peto which depends on details we ignore. If we instead estimate it from the 'at-risk' function, we derive Gehan's version of the test, for which the test statistic is

$$\int_0^\infty \frac{Y_+(t)}{n+m}\left(dN_n(t) - Y_n(t)\frac{dN_+(t)}{Y_+(t)}\right).$$

In the sequel, when we refer to the Wilcoxon test for censored data, we mean this version.

The (partial) log-rank test (up to time t) can be written as

$$N_n(t) - \int_0^t \hat{p}(s)dN_+(s), \quad \hat{p}(t) = \frac{Y_n(t)}{Y_+(t)}.$$

The entity $\hat{p}(t)$, which is a predictable process, is an estimate of the conditional probability that an event which we know occurs at time t, occurs in the first group. The test is therefore simply the difference between the number of events that have occurred in the first group and our prediction of what should happen, conditional on the situation just before each event. With this interpretation we see that the variance of the log-rank test can be derived from the variance of the binomial distribution as

$$\int_0^t \hat{p}(s)(1-\hat{p}(s))dN_+(s).$$

We can use this to compute the p-value for the test of the null hypothesis, at least if there are no ties. In the presence of ties we need to split these and, referring to the observation on page 158, we have the adjusted formula:

$$\int_0^t \hat{p}(s)(1-\hat{p}(s))\frac{Y_+(s)-\Delta N_+(s)}{Y_+(s)-1}dN_+(s).$$

The only reason we mention this correction is that it helps us to understand how Nathan Mantel arrived at the log-rank test. For this we write the integral explicitly as a sum. The integral in equation (12.3), with $\hat{w}(t)=1$, is a sum over event times t_j, namely

$$\sum_j \left(d_{1j} - \frac{n_{1j}d_j}{n_j}\right),$$

where n_{ij} is the number at risk in group i at time t_j, d_{ij} the corresponding number of events in the respective groups and n_j, d_j the total number at risk and number of events, respectively. In this notation the variance above is given by

$$\sum_j \frac{n_{1j}n_{2j}d_j(n_j-d_j)}{n_j^2(n_j-1)}.$$

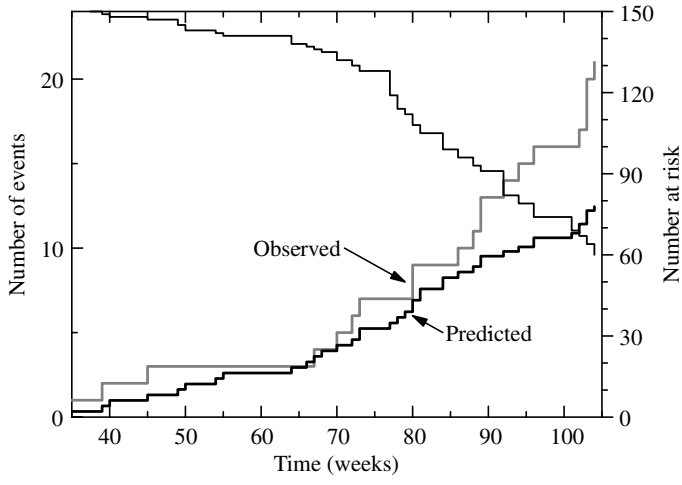

Figure 12.3 The observed and predicted number of deaths in the drug-treated rat group, as well as the number 'at risk' (the decreasing curve in the upper half of the graph) at each time point.

This means that the test statistic is formally the same as the Mantel–Haenszel test in Section 5.5. The strata are the 2×2 tables we obtain at the event times, with the analysis within each table done conditionally on margins.

Example 12.2 Figure 12.3 shows the key information for the log-rank test and the Wilcoxon test for the rat data in Example 12.1. The three curves represent the number of cancer deaths, $N_n(t)$, in the drug-treated group (Observed), the predicted number of cancer deaths in that group based on the combined sample, assuming no difference between the groups (Predicted), and the total number at risk, $Y_+(t)$ (the decreasing step function with y-axis to the right). One summary of the log-rank test is as follows:

Group	N	Observed	Expected
Control	100	19	27.55
Drug-treated	50	21	12.45

This test compares $N_n(\infty) = N_n(104)$ to the corresponding predicted value, and gives us the p-value 0.0034. The Wilcoxon test is different: it weights the differences between observed and predicted at each time point, with the weights given by the number of subjects at risk at the same time. In Figure 12.3 we see that this down-weights the parts of the data where the difference is the largest, so it should come as no surprise that the p-value in this case is larger than that for the log-rank test, namely 0.026.

So far we have discussed the time to cancer death as an isolated phenomenon, when all other causes of death are eliminated. If we wish to understand the effect of the drug in the presence of these competing risks, we need to take a different approach. The analysis should

now focus on the CIF $G(t)$ (see Section 11.4) for cancer death, as we did in Example 12.1. In general this function may also be prone to censoring; subjects may be lost to follow-up or subject to other censoring mechanisms that are not considered to be real competing risks. This means that we want to apply the survival analysis methods to $G(t)$, despite the fact that it is not a proper distribution function. The methodology only uses the fact that we can write $G(t) = e^{-\Lambda(t)}$, with a corresponding expression as a product-limit operator at jump points. We can therefore perform any of the tests discussed above, based on this $\Lambda(t)$, in order to derive a comparison of the event rate in the presence of competing risks. It means that $\Lambda(t)$ is defined through the relation $d\Lambda(t) = dG(t)/G^c(t-)$, and we arrive at what is called Gray's test. If there are no other censoring mechanisms present, this test can be computed by redefining the stochastic variable, so that observations that are censored due to a competing risk are replaced by infinitely large values. We then analyze this modified variable with a log-rank or Wilcoxon test.

Example 12.3 If we perform the log-rank test on the modified variable for the toxicological rat data in Example 12.1, we get the p-value 0.00445. This is larger than the p-value we found when we compared survival with other causes of death removed, but still statistically significant at the conventional 5% level. The conclusion is that the drug also has an effect on cancer survival in the presence of competing causes of death.

When we apply Gray's test we must be careful not to conclude that an intervention is beneficial for one event type, when it instead increases the incidence of a competing event. To draw conclusions in a competitive environment is more complicated than in the non-competing world targeted by the Kaplan–Meier approach, which helps to explain the popularity of the latter approach.

