Chapter 6
Survival Analysis

Children receiving a kidney transplant may be followed to identify predictors of mortality. Specifically, is mortality risk lower in recipients of kidneys obtained from a living donor? If so, is this effect explained by the time the transplanted kidney is in transport or how well the donor and recipient match on characteristics that affect immune response? Similarly, HIV-infected subjects may be followed to assess the effects of a new form of therapy on incidence of opportunistic infections. Or patients with liver cirrhosis may be followed to assess whether liver biopsy results predict mortality.

The common interest in these studies is to examine predictors of time to an event. The special feature of the survival analysis methods presented in this chapter is that they take time directly into account: in our examples, time to transplant rejection, incidence of opportunistic infections, or death from liver failure. Basic tools for the analysis of such *time-to-event* data were reviewed in Sect. 3.5. This chapter covers multipredictor regression techniques for the analysis of outcomes of this kind.

6.1 Survival Data

6.1.1 Why Linear and Logistic Regression Would not Work

In Sect. 3.5, we saw that a defining characteristic of survival data is *right-censoring:*

> *Definition*: A survival time is said to be *right-censored* at time t if it is only known to be greater than t.

Because of right-censoring, survival times cannot simply be analyzed as continuous outcomes. But survival data also involve an outcome *event*, so why is logistic regression not applicable? The reason is variable lengths of follow-up. In Chap. 5, the logistic model was used to study CHD events among men in the WCGS (Rosenman et al. 1964). But in that study, the investigators were able to determine

E. Vittinghoff et al., *Regression Methods in Biostatistics*, Statistics for Biology and Health, DOI 10.1007/978-1-4614-1353-0_6,

whether each one of the study participants experienced the outcome event at any time in the well-defined 10-year follow-up period; follow-up was constant across participants.

In contrast, follow-up times were quite variable in ACTG 019 (Volberding et al. 1990), a randomized double-blind placebo-controlled clinical trial of zidovudine (ZDV) for prevention of AIDS and death among patients with HIV infection. Between April 1987 and July 1989, 453 patients were randomized to ZDV and 428 to placebo. When the data were analyzed in July 1989, some had been in the study for less than a month, while others had been observed for more than 2 years. Simply applying logistic regression to the binary indicator of mortality in this example would ignore the broad variation between patients in length of follow-up. Regression adjustment for duration of follow-up would address this partially, but impose unnecessary assumptions about the relationship between event risk and duration. Although the pooled logistic regression model introduced in Sect. 5.5.2 addresses some of these concerns, that approach is more appropriate when follow-up and event time information is restricted to intervals corresponding to regular study visits. The concepts and methods introduced in this chapter offer a more complete approach to regression for survival data including observations of actual event times.

6.1.2 Hazard Function

In Sect. 3.5, we introduced the survival function and its complement, the cumulative event function, as useful summaries of the distribution of a survival time.

> *Definition*: The *survival function* at time t, denoted $S(t)$, is the probability of being event-free at t. The *cumulative event function* at time t, denoted $F(t) = 1 - S(t)$, is the complementary probability that the event has occurred by time t.

Another useful summary is the hazard function $h(t)$.

> *Definition*: The *hazard function* $h(t)$ is the short-term event rate for subjects who have not yet experienced the outcome event.

The hazard function is systematically related to both the survival and cumulative event functions.

Table 6.1 shows mortality rates for children who have recently undergone kidney transplantation, on each of the first ten days after surgery, using data from the united network for organ sharing (UNOS). At the beginning of fifth day after surgery, for example, 9,651 children remained alive and in the study, and of these, 3 died during the next 24 h, yielding an estimated death rate of 0.31 deaths per 1,000 subjects per day. From the rightmost column of the table, it appears that the mortality rate declines over the first 10 days, although the estimates spike on days 8 and 10.

Table 6.1 Mortality among pediatric kidney transplant recipients

Days since transplant	No. in follow-up	No. died	No. censored	Death rate per 1,000 subject-days
1	9,750	7	14	$7/9{,}750 \times 1{,}000 = 0.72$
2	9,729	5	8	$5/9{,}729 \times 1{,}000 = 0.51$
3	9,716	5	12	$5/9{,}716 \times 1{,}000 = 0.51$
4	9,699	7	41	$7/9{,}699 \times 1{,}000 = 0.72$
5	9,651	3	54	$3/9{,}651 \times 1{,}000 = 0.31$
6	9,594	2	57	$2/9{,}594 \times 1{,}000 = 0.21$
7	9,535	0	50	$0/9{,}535 \times 1{,}000 = 0.00$
8	9,485	4	49	$4/9{,}485 \times 1{,}000 = 0.42$
9	9,432	1	49	$1/9{,}432 \times 1{,}000 = 0.11$
10	9,382	3	28	$3/9{,}382 \times 1{,}000 = 0.32$

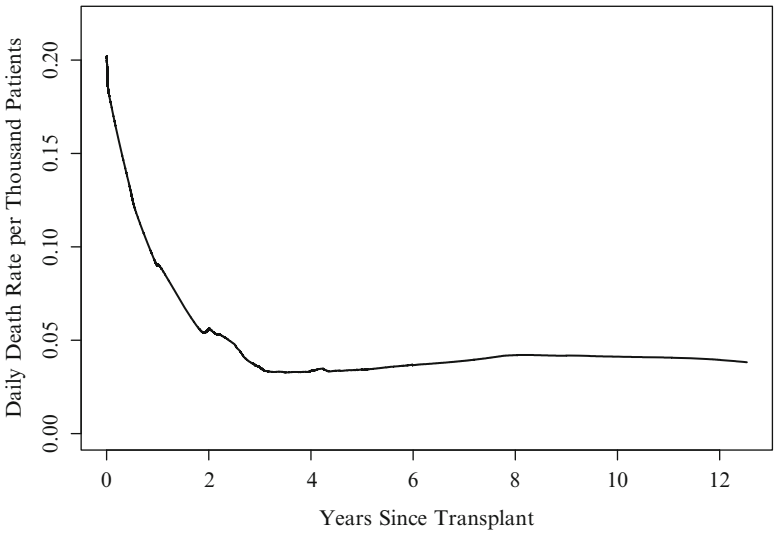

Fig. 6.1 Mortality rate for pediatric kidney transplant recipients

In Fig. 6.1, daily death rates, smoothed by LOWESS, are used to estimate the mortality hazard for a much longer time period, the first 12 years after transplantation. The mortality hazard declines rapidly over the course of the first 2 years, reaching a plateau approximately 3 years after transplantation.

6.1.3 Hazard Ratio

We now compare the hazard functions for children whose transplanted kidney was provided by a living donor, commonly a family member, and those for whom the

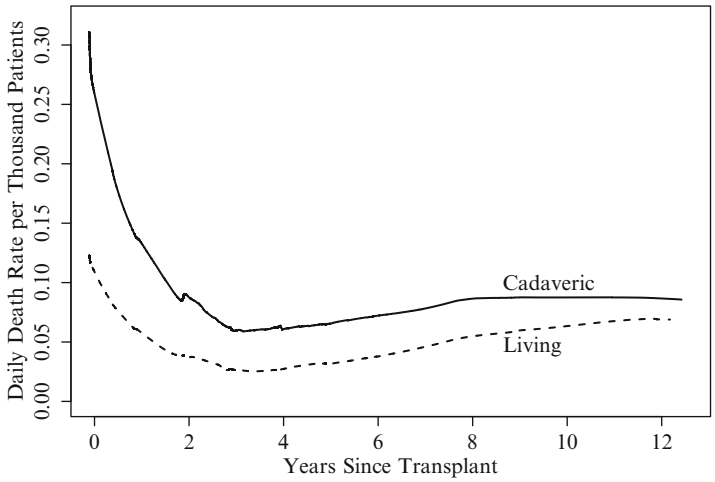

Fig. 6.2 Smoothed mortality rates for recipients by kidney donor type

Table 6.2 Smoothed death rates (per 1,000 days) by donor type

Years since transplantation	Smoothed rates		Death rate ratio
	Cadaveric	Living	
0.25	0.235	0.098	2.40
0.50	0.193	0.082	2.36
1.00	0.138	0.061	2.27
2.00	0.088	0.038	2.30
3.00	0.061	0.027	2.25
4.00	0.063	0.026	2.37
5.00	0.065	0.032	2.03

source was recently deceased. Figure 6.2 shows LOWESS-smoothed death rates for the recipients of kidneys from living and recently deceased donors. The mortality rate is considerably lower among the recipients of kidneys from living donors at all time points, but the curves are similar in shape.

Table 6.2 gives the values of the LOWESS-smoothed death rates shown in Fig. 6.2 for selected time points, which estimate the hazard functions in each group, as well as the death rate ratio, an estimate of the *hazard ratio*. We could write the hazard ratio as

$$\mathrm{HR}(t) = h_{\mathrm{c}}(t)/h_{\mathrm{l}}(t), \tag{6.1}$$

where $h_{\mathrm{c}}(t)$ is the hazard function in the recipients of kidneys from cadaveric donors, and $h_{\mathrm{l}}(t)$ is the corresponding hazard function in the reference group, the recipients of kidneys from living donors.

6.1.4 Proportional Hazards Assumption

The results in Table 6.2 show that while the mortality hazards decline over time in both groups of pediatric kidney transplant recipients, the hazard ratio is roughly constant. In other words, the hazard in the comparison group is a constant proportion of the hazard in the reference group.

> *Definition*: Under the *proportional hazards assumption*, the hazard ratio does not vary with time. That is, $\text{HR}(t) \equiv \text{HR}$.

Provided the hazards are proportional in this sense, the effect of donor source on post-transplant mortality risk can be summarized by a single number. This simplification is useful, but not necessary for the Cox proportional hazards model described in the next section. We can generalize the model by including an interaction between the predictor and time; this allows the hazard ratio for that predictor to change with time. In Sect. 6.4.2, we show how this strategy can be used to check and model nonproportional hazards with respect to a variable. This is implemented using time dependent covariates (*TDCs*), an extension of the basic Cox model introduced in Sect. 6.3.1.

6.2 Cox Proportional Hazards Model

The Cox proportional hazards regression model is a flexible tool for assessing the relationship of multiple predictors to a right-censored, time-to-event outcome, and has much in common with linear and logistic models. To understand how the Cox model works, we first consider the broader class of proportional hazards models.

6.2.1 Proportional Hazards Models

In the linear model for continuous outcomes, covered in Chaps. 4 and 10, the linear predictor $\beta_1 x_1 + \ldots + \beta_p x_p$, which captures the effects of predictors, is linked directly to the conditional mean of the outcome, $\text{E}[y|\mathbf{x}]$:

$$\text{E}[y|\mathbf{x}] = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p. \tag{6.2}$$

In the logistic model for binary outcomes, covered in Chap. 5, the linear predictor is linked to the conditional mean through the logit transformation:

$$\log \frac{p(\mathbf{x})}{1 - p(\mathbf{x})} = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p. \tag{6.3}$$

In (6.3), $p(\mathbf{x}) = \mathrm{E}[y|\mathbf{x}]$ is the probability of the outcome event for a observation with predictor values $\mathbf{x} = (x_1, \ldots, x_p)$.

In proportional hazards regression models, the linear predictor is linked through the log-transformation to the hazard ratio introduced in Sect. 6.1.3. If the hazard ratio obeys the proportional hazards assumption, and thus does not depend on time, we can write

$$\log[\mathrm{HR}(\mathbf{x})] = \log \frac{h(t|\mathbf{x})}{h_0(t)} = \beta_1 x_1 + \ldots + \beta_p x_p. \tag{6.4}$$

In (6.4), $h(t|\mathbf{x})$ is the hazard at time t for an observation with covariate value $\mathbf{x}$, and $h_0(t)$ is the *baseline hazard function*, defined as the hazard at time t for observations with all predictors equal to zero. As with the intercept in linear and logistic regression, this may mean that the baseline hazard does not apply to any possible observation, and argues for centering continuous predictors.

Solving (6.4) for $h(t|\mathbf{x})$ gives

$$\begin{aligned} h(t|\mathbf{x}) &= h_0(t) \exp(\beta_1 x_1 + \ldots + \beta_p x_p) \\ &= h_0(t)\mathrm{HR}(\mathbf{x}). \end{aligned} \tag{6.5}$$

Note that exponentiating the linear predictor ensures that $\mathrm{HR}(\mathbf{x})$ cannot be negative, as required. Furthermore, taking the log of both sides of (6.5), we obtain

$$\log[h(t|\mathbf{x})] = \log[h_0(t)] + \beta_1 x_1 + \ldots + \beta_p x_p. \tag{6.6}$$

This shows that the log baseline hazard plays the role of the intercept in other regression models, though in this case it can change over time. Furthermore, (6.6) defines a *log-linear* model, which implies that the log of the hazard is assumed to change linearly with any continuous predictors.

Note also that (6.5) defines a *multiplicative* model, in the sense that the predictor effects act to multiply the baseline hazard. This is like the logistic model, where the linear predictor acts multiplicatively on the baseline odds. In contrast, (6.2) shows that in the linear model the predictor effects are *additive* with respect to the intercept β_0.

6.2.2 Parametric Versus Semi-parametric Models

We have two options in dealing with the baseline hazard $h_0(t)$. One is to model it with a parametric function. For instance, the exponential survival model specifies that the hazard is a constant while the Weibull regression model has a hazard which is a polynomial in time. In both of these models, the baseline hazard $h_0(t)$ is specified by a small number of additional parameters, which are estimated along

Table 6.3 Cox model for type of donor

```
stcox i.txtype

No. of subjects =         9750                 Number of obs    =      9750
No. of failures =          461
Time at risk    =  38004.90961
                                               LR chi2(1)       =     44.82
Log-likelihood  =   -3952.3735                 Prob > chi2      =    0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    1.txtype |   1.879674   .1801323     6.59   0.000     1.557795     2.26806
------------------------------------------------------------------------------
```

with $\beta_1, \beta_2, \ldots, \beta_p$. If the baseline hazard is specified correctly, this approach is efficient, handles right-censoring as well as more complicated censoring schemes with ease, and makes it simple (though still risky) to extrapolate beyond the data. Of course the adequacy of the model for the baseline hazard has to be checked.

In contrast to parametric models, the *Cox model*, or Cox proportional hazards model, does not require us to specify a parametric form for the baseline hazard, $h_0(t)$. Because we still specify (6.4) as the model for the log-hazard ratio, the Cox model is considered semi-parametric. Nonetheless, estimation of the regression parameters $\beta_1, \beta_2, \ldots, \beta_p$ is done without having to estimate the baseline hazard function. Note that estimates of this function can be useful in summarizing hazards associated with particular predictor values, and can be obtained once the regression parameters are estimated (Kalbfleisch and Prentice 1980). The Cox model is more robust than parametric proportional hazards models because it is not vulnerable to misspecification of the baseline hazard. Furthermore, the robustness is commonly achieved with little loss of precision in the estimated predictor effects.