12.4 Parameter estimation in hazard models

The methodology we used in the previous section has an immediate extension to an estimation method for the appropriate hazard model. Among the models previously discussed, the shift model in equation (8.1) is not really relevant in this situation, in contrast to the model in equation (8.2) which can be expressed in terms of hazards as

$$\Lambda_G(t) = \Lambda_F(t/\theta). \tag{12.4}$$

This model is called the accelerated failure time (AFT) model, and we will discuss it further in Section 12.5. Since the right-hand side is the cumulative hazard for a process with time parameter θT, we call θ an acceleration factor. It describes how much faster the biological clock runs in the second group compared to the first. The most popular model for hazards, however, is arguably the proportional hazards model

$$\Lambda_G(t) = \theta\Lambda_F(t), \tag{12.5}$$

which we encountered in our discussion on frailty in Section 11.5. Here we will focus on this model and discuss how to estimate the model parameter θ. (We may note in passing that, because of the relation between the log-rank test and the Mantel–Haenszel test, this test is

really a test of proportional odds in each 2×2 table of events

$$\frac{d\Lambda_G(t)}{1 - d\Lambda_G(t)} = \theta \frac{d\Lambda_F(t)}{1 - d\Lambda_F(t)}.$$

The denominators are one here, because the terms subtracted are zero at a continuity point, which would not be the case if we have truly discrete distributions, instead of continuous ones.)

The problem here is how to estimate $d\Lambda(t)$ from the combined sample, when we assume that the proportional hazards model holds. The way to do this was given in Section 11.5, if we note that the frailty distribution here is the $\theta \text{Bin}(1, 1 - r)$ distribution, which takes values 0 and θ with probability r and $1 - r$, respectively. This means that

$$d\Psi(t, \theta) = (rF^c(t-) + (1 - r)\theta G^c(t-))d\Lambda(t),$$

and we can estimate $d\Lambda(t)$ from the combined sample by

$$\frac{dN_+(t)}{S^0(t, \theta)}, \quad \text{where } S^0(t, \theta) = Y_n(t) + \theta Y_m(t).$$

(Intuitively, if we use the combined sample and there is a twofold increased hazard for group 2, each individual in that group counts as two when we compute the probability, which is why we multiply $Y_m(t)$ by θ in the denominator.) The log-rank test corresponds to the observation that the mean of $N_n(\infty)$ is equal to the mean of

$$\int_0^\infty \hat{p}(t, \theta) dN_+(s), \quad \text{where } \hat{p}(t, \theta) = \frac{Y_n(t)}{S^0(t, \theta)}.$$

If we choose a weight process $\hat{w}(t)$ and apply the discussion above, we arrive at the estimating equation $U(\theta) = 0$ for θ, where

$$U(\theta) = \int_0^\infty \hat{w}(t)(dN_n(t) - \hat{p}(t, \theta)dN_+(t)).$$

Different choices of statistical tests (which means weight function $\hat{w}(t)$) produce different estimates for θ. In order to apply a test to real data and obtain confidence information about θ, we need to have an estimate of the variance of $U(\theta)$. Such an estimate is

$$\hat{V}(\theta) = \int_0^\infty \hat{w}(t)^2 \hat{p}(\theta, t)(1 - \hat{p}(\theta, t))dN_+(t),$$

provided there are no ties. (The true variance depends on the exact censoring mechanism and is therefore seldom possible to compute.) An approximative (two-sided) confidence function for θ is now given by

$$C(\theta) = \chi_1(U(\theta)^2 / \hat{V}(\theta)).$$

Example 12.4 In Example 12.2 we applied both the log-rank test and a Wilcoxon test to the rat data; now we wish to estimate the corresponding parameter for the proportional hazards model. The two confidence functions for these tests are shown in Figure 12.4. They are similar in shape, with the one for the Wilcoxon test lying to the left of that for the log-rank test, and provide the following hazard ratio estimates:

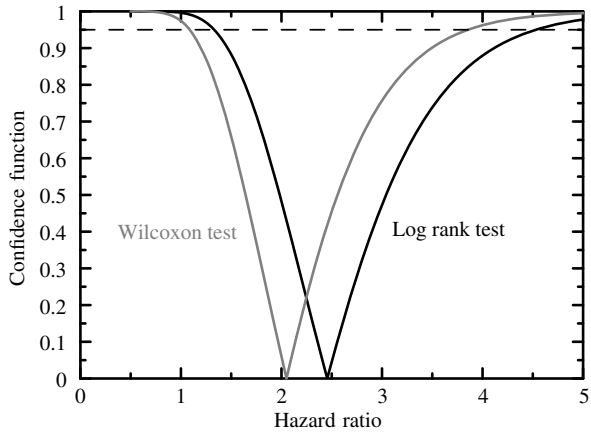

Figure 12.4 Two-sided confidence functions for the hazard ratio parameter using the log-rank and Wilcoxon test statistics.

Test	Estimate	95% CI
Log-rank	2.46	(1.33, 4.53)
Wilcoxon	2.05	(1.09, 3.86)

The fact that the estimate of θ is smaller with the Wilcoxon test is consistent with the observation in Example 12.2 that this test down-weights the parts where the group difference happens to be largest.

So far we have discussed non-parametric group comparison methods under the proportional hazards model. What about parametric analysis? For such an analysis we prefer to use a family of distributions that is closed under the proportional hazards model. This means that if we take one member from such a family and define a new distribution by equation (12.4), this new distribution will belong to the same family. An important example is the Weibull family discussed in Section 11.3, which includes the exponential distributions as a special subfamily. We will use the maximum likelihood method for estimation, and therefore pick up the discussion where we left off in Section 11.3, allowing for individual hazards $\lambda_i(t, \psi)$ including some unknown parameter vector ψ. In this notation, the estimating equation is written as

$$\sum_i \int_0^\infty \frac{\partial_\psi \lambda_i(t, \psi)}{\lambda_i(t, \psi)} (dN_i(t) - Y_i(t) d\Lambda_i(t, \psi)) = 0.$$

For a family closed under the proportional hazards model there is a baseline hazard density $\lambda_0(t, \alpha)$, defined by some parameters α. The complete model is then that $\psi = (\alpha, \theta)$ and $\lambda_i(t, \psi) = \lambda_0(t, \alpha)$ for subjects in the first group, and $\lambda_i(t, \psi) = \theta\lambda_0(t, \alpha)$ in the second group. This means that the estimating equation for α is

$$\int_0^\infty \frac{\partial_\alpha \lambda_0(t, \alpha)}{\lambda_0(t, \alpha)} (dN_+(t) - S^0(t, \theta) d\Lambda_0(t, \alpha)) = 0,$$

Box 12.2 Power analysis of the log-rank test

The log-rank test works conditionally on information about when each event happens, which means that for power calculations we need to make some simplifying assumptions. One such assumption is that the ratio $R = Y_n(t)/Y_m(t) = r/(1-r)$ does not depend on t. This should be approximately true if either the fraction of individuals with events is small, or θ is close to one. This approximation means that $\hat{p}(\theta, t) = R/(R+\theta)$ at all time points, and that the test statistic is $Z(\theta) = U(\theta)/\sqrt{V(\theta)}$, where $V(\theta) = DR\theta/(\theta + R)$ and D is the total number of events observed. To test the hypothesis $\theta = 1$ we use

$$Z(1) = Z(\theta)\sqrt{\frac{V(\theta)}{V(1)}} + \frac{e(\theta) - e(1)}{\sqrt{V(1)}}$$

where

$$e(\theta) = \int_0^\infty \hat{p}(\theta, t) dN_+(t) = \frac{DR}{R+\theta}.$$

For most relevant θ we have that $V(\theta) \approx V(1)$, and with this approximation we see that

$$Z(1) = Z(\theta) - DV(1)^{-1/2}\frac{R(\theta-1)}{(\theta+R)(1+R)} = Z(\theta) - \sqrt{DR}\frac{\theta-1}{\theta+R}.$$

From this we can compute the power function for a one-sided test:

$$\beta(\theta) = P_\theta(Z(1) \leq -z_\alpha) = \Phi\left(-z_\alpha + \frac{\sqrt{DR}}{\theta+R}(\theta-1)\right).$$

A further approximation is $(\theta-1)/(\theta+R) \approx (\ln\theta)/(1+R)$, which gives the power function

$$\beta(\theta) = \Phi(-z_\alpha + \sqrt{Dr(1-r)}\ln\theta).$$

In order to find out what θ we are looking for, it may be helpful to note that under the proportional hazards model we have that

$$\theta = \frac{\ln(G^c(\infty))}{\ln(F^c(\infty))},$$

so we reengineer θ from our perception of the percentage of individuals that should not experience the event during the study for each group.

One remaining question is how many patients we need to study in order to get the number of events this calculation assumes. Many clinical trials on survival have an accrual period a, during which patients enter the study, and a follow-up period, f, from the end of the accrual period until the end of the study. In order to assess the number of patients needed, previous information on the survival distribution for one of the treatments from a similar protocol is needed (or some qualified guess). The proportion of patients who will survive is the average of this survival function in the interval $(f, a+f)$, provided the patients enter the trial at a constant rate.

with the same $S^0(t, \theta)$ as above, whereas the estimating equation for the proportionality constant θ is simpler:

$$\theta^{-1} \int_0^\infty (dN_m(t) - \theta Y_m(t) d\Lambda_0(t, \alpha)) = 0.$$

This is essentially the same equation as we have for the log-rank test, except that it is written for the other group. If our parametric model is flexible enough, so that $\Lambda_0(t, \alpha)$ is a reasonable approximation of the true hazard for some α, we therefore do not expect much difference in the result between the parametric model and the log-rank test. (Another observation is that the estimating equation for θ indicates that it may be worthwhile to parameterize in $\beta = \ln \theta$ instead of in θ. This discussion will be picked up again in Chapter 13.) If we know the baseline hazard we can solve this equation and get the estimate

$$\hat{\theta} = \frac{N_m(\infty)}{\int_0^\infty Y_m(t) d\Lambda_0(t, \alpha)},$$

which is the ratio of the number of events we see in the second group, compared to what we expect, based on the known hazard. In other words, it is the standardized mortality ratio.

Example 12.5 We have seen that the Weibull family supports a proportional hazards model, and that the exponential family is a subfamily of it. We therefore compare the two groups of rats in Example 12.1, using these distributions:

Baseline distribution	Estimate	95% CI	p-value
Exponential	2.30	(1.23, 4.27)	0.0086
Weibull	2.47	(1.32, 4.62)	0.0049

Comparing with the result in Example 12.4, we find that for the Weibull distribution we get a result which is very similar to what we obtained with the log-rank test, whereas the result for the exponential distribution is somewhat different. The reason why these two analyses differ can be inferred from Figure 12.2, where we see that if we were to approximate the baseline hazard by a power function (as for the Weibull distribution), the exponent must be greater than one. The shape forced by the exponential distribution, a straight line through the origin, is therefore a poor fit to it.

12.5 The accelerated failure time model

Analysis of time-to-event data under the AFT model in equation (12.4) uses the observation that $\ln T$ fulfills the shift model with shift $\ln \theta$. We can therefore apply the Wilcoxon test to the logged data when there is no censoring. If we have censored data, we can potentially still estimate $\ln \theta$ by the Hodges–Lehmann estimate, if we use the Kaplan–Meier form of the e-CDF and can estimate sufficiently large parts of the distributions. An alternative is to construct an estimation equation $U(\theta) = 0$ as follows. Consider the rat data. Given θ, multiply the observed times in the drug-treated group by it. Under the assumption that the two groups have the same survival distribution for this modified variable, compute the test function for one

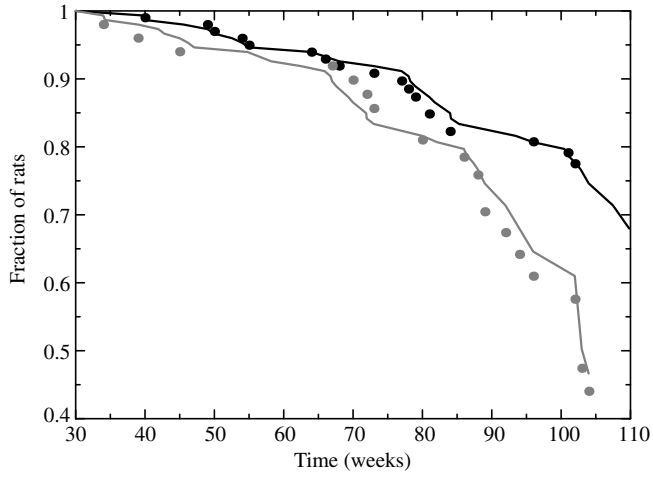

Figure 12.5 Non-parametric AFT models fitted to the rat data. The symbols represent the Kaplan–Meier estimates for the two groups (black is control, gray drug-treated), and the curves are obtained from a Kaplan–Meier estimate of a common survival function, using model adjusted times.

of the tests discussed in Section 12.3. If we choose the Wilcoxon test, which seems natural, and solve the corresponding estimating equation, we find that 'true' time for drug-treated rats is really the clock time multiplied by $\theta = 1.17$ with 95% confidence interval (1.01, 1.55). This means that time (to death) runs 17% faster for drug-treated rats than for the controls. To see how this model fits the data, consider Figure 12.5, in which we have estimated the survival functions for the two groups from the combined Kaplan–Meier estimate, with survival times adjusted by the acceleration factor.

Alternatively, we may use some parametric form for the distributions involved. This is the most common approach to the AFT model, using distribution families that are closed under this model. Again one example is the family of Weibull distributions, since $F^c(t/\theta) = e^{-\lambda\theta^{-\gamma}t^\gamma}$ defines the CDF for a Wei$(\lambda\theta^{-\gamma}, \gamma)$ distribution. The Weibull distribution is therefore closed under both of the most important models for time-to-event data – the proportional hazards model and the AFT model – and if we fit data to this distribution we can choose between a proportional hazards model and an AFT model interpretation. We may note that if θ_{PH} is the proportionality constant in the proportional hazards model, and θ_{AFT} the constant in the AFT model, we have the relation $\theta_{\text{AFT}}^{\gamma} = \theta_{\text{PH}}$.