6.2.2.1 Proportionality and Multiplicativity

Figure 6.2 and the summary statistics in Table 6.2 showed that the two mortality hazards for pediatric recipients of kidney transplants from living and recently deceased donors were very nearly proportional over time, in the sense that the ratio of the LOWESS-smoothed death rates was approximately constant. So the Cox model appears appropriate for these data, because the proportional hazards assumption appears to be met for this important predictor. Table 6.3 shows the unadjusted Cox model hazard ratio estimate for `txtype`, a binary indicator identifying the group receiving transplants from recently deceased donors. The estimated hazard ratio of 1.9 (95% CI 1.6–2.3 $P < 0.0005$) is consistent with the estimates shown in Table 6.2, and suggests that receiving a transplant from a recently deceased donor roughly doubles the mortality risk at every point over the 12 years of follow-up.

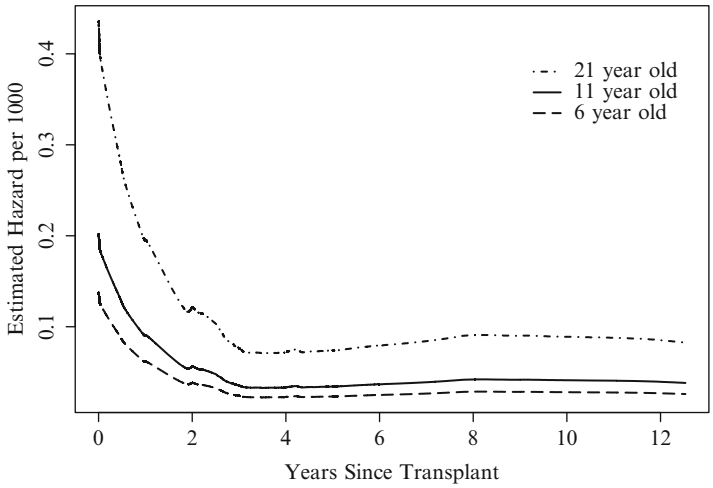

Fig. 6.3 Hazard functions for 6-, 11-, and 21-year-old transplant recipients

Another important determinant of mortality after kidney transplant is the age of the recipient. Using results from a Cox model with age as continuous (results not shown), Fig. 6.3 shows fitted hazards for 6-, 11-, and 21-year-olds. The hazards for the three groups differ proportionally. However, it is important to point out that the perfect proportionality of the hazard functions plotted in Fig. 6.3 is imposed under the fitted model, like the perfectly parallel regression lines for the additive linear model without interaction terms shown in Fig. 4.2. This is in contrast to the apparently proportional relationship between the independently smoothed death rates in Fig. 6.2, which are based only on the data.

While the hazard ratio is assumed to be constant over time in the basic Cox model, under this multiplicative model the between-group *differences* in the hazard can easily be shown to depend on $h_0(t)$ and thus on time. This is reflected in the fact that the hazard functions in Fig. 6.3 are considerably farther apart immediately after transplant when the baseline hazard is higher.

6.2.2.2 DPCA Study of Primary Biliary Cirrhosis

To illustrate interpretation of Cox model results, we consider a cohort of 312 participants in a placebo-controlled clinical trial of D-penicillamine (DPCA) for primary biliary cirrhosis (PBC) (Dickson et al. 1989). PBC destroys bile ducts in the liver, causing bile to accumulate. Tissue damage is progressive and ultimately leads to liver failure. Time from diagnosis to end-stage liver disease ranges from a few months to 20 years. During the approximate 10-year follow-up period, 125 study participants died.

Table 6.4 Cox model for treatment and bilirubin

```
stcox i.rx bilirubin

No. of subjects =          312                 Number of obs    =       312
No. of failures =          125
Time at risk    =  1713.853528
                                               LR chi2(2)       =     85.79
Log-likelihood  =   -597.08411                 Prob > chi2      =    0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|   [95% Conf. Interval]
-------------+----------------------------------------------------------------
        1.rx |   .8181612   .1500579    -1.09   0.274   .5711117    1.172078
   bilirubin |   1.163459   .0154566    11.40   0.000   1.133556    1.194151
------------------------------------------------------------------------------
```

Predicting survival in PBC patients is important for clinical decision making. The investigators collected data on age as well as baseline laboratory values and clinical signs including serum bilirubin levels, enlargement of the liver (hepatomegaly), accumulation of water in the legs (edema), and visible veins in the chest and shoulders (spiders)—all signs of liver damage.

In the sections that follow, we will illustrate use of the Cox model for testing and interpretation. This will present a series of largely unrelated models. The objective will not be to illustrate a model selection strategy.

6.2.3 Hazard Ratios, Risk, and Survival Times

Table 6.4 displays a Cox model for the effects of treatment with DPCA (`rx`) and bilirubin (`bilirubin`) on mortality risk in the PBC cohort.

The hazard ratio for treatment, 0.82, means that estimated short-term mortality risk among patients assigned to DPCA was 82% of the risk in the placebo group. This ratio is assumed to be constant over the 10 years of follow-up. Likewise, the hazard ratio for bilirubin levels means that for each mg/dL increase in bilirubin, short-term risk is increased by a factor of 1.16.

More broadly, (6.6) implies that in a model with predictors $x_1, x_2, \ldots, x_p$, coefficient β_j is the increase in the log-hazard ratio for a one-unit increase in predictor x_j, holding the values of the other predictors constant. It follows that $\exp(\beta_j)$ is the hazard ratio for a one-unit increase in x_j. Below, we show how this applies to continuous as well as binary and categorical predictors. Furthermore, for predictors with hazard ratios less than 1 ($\beta < 0$), increasing values of the predictors are associated with lower risk and longer survival times. Conversely, when hazard ratios are greater than 1 ($\beta > 0$), increasing values of the predictor are associated with increased risk and shorter survival times. In using the term *risk* in this context, it is important to keep in mind the definition of the hazard as a short-term rate and distinguish risk in this sense from cumulative risk over a defined follow-up period.

Table 6.5 Cox model for treatment and bilirubin showing coefficients

```
stcox i.rx bilirubin, nohr
                                                LR chi2(2)      =      85.79
Log-likelihood  =   -597.08411                  Prob > chi2     =     0.0000

------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        1.rx |  -.2006959   .1834088    -1.09   0.274    -.5601705    .1587787
   bilirubin |   .1513976   .0132851    11.40   0.000     .1253594    .1774358
------------------------------------------------------------------------------
```

6.2.4 Hypothesis Tests and Confidence Intervals

In the Cox model, as in the logistic model, the estimated coefficients have an approximate normal distribution when there are adequate numbers of events in the sample. The normal approximation is better for the coefficient estimates than for the hazard ratios, so hypothesis tests and confidence intervals are based on calculations involving the coefficients and their standard errors. If there are fewer than 15–25 events, the normal approximation is suspect and bootstrap CIs may work better; see Sect. 6.6.1. Table 6.5 displays the Cox model for the effects of DPCA and bilirubin on mortality risk with results on the coefficient rather than the hazard ratio scale.

For each predictor in the model, Wald Z-tests are the default used by Stata to test the null hypothesis H_0: $\beta = 0$, or equivalently that the hazard ratio equals 1. Under the null, the ratio of the coefficient estimate to its standard error tends to a standard normal, or Z, distribution with mean 0 and standard deviation 1. In Table 6.5, the Z-statistics and associated P-values for `rx` and `bilirubin` appear in the columns headed `|z|` and `P > |z|`, respectively. The evidence for the efficacy of DPCA is not persuasive ($P = 0.27$), but there is strong evidence that bilirubin levels are associated with mortality risk ($P < 0.0005$). You can verify that the test results in Table 6.4 are identical to those in Table 6.5 and refer to the Z-test involving the actual coefficients and their standard errors, and not to a Z-test involving the ratio of the hazard ratio to its standard error (Problem 6.1).

Since Cox regression is a likelihood-based method, tests for predictors can also be obtained using the LR tests introduced in Sect. 5.2.1 for the logistic regression model. The procedure is the same in this setting, comparing twice the difference in log-likelihoods for nested models to a χ^2 distribution with degrees of freedom equal to the between-model difference in the number of parameters. For instance, to obtain an LR test of the null hypothesis that the hazard ratio for treatment is 1, we would compare the log-likelihood for the model in Table 6.4 to the log-likelihood for a model with `bilirubin` as the only predictor. These log-likelihoods are –597.1 and –597.7, yielding a LR test statistic of $2[(-597.1) - (-597.7)] = 1.2$, with an associated p-value of 0.27.

In this case, the Wald and LR results are essentially identical. In most situations, these tests give results which are similar but not exactly the same. The results

be will closest when the sample size is large or the estimated hazard ratio is near 1. However, in datasets with few events, the LR test gives more accurate p-values, and so is recommended in that context. As noted in Sect. 10.4.2, qualitative discrepancies between the two test results may indicate that the model includes too many predictors for the number of events.

A 95% CI for each β is obtained by computing $\hat{\beta} \pm 1.96\text{SE}(\hat{\beta})$. Stata and other packages usually make it possible to compute CIs with other significance, or α, levels. In Stata, this can be done by using the `level()` option.

In turn, CIs for the hazard ratios are obtained by exponentiating the upper and lower limits of the CIs for the coefficients, again because the normal approximation is better on the coefficient scale. From Table 6.4, the CI for `rx`, the indicator for treatment with DPCA, shows that the data are consistent with risk reductions as large as 43%, but also with risk increases of 17%. It is also clear that the increase in risk associated with each mg/dL increase in bilirubin is rather precisely estimated (95% CI for the hazard ratio 1.13–1.19).

You can also verify that the CIs in Table 6.4 are *not* equal to the estimated hazard ratio plus or minus 1.96 times *its* standard error (Problem 6.1). For `rx`, that calculation would yield (0.52–1.11) rather than (0.57–1.17). In reasonably large samples like this one, the two intervals are usually very similar. However, since the intervals based on exponentiating the confidence limits for the coefficients are more accurate in small samples, they are the ones used in Stata.

6.2.5 *Binary Predictors*

Binary predictors can be coded as 1 and 0 and entered as numeric predictors, as opposed to categorical. For example, we could code `rx` as 1 for the DPCA arm and 0 for placebo. Then the exponentiated coefficient gives the hazard ratio for treatment versus placebo (and retains its literal interpretation as the hazard ratio for a one-unit increase in the predictor). Some alternative codings, (e.g., placebo $=$ 1 and treatment $=$ 2) would give the same results in this instance, but would complicate interpretation in the presence of an interaction involving the binary predictor. This would also make the baseline hazard harder to interpret; in the DPCA example, the baseline hazard would not refer to either the placebo or the treatment group. Thus, if binary predictors are treated as numeric, we recommend the 0/1 coding in this context as well (Problem 6.2).

6.2.6 *Multilevel Categorical Predictors*

Patients in the PBC study underwent a liver biopsy to determine their level of tissue damage. The scores ranged from 1 to 4, with increasing values reflecting

Table 6.6 Categorical fit for histology

```
. stcox i.histol

Cox regression -- Breslow method for ties

No. of subjects =          312                 Number of obs    =       312
No. of failures =          125
Time at risk    =  1713.853528
                                               LR chi2(3)       =     52.72
Log-likelihood  =   -613.62114                 Prob > chi2      =    0.0000
------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      histol |
           2 |   4.987976   5.143153    1.56   0.119     .6610611    37.63631
           3 |   8.580321   8.685371    2.12   0.034     1.179996    62.39165
           4 |   21.38031   21.57046    3.04   0.002     2.959663    154.4493
------------------------------------------------------------------------------

 testparm i.histol
 ( 1)  2.histol = 0
 ( 2)  3.histol = 0
 ( 3)  4.histol = 0
           chi2(  3) =   43.90
         Prob > chi2 =    0.0000

lincom -3.histol + 4.histol, hr
 ( 1)  - 3.histol + 4.histol = 0
------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   2.491785   .4923268    4.62   0.000     1.691727     3.67021
------------------------------------------------------------------------------
```

greater damage. When we model a multiple category variable, a series of new variables are created to represent group membership with one group serving as the reference. Results are shown in Table 6.6. By default, Stata has chosen the group with the lowest score as the reference category. Estimated hazard ratios with respect to the reference group are 5.0, 8.6, and 21.4 for the groups with ratings of 2, 3, and 4, respectively, suggesting a steady increase in the hazard with higher ratings.

In addition to the default comparisons with the selected reference group, pairwise comparisons between any two categories can be obtained using the `lincom` command, as shown in Table 6.6 for groups 3 and 4. The hazard in group 4 is 2.5 times higher than in group 3 (95% CI 1.7–3.7, $P < 0.0001$).

6.2.6.1 Categories with No Events

In our example, the default reference category is sensible and does not cause problems. However, categories may sometimes include no events, because the group is small or cumulative risk is low. Hazard ratios with respect to a reference category with no events are infinite, and the accompanying hypothesis tests and CIs are hard to interpret. In this case, selecting an alternative reference group can

correct the problem, although the hazard ratio, Wald test, and CI for the category without events, with respect to the new reference category, will remain difficult to interpret.

6.2.6.2 Global Hypothesis Tests

As in logistic models, global hypothesis tests for the overall effect of a multilevel categorical predictor can be conducted using Wald or likelihood ratio (LR) χ^2 tests, with degrees of freedom equal to the number of categories minus 1. The Wald test result ($\chi^2 = 43.9, P < 0.00005$), obtained using the `testparm` command, is displayed in Table 6.6. The LR test result ($\chi^2 = 52.7, P < 0.00005$) also appears in the upper right corner of the table. Note that if covariates were included in the model, this default Stata output would refer to a test of the overall effect of *all* covariates in the model, not just `histology`; thus a LR test focused on the overall effect of `histology` would require combining the results of models with and without this predictor. Finally, a logrank test, as in Sect. 3.5.6, is available; this yields a χ^2 of 53.8 ($P < 0.0001$). The tests agree closely and all show that the groups with different histology scores do not have equal survival.

The statistical significance of pairwise comparisons should be interpreted with caution, especially if the global hypothesis test is not statistically significant, as discussed in Sect. 4.3.4. With a large number of categories, multiple comparisons can lead to inflation of the familywise type-I error rate (FER); Bonferonni, Sidak, and Scheffé adjustments are implemented in the `contrast` command, as explained in Sect. 4.3.4. In addition, some comparisons may lack power due to small numbers in either of the categories being compared.

6.2.6.3 Ordinal Predictors and Tests for Trend

The histology score is ordinal, suggesting a more specific question: does the log mortality hazard increase linearly with higher histology ratings? This question can be addressed using tests for trend across categories like those introduced in Sect. 4.3.5. Note that these tests, like other hypothesis tests for the Cox model, are conducted using the coefficients and their standard errors, rather than the relative hazards. Thus for the Cox model, these linear trend tests assess log-linearity of the hazard ratios. From Table 4.8, the trend test for a four-category variable such as `histol` is

$$-\beta_2 + \beta_3 + 3\beta_4 = 0. \tag{6.7}$$

In Stata, the test for linear trend can be obtained using the `test` or `contrast` commands—see Sect. 4.3.5 for an explanation of use of contrast command. The three equivalent tests presented in Table 6.7 confirm an increasing linear trend across the four histologic categories ($\chi^2 = 10.23, P = 0.0014$).