Another important distribution which defines a family closed under the AFT model is the log-logistic distribution, where $\ln T$ follows a logistic distribution. Its survival function is given by

$$F^c(t) = P(\ln T > \ln t) = \frac{1}{1 + \lambda e^{\gamma \ln t}} = \frac{1}{1 + \lambda t^\gamma}.$$

That this family is closed under the AFT model follows from the observation that $F^c(t/\theta) = 1/(1 + \lambda\theta^{-\gamma}t^\gamma)$, so we do the same replacement as for the Weibull distribution.

A special case of the family of Weibull distributions is the family of exponential distributions, and we have already noted that this family can be generalized in another direction

as well, to the gamma distribution. This too defines a family closed under the AFT model. In fact, any family for which the CDF is a function not of t but of λt for some parameter λ will be closed under the AFT model.

Example 12.6 The following table shows three different parametric AFT analyses of the rat data in Example 12.1:

Distribution	Estimate	95% CI	p-value
Weibull	1.27	(1.06, 1.51)	0.0083
log-logistic	1.26	(1.04, 1.52)	0.018
gamma	1.28	(1.05, 1.56)	0.016

We see here a consistent message from the different models: time to death runs about 25% faster for drug-treated rats than for controls. This conclusion is independent of which family of distributions we analyze, but the estimate is larger than that found in the non-parametric analysis above.

Figure 12.6 show the survival functions for these models. The individual models are not labeled, because the choice of model does not make much of difference in this case. The data points are the Kaplan–Meier estimates from Figure 12.2, and we see that none of the models provides a very good fit to them.

As already noted, the fit to the data for the Weibull distribution is the same whether we consider an AFT model or a proportional hazards model. It is the same function; it is only a matter of how we parameterize it. In fact, the estimate of γ in the proportional hazards model previously analyzed was 3.79, which means that $\theta_{\text{AFT}} = \theta_{\text{PH}}^{1/\gamma} = 2.47^{1/3.79} = 1.27$, in agreement with the analysis above.

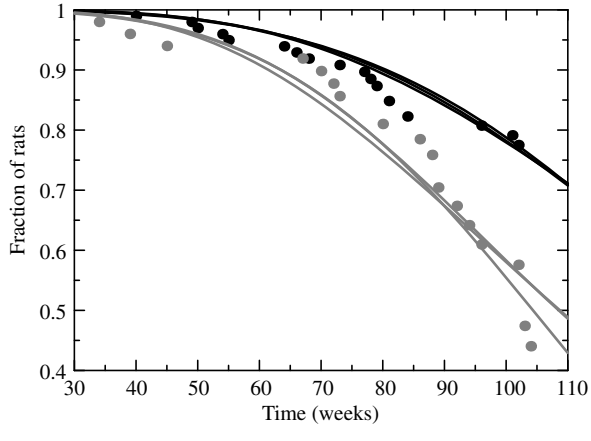

Figure 12.6 Parametric AFT models fitted to rats data. Curves represent estimated group CDFs for the Weibull, log-logistic and gamma distributions, points represent empirical Kaplan–Meier estimates of the CDFs (from Figure 12.2).

Box 12.3 How to estimate the parameter in a failure time model

Parameter estimation in the nonparametric proportional hazards (PH), accelerated failure time (AFT) and accelerated hazards (AH) models is very similar. In all cases the parameter θ is estimated by the equation $U(\theta) = 0$, where

$$U(\theta) = \int_0^\infty a(Y_+(t))(dN_n(t) - \hat{p}(\theta, t)dN_+(t, \theta)), \quad \hat{p}(\theta, t) = \frac{Y_n(t)}{S^0(t, \theta)}.$$

What differ are the functions $S^0(t, \theta)$ and $N_+(t, \theta)$:

PH: $S^0(t, \theta) = Y_n(t) + \theta Y_m(t)$ and $N_+(t, \theta) = N_n(t) + N_m(t)$;
AFT: $S^0(t, \theta) = Y_n(t) + Y_m(t/\theta)$ and $N_+(t, \theta) = N_n(t) + N_m(t/\theta)$;
AH: $S^0(t, \theta) = Y_n(t) + Y_m(t/\theta)/\theta$ and $N_+(t, \theta) = N_n(t) + N_m(t/\theta)$.

The expressions $Y_m(t/\theta)$ and $N_m(t/\theta)$ are obtained from an analysis of the variable θT for group 2.

For all models the variance of $U(\theta)$ is given by

$$V(U(\theta)) = \int_0^\infty a(Y_+(t))^2 \hat{p}(\theta, t)(1 - \hat{p}(\theta, t)dN_+(t, \theta)),$$

and we can obtain confidence intervals and p-values by using the confidence function

$$C(\theta) = \Phi\left(\frac{U(\theta)}{\sqrt{V(U(\theta))}}\right),$$

or its two-sided counterpart.

The proportional hazards and accelerated failure time models are not the only models available for survival data. In particular, there is a compromise between the two which should be mentioned. The assumption of the AFT model is a time acceleration of the integrated hazard, so that $\Lambda_G(t) = \Lambda_F(t/\theta)$. This means that the instantaneous hazard is a mixture of proportional hazards and accelerating time, because $d\Lambda_G(t) = d\Lambda_F(t/\theta)/\theta$. This leads naturally to the alternative suggestion that the difference between the hazards is a difference in time scale only, $d\Lambda_G(t) = d\Lambda_F(t/\theta)$, a model which is called the accelerated hazards model. We can write down the estimating equation for this model (see Box 12.3), but the parameter estimate can also be obtained as follows: for a given θ, perform a log-rank (or Wilcoxon) test on time-adjusted data and estimate a proportional hazards constant θ^* for that data. It is then the case that $d\Lambda_G(t) = \theta^* d\Lambda_F(t/\theta)/\theta$, so we seek out the θ for which we have that $\theta^* = \theta$. For our rat data and the log-rank test, this acceleration factor is estimated to be $\theta = 1.30$.

12.6 The Cox proportional hazards model

The log-rank test is a special case of one of the most celebrated models in biostatistics, the Cox proportional hazards model for survival data. In order to understand the relation between them we will first rederive the former. It is the same derivation as before, but in new

notation adapted to a more general situation. The starting point is the repeated means formula $E(Z) = E(E(Z|T))$ which is valid for all stochastic variables. In our application T will be the time-to-event variable, and we can introduce censoring into this by a censor process which is independent of T. We then have that $E(C(T)Z) = E(C(T)E(Z|T))$, which we can write as

$$E(C(T)Z) = \int_0^\infty E(Z|T = t)d\Psi(t).$$

The left-hand side here is the expected value of Z among those individuals for whom we observe an event, multiplied by the fraction of these among all. $\Psi(t)$ is the sub-distribution function describing observed events, and if we have a model from which we can deduce the conditional means $E(Z|T = t)$, the right-hand side is what the model predicts about Z in individuals with an event. Replacing the left-hand side with what we observe, and $d\Psi(t)$ with the Nelson–Aalen estimator, this gives us a relation that can be used to fine-tune the model that defines the conditional means. The log-rank test corresponds to the case where Z is one for those in group 1, and zero for those in group 2. The left-hand side is then the number of events, and the proportional hazards model tells us how to compute the conditional means. The relation is therefore exactly what we use to estimate the hazard ratio parameter from data (the fine-tuning referred to above). Note that this is the same interpretation as we had for the estimating equation for the logistic equation, as was discussed in Section 9.4.

However, the derivation above is more general than the log-rank test, and we can make it even more general by replacing Z with a predictable stochastic process. For our purposes we settle for less, and replace Z with $a(Y(T))Z$, where $Y(t)$ is the fraction at risk at time t, $a(u)$ a function, and Z a stochastic variable (actually a vector). Suppose that we have a model which depends on a parameter β, such that we can compute the function $\bar{z}(t, \beta) = E(Z|T = t)$. This gives us the stochastic variable $U(T, \beta) = a(Y(T))(Z - \bar{z}(T, \beta))$ about which we know that $E_\beta(U(T, \beta)) = 0$. If we have a sample of n from the population with observed event times t_i, we can estimate this mean with the average of the observations. This gives us the following estimating equation for β:

$$U_n(\beta) = \frac{1}{n}\sum_i \int_0^\infty a(Y(t))(z_i - \bar{z}(t, \beta))dN_i(t) = 0. \tag{12.6}$$

Since $a(Y(t))$ is a predictable process the variance of $U_n(\beta)$ is estimated by

$$\frac{1}{n^2}\sum_{i=1}^n \int_0^\infty a(Y(t))^2(z_i - \bar{z}(\beta, t))^2 dN_i(t),$$

a fact we need when we want to derive a confidence function for β.

It remains to compute $\bar{z}(t, \beta)$, for which we need a specific model. The log-rank test was derived under the assumption of a proportional hazards model, so we assume it here as well. This model will explain the frailty θ in terms of the covariate vector Z, so that there is a vector of regression coefficients β such that $\theta = e^{Z\beta}$. The choice of the exponential link here is convenient, but not necessary. It simplifies some calculations and it is the assumption of the Cox model, so we stick to it. Equation (11.4) means that this model estimates the conditional mean $\bar{z}(t, \beta)$ by

$$\frac{\hat{S}^1(t, \beta)}{\hat{S}^0(t, \beta)} = \partial_\beta \ln \hat{S}^0(t, \beta),$$

where

$$\hat{S}^1(t, \beta) = \frac{1}{n} \sum_i z_i e^{z_i \beta} Y_i(t), \quad \hat{S}^0(t, \beta) = \frac{1}{n} \sum_i e^{z_i \beta} Y_i(t).$$

If we plug this into equation (12.6) we get our final estimating equation for β. Varying the weight function defines a whole family of proportional hazard estimating equations (and therefore tests), of which the original Cox model used $a(u) = 1$, and we get a Wilcoxon-type test if we choose $a(u) = u$. We will return to these models in Chapter 13, where we will further discuss the confidence function and how it can be used in situations where the assumption of independence between event times is not fulfilled, as is the case when we analyze recurrent events.

Cox originally derived this model in a different way. To see how he did it, we write the estimating equation that determines β (which we have written as an integral above) as a sum:

$$U_n(\beta) = \sum_i (z_i - \partial_\beta \ln \hat{S}^0(t_i, \beta)) = 0.$$

The sum is over observed event times. It follows that $U_n(\beta)$ is the derivative of $\ln PL(\beta)$, where

$$PL(\beta) = \prod_{t_i} \frac{e^{z_i \beta}}{\hat{S}^0(t_i \beta)}.$$

This is called the *Cox partial likelihood*. The factors in this product are the conditional probabilities that an event, observed at time t_i, is from the individual with covariate value z_i, among those who are still at risk. This follows the idea of survival analysis in general, that observed times are analyzed conditionally on the state of the world at that time. Technically this is not the model likelihood, but Cox treated it as if it was, from an analysis point of view.

In order to see how the partial likelihood above relates to the true likelihood, we need to write down the latter explicitly. For this purpose we recall from Section 12.4 that the log-likelihood for this type of data is given by

$$\sum_i \left(\int_0^\infty \ln \lambda_i(t, \beta) dN_i(t) - \int_0^\infty Y_i(t) \lambda_i(t, \beta) dt \right).$$

The Cox model corresponds to the assumption that $\lambda_i(t, \beta) = \lambda_0(t) e^{z_i \beta}$ for some baseline hazard $\lambda_0(t)$. If we insert this expression for $\lambda_i(t, \beta)$ into the log-likelihood we get

$$\int_0^\infty \left(\sum_i (z_i \beta + \ln \lambda_0(t)) dN_i(t) - \int_0^\infty \lambda_0(t) \hat{S}^0(\beta, t) dt \right).$$

If we know β, we can use $dN_+(t)/\hat{S}^0(\beta, t)$ to estimate $\lambda_0(t)dt$, and if we insert this into the expression above, we see that the log-likelihood is

$$\sum_i \int_0^\infty (z_i \beta - \ln \hat{S}^0(\beta, t)) dN_i(t)$$

plus a term that does not involve β. The derivative of this is the $U_n(\beta)$ above, which means that the Cox partial likelihood is essentially a profiled likelihood, where we have profiled out the unknown baseline hazard by estimating it with the Nelson–Aalen estimator.

There is one more important question we need to address. It has to do with what will happen when the Cox model is true, but we have omitted to include one of the covariates in the analysis. This is the same discussion as we had in Section 9.6, but for this type of data/model. Let Z represent the observed covariates and ξ the omitted one. The explicit assumption is that $\theta = e^{Z\beta+\xi} = \eta e^{Z\beta}$, where $\eta = e^{\xi}$. Estimation in the Cox model is based on the functions $\hat{S}^0(t, \beta)$ and $\hat{S}^1(t, \beta)$ previously defined. In the presence of heterogeneity the expected values of these are given by

$$S^k(t, \beta) = \int z^k \left(\int F^c(t-, \eta e^{z\beta}) \eta e^{z\beta} dP(\eta) \right) dF(z) = - \int z^k \mathcal{L}'(e^{z\beta} \Lambda(t)) e^{z\beta} dF(z),$$

where we assume that η is independent of Z in the population. ($\mathcal{L}(u)$ is the Laplace transform of the frailty distribution $P(\theta)$.) The true conditional expectation above is therefore

$$E(Z|T = t) = \frac{\int w(t, e^{z\beta}) z e^{z\beta} dF(z)}{\int w(t, e^{z\beta}) e^{z\beta} dF(z)}, \quad w(t, \theta) = \mathcal{L}'(\theta \Lambda(t)).$$

With no missing covariates, we have that $w(t, \theta) = F^c(t, \theta)$, but if we have missed some important predictor for the event in our model, and we analyze the data using a Cox model, we will underestimate the true effect of the covariate on survival time. The next example investigates this in more detail for the log-rank test. It may be skipped, as one can read the next section without these details. For a heuristic explanation of this material, see Box 12.4.