Table 6.7 Linear trend test for histology

```
test -1*2.histol + 3.histol +3*4.histol=0
 ( 1)  - 2.histol + 3.histol + 3*4.histol = 0
           chi2(  1) =   10.23
         Prob > chi2 =    0.0014

contrast {histol -3 -1 1 3}, noeffects
Contrasts of marginal linear predictions
Margins      : asbalanced
------------------------------------------------
             |         df        chi2     P>chi2
-------------+----------------------------------
      histol |          1       10.23     0.0014
------------------------------------------------

contrast q(1).histol, noeffects
Contrasts of marginal linear predictions
Margins      : asbalanced
------------------------------------------------
             |         df        chi2     P>chi2
-------------+----------------------------------
      histol |          1       10.23     0.0014
------------------------------------------------
```

Table 6.8 Test of departure from linear trend

```
. quietly stcox histol i.histol

. testparm i.histol
 ( 1)  2.histol = 0
 ( 2)  3.histol = 0
           chi2(  2) =    1.24
         Prob > chi2 =    0.5385

. contrast q(2/3).histol, noeffects
Contrasts of marginal linear predictions
Margins      : asbalanced
------------------------------------------------
             |         df        chi2     P>chi2
-------------+----------------------------------
      histol |
(quadratic)  |          1        0.44     0.5085
    (cubic)  |          1        1.15     0.2832
      Joint  |          2        1.24     0.5385
------------------------------------------------
```

It is also possible to check whether the linear trend adequately captures the pattern of the coefficients across categories, or whether there are also important departures from this trend. To do this, we use a model with both categorical and log-linear terms for `histol`. Then a Wald test for the joint effect of the categorical terms, obtained using the `testparm` command, can be used to assess the departure from log-linearity. We also implement this test using the `contrast` command; see Sect. 4.3.5 for the rationale for these tests.

The result ($\chi^2 = 1.24$, $P = 0.54$) suggests that a linear trend across categories is an adequate description of the association between histology score and mortality risk. However, it is not uncommon for both trend and departure from trend to be statistically significant, signaling a more complex pattern in risk (Table 6.8).

Table 6.9 Cox model for age in 1-year units

```
stcox age
                                                LR chi2(1)      =      20.51
Log-likelihood  =   -629.72592                  Prob > chi2     =     0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |    1.04081   .0091713    4.54   0.000     1.022989    1.058941
------------------------------------------------------------------------------
```

6.2.7 Continuous Predictors

Age at enrollment of participants in the PBC study was recorded in years. The Cox model shown in Table 6.9 shows that the hazard ratio for a 1-year increase in `age` is 1.04 (95% CI 1.02–1.06, $P < 0.0005$). The hazard ratio for continuous predictors is affected by the scale of measurement. In the PBC study, ages range from 26 to 78; thus, a 1-year difference in age is small compared to the range of values. A 5-year increase in age might provide a more clinically interpretable result (Problem 6.5).

Using (6.5), we can write down the ratio of the hazards for any two patients who differ in age by k years—that is, for a patient at age $x + k$ compared with another at age x:

$$\begin{aligned}\frac{h_0(t)\exp(\beta(x+k))}{h_0(t)\exp(\beta x)} &= \frac{\exp(\beta(x+k))}{\exp(\beta x)} \\ &= \exp(\beta(x+k)-\beta x) \\ &= \exp(\beta k). \end{aligned} \tag{6.8}$$

Thus a k-unit change in a predictor multiplies the hazard by $\exp(\beta k)$, no matter what reference value x is considered.

Applying (6.8), with $\hat{\beta} = \log(1.04081)$ being the log of the hazard ratio for `age` from Table 6.9, the hazard ratio for an increase in age of 5 years is $\exp(\hat{\beta}5) = 1.22$. The same transformation can be applied to the confidence limits for `age` giving a 95% CI for a 5-year increase in age of 1.12–1.33. Equivalently, we could raise the hazard ratio estimate for an increase of one unit to the fifth power, that is, $[\exp(\beta)]^k$, and apply the same operation to the confidence limits (Problem 6.6).

The hazard ratio for a five-unit change can also be obtained by defining a new variable `age5` equal to age in years divided by 5. The Cox model for `age5` appears in Table 6.10. Note that the Wald and LR test results are identical in Tables 6.9 and 6.10; changes in the scale of a continuous variable do not affect these tests.

Hazard ratios can be interpreted in terms of percent changes in risk. It is easy to see from Table 6.9 that estimated mortality risk among PBC patients increases about 4% for every year increase in age. We could also compute the percent increase risk associated with larger increases in age. A k-unit increase in the predictor implies a

Table 6.10 Cox model for age in 5-year units

```
stcox age5
                                                LR chi2(1)      =      20.51
Log-likelihood  =   -629.72592                  Prob > chi2     =     0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        age5 |   1.221397   .0538127    4.54   0.000     1.120352    1.331556
------------------------------------------------------------------------------
```

Table 6.11 Unadjusted Cox model for bilirubin

```
stcox bilirubin
                                                LR chi2(1)      =      84.59
Log-likelihood  =    -597.6845                  Prob > chi2     =     0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   bilirubin |   1.160509   .0151044   11.44   0.000     1.131279    1.190494
------------------------------------------------------------------------------
```

$100(\exp\hat{\beta}k-1)\%$ change in risk. Note that this is the back transformation presented in Sect. 4.7.5 for linear regression models with log-transformed outcomes. Using the log of the hazard ratio estimate from Table 6.9 in place of $\hat{\beta}$, this calculation gives 22% for the increase in mortality risk associated with a 5-year increase in age, a result we could get more directly from Table 6.10.

6.2.8 Confounding

The definition of confounding in Sect. 4.4 is not specific to the linear regression model. The conceptual issues and statistical framework for dealing with confounding are similar across all regression models and discussed in more depth in Chap. 9. To illustrate regression adjustment to control confounding in the Cox model, we examined the association between bilirubin levels and survival among patients in the DPCA trial. We first fit the simple Cox model which appears in Table 6.11. For each one-point increase in baseline bilirubin, the hazard is increased by 16%.

However, patients with higher bilirubin may also be more likely to have hepatomegaly, edema, or spiders—other signs of liver damage which are correlated with elevated bilirubin levels but not mediators of its effects, and all associated with higher mortality risk. Table 6.12 shows the estimated effect of bilirubin on mortality risk adjusted for hepatomegaly, edema, and spiders.

The adjusted hazard ratio for a one-point increase in bilirubin is 1.12 (95% CI 1.09–1.15, $P < 0.0005$). This coefficient represents the effect of a one-unit change in bilirubin while holding edema, hepatomegaly, and spiders constant. The other predictors, which may reflect other aspects of PBC-associated damage to the liver,

Table 6.12 Adjusted Cox model for bilirubin

```
stcox bilirubin i.edema i.hepatom i.spiders

                                                LR chi2(4)      =    118.82
Log-likelihood  =   -580.56805                  Prob > chi2     =    0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   bilirubin |   1.118276   .0166316     7.52   0.000     1.086149    1.151353
     1.edema |   2.126428   .4724983     3.40   0.001     1.375661    3.286927
   1.hepatom |   2.050617    .434457     3.39   0.001     1.353765    3.106173
   1.spiders |   1.474788     .28727     1.99   0.046      1.00676    2.160393
------------------------------------------------------------------------------
```

account for a modest proportion of the unadjusted effect of bilirubin, and clearly contribute independent information about mortality risk. The attenuation of the unadjusted hazard ratio for bilirubin in the adjusted model is typical of confounding.

6.2.9 Mediation

Mediation can also be addressed with the Cox model, using the strategies outlined in Sect. 5.2.3. Here, we use data from the FIT trial Black et al. 1996b, which showed that treatment with alendronate can reduce the risk of fracture in the spine. The relative hazard of fracture of participants on alendronate was 0.52 compared with placebo with a 95% CI from 0.41 to 0.66 ($p < 0.001$). Measures of BMD were also increased by alendronate—the placebo arm showed a 0.8% decrease from baseline while the treated group had a 3.8% increase in BMD from baseline, yielding a net increase in BMD due to alendronate of 4.5% with 95% CI from 4.2% to 4.8%. We can reject a null hypothesis that change in BMD is equal for the two arms ($p < 0.001$). This raises the natural question as to whether the reduction in fracture risk is mediated, or captured by, the observed changes in BMD. Whenever we approach an analysis of mediation, a causal role of the primary predictor is implied. Hence, we should believe that the association between the primary predictor and the possible mediator is a causal one. Here, we have a randomized trial and can comfortably make such an assumption.

As we showed in Sect. 5.2.3, we establish mediation by requiring an association between the predictor of interest (treatment by alendronate in this example) with the mediator (BMD here) *and* the outcome (time to fracture here). The statistical test to establish mediation requires that we test each of these associations at the 0.05 level. Both null hypotheses are rejected with $p < 0.01$, establishing that BMD plays some mediating role in the effect of alendronate on fracture risk.

A fuller picture emerges when we examine the magnitude of the direct effect of alendronate on fracture risk. We can approach this by examining hazard ratios for treatment group in a Cox model which includes an adjustment for BMD. Because

Table 6.13 Cox model for FIT data assessing mediating value of changes in BMD due to alendronate

```
. stcox i.treat i.smoking age bmd_diff bmd_base, strata(frac_base)

No. of subjects =         5324                  Number of obs    =      5324
No. of failures =          294
Time at risk    =  20494.62287
                                                LR chi2(6)       =    123.14
Log-likelihood  =   -1911.6879                  Prob > chi2      =    0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     1.treat |   .6237068    .082513   -3.57   0.000     .4812505    .8083319
             |
     smoking |
           1 |   1.107652   .1422723    0.80   0.426     .8611343    1.424741
           2 |   1.391522    .254672    1.81   0.071     .9720888    1.991931
             |
         age |   1.069186   .0116993    6.11   0.000       1.0465    1.092364
    bmd_diff |   .8274497   .0558578   -2.81   0.005      .724904    .9445018
    bmd_base |    .004533   .0082887   -2.95   0.003     .0001259    .1632403
------------------------------------------------------------------------------
                                                      Stratified by frac_base
```

of the possibility of confounding between the outcome and the mediator, we recommend including potential confounders of the outcome/mediator relationship in a Cox model which examines direct effects. Table 6.13 fits a Cox model to the risk of a spinal fracture to examining the mediating value of change in BMD after adjustment for baseline BMD, smoking, age, and history of fractures at baseline. The latter variable used as a stratification variable in the Cox model because direct adjustment yields an infinite hazard ratio. The Cox model shows that there is clearly a statistically and clinically important benefit of treatment even after adjustment for BMD.

There is a temptation to compare the effect of the treatment prior to and after adjustment for the mediator. A model with treatment alone yields a hazard ratio of 0.52 with 95% CI of 0.41 to 0.66. However, it is not straightforward to compare hazard ratios for treatment across the models. Methods that compare these coefficients directly using "proportion of the treatment effect explained" are problematic. For instance, a variable which is strongly associated with the outcome but not a mediator can change the coefficient for the treatment effect in a Cox model. Hence, we do not recommend methods which calculate the "proportion of effects explained" for examining mediation.

6.2.10 Interaction

The concept of interaction presented in Sect. 4.6 is also common to other multipredictor models. To illustrate its application to the Cox model, we examined

Table 6.14 Cox model with interaction

```
stcox rx##hepatom

Log-likelihood  =    -619.7079                     Prob > chi2     =    0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       1.rx |   .8365301   .2778607    -0.54   0.591     .4362622    1.604041
  1.hepatom |    3.15151   .8380138     4.32   0.000     1.871444     5.30714
            |
 rx#hepatom |
       1 1  |   1.099791   .4343044     0.24   0.810     .5071929    2.384775
------------------------------------------------------------------------------

. lincom 1.rx+1.rx#1.hepatom, hr
 ( 1)  1.rx + 1.rx#1.hepatom = 0

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   .9200085   .1963396    -0.39   0.696     .6055309    1.397807
------------------------------------------------------------------------------
```

Table 6.15 Cox model with interaction

Group	rx	hepatom	$h(t\|\mathbf{x})$
1	Placebo	No	$h_0(t)$
2	DPCA	No	$h_0(t)\exp(\beta_1)$
3	Placebo	Yes	$h_0(t)\exp(\beta_2)$
4	DPCA	Yes	$h_0(t)\exp(\beta_1+\beta_2+\beta_3)$ $= h_0(t)\exp(\beta_1)\exp(\beta_2)\exp(\beta_3)$

interaction between two binary variables in the PBC data, treatment with DPCA (`rx`), and the presence of liver enlargement or hepatomegaly (`hepatom`). This analysis examines the hypothesis that the effect of treatment is modified by the presence of hepatomegaly. As in linear and logistic models, interaction is handled by including additional terms in the model. In Stata, interaction terms are created by including the # operator between the two interacting variables. Including the ## operator between the two variables is shorthand for the interaction term and each of the two variables themselves. The interaction model is shown in Table 6.14.

Column 4 of Table 6.15 shows the hazard functions for the four groups defined by treatment and hepatomegaly (Problem 6.7). The coefficients β_1, β_2, and β_3 correspond to the predictors `rx`, `hepatom`, and `rx#hepatom`, where the latter is the interaction term. We obtain the hazard ratios of interest by dividing the hazard functions for the different rows. Specifically, the comparison of the hazard for group 2 to the hazard for group 1 gives the effect of DPCA in the absence of hepatomegaly. The model specifies that the ratio of these is $\exp(\beta_1)$. In Table 6.14, the estimated hazard ratio for `rx` is 0.84 (95% CI 0.44–1.60, $P = 0.59$).

Similarly, the ratio of the hazard for group 4 to the hazard for group 3, or $\exp(\beta_1)\exp(\beta_3)$, gives the effect of DPCA in the presence of hepatomegaly. From Table 6.14, the estimated effect is then the product of the estimated hazard ratios for `rx` and `rx#hepatom`, or $0.84 \times 1.1 = 0.92$. This estimate, along with a 95% CI (0.61–1.40) and p-value (0.70), can also be obtained using the `lincom` command shown in Table 6.14.

It follows that the interaction hazard ratio $\exp(\beta_3)$ gives the ratio of the DPCA treatment effects among patients with and without hepatomegaly. In Table 6.14, the estimated hazard ratio for `rx#hepatom` is 1.1 (95% CI 0.51–2.4, $P = 0.81$). The Z-test of H_0: $\beta_3 = 0$ assesses the equality of the effects of DPCA in the two groups.

To interpret these negative findings fully, as discussed in Sect. 3.7, both the point estimates and CIs need to be considered. The stratum-specific treatment effect estimates as well as the interaction are weakly negative, in the sense that the point estimates represent almost no effect or interaction, but the confidence limits include fairly large effects. In view of the weak evidence for interaction, the overall—also negative—finding for treatment with DPCA is the more sensible summary. Similar methods can be used to obtain estimates of the effect of hepatomegaly stratified by treatment assignment: that is, by comparing groups 3 and 1, then 4 and 2.

Interactions involving continuous or multilevel categorical predictors can also be set up using the # and ## operators, but as Sect. 4.6 explains, care must be taken with these more complex cases.

6.2.11 Model Building

Model building with the Cox model is similar to other regression models. Chapter 10 discusses the issues and makes recommendations. To prevent erosion of efficiency as well as bias, models should avoid including too many predictors for the number of observed events. A familiar guideline (Peduzzi et al. 1995, 1996; Concato et al. 1995) prescribes at least ten events per predictor. Vittinghoff and McCulloch (2007) show that as few as five events per predictor may give consistent results in cases where the additional covariates are needed to rule out confounding, but point out that precision in this case may often be poor.

6.2.12 Adjusted Survival Curves for Comparing Groups

Suppose we would like to examine the survival experience of pediatric recipients of kidney from living as compared to recently deceased donors, using the UNOS data. Kaplan–Meier curves, introduced in Sect. 3.5.2, would be a good place to start and are shown in Fig. 6.4.

In accord with the hazard ratio of 2.1 estimated by the unadjusted Cox model shown in Table 6.3, the curves show superior survival in the group with living

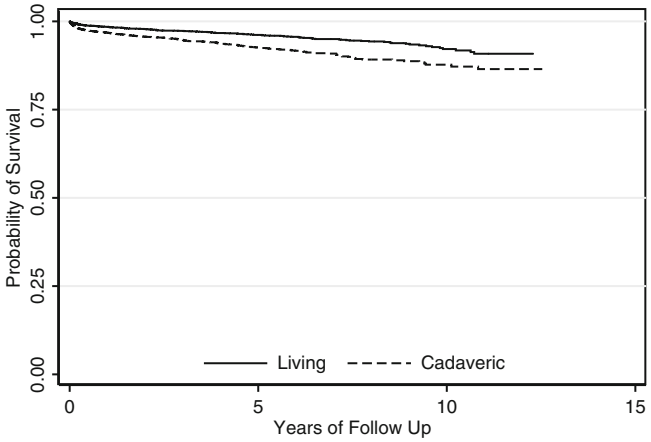

Fig. 6.4 Kaplan–Meier curves for transplant recipients by donor type

donors. However, there are two potentially important confounders of this effect. First, living donors are more likely to be related and thus are closer tissue matches, as reflected in the number of matching human leukocyte antigen (HLA) loci (range 0–6). Second, cold ischemia time (essentially the time spent in transport) is shorter for kidneys obtained from living donors. After adjustment for these two factors, the hazard ratio for donor type is reduced to 1.3 (95% CI 0.9–1.9, $P = 0.19$).

To see how adjusted survival curves might be constructed, first recall that adjustment for these covariates implies that adjusted curves for the two groups should differ only by donor type, with the other covariates being held constant. Curves meeting these criteria can be obtained using the coefficient estimates from the Cox model and an estimate of the baseline survival function, $\hat{S}_0(t)$, based on the Breslow baseline hazard estimate described earlier. Like the baseline hazard, the baseline survival function refers to observations with all predictor values equal to zero. If we assume a proportional hazard model, then a formula which links hazard and survival functions implies, the survival function follows:

$$\left\{\hat{S}_0(t)\right\}^{\exp(\hat{\beta}_1 x_1 + \ldots + \hat{\beta}_p x_p)}. \tag{6.9}$$

That is, we raise the baseline survival to the $\exp(\hat{\beta}_1 x_1 + \ldots + \hat{\beta}_p x_p)$ power. To evaluate (6.9), we need to specify a value for each of the predictors. In our example with three predictors, we would need to choose and hold constant values for x_2 (cold ischemia time) and x_3 (number of matching HLA loci), then generate the two curves by varying the predictor x_1 (recently deceased versus living donor).

It is conventional to use values for the adjustment variables which are close to the "center" of the data. Thus we centered cold ischemia time at its mean value of 10.8 h and number of matching variable HLA loci at its median, three. With

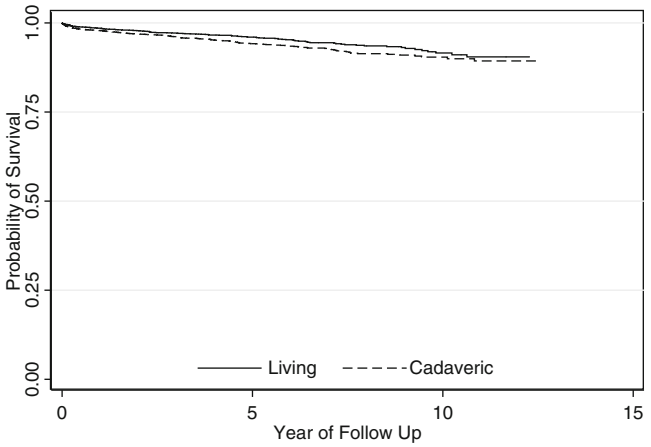

Fig. 6.5 Adjusted survival curves for transplant recipients by donor type

this centering, the baseline hazard and survival functions now refer to observations with cold ischemia time of 10.8 h, three matching HLA loci, and a living donor. Then our adjusted estimate of the survival function for the group with living donors, holding the covariates constant at the chosen values, is $\hat{S}_0(t)$, while the corresponding estimate for the group with recently deceased donors is $\{\hat{S}_0(t)\}^{\exp(\hat{\beta}_1)}$. These adjusted curves, obtained in Stata using the `stcurve` command, are shown in Fig. 6.5. The differences between the survival curves are, as expected, narrower after adjustment. Note that the adjusted survival curves could also be estimated using a stratified Cox model, as discussed in Sect. 6.3.2.

6.2.13 Predicted Survival for Specific Covariate Patterns

The estimated survival function (6.9) is also useful for making predictions for specific covariate patterns. For example, consider predicting survival for a PBC patient based on hepatomegaly status and bilirubin level, the two strongest predictors in the model shown in Table 6.12. Figure 6.6 displays the predicted survival curve for a PBC patient with hepatomegaly and a bilirubin level of 4.5 mg/dL. From the curve, the median survival function for this covariate pattern is 6.3 years. Survival probabilities at key time points can likewise be read from the plot: at 5 years, predicted survival for this covariate pattern is below 60%, and by 10 years, it has dropped to less than 20%. However, mean survival cannot be estimated in this case, because the longest follow-up time in the PBCA data is censored (Sect. 3.5).

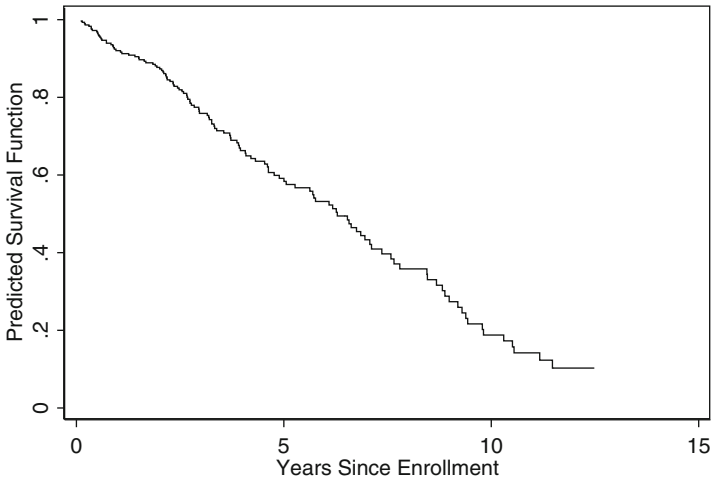

Fig. 6.6 Predicted survival curve for PBC covariate pattern

6.3 Extensions to the Cox Model

6.3.1 Time-Dependent Covariates

So far we have only considered fixed predictors measured at study baseline, such as bilirubin in the DPCA study. However, multiple bilirubin measurements were made over the 10 years of follow-up, and these could provide extra prognostic information. A special feature of the Cox model is that these valuable predictors can be included as TDCs.

Definition: A *time-dependent covariate* in a Cox model is a predictor whose values may vary with time.

In some cases, use of TDCs is critical to obtaining reasonable effect estimates. For example, Aurora et al. (1999) followed 124 patients to study the effect of lung transplantation on survival in children with cystic fibrosis. The natural time origin in this study is the time of *listing* for transplantation, not transplantation itself, because the children are most comparable at that point. However, waiting times for a suitable transplant can be long, and there is considerable mortality among children on the waiting list.

In this context, lung transplantation has to be treated as a TDC. To see this, consider the alternative in which transplantation is modeled as a fixed binary covariate, in effect comparing mortality risk in the group of children who undergo transplantation during the study to risk among those who do not. This method can

make transplantation look more protective than it really is. Here is how the artifact, sometimes called *immortal time bias* (Suissa 2008), comes about:

- Because transplanted patients must survive long enough to undergo transplantation, and waiting times can be long, the survival times measured from listing forward will on average be longer in the transplanted group even if transplantation has no protective effect.
- Because of this, children in the transplanted group are selected for better prognosis. So the randomization assumption discussed in Sect. 9.1.4 does not hold.
- Children are counted as having received a transplant from the time of listing forward, in many cases well before transplantation occurs. As a result they appear to be protected by a procedure that has not yet taken place. This illustrates the general principle that we can get into trouble by using information from the future to estimate current risk.

Treating transplantation as a TDC avoids this artifact. For each child, we define an indicator of transplantation $X(t)$, which takes on value 0 before transplantation and 1 subsequently. For children who are not observed to undergo transplantation, $X(t)$ retains its original value of 0. Thus in an unadjusted model, the hazard at time t can be written as

$$\begin{aligned} h(t|x) &= h_0(t)\exp\{\beta X(t)\} \\ &= \begin{cases} h_0(t) & \text{before transplantation} \\ h_0(t)\exp(\beta) & \text{at or after transplantation.} \end{cases} \end{aligned} \tag{6.10}$$

So now, all children are properly classified at t as having undergone transplantation or not, and we avoid the artifact that comes from treating transplantation as a fixed covariate. Note that Kalbfleisch and Prentice (1980) cite additional conditions concerning the allocation of transplants that must be met for the randomization assumption to hold and an unbiased estimate of the effect of transplantation to be obtained.

The transplantation TDC is relatively simple, because it is binary and cannot change back in value from 1 to 0. In practice, however, use of TDCs in Cox models is often more complicated. Some additional considerations include the following:

- In most prospective studies, predictors like bilirubin will only be measured occasionally, but we need a value at each event time. A commonly used approach is to evaluate $X(t)$ using the most recent measurement before t, but this so-called *last observation carried forward* (LOCF) approach is susceptible to bias; we return to this in Chap. 11. More difficult is a two-stage approach in which we first model the mean trajectory of the TDC for each subject. Then in the second stage we can set $X(t)$ equal to its expected value at t, based on the first-stage model. However, fitting and inference are both complicated in this procedure (Self and Pawitan 1992; DeGruttola and Tu 1994; Wulfsohn and Tsiatis 1997; Tsiatis and Davidian 2004).

- While $X(t)$ cannot legitimately be evaluated using information from the future, it often should be evaluated using all available information up until t. Consider two PBC patients, one with bilirubin values of 0.8 and 3.5 at baseline and year two, and the other with values of 2.5 and 3.5 at those times. In evaluating a TDC for bilirubin at year two, it might not be adequate to account only for the most recent values. A commonly used approach is to include the baseline value as a fixed covariate along with the change since baseline as a TDC. But other combinations of baseline and TDCs summarizing history up to t may be more appropriate.
- Mediation can be evaluated using TDCs, extending the analysis of mediation of the effect of alendronate on fracture by first year changes in BMD, treated as a fixed covariate, as discussed in Sect. 6.2.9. For example, we could examine mediation of the effects of ZDV via its effects on CD4 counts in the ACTG 019 trial by assessing both links in the hypothesized indirect pathway. Specifically, we might use a model for repeated measures, covered in Chap. 7, to assess ZDV effects on CD4 counts over time, and then assess the independent effects of post-randomization CD4 values in a Cox model for AIDS-free survival, controlling for treatment. Finally, we might informally compare the effect estimates for ZDV before and after adjustment for post-randomization CD4 counts.
- Special methods are needed if a TDC both confounds and mediates the effects of a time-dependent exposure or treatment. Suppose we wanted to evaluate the overall effect of highly active anti-retroviral therapy (HAART) on progression to AIDS, using data from an observational cohort. To avoid immortal time bias, HAART would need to be modeled as a TDC. Now suppose we attempt to control confounding by disease severity at treatment initiation by adjusting for time-dependent prognostic measures including CD4 count. The problem is that the effects of HAART on progression to AIDS are also *mediated* via its effects on CD4 count, so this would adjust away some of the protective effect of treatment. As a result, we would not obtain an estimate of the *overall* treatment effect. In Sect. 9.5, we discuss a solution to this problem using *IPW*.
- Ideally TDCs are measured at regularly scheduled visits, so ascertainment does not depend on prognosis. Missing visits can induce bias if the missingness is related to the value of the TDC that would have been obtained. Likewise, ascertainment of TDCs by clinical chart review can be fraught with pitfalls.
- In Stata, accommodating TDCs like the post-randomization CD4 counts in the ACTG 019 example requires a specially constructed dataset with multiple records for each unit. The `stsplit` and `stjoin` commands make this straightforward. In Sect. 6.4.2, we also show how the `stcox` option `tvc` accommodates a different kind of TDC, specifically interactions between a fixed covariate and time, which are useful in dealing with violations of the *proportional hazards assumption*.

6.3.2 *Stratified Cox Model*

Suppose we want to model the effect of edema (coded 1 for patients with edema and 0 for others) among patients with PBC in the DPCA cohort. Then in an unadjusted model, the hazard for patients with edema is $h(t|x) = h_0(t)\exp(\beta)$, while for other patients it is just $h_0(t)$. So the hazard for patients with edema is modeled as a constant proportion $\exp(\beta)$ of the baseline hazard $h_0(t)$.

However, we will show in Sect. 6.4.2 that the proportional hazards assumption does not hold for `edema`. We can accommodate the violation by fitting a stratified Cox model in which a separate baseline hazard is used for patients with and without edema. Specifically, we let

$$h(t|\texttt{edema} = 1) = h_{01}(t) \tag{6.11}$$

for patients with edema, and

$$h(t|\texttt{edema} = 0) = h_{00}(t) \tag{6.12}$$

for other patients. Now the hazards for the two groups can differ arbitrarily.