Example 12.7 Assume that we have continuous distributions, that the length of the study is τ, and that there is no other censoring. The hazard factor θ^* in a log-rank test is the solution to the equation which equates the expected number of events observed in the first group to what is expected using the combined hazard, which is the equation

$$rF(\tau) = \int_0^\tau \frac{rF^c(t)}{rF^c(t) + (1-r)\theta^* G^c(t)} d\Psi(t).$$

Assume there is heterogeneity in the population such that if the frailty for a patient is η without treatment, it becomes $\theta\eta$ with treatment. This means that θ is the individual hazard ratio, assumed constant. In that case the survival functions in the formula above are obtained as Laplace transforms of the frailty distribution, computed for the values of the cumulative hazards. This defines a relation between θ^* and θ. If we change the variable in this integral to $u = \Lambda(t)$, this relation is

$$1 - \mathcal{L}(\Lambda(\tau)) + \int_0^{\Lambda(\tau)} \frac{\mathcal{L}(u)(r\mathcal{L}'(u) + (1-r)\theta\mathcal{L}'(\theta u))}{r\mathcal{L}(u) + (1-r)\theta^*\mathcal{L}(\theta u)} du = 0.$$

This relation is illustrated in Figure 12.7, assuming gamma frailty and that $\theta^* = 0.91$. We see that if there is considerable heterogeneity, expressed as a large variance for the frailty distribution, we may have a true treatment effect that is as large as a reduction of almost 20%.

Box 12.4 Bias due to heterogeneity: a heuristic explanation

The proportional hazards model with frailty for a time-to-event variable T says that, conditionally on the frailty η, $\Lambda(t|\eta) = \eta e^{z\beta}\Lambda_0(t)$. Assume that $\Lambda_0(t) = \lambda t^{\gamma}$ is from a Weibull distribution. This can also be expressed as an AFT model, which means a linear model in $\ln T$:

$$\ln T = \gamma^{-1}(-\ln\lambda - z\beta + X + Y),$$

where X has the smallest extreme value (SEV) distribution and Y has the distribution of $-\ln\eta$. It is convenient to assume that the frailty has a lognormal distribution with mean one, which means that $Y \in N(\sigma^2/2, \sigma^2)$ and therefore that the variance of $\ln T$ is $(\eta^2 + \sigma^2)/\gamma^2$, where η^2 is the variance of the SEV distribution.

Next suppose that we ignore the frailty, but assume that $\Lambda_0(t) = \mu t^{\nu}$. This means that

$$\ln T = \nu^{-1}(-\ln\mu - x\beta^* + X),$$

and the variance of $\ln T$ is η^2/ν^2. Solving for ν gives $\nu^2 = \gamma^2\eta^2/(\eta^2 + \sigma^2)$. If we estimate the regression coefficients by least squares, we should have the same estimate in the two situations: $\beta/\gamma = \beta^*/\nu$, which implies that

$$\beta^* = \frac{\beta}{\sqrt{1 + \sigma^2/\eta^2}}.$$

This shows that what we consider the treatment effect moves toward no effect ($\beta^* = 0$) in the presence of frailty. How much depends on the magnitude of heterogeneity.

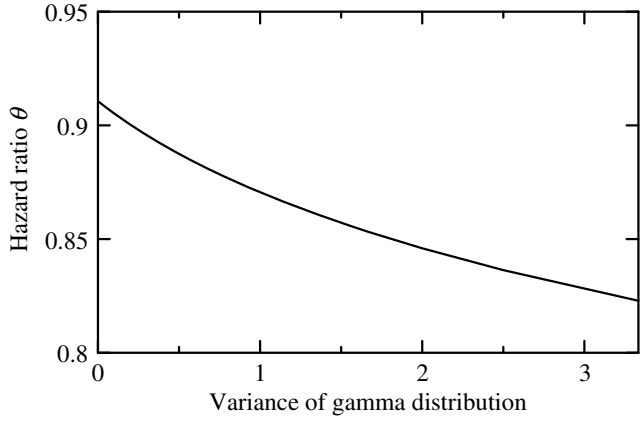

Figure 12.7 The relation between a subject-specific hazard ratio θ and the degree of heterogeneity in the population, when the population hazard ratio (obtained by a log-rank test) is 0.91. The assumptions are a gamma frailty and no censored data except for a finite study length.

What we measure with the log-rank test (θ^*) is not necessarily what matters to the individual patient, who is probably more interested in θ. That parameter is what determines the effect on him, although on a relative scale, so it might not be easily translated to entities such as number of years of added life. But this discussion shows that it is important to find prognostic variables that can explain as much as possible of the heterogeneity when we analyze time-to-event data.

12.7 On omitted covariates and stratification in the log-rank test

In this section we will illustrate the consequences of omitting covariates in the Cox model, using as background a real-life example. For various reasons, the exact details of the study in question will not be described, and they are not important for this discussion anyway. Suffice it to say that it was a placebo-controlled study, where the outcome was survival after start of treatment; it was a two-armed parallel group study which was randomized 2 : 1 between the active drug and placebo. The overall log-rank analysis table comparing the two treatments gives us the key outcome data:

	N	Observed	Expected
Active	1129	634	654.4
Placebo	563	342	321.6

The hazard ratio was estimated to be 0.91 with 95% confidence interval (0.80, 1.04) and with $p = 0.16$ for the hypothesis of no treatment effect. Although there were fewer deaths than expected (assuming no treatment effect) in the active group, there is not enough evidence to claim that the drug has an effect on survival.

However, we cannot rest with this. The effect we are looking for gets attenuated in the presence of heterogeneity, and we wish to explain as much of the heterogeneity as possible, in order to home in on the hazard ratio we are interested in. At our disposal we have six covariates, each of which is dichotomous in nature. The one which on its own is the most predictive of survival is related to the patient's performance status according to a WHO scale. We adjust for this variable in the analysis by carrying out a Cox regression with two factors, treatment and the WHO scale, both dichotomous. Now the estimated hazard ratio is 0.855 with 95% confidence interval (0.75, 0.98), which gives us $p = 0.020$ for the hypothesis of no effect. With this single adjustment we have decreased the hazard ratio so much that we now have sufficient evidence at the conventional (two-sided) 5% significance level that the drug has an effect on survival. There is no reason to believe that we have captured all the heterogeneity, but all we can do with available data is see what the effect is when we include all the covariates (additively) into a Cox regression model. The result is that the treatment hazard ratio is estimated as 0.863 with 95% confidence interval (0.76, 0.99) and $p = 0.029$. Although most of these individually have an effect on survival, including them all seems not to explain any more heterogeneity than is explained already by the first covariate.

There is more to say about this. The original model, the log-rank test for the two treatment groups, corresponds to a model in which the placebo group has hazard $d\Lambda(t)$ and

the active group has hazard $\theta d\Lambda(t)$. The model with the WHO covariate is such that the hazard is

WHO scale	Placebo	Active
0 or 1	$d\Lambda(t)$	$\theta d\Lambda(t)$
2 or 3	$rd\Lambda(t)$	$r\theta d\Lambda(t)$

where r is the proportionality factor for the covariate. This is a stronger assumption than assuming that each WHO scale subgroup has its own hazard, which is the table

WHO scale	Placebo	Active
0 or 1	$d\Lambda_0(t)$	$\theta d\Lambda_0(t)$
2 or 3	$d\Lambda_1(t)$	$\theta d\Lambda_1(t)$

The first model corresponds to the assumption $d\Lambda_1(t) = rd\Lambda_0(t)$. The second model can be analyzed by a stratified log-rank test, in which we compute the estimating function for each stratum (i.e., subgroup defined by the WHO scale variable) and from which an estimating equation for θ is obtained by equating a weighted average of these to zero. The convention here is to take equal weights for the strata, and the result of this analysis is summarized in the following table:

WHO scale	Treatment	N	Observed	Expected
0 or 1	Active	387	283	299.3
	Placebo	174	135	118.7
2 or 3	Active	739	351	365.5
	Placebo	389	207	192.5

The corresponding hazard estimate is 0.867 with 95% confidence interval (0.76, 0.99) and $p = 0.034$. We see that the result is very similar to the Cox regression result presented above. However, if we instead stratify on all six covariates the results of the two methods differ more, and in a crucial way: the treatment hazard ratio is estimated to be 0.89 with 95% confidence interval (0.77, 1.02) and $p = 0.087$. Numerically the difference is not large, but the p-value moves over to the other side of the conventional cut-off limit of 5%. It therefore becomes important to understand whether the first analysis was based on faulty assumptions, or whether the explanation for the discrepancy is to be found elsewhere. In exploring this we will highlight an important risk with a stratified analysis.

Figure 12.8 shows the estimates and confidence intervals for the two models (Cox regression in black, stratified test in gray) at different degrees of stratification, in such a way that we start with the WHO scale variable as a single covariate, and then add one new covariate at each step according to how predictive they are for survival on their own. Not much happens, with one important exception, which is when we add the last covariate to the stratified test (a covariate that is not even shown to be predictive on its own). In this test we have stratified on six dichotomous variables, which means that we divide the population into $2^6 = 64$ cells. Four of these are empty, and 23% of the remaining cells have at most 3 patients. In such small

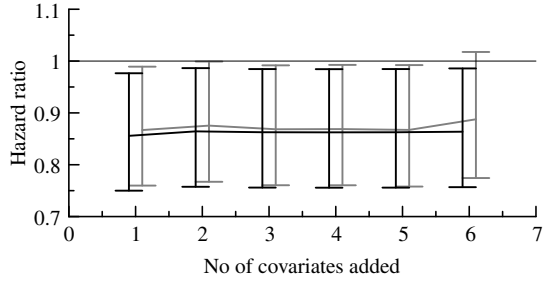

Figure 12.8 Illustration of how confidence intervals for hazard ratio change as we increase the number of variables to stratify on. The gray data are for the stratified test, the black data for the (additive) Cox regression model.

cells it is not unlikely that only one treatment will be represented, a risk that is augmented by the fact that we had a 2 : 1 randomization. In fact, 11 cells have only one treatment and of these, 9 contain only active drug. These cells do not contribute to the stratified test, which means that we effectively loose 27 patients on active drug and 2 on placebo in the analysis. Among these there are in total 14 deaths (one on placebo) that no longer contribute to the analysis. This is a loss of power and explains much of the effect we see when we add the sixth covariate to the analysis. It means that drawing conclusions from this test is not a sensible thing to do, however prespecified the analysis may have been.

The lesson is simple. Do not over-stratify! You must make certain that no cells are too small. This is of course true for all stratified tests. The idea behind stratification is this: we have a heterogeneous population, and in order to apply a test which assumes a homogeneous population we divide the population up into strata, such that within each stratum the population is homogeneous. Thereafter we pool the strata. However, the quest for homogeneity strives toward small strata, and in a small stratum there is a severe risk that the treatment groups are unbalanced. Unbalanced comparisons are less effective than balanced ones, so we loose power as the number of strata increases. On the other hand, if we take fewer strata, these may be heterogeneous with treatment bias as a result. Note that if we stratify when we do not need to, when our population is actually homogeneous, we may have a substantial loss of power, at least if some cells become small. In all, this makes the application of stratified tests problematic if one is forced, as is often the case in the pharmaceutical industry, to prespecify in detail the analysis to be performed, in order to gain credibility (in the eyes of the regulatory agencies).

12.8 Comments and further reading

The rat data we have used in this chapter to illustrate different methods was originally given in Mantel et al. (1977,Table 1), and is reproduced in Hougaard (2000,Table 1.5). The original paper is of independent interest, because it illustrates how the log-rank test is related to the Mantel–Haenszel technique in a very explicit way.

Much of the material in this chapter is covered in major books on the statistical analysis of survival data, some of which were referenced in the last chapter. An overview of how traditional non-parametric tests are expressed and analyzed in a counting process theory

context is given by Gill (1983). The accelerated hazards model is less used, but is discussed by Chen and Wang (2000). Gray's analysis in a competitive environment is discussed in Gray (1988).

The power calculation in Box 12.2 is based on the validity of the proportional hazards model. In the design stage we assume a certain (subject-specific) hazard ratio, but when we do the analysis, in order to achieve this assumption, we may need to include a series of predictive variables in the analysis model (Schoenfeld, 1983). In other words, we use the formula for the log-rank test when we compute the number of patients needed, and also if we plan for a more extensive Cox regression model. If we apply the log-rank test and ignore the predictors, the loss of power comes from the fact that the treatment effect is time-dependent and we estimate a parameter which corresponds to a smaller effect than the true one.