Generalizing from `edema` to a stratification variable with two or more levels, and to a model with covariates $(x_1, \ldots, x_p)$, the hazard for an observation in stratum j would have the form

$$h_{0j}(t)\exp(\beta_1 x_1 + \ldots + \beta_p x_p). \tag{6.13}$$

Note that in this model we assume that the effect of each of the covariates is the same across strata; later, we examine methods for relaxing this assumption. It is also important to point out that while the stratified, adjusted survival curves presented in Sect. 6.2.12 above can give a clear visual impression of the effect of the stratification variable after adjustment, current methods for the stratified Cox model do not allow us to estimate or test the statistical significance of its effect. Thus stratification could be used in our example to adjust for edema, but might be less useful if edema were a predictor of primary interest. In Sect. 6.4.2, we show how TDCs can be used to obtain valid estimates of the effects of a predictor which violates the proportional hazards assumption. Stratification is also useful in the analysis of stratified randomized trials. We pointed out in Sect. 10.2.6 that we need to take account of the stratification to make valid inferences. But we also need to avoid making an unwarranted assumption of proportional hazards for the stratification variable that could potentially bias the treatment effect estimate. The stratified Cox model is easy to implement in Stata as well as other statistical packages. In ACTG 019, participants were randomized within two strata defined by baseline CD4 count. To conduct the stratified analysis, we defined `strcd4` as an indicator coded 0 for the stratum with baseline CD4 count of 200–499 cells/mm^3 and 1 for the stratum with baseline CD4 of less than 200. The stratified model for the

Table 6.16 Cox model for treatment with ZDV, stratified by baseline CD4

```
stcox i.rx, strata(strcd4)
                                                LR chi2(1)      =       7.36
Log-likelihood  =   -276.45001                  Prob > chi2     =     0.0067

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        1.rx |   .4646665   .1362697   -2.61   0.009     .2615261    .8255963
------------------------------------------------------------------------------
                                                      Stratified by strcd4
```

effect of ZDV treatment (`rx`) is shown in Table 6.16. In this instance, the estimated 54% reduction in risk for treatment with ZDV is the same as an estimate reported below in Sect. 6.6.3, which was adjusted for rather than stratified on CD4.

6.3.2.1 Number of Strata

Stratification is a flexible approach to adjustment for a categorical variable even when it has a large number of levels. An example is in a multicenter randomized trial with many centers. For stratification to work well, there do need to be a reasonable number of events (about 5 to 7) in each stratum. When the number of strata gets large, there can be some loss of efficiency in estimation of the treatment or other covariate effects, since the stratified model does not "borrow strength" across strata. Nonetheless, Glidden and Vittinghoff (2004) showed that in this situation, the stratified Cox model performs better than an unstratified model in which the strata are entered into the model as a nominal categorical predictor.

6.3.2.2 Interaction Between Stratum and a Predictor of Interest

In Table 6.16, the model assumes that the ZDV effect is the same in both strata. It is possible, however, that patients with less severe HIV disease, as reflected in higher CD4 counts, may respond better to ZDV. Such an interaction between stratum and treatment can be examined by including a product term between the treatment and stratum indicators. Note that in the stratified model only the product term `i.rx#strcd4` and the treatment indicator `rx` term are entered as predictors. The predictor `strcd4` is dropped automatically by Stata, because it has already been incorporated as a stratification factor. In Table 6.17, we find persuasive evidence of an effect of ZDV (rx = 1) in the higher CD4 stratum (strcd4 = 0) with hazard ratio of 0.32 (95% CI 0.14–0.74, $P = 0.008$). However, from the `lincom` result, we derive the effect of ZDV in the lower CD4 stratum (strcd4 = 1) where there is weak evidence for a protective effect of ZDV (hazard ratio 0.71, 95% CI 0.32–1.65, $P = 0.43$). There is the suggestion for interaction (hazard ratio 0.45, 95% CI 0.14–1.48, $P = 0.19$), given by the product term `rx#strcd4`, although this is not statistically significant.

Table 6.17 Stratified fit with interaction term

```
. stcox i.rx##i.strcd4, strata(strcd4)

No. of subjects =          880                     Number of obs    =       880
No. of failures =           55
Time at risk    =       354872
                                                    LR chi2(2)       =      9.14
Log-likelihood  =   -275.56324                      Prob > chi2      =    0.0104

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        1.rx |   .3211889    .136976   -2.66   0.008     .1392362    .7409156
    1.strcd4 |  (omitted)
             |
   rx#strcd4 |
        1 1  |   2.218026   1.342113    1.32   0.188      .677501    7.261448
------------------------------------------------------------------------------
                                                              Stratified by strcd4
. lincom 1.rx+1.rx#1.strcd4, hr

 ( 1)  1.rx + 1.rx#1.strcd4 = 0

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   .7124052    .305808   -0.79   0.430      .307142    1.652399
------------------------------------------------------------------------------
```

6.3.2.3 Stratified and Adjusted Survival Curves

In Sect. 6.2.12, we presented adjusted survival curves for pediatric kidney transplant recipients according to donor type, based on an adjusted model in which the effect of donor type was modeled as proportional. We can also obtain adjusted survival curves according to the levels of a stratification factor. We will show in Sect. 6.4.2 that the effects of baseline edema on mortality risk among PBC patients in the DPCA cohort were not proportional. Suppose we would like to compare the survival curves according to edema, adjusting for age. As in the earlier example, we need to specify a value for age in order to estimate the survival curves, and make a similar choice in centering age on its mean of 50. Under the stratified Cox model, the survivor function for a PBC subject with centered `agec` is given by

$$[S_{0j}(t)]^{\exp(\beta \texttt{agec})}. \tag{6.14}$$

The adjusted survival curves for the edema ($j = 1$) and no edema ($j = 0$) strata, adjusted to age 50 (i.e., $\texttt{agec} = 0$), are therefore $S_{01}(t)$ and $S_{00}(t)$, respectively. Figure 6.7 shows shorter survival in patients with edema at baseline. However, these stratum-specific survival functions also suggest that the multiplicative effect of `edema` on the mortality hazard is not constant over time. We examine this more carefully in Sect. 6.4.2.

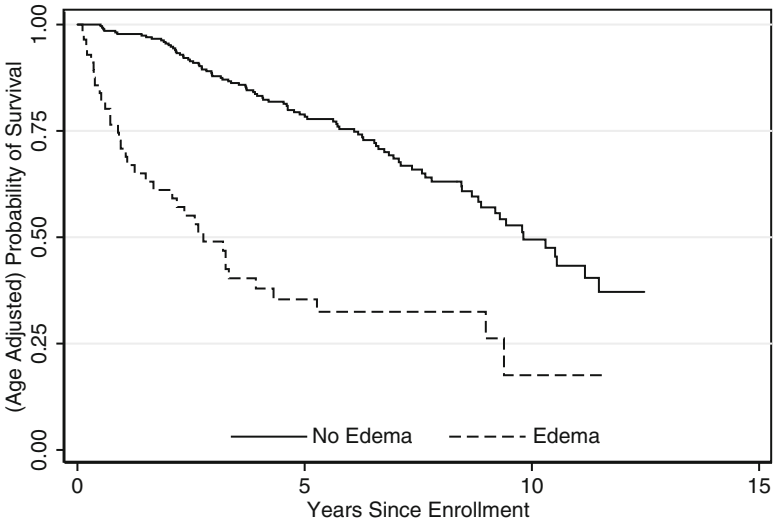

Fig. 6.7 Stratified survival curves for edema adjusted for age

6.4 Checking Model Assumptions and Fit

Two basic assumptions of the Cox model are *log-linearity* and *proportional hazards*. Just as with other regression models, these assumptions can be examined, and extensions of the model can deal with violations and model more complex effects.

6.4.1 Log-Linearity of the Hazard Function

In Sect. 6.2.1, (6.6) specifies that each unit change in a continuous predictor has the same effect on the log of the hazard. This implies that the hazard ratio is log-linear in continuous predictors.

Unlike the linear model, but like the logistic, diagnostics for violations of log-linearity using plots of residuals do not work very well for the Cox model. However, violations of this assumption are easy to detect and accommodate with the tools covered in Sect. 4.7.1 for the linear model. The approach is simple: attempt more general models and examine improvements in fit.

Like other models, the Cox model can be generalized by adding polynomial terms for the predictor in question to the model. Effect sizes and p-values are then checked to determine whether the higher order terms are important; or the predictor can be log-transformed and the log-likelihoods informally compared (Problem 6.4). Alternatively, the continuous predictor can be categorized using well-chosen cut-points; then log-linearity is checked using the methods outlined above in Sect. 6.2.2

for assessing both trend and departures from trend in ordinal predictors. These approaches have limitations: a susceptibility to outliers for polynomial models and sensitivity to the number and placement of the cutpoints for categorizations.

Restricted cubic splines are an alternative approach offering flexibility with relative parsimony. These methods, discussed in Sect. 4.7.1, lay down a series of "knots" along the values of the predictor and fit a polynomial curve between them—allowing for a wide variety of shapes. Consider the relationship between age and hazard of death for the PBC dataset. Using a spline fit, we could detect a nonlinear pattern between age and mortality. Unlike categorization of a continuous predictor, splines are not greatly sensitive to number and placement of knots. Three to five knots provide a great deal of flexibility. The choice of the number of knots is a balance between the sample size and the degree of flexibility desired. A further advantage is the similarity of implementation across diverse regression models. First, a spline *basis* is derived—in Stata this uses the `mkspline` command. This basis comprises $k - 1$ predictors, where k is the number of knots. These predictors then take the place of the continuous variable in the regression model.

Table 6.18 uses the commands and output for splines in a Cox model. Note, the similarity to the application of splines linear model in Sect. 4.7.1. Two `test` statements appear in Table 6.18. The tests suggest strong support for an overall effect of age on survival (given by the $p = 0.0004$) but find no evidence that the spline model fits the data better than a log-linear term in age (given by the $p = 0.99$). A *p*-value alone should not be used for model selection; hence, we compare the log-linear and spline fits graphically to examine the magnitude of the differences. Figure 6.8 shows the fits compared with a categorical fit placing cutpoints at the knots. The linear and restricted spline fit agree closely, suggesting a log-linear model fit age reasonable well.

6.4.2 Proportional Hazards

The adjusted Cox model shown in Table 6.12 shows that mortality risk is increased about twofold in PBC patients with edema at baseline. However, Fig. 6.7 suggests that edema may violate the proportional hazards assumption: specifically, the hazard ratio in edema is greatest in the first few years and then diminishes. Thus the effect of edema on the hazard is time-dependent. A transformed version of Fig. 6.7 turns out to be more useful for examining violations of the proportional hazards assumption.

6.4.2.1 Log-Minus-Log Survival Plots

To illustrate the use of transformed survival plots for assessing proportionality for binary or categorical predictors, we consider the treatment indicator (`rx`) in the DPCA trial. This method exploits the relationship between the survival and hazard functions. If proportional hazards hold for `rx`, then by (6.9)

Table 6.18 Restricted cubic spline Cox model for the effect of age on mortality

```
. mkspline age_sp = age, cubic

. stcox age_sp1 age_sp2 age_sp3 age_sp4

 Cox regression -- Breslow method for ties

No. of subjects =          312                  Number of obs    =       312
No. of failures =          125
Time at risk    =  1713.853528
                                                LR chi2(4)       =     20.59
Log-likelihood  =   -629.68657                  Prob > chi2      =    0.0004

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     age_sp1 |   1.049751   .0702359     0.73   0.468     .9207355    1.196845
     age_sp2 |   .9849075   .3583548    -0.04   0.967      .482713    2.009564
     age_sp3 |   .9746888   1.345662    -0.02   0.985     .0651166     14.5895
     age_sp4 |    1.17647   2.203381     0.09   0.931     .0299493    46.21414
------------------------------------------------------------------------------

. * test for departure from linearity

. test age_sp2 age_sp3 age_sp4

 ( 1)  age_sp2 = 0
 ( 2)  age_sp3 = 0
 ( 3)  age_sp4 = 0

           chi2(  3) =    0.08
         Prob > chi2 =    0.9943

. * test for overall effect

. test age_sp1 age_sp2 age_sp3 age_sp4

 ( 1)  age_sp1 = 0
 ( 2)  age_sp2 = 0
 ( 3)  age_sp3 = 0
 ( 4)  age_sp4 = 0

           chi2(  4) =   20.54
         Prob > chi2 =    0.0004
```

$$S_1(t) = [S_0(t)]^{\exp(\beta)}, \tag{6.15}$$

where $S_0(t)$ is the survival function for placebo patients and $S_1(t)$ is the corresponding survival function for the DPCA-treated patients. Then, the *log-minus-log* transformation of (6.15) gives

$$\log\{-\log[S_1(t)]\} = \beta + \log\{-\log[S_0(t)]\}. \tag{6.16}$$

Thus when proportional hazards hold, the two transformed survival functions will be a constant distance β apart, where β is the log of the hazard ratio for treatment with DPCA. This approach assumes a categorical variable but can be adapted to a continuous variable by, for instance, factoring a continuous variable into quartiles.

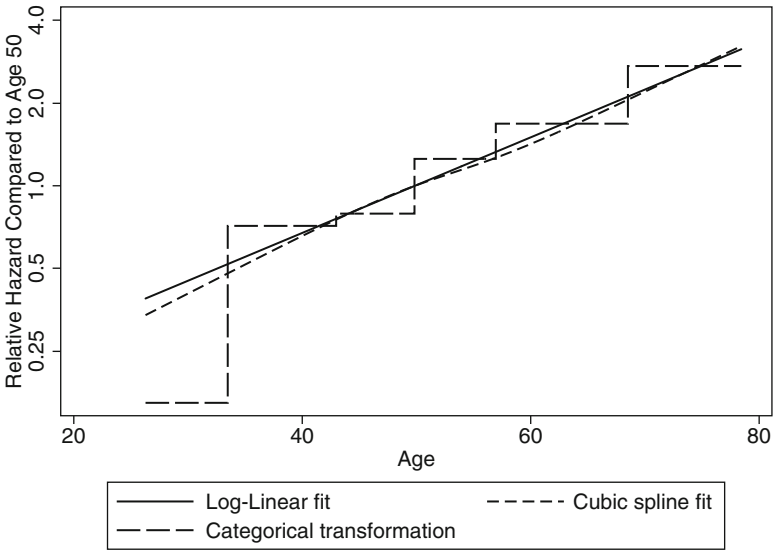

Fig. 6.8 Cox model fit to PBC data using a log-linear fit, restricted cubic spline, and categorical transformation of age

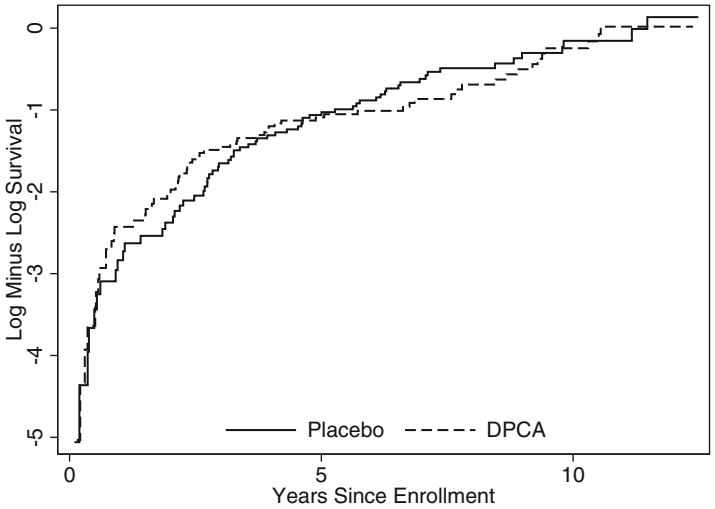

Fig. 6.9 Log-minus-log survival plot for DPCA treatment

This result enables us to use a simple graphical method for examining the proportional hazards assumption. Specifically, log-minus-log-transformed Kaplan–Meier estimates of the survival functions for the placebo and DPCA groups are plotted against follow-up time. In Stata, this plot is implemented in the `stphplot` command. The log-minus-log survival plot for DPCA is shown in Fig. 6.9.

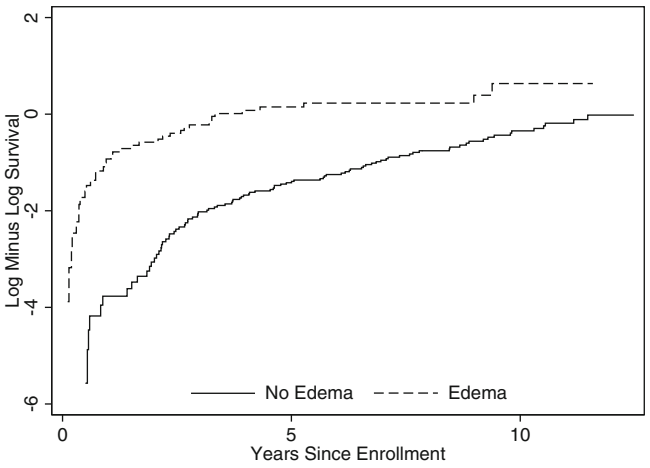

Fig. 6.10 Log-minus-log survival plot for edema

In assessing the *log-minus-log survival* plot for evidence of nonproportional hazards, the patterns to look for are convergence, divergence, or crossing of the curves. Converging curves suggest that the difference between the groups decreases with time; diverging curves suggest that differences increase with time. If the curves show pronounced crossing, then the nonproportionality may be more important; for example, this might indicate that treatment is harmful early on but protective later. In Fig. 6.9, however, the curves for DPCA and placebo remain close over the entire follow-up period and do not suggest nonproportionality.

In contrast, the log-minus-log survival plot for edema in Fig. 6.10 shows rather clear evidence of a violation of proportionality. While there is a pronounced difference between the groups at all time points, showing that patients with edema have poorer survival, the difference between the groups diminishes with follow-up. Specifically, the distances between the curves—that is, the implied log-hazard ratios—are 4.7, 1.8, 1.1, and 1.0 at years 1, 4, 7, and 10, respectively.

6.4.2.2 Smoothing the Hazard Ratio

Log-minus-log survival plots are good diagnostic tools for violations of the proportional hazards assumption. To address such a violation, however, we may need more information about how the log-hazard ratio changes with follow-up time. We can do this using a nonparametric, smoothed estimate of the hazard ratio against time, analogous to the LOWESS estimates of the regression function used in diagnosing problems in linear models in Sect. 4.7. If the smoothed estimate of the hazard ratio is nearly constant, then the assumption of proportional hazards is approximately satisfied. Conversely, when curvature is pronounced, the shape of the smooth line helps us determine how to model the hazard ratio as a function of time. In Stata,

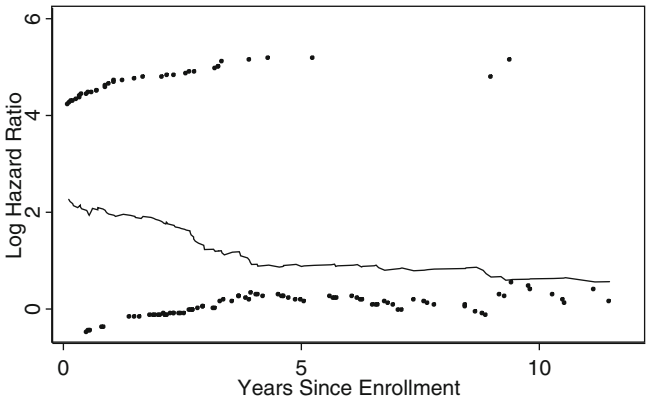

Fig. 6.11 Smoothed estimate of log-hazard ratio for edema

the plot can be generated using the `estat phtest` command with the `plot` option. Figure 6.11 shows the smoothed estimated plot of the hazard ratio over time for `edema`. A nonconstant trend is readily apparent: the log-hazard ratio decreases steadily over the first 4 years and then remains constant. This works by smoothing a specialized type of residual, *scaled Schoenfeld residuals*, for each predictor against time using LOWESS. The residuals appear as points in the plot. Smoothing them against time provides a nonparametric estimate of the log-hazard ratio for that predictor as it changes over time. An advantage of this approach is that it works for both categorical and continuous variables.

Relatively influential points are identifiable from the plots of the Schoenfeld residuals. DFBETA statistics, a measure of how much coefficients are changed by the deletion of individual observations (see Sect. 4.7.4 for an illustration of their applications in linear models), can be obtained for the Cox model in Stata by using the `predict` command.

6.4.2.3 Schoenfeld Test

Schoenfeld (1980) provides a test for violation of proportional hazards which is closely related to the diagnostic plot using LOWESS smooths of scaled Schoenfeld residuals just described. The test assesses the correlation between the scaled Schoenfeld residuals and time. This is equivalent to fitting a simple linear regression model with time as the predictor and the residuals as the outcome, and the parametric analog of smoothing the residuals against time using LOWESS. If the hazard ratio is constant, the correlation should be zero.

The Schoenfeld tests for `rx` and `edema` are shown in Table 6.19. Positive values of the correlation `rho` suggest that the log-hazard ratio increases with time and vice versa. In accord with the graphical results, the Schoenfeld test finds strong evidence for a declining log-hazard ratio for `edema` (`rho` $= -0.36$, $P = 0.0001$), but does not suggest problems with `rx` (`rho` $= -0.07$, $P = 0.5$).

Table 6.19 Schoenfeld tests of proportional hazards assumption

```
. estat phtest, detail

      Test of proportional-hazards assumption

      Time:  Time
      ----------------------------------------------------------------
                  |       rho            chi2       df       Prob>chi2
      ------------+---------------------------------------------------
      rx          |     -0.05862          0.43        1          0.5129
      edema       |     -0.36107         14.63        1          0.0001
      ------------+---------------------------------------------------
      global test |                      14.71        2          0.0006
      ----------------------------------------------------------------
```

The Schoenfeld test is most sensitive in cases where the log-hazard ratio is linearly increasing or decreasing with time. However, because the test is based on a linear regression model, it is sensitive to a few large residual values. Such values should be evident on the scatterplot of the scaled Schoenfeld residuals against time. Useful examples and discussion of the application of the Schoenfeld test appear in Sect. 6.5 of Therneau and Grambsch (2000).

6.4.2.4 Graphical Diagnostics Versus Testing

We have described both graphical and hypothesis testing methods for examining the proportional hazards assumption. The Schoenfeld test is widely used and gives two easily interpretable numbers that quantify the violation of the proportional hazards assumption. However, as pointed out in Sect. 4.7, such tests may lack power to detect important violations in small samples, while in large samples they may find statistically significant evidence of model violations which do not meaningfully change the conclusions. While also lacking sensitivity in small samples, graphical methods give extra information about the magnitude and nature of model violation, and should be the first-line approach in examining the fit of the model.

6.4.2.5 Stratification

The stratified Cox model introduced in Sect. 6.3.2 is an attractive option for handling binary or categorical predictors which violate the proportional hazards assumption. We explained there that no assumption is made about the relationships between the stratified hazard functions specific to the different levels of the predictor. Because the resulting fit to the stratification variable is unrestricted, this is a particularly good way to rule out confounding of a predictor of interest by a covariate that violates the proportional hazards assumption. However, because no estimates, CIs, or p-values are obtained for the stratification variable, this approach is less useful for any predictor of direct interest.

Note that we can apply this approach to a continuous variable by first categorizing it. How many categories to use involves a trade-off (Problem 6.9). Using more strata more effectively controls confounding, but as we suggested in Sect. 6.3.2, precision and power can suffer if the confounder is stratified too finely, because strength is not borrowed across strata. Five or six strata generally suffice, but there should be at least 5–7 events per stratum.

6.4.2.6 Modeling Interactions with Time

In this section, we briefly outline a widely used approach to addressing violations of the proportional hazards assumption using interactions with time, and implemented using TDCs, as described above in Sect. 6.3.1. We return to the edema example and show how the declining hazard ratio can be modeled. To begin, let $h_1(t)$ and $h_0(t)$ denote the hazard functions for PBC patients with and without edema. Because proportional hazards does not hold, the hazard ratio

$$\text{HR}(t) = \frac{h_1(t)}{h_0(t)} \tag{6.17}$$

is a function of t. To address this, we define $\beta(t) = \log\{\text{HR}(t)\}$ as a coefficient for edema which changes with time. This is equivalent to a hazard function of the form

$$h(t|\texttt{edema}) = h_0(t)\exp\{\beta(t)\texttt{edema}\}, \tag{6.18}$$

where as before `edema` is a 0/1 indicator of the presence of edema. This can be modeled in one of two ways.

- We can model the log-hazard ratio for edema as a linear function of time. This is implemented using a main effect, `edema`, plus an interaction term, `edemat`, defined as a TDC, the product of `edema` and t. That is, we set

$$\begin{aligned}\beta(t)\texttt{edema} &= (\beta_0 + \beta_1 t)\texttt{edema} \\ &= \beta_0\texttt{edema} + \beta_1 t\texttt{edema} \\ &= \beta_0\texttt{edema} + \beta_1\texttt{edemat}.\end{aligned} \tag{6.19}$$

  Alternatively, we could model the log-hazard ratio as linear in log time, defining the product term with $\log(t)$ in place of t; this might be preferable in the edema example, since the decline in the log-hazard ratio shown in Fig. 6.11 grows less steep with follow-up (Sect. 4.7.1).
- We can split follow-up time into sequential periods and model the log-hazard ratio for edema as a step function with a different value in each period. For example, we could estimate one log-hazard ratio for edema in years 0–4, and

another in years 5–10, again motivated by Fig. 6.11. We could do this by defining two TDCs:

- `edema04`, equal to 1 during the first 4 years for patients with edema, and 0 otherwise.
- `edema5on`, equal to 1 during subsequent follow-up for patients with edema, and 0 otherwise.

Then we set

$$\beta(t)\texttt{edema} = \beta_1\texttt{edema04} + \beta_2\texttt{edema5on}. \tag{6.20}$$

This approach is analogous to categorizing a continuous predictor to model nonlinear effects (Sect. 4.7.1).

The first alternative is more realistic because it models the hazard ratio for edema as a smooth function of time. But it is harder to implement because the TDC `edemat` changes continuously for patients with edema from randomization forward; up to one record for every distinct time at which an outcome event occurs would be required for these patients in the "long" dataset used for the analysis in Stata, now easily obtained using the `stsplit` and `stjoin` commands. In contrast, the second alternative is less realistic but easier to implement, only requiring two records for patients with edema and more than 4 years of follow-up, and one record per patient otherwise. See Sect. 6.9 for discussion of another flexible approach.

6.5 Competing Risks Data

6.5.1 What Are Competing Risks Data?

The MrOS study (Orwoll et al. 2005) followed 5,993 men over the age 65 and examined predictors of bone fracture and low BMD (evidence of subclinical bone loss). At enrollment, all men underwent a dual X-ray absorptiometry (DEXA) scan to determine their BMD and were followed for an average of 5 years for risk of bone fracture. At the conclusion of follow-up, 531 participants had developed fracture, 4,805 remained alive without fracture and 657 had died prior to fracture. An important question is how well a baseline BMD measure predicts fracture risk over the follow-up period.

There are two possible sources of incomplete follow-up: (1) the end of the observation (due to loss to follow-up or short observation times due to staggered entry) and (2) death. To understand why it is important to distinguish between them, consider how our methods have handled incomplete follow-up. The approach that has been used so far in this chapter attempts to project forward the experience of a censored observation by representing their experience with those followed longer. Embedded in this approach is an assumption and an objective. The assumption

is the *independent censoring* assumption (see Sect. 6.6.4) that the future risk of those whose follow-up has ended can be represented by those who are followed longer (see Sect. 3.5.2 for a discussion of this in the context of the Kaplan–Meier survivor function). The implicit objective is to make an extrapolation to a setting in which the source of incomplete follow-up is eliminated. For incomplete follow-up due to patient dropout, the assumption may be suspect but the objective is highly relevant since we would like to estimate what would have happened if people had been followed completely. For death, both the assumption and the objective are in question. To extrapolate to a setting where death is not possible would be to project a new population or the ability to extend lives—altering the underlying conditions of the study. Instead we could acknowledge death as another possible outcome which can cut short the observation of fracture without attempting to project fracture experience beyond participant lifetimes. Thus, objectives and approaches to incomplete follow-up may differ depending on whether it is due to death or the inability of a study to retain or follow participants.

> *Definition*: *Competing risks data* arise when multiple events can occur and follow-up can end due to occurrence of one or more of those types of events, precluding observation of at least one of the other event types.

The definition given is the most expansive possible definition. It could cover a situation in which observations are cut short due to patient dropout or due to end-of-study censoring. In the analysis of competing risks data, two major approaches can be taken—one which seeks to extrapolate to a scenario in which a type of event is not possible (typically, loss to follow-up or incomplete follow-up due to staggered entry). We call this approach *elimination*. Another family of methods is based on acknowledging and allowing for the competing risks in the analysis. This approach will be called *accommodation*.

In our example, we can observe fractures, death, or losses to follow-up. The objective of the analysis is to estimate the risk of fracture (in the presence of death) where there is no loss to follow-up.

6.5.2 Notation for Competing Risks Data

We denote competing risks outcome data using two variables: one which denotes the time of the first event and the other which denotes the type of event. Let

- Y: be the time of the first observed event of any type
- $\Delta = k$ if the kth event type occurs first

where each of the K possible types of failure are denoted by a numerical code. In the MrOS dataset, the failure types were coded as 0: loss to follow-up, 1: fracture, and 2: death. Using this, a participant who is followed for 18 months and dies (prior to fracture) will have $Y = 18$ months and $\Delta = 2$. Note, that this same notation is standard for ordinary survival analysis where there are two possible events: censoring and failure.

6.5.3 Summaries for Competing Risk Data

Two important summaries are typically available for competing risk data: these are analogs of the hazard and survival function.

6.5.3.1 Cause-Specific Hazard Functions

Definition: The *cause-specific hazard function* for event type k, $h_k(t)$, is the short-term rate at which subjects experience the onset of the kth event among those who have not yet experienced the event of interest (e.g., fracture) or a competing event (e.g., death) prior to t.