The original article by David R. Cox (1972) on the proportional hazards model has had a huge number of citations and its author has received a large number of honors. Our derivation of his model is not the traditional one and is deliberately sketchy; missing details may be found in papers by Sasieni (1993) and Tsiatis (1981). The value of this derivation is that it emphasizes the underlying connection between the model and the problem of explaining heterogeneity. It emphasizes that on an individual level we may well have proportional hazards, even when it does not appear so from the overall population (Kaplan–Meier) perspective. The traditional derivation can be found in most books on survival analysis, many of which contain numerous applications. There are different ways to extend the Cox model (Therneau and Grambsch, 2000) to situations where its basic assumptions are not fulfilled, some of which will be touched upon in the next chapter.

The heuristic idea for the bias (if that is the proper word) in the presence of frailty, or omitted covariates, in the Cox model, described in Box 12.4, is essentially taken from Keiding et al. (1997). A fuller discussion of this bias is given by Henderson and Oman (1999). The amount of bias depends on the frailty distribution, and is actually more pronounced with complete data than if there are censored data. Another discussion about the balancing act between stratification with small cells versus the problem of heterogeneity can be found in Akazawa et al. (1997) with a related discussion in Stavola and Cox (2008) for a Poisson process setting.

References

Akazawa, K., Nakamura, T. and Palesch, Y. (1997) Power of logrank test and Cox regression model in clinical trials with heterogeneous samples. *Statistics in Medicine*, **16**, 583–597.

Chen, Y.Q. and Wang, M.C. (2000) Analysis of accelerated hazards models. *Journal of the American Statistical Association*, **95**(450), 608–618.

Cox, D.R. (1972) Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society, Series B*, **34**, 187–220.

Gill, R.D. (1983) *Censoring and Stochastic Integrals* vol. Mathematical Centre Tracts 124. Amsterdam: Mathematisch Centrum.

Gray, R.J. (1988) A class of K-sample tests for comparing the cumulative incidence of competing risks. *Annals of Statistics*, **16**, 1141–1154.

Henderson, R. and Oman, P. (1999) Effect of frailty on marginal regression estimates in survival analysis. *Journal of the Royal Statistical Society, Series B*, **61**(2), 367–379.

Hougaard, P. (2000) *Analysis of Multivariate Survival Data* Statistics for Biology and Health. New York: Springer.

Keiding, N., Andersen, P.K. and Klein, J.P. (1997) The role of frailty models and accelerated failure time models in describing heterogeneity due to omitted covariates. *Statistics in Medicine*, **16**, 215–224.

Mantel, N., Bohidar, N. and Ciminera, J. (1977) Mantel-Haenszel analyses of litter-matched time-to-response data, with modifications for recovery of interlitter information. *Cancer Research*, **37**, 3863–3868.

Sasieni, P. (1993) Some new estimators for Cox regression. *Annals of Statistics*, **21**(4), 1721–1759.

Schoenfeld, D.A. (1983) Sample-size formula for the proportional-hazards regression model. *Biometrics*, **39**, 499–503.

Stavola, B.L.D. and Cox, D.R. (2008) On the consequence of overstratification. *Biometrika*, **95**(4), 992–996.

Sun, J. (2006) *The Statistical Analysis of Interval-Censored Failure Time Data* Statistics for Biology and Health. New York: Springer.

Therneau, T.M. and Grambsch, P.M. (2000) *Modeling Survival Data: Extending the Cox Model* Statistics for Biology and Health. New York: Springer.

Tsiatis, A.A. (1981) A large sample study of Cox's regression model. *Annals of Statistics*, **9**(1), 93–108.

12.A Appendix: Comments on interval-censored data

With interval-censored data there are some adjustments that need to be made to the way we compute things. The Kaplan–Meier e-CDF can be computed only when we know the situation at each time point, so we need to find another way to obtain an e-CDF. Suppose, then, that we have n patients, for each of whom we have an interval $(l_i, r_i]$, such that the event has occurred somewhere within that interval. By going to the limit $r_i - l_i \to 0$ we can include exact observations, and by taking $r_i = +\infty$ we can also include right-censored events. For this discussion we assume that all intervals are proper, finite intervals. Let $t_1 < t_2 < \ldots < t_m$ denote the unique elements from the list of left and right interval limits.

An e-CDF $F_n(t)$ will be a step function with jumps at the t_j of magnitude $\Delta_j = F_n(t_j) - F_n(t_{j-1})$. To determine Δ_j, let I_j^i be the indicator variable which is one if the censor interval for subject i contains the point t_j (i.e., if $(t_{j-1}, t_j] \subset (l_i, r_i]$), otherwise zero. The contribution of subject i to the e-CDF at point $t_j \in (l_i, r_i]$ is then given by $\Delta_j/(F(r_i) - F(l_i)$. But the average over all individuals at that point is the actual jump size, so we have the relation

$$\Delta_j = \frac{1}{n} \sum_{i=1}^{n} \frac{I_j^i \Delta_j}{\sum_k I_k^i \Delta_k}.$$

This defines the jump sizes and therefore what is called the Turnbull e-CDF for interval-censored data. We may note that Δ_j can only be non-zero if t_{j-1} is a left end point of the original data and t_j a right end point, but not necessarily from the same censored interval. (The intervals $(t_{j-1}, t_j]$ are called Turnbull intervals, and identifying them first is helpful for computational reasons.)

Given the Turnbull e-CDF $\Psi_{mn}(t)$, we can define the log-rank test for interval-censored data as follows. Instead of observed event times, use predicted event times, so that

$$N_+(t) = \sum_{i=1}^{n+m} \frac{I^i(t) \Delta \Psi_{mn}(t)}{\sum_k I^i(t_k) \Delta \Psi_{mn}(t_k)},$$

where $I^i(t)$ is an indicator for the interval $(l_i, r_i]$. We can also define the predicted number at risk by

$$Y_+(t) = n \sum_{t_k \geq t} \Delta \Psi_{mn}(t_k).$$

Together with similar estimates for one group alone we derive a log-rank test, or Wilcoxon test, for interval-censored data that is analogous to what they are for right-censored data, except that we use these predicted entities instead of observed ones. The extension to parameter estimation is immediate.

We can alternatively construct a generalized log-rank test for interval-censored data based on the expression $1 + \int_0^\infty \ln(\Psi^c(t)) dF(t)$, which underlies the log-rank test. It is estimated by

$$1 + \frac{1}{n} \sum_i \ln \Psi_{mn}^c(t_i) = \frac{1}{n} \sum_i (1 + \ln \Psi_{mn}^c(t_i)).$$

A primitive function of $1 + \ln x$ is $x \ln x$, which means that a generalization from complete data to interval-censored data can be done by using

$$\frac{1}{n} \sum_i \frac{\Psi_{mn}^c(l_i) \ln(\Psi_{mn}^c(l_i)) - \Psi_{mn}^c(r_i) \ln(\Psi_{mn}^c(r_i))}{\Psi_{mn}^c(l_i) - \Psi_{mn}^c(r_i)}.$$

In the limit this reduces to the previous expression. For the details necessary for practical implementation of this, see Sun (2006), for example.