A hazard function can be thought of as a short-term rate of failure. Rate functions are ratios defined by which events get counted (in the numerator) and who is included in the "at risk" population (in the denominator). The kth cause-specific hazard only counts events of type k in the numerator (e.g., number of fractures). The denominator includes follow-up for all people who could have developed the event by time t. In the MrOS example, the cause-specific hazard function for fracture at a given time t calculates the rate of fracture among all people who are alive, without fracture, and uncensored prior to time t. Follow-up is counted in the denominator (the "at risk" population) until fracture, death, or loss to follow-up. Note that the cause-specific hazard reduces to the ordinary hazard function in the case that there is only one type of failure.

Estimating and modeling cause-specific hazard functions is straightforward. Simply set up the data as ordinary survival data with the kth failure type as the only type of failure and treat competing causes (even death) as "censored." Counterintuitively, the cause-specific hazard function's calculation does not make a distinction between events which are accommodated versus those which we attempt to eliminate. This will be in contrast to the cumulative incidence function for which this distinction will be very important.

If the data are analyzed this way, it is possible to examine the effect of predictor like bone-mineral density on the cause-specific hazard of fracture using a standard Cox model-type formulation. This will be discussed further in Sect. 6.5.3.3.

6.5.3.2 Cumulative Incidence Functions

The extension of the hazard function to competing risks data is given by the cause-specific hazard functions. The extension of the survivor function is given by what is called the *cumulative incidence function*. The cumulative incidence function at time t for the event type k is the proportion of the sample who have experienced the kth event by time t. For instance, at 5 years the cumulative incidence estimate of fracture is 0.080 and for death it is 0.093. This implies that after 5 years, about 8.0% of the study population developed a fracture prior to death, that 9.3% died without fracture and the remaining 82.7% are alive without fracture.

Table 6.20 Deaths and fractures in the MrOS cohort

Months in cohort	No. in follow-up	No. fractured	No. died	No. lost to FU	Pr event-free
1	5,993	7	2	1	0.9985
2	5,983	7	5	0	0.9965
3	5,971	10	1	0	0.9947
4	5,960	5	3	0	0.9933
5	5.952	4	6	0	0.9917
6	5,942	8	6	1	0.9893
7	5,927	10	1	1	0.9875
8	5,915	11	4	1	0.9850
9	5,899	6	5	1	0.9831
10	5,887	4	3	1	0.9818

Definition: The *cumulative incidence function* for cause type k at time t, $F_k(t)$, is the proportion who have developed the kth event prior to t.

The cumulative incidence function is a measure of prevalence of a particular event at each time t for a population which started with none.

This would be easy to estimate if there are no competing causes which we plan to eliminate. If we wanted to calculate the estimated cumulative incidence function at 1 year, we would simply count the number of people who at 1 year were alive without fracture (n_0), had experienced a fracture (n_1), or had died without experiencing a fracture (n_2). The cumulative incidence of fracture at 1 year would be $n_1/(n_0 + n_1 + n_2)$.

However, with incomplete follow-up, estimation requires more care. It requires that we consider increments of time as we did in the calculation of the Kaplan–Meier survivor estimate in Table 6.1.

Table 6.20 traces the first 10 months of the MrOS follow-up period and gives a summary by month. The second column is the number of participants who start the month alive, unfractured and under follow-up. The third through fifth column gives the number of fractures, losses to follow-up and deaths in that month, respectively. The final column is the estimated fraction of the proportion of the cohort who are alive and free of fracture at the end of each month. This quantity can be calculated by the Kaplan–Meier method by combining death and fracture into a single event.

Table 6.21 shows the cumulative incidence of fracture for the MrOS cohort during the first 10 months of follow-up. The cumulative incidence function for each month is the cumulative sum over the time-specific probabilities of a new fracture. This probability of a new fracture in a given month is the probability that someone is alive and fracture free at the beginning of the month (note that this is the probability from the end of the prior month) and develops a fracture during the month. For instance, the probability of an incident fracture in month seven is the chance of being alive and unfractured at the end of month 6 (0.9893) times the rate of failure in the seventh month. This rate is simply the number of fractures during the month divided by the number of participants under follow-up during the seventh month—

Table 6.21 Estimating the cumulative incidence of fracture in the MrOS cohort

Months in cohort	Event-free start of mo	Rate new fracture	Pr new fracture	Cum. incidence fracture
1	1.0000	7/5,993	0.0012	0.0012
2	0.9985	7/5,983	0.0012	0.0023
3	0.9965	10/5,971	0.0017	0.0040
4	0.9947	5/5,960	0.0008	0.0048
5	0.9933	4/5,952	0.0007	0.0055
6	0.9917	8/5,942	0.0013	0.0068
7	0.9893	10/5,927	0.0017	0.0085
8	0.9875	11/5,915	0.0018	0.0103
9	0.9850	6/5,899	0.0010	0.0113
10	0.9831	4/5,887	0.0005	0.0118

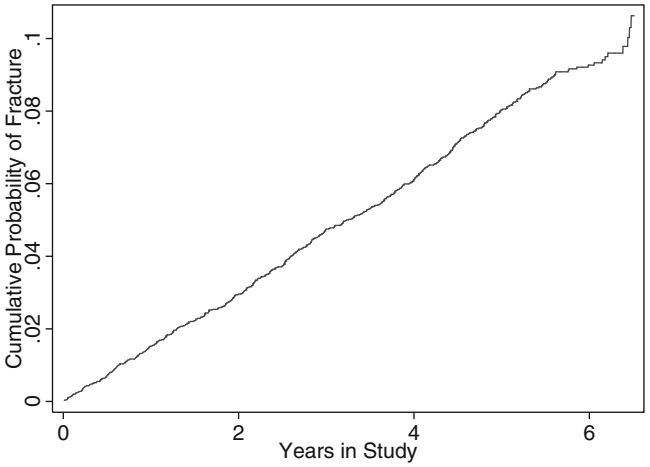

Fig. 6.12 Cumulative incidence of fracture in the MrOS cohort

10 of out 5,927 followed. The product of these two numbers is 0.0017 and means that about 0.17% of the cohort develops a new fracture in the seventh month. The cumulative incidence function is the sum over all the previous months. It estimates that the probability of developing a fracture by the end of the seventh month of the study is 0.085. Figure 6.12 graphs the cumulative incidence of fracture in MrOS cohort over a 6 year period.

The cumulative incidence function for the kth event at time t_j, $F_k(t_j)$, equals

$$F_k(t_j) = F_k(t_{j-1}) + \tilde{S}(t_{j-1})h_k(t_j), \tag{6.21}$$

where $\tilde{S}(t_{j-1})$ is the chance of being free of events at time t_{j-1} (just prior to time t_j). In the MrOS data, it is the chance of being both alive and unfractured. This can

be calculated by combining death and fracture into a composite and calculating the usual Kaplan–Meier estimator of being event-free and is given by values in the second column of Table 6.21. The hazard $h_k(t_j)$ denotes the cause-specific hazard for the kth event at time t_j which is given by the values in the third column of Table 6.21. The risk of a new event of type k at time t_j is the product of the kth cause-specific hazard at time by the cumulative incidence of the kth event at time t_{j-1} and appears as a fourth column. The cumulative incidence is the total of new events over all time periods and in the final column.

The cause-specific *hazard* function can be obtained by censoring competing events. However, censoring competing causes and calculating a *survival* function lead to estimates which have an awkward interpretation. It can only be interpreted as probability of event type k if (1) all competing causes have been eliminated and (2) the competing events can be assumed to be independent of the kth event. Note, this is typically the assumption made with people who are lost to follow-up. However, it would be highly speculative to extrapolate the likelihood of fracture if death could be eliminated from the MrOS cohort and contrary to the objective of accommodating competing causes.

From the calculations in Table 6.21 and (6.21), it is evident that the cumulative incidence function for the kth event depends on two things—the kth cause-specific hazard function and the Kaplan–Meier curve for remaining event-free. The event-free curve combines those who have not had the kth event or any other event. Hence, the cumulative chance of developing a fracture can be decreased (holding the k*th* cause-specific hazard constant) by increasing the rate of other types of failures, putting far fewer participants at risk for a fracture.

For instance, if age has no effect on the risk of fracture but does increase the risk of death, then a comparison of the cumulative incidence function by age would show a lower cumulative incidence of fracture among older men. This follows from the fact that older men are less likely to develop a fracture over the follow-up period. However, the lower number of fractures is due to the fact that fewer older men live long enough to develop fractures. This effect, while real, happens through age's effect on death rather than on fracture.

In such a scenario, modeling the age effect on the cause-specific hazard of fracture will show no effect. A model for the effect of age on the cumulative incidence of fracture would show a lower incidence of fracture with increased age. Both descriptions are faithful to the situation but reflect different aspects. Hence, the analyses are complementary and we discuss regression models for both.

6.5.3.3 Cox Model for Cause-Specific Hazard Functions

One approach to allowing predictors to affect the onset of competing risks is to model the cause-specific hazard function using proportional hazard formulation. Covariate effects can be interpreted as ratios of cause-specific hazards. Let $h_k(t)$

Table 6.22 Cox model for effect of BMD on fracture risk adjusted for body weight

```
stset time, failure(status==1)

stcox i.bmd3 weight

No. of subjects =         5993                  Number of obs    =      5993
No. of failures =          531
Time at risk    =  30483.46339
                                                LR chi2(3)       =    121.22
Log-likelihood  =   -4442.2904                  Prob > chi2      =    0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        bmd3 |
          2  |   .4193745   .0447953   -8.14   0.000     .3401585    .5170384
          3  |   .3290229   .0396476   -9.23   0.000     .2598098    .4166743
      weight |   1.004146     .00362    1.15   0.251      .997076    1.011266
------------------------------------------------------------------------------
```

be the kth cause-specific hazard and let $x_1, \ldots, x_p$ be a set of predictors. The model has the form

$$h_k(t|\mathbf{x}) = h_{0k}(t)\exp(\beta_1 x_1 + \ldots + \beta_p x_p), \tag{6.22}$$

where $h_{0k}(t)$ is a baseline hazard function. The model incorporates covariates into the model just as they appear in (6.5). The only difference is the hazard ratios apply to the kth cause-specific hazard while the interpretation of predictor effects are identical.

Consider a model for baseline BMD as a predictor of fracture in the MrOS cohort. The variable `time` in this dataset is the time to event variable Y and the type of failure is given by `status` which is coded as 0 if a person is free of death and fracture at the end of observation, 1 if the follow-up ends with fracture, and 2 if follow-up ends with a death prior to the onset of fracture. The predictors are BMD `bmd3` categorized into levels <0.895 g/cm^2, 0.895 to 1.01 g/cm^2, and >1.01 g/cm^2 and baseline weight (`weight`) measured in kilograms.

In Table 6.22, we see that compared to the group with BMD < 0.895 g/cm^2, those with BMD 0.895 to 1.01 g/cm^2 and >1.01 g/cm^2 have relative hazards of fracture of 0.42 (a 58% reduction) and 0.33 (a 67% reduction) adjusting for body weight at enrollment.

6.5.3.4 Fine–Gray model for Cause-Specific Hazard Functions

The result of the cause-specific regression in Table 6.22 shows a higher rate of fractures with lower BMD, but what if BMD is correlated with a series of unmeasured factors which make death more likely? If the death rate was high enough, men with low BMD might develop fewer fractures after 2 years than high-BMD men simply because the low-BMD men are more likely to die before fracture. This would be apparent from a comparison of the cumulative incidence function

Table 6.23 Fine and Gray model for effect of BMD on cumulative incidence of fracture adjusted for body weight

```
stset time, failure(status==1)

stcrreg i.bmd3 weight, compete(status==2)


Competing-risks regression                    No. of obs       =       5993
                                              No. of subjects  =       5993
Failure event  : status == 1                  No. failed       =        531
Competing event: status == 2                  No. competing    =        657
                                              No. censored     =       4805

                                              Wald chi2(3)     =     119.64
Log pseudolikelihood = -4472.9261             Prob > chi2      =     0.0000

------------------------------------------------------------------------------
             |               Robust
          _t |        SHR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        bmd3 |
          2  |   .4219663   .0443983    -8.20   0.000     .3433337    .5186079
          3  |   .3369128     .03982    -9.20   0.000     .2672472    .4247387
             |
      weight |   1.004669   .0036759     1.27   0.203     .9974896    1.011899
------------------------------------------------------------------------------
```

but not from the cause-specific hazard function. This is the disconnect between the hazard and cumulative incidence scale—it is helpful to have a way to describe regression effects on the cumulative incidence scale.

The Fine and Gray model (Fine and Gray 1999) adapts the spirit of a proportional hazards model to the cumulative incidence formulation. The idea is to model a different kind of rate function. The kth cause-specific hazard function is the rate of event among those who have experienced no event. For the MrOS data, this means that those who die are no longer counted in the rate (or hazard). The type of hazard that Fine and Gray construct retains cohort members who succumb to the competing risk in the denominator of their rate. For the MrOS data, this can be thought of as the rate of developing fracture among those without a previous fracture who are not lost to follow-up and, in particular, including those who have died. Maintaining those who die in the risk set acknowledges that someone who succumbs to a competing risk will not develop the event of interest and does not require the kind of extrapolation used for someone who is lost to follow-up.

Denote this new rate function for the kth type of failure as $f_k(t)$

$$f_k(t|\mathbf{x}) = f_{0k}(t)\exp(\beta_1 x_1 + \ldots + \beta_p x_p), \tag{6.23}$$

where $f_{0k}(t)$ takes the place of the usual baseline (or cause-specific) hazard function. The model incorporates covariates just as they appear in (6.22). Fitting the Fine and Gray model in Stata is done with the `stcrreg` command which is illustrated in Table 6.23.

The result of the Fine and Gray regression in Table 6.23 shows the hazard ratios for the model in (6.23) under the column marked "SHR" and the standard errors are labeled as "robust" because it is calculated in a way which is not model-based. The results are striking similar to the regression based on the cause-specific hazard function in Table 6.22. This is not surprising—the methods differ on whether people who die are retained in the risk set. The risk of death is not large and hence the two approaches give very similar results.

6.6 Some Details

In this section, we discuss some useful additional topics.

6.6.1 Bootstrap Confidence Intervals

The ACTG 019 dataset includes 880 observations but only 55 failures. Stata provides Wald-based CIs for the Cox model which require sample size which are "large." The effective sample size is determined by the number of failures rather than the number of observation. Hence, it can be useful to check the validity of the Wald-based CIs for the Cox model for ZDV treatment (`rx`) and baseline CD4 cell count (`cd4`) using the bootstrap (Sect. 3.6). The results are reported on the coefficient scale in Table 6.24.

The standard and bias-corrected bootstrap CIs, based on 1,000 resampled datasets, yield very similar results, confirming that the semi-parametric model works well in this case, even though there are only moderate numbers of events.

Table 6.24 Cox model for ZDV and CD4 with bootstrap confidence intervals

```
stcox i.rx cd4, vce(bootstrap, bca reps(1000) nodots seed(881) )

Cox regression -- Breslow method for ties

No. of subjects =          880                     Number of obs    =       880
No. of failures =           55
Time at risk    =       354872
                                                    Wald chi2(2)     =     32.34
Log likelihood  =   -314.17559                      Prob > chi2      =    0.0000

------------------------------------------------------------------------------
             |   Observed   Bootstrap                         Normal-based
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        1.rx |   .4560671    .138132    -2.59   0.010     .2518951    .8257293
         cd4 |   .9934464   .0013741    -4.75   0.000     .9907569    .9961432
------------------------------------------------------------------------------
```

6.6.2 Prediction

Evaluating prediction error using some form of cross-validation, as described in Sect. 10.1, is more complicated with time-to-event outcomes. Comparing observed to expected survival times is ruled out for censored observations in the test set; moreover, as we explained above in Sect. 6.2.13, expected—that is, mean—survival times are usually undefined under the Cox model. Comparing the *occurrence* of events in the test set with predictions based on the learning set, as with binary outcomes analyzed using a logistic model, is relatively tractable, but complicated by variations in follow-up time, in particular extrapolations for any follow-up times in the test set that exceed the longest times in the learning set.

Dickson et al. (1989) give one way in which predictions based on a Cox model can be cross-validated using a test dataset. The basic idea is to use coefficients estimated from a development dataset to classify observations in a test dataset. To see how this works, first note that the predictors are associated with the hazard ratio only through what is called the *linear predictor*, $\beta_1 x_1 + \ldots \beta_p x_p$, as demonstrated in (6.5). The larger the value of the linear predictor, the larger the hazard and the shorter survival times tend to be. Obtaining the estimated coefficients from a development dataset, it is possible to calculate what is called a *risk score*, namely $\hat{\beta}_1 x_1 + \ldots \hat{\beta}_p x_p$, for each observation in either the development or test datasets. The investigators grouped the patients in the development dataset into four predicted survival categories on the basis of the risk score, with the cutpoints determined to give approximately equal numbers of events in each category. They then demonstrated the models ability to discriminate by calculating the Kaplan–Meier survival curves for the test set using the groups defined by cutpoints from the development set. The pronounced separation of the survival curves in the test set is evidence of the ability of the model to stratify by risk.

6.6.3 Adjusting for Nonconfounding Covariates

If a covariate is strongly predictive of survival but uncorrelated with a predictor of interest, omitting it from a Cox model will nonetheless attenuate the estimated hazard ratio for the predictor of interest, as discussed in Sect. 10.2.6 (Gail et al. 1984; Schmoor and Schumacher 1997; Henderson and Oman 1999). Omitting important covariates from logistic models also induces such attenuation. Although the gain in precision is usually modest at best, it can be advantageous to include such a prognostic factor in order to avoid the attenuation.

A compelling example is provided by ACTG 019, the randomized clinical trial of ZDV for prevention of AIDS and death in HIV infection discussed in Sect. 6.1. As expected in a clinical trial, there was no between-group difference in mean baseline CD4 count, known to be an important prognostic variable. Thus by definition, baseline CD4 count could not have confounded the effect of ZDV. However, when

CD4 count is added to the model, the estimated reduction in risk of progression to AIDS or death afforded by ZDV goes from 49% to 54%, an increase of about 12%. More discussion of whether to adjust for covariates in a clinical trial is given in Sect. 10.2.6.

6.6.4 Independent Censoring

To deal with right-censoring, we have made the assumption of *independent censoring*. The essence of this assumption is that after adjustment for covariates, future event risk for a censored subject does not differ from the risk among other subjects who remain in follow-up and have the same covariate values. Under this assumption, subjects are censored independent of their future risk.

To see how this assumption may be violated, consider a study of mortality risk among patients followed from admission to the intensive care unit until hospital discharge. Suppose no survival information is available after discharge, so subjects have to be censored at that time. In general, subjects are likely to be discharged because they have recovered and are thus at lower risk than patients who remain hospitalized. Unless we can completely capture the differences in risk using baseline and TDCs, the assumption of independent censoring would be violated.

Dependent censoring can also arise from informative loss to follow-up. In prospective cohorts, it is not unlikely that prognosis for dropouts differs from that for participants remaining in follow-up in ways that can be difficult to capture with variables routinely ascertained.

It can also be difficult to diagnose dependent censoring definitively, because that would require precisely the information that is missing—for example, mortality data after discharge from the ICU. But that is a case where an experienced investigator might recognize on substantive grounds that censoring is likely to be dependent. Furthermore, the problem could be addressed in that study by ascertaining mortality for a reasonable period after discharge. Similarly, losses to follow-up are best addressed by methods to maximize study retention; but it also helps to collect as much information about censored subjects as possible. Inverse weighting methods can be used in situations where the dependence between failure and censoring is explained by a series of measured variables and where a model for censoring can be specified in terms of these measured (possibly time-dependent) covariates (see Sect. 9.5).

6.6.5 Interval Censoring

We also assume that the time of events occurring during the study is known more or less exactly. This is almost always the case for well-documented events like death, hospitalization, or diagnosis of AIDS. But the timing of many events is not

observed with this level of precision. For example, in prospective cohort studies of people at risk for HIV infection, it is common to test participants for infection at semi-annual visits (Buchbinder et al. 1996). Thus the actual time of an incident infection is only known up to an interval of possible values; in technical terms, it is *interval-censored* between the last visit at which the participant tested negative and the first at which the result was positive. Another example is development of abnormal cellular changes in the cervix, which must be assessed by clinical exam. These exams may be performed periodically, perhaps months or even years apart. As with HIV infection, newly observed changes may have occurred at any time since the last exam. In settings where intervals arise because of the study follow-up schedule and are regularly spaced, pooled logistic regression Sect. 5.5.2 can be used to handle the interval censoring. Interval censoring becomes more complex when the time between intervals is unequal and/or vary by individual requiring specialized methods beyond the scope of this book.

6.6.6 Left-Truncation

Survival times are measured from some initial time with more than one possible choice of origin. In the PBC study, we defined the survival time as the time from cohort enrollment until death. We could, instead, chose to measure survival time from the diagnosis of disease. Diagnosis is a more meaningful event biologically and easily aligning the time scales on this initial time will lead to more interpretable results.

The PBC study recruited patients from a referral center and months or years may have elapsed between diagnosis and entry into the cohort. A patient with a rapid disease course is less likely to be enrolled simply because they may die prior to referral to the center or prior to recruitment into the study. When this type of selection is active, there can be an undercounting of short survival times. The setting where some survival times are not observed because the sampling scheme tends to miss short survival times is known as *left-truncation*. To avoid bias, we need to consider the length of the period between the time origin and entry into the cohort. We denote this *truncation time* by V. Staggered entry into a cohort does not imply left-truncation; the key feature of left-truncation is the truncation time, V—there must be some time delay between the event which defines the origin event and entry into the cohort. For instance, if a PBC cohort was able to enroll participants at the time of their diagnosis, truncation times would be 0 and there would be no left-truncation. However, because patients have their diagnoses at different times, the cohort will still exhibit staggered entry.

The nature of the incomplete data from truncation is different from censored data. For censoring, it is incomplete because the event time falls outside of follow-up. Truncated values outside the follow-up period are not merely incomplete; rather, they are not observed at all. Because a left-truncated patient dies before enrollment, they leave no trace in the study—truncation is said to result in "ghosts."

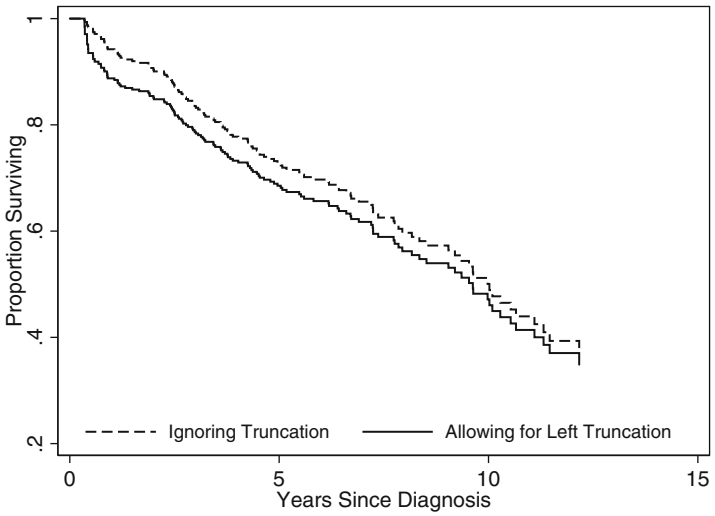

Fig. 6.13 Kaplan–Meier curves for the PBC data incorporating and ignoring left-truncation

Right-truncation can also arise if a study recruits based on an endpoint and people with large event times (or who never had the event) are not recruited. An example of right-truncation is a fecundability study that excludes couples who never conceive.

Survival analysis of risk factors can be conducted on the natural time-scale with origin at HIV infection under an assumption of *independent truncation*. This assumption is that the time of delayed entry and subsequent survival are independent and it is satisfied when the incidence of a disease and survival post-diagnosis are independent. Under independent truncation, the analysis uses the truncation time V along with the time and censoring indicators (X, Δ), where X is the follow-up time relative to diagnosis. The PBC dataset does not include truncation times but we created the truncation time `disease_dur` for illustrative purposes.

In Stata, we introduce the censoring and truncation into the analysis using the stset command in Stata as follows:

```
stset years_since_diag, failure(status) entry
(disease_dur)
```

Figure 6.13 graphs two survival curves based on time since diagnosis—one which uses the `stset` statement to account for left-truncation and naive calculation which ignores truncation. The estimator which ignores truncation estimates higher post-diagnosis survival probabilities. By ignoring the truncation, the analysis fails to account for undersampling of short survival times and, thus, overestimates survival. The effect of ignoring truncation on hazard ratios in a Cox model is less predictable but can often attenuate them.

Note that both survival estimators in Fig. 6.13 make drops at the same event times but the size of the drops for the truncation-based estimator are larger at earlier time points. This reflects the importance of short event times under left-truncation just as

long event times are important under right-censoring. A key assumption under left-truncation is that there is a positive probability that even the shortest failures could make it into the sample. If they are completely excluded, only strong parametric assumptions can account for their absence.

Fortunately, Stata can handle (independent) censoring and truncation simultaneously and once the stset command has been used it is possible to use the full set of survival analysis techniques without taking further account of the nature of the incomplete data.

6.7 Sample Size, Power, and Detectable Effects

Sections 4.8 and 5.7 provide formulas for calculating sample size, power, and minimum detectable effects for the linear and logistic models. Analogous results hold for the Cox model. To compute the sample size that will provide power of γ in two-sided tests with type-1 error of α to reject the null hypothesis $\beta_j = 0$ for the effect of a predictor X_j, accounting for the loss of precision due to adjustment for covariates, we can use

$$n = \frac{(z_{1-\alpha/2} + z_\gamma)^2}{(\beta_j^a \sigma_{x_j})^2 \psi (1 - \rho_j^2)}, \tag{6.24}$$

where β_j^a is the hypothesized value of β_j under the alternative, $z_{1-\alpha/2}$ and z_γ are the quantiles of the standard normal distribution corresponding to the specified type-1 error and power, σ_{x_j} is the standard deviation of X_j and ρ_j is its multiple correlation with the other covariates, and ψ is the probability that an observation is uncensored, so that the expected number of events $d = n\psi$ (Hsieh and Lavori 2000; Schmoor et al. 2000; Bernardo et al. 2000). The variance inflation factor $1/(1 - \rho_j^2)$ in (6.24) accounts for the potential loss of precision due to the inclusion of other predictors in the model (Hsieh et al. 1998). For problems with fixed values of n and ψ, power is given by

$$\gamma = 1 - \Phi\left[z_{1-\alpha/2} - |\beta_j^a| \sigma_{x_j} \sqrt{n\psi(1 - \rho_j^2)}\right]. \tag{6.25}$$

Finally, the minimum detectable effect (on the log-hazard scale) is

$$\pm \beta_j^a = \frac{z_{1-\alpha/2} + z_\gamma}{\sigma_{x_j} \sqrt{n\psi(1 - \rho_j^2)}}. \tag{6.26}$$

Some additional points:

- Sample size (6.24) and minimum detectable effect (6.26) calculations simplify considerably when we specify $\alpha = 0.05$ and $\gamma = 0.8$, β_j^a is the effect of a one standard deviation increase in continuous x_j, and we do not need to penalize for covariate adjustment. In that case,

$$n = \frac{7.849}{(\beta_j^a)^2 \psi}. \tag{6.27}$$

For the minimum detectable effect, we have

$$\pm \beta_j^a = \frac{2.802}{\sqrt{n\psi}}. \tag{6.28}$$

For 90% power, substitute 10.51 for 7.849 and 3.242 for 2.802.

- Similarly, for a two-arm clinical trial with equal allocation to arms, so that β_j^a is the log-hazard ratio for treatment and $s_{x_j}^2 = 0.25$, we can calculate

$$n = \frac{4 \times 7.849}{(\beta_j^a)^2 \psi}. \tag{6.29}$$

For the minimum detectable effect, we have

$$\pm \beta_j^a = \frac{2 \times 2.802}{\sqrt{n\psi}}. \tag{6.30}$$

- Power calculations using (6.25) simplify analogously, but still require a statistical calculator or computer package to evaluate the normal cumulative distribution function $\Phi(\cdot)$.
- Power in the Cox model is driven by the expected number of events $d = n\psi$, with little or no independent influence of n once d is fixed. For the same reason, early censoring has relatively little influence. Some calculators may return d rather than n, or require d rather than n and ψ as inputs.
- Sample size, power, and minimum detectable effects can be calculated using the `stpower cox` command in Stata as well as many other statistical packages. Alternatively, (6.24)–(6.26) can easily be programmed in Stata, R, or Excel, or evaluated by hand if values of $z_{1-\alpha/2}$, z_γ, and $\Phi(\cdot)$ are available.
- When X_j is a binary predictor with prevalence f_j, $\sigma_{x_j} = \sqrt{f_j(1-f_j)}$ in (6.24)–(6.26).
- When X_j is a continuous predictor with standard deviation σ_{x_j}, it is important to recognize that sample size, power, and minimum detectable effects do not depend in any real way on the units in which X_j is measured. This is most clearly seen in (6.26). Suppose X_j is usually measured in grams. Changing the unit to milligrams increases σ_{x_j} by a factor of 1,000, and shrinks β_j^a by the same factor. But of course the effect on the outcome of a 1-mg increase in the predictor is 1,000 times smaller than the effect of a 1-g increase. One way to avoid confusion is to consider the minimum detectable effect size for a one standard deviation change in X_j, which is often a reasonable-sized change to consider. That effect size is obtained by setting $\sigma_{x_j} = 1$ in (6.26).

- As in calculations for the linear and logistic model, we need to use $|\beta_j^a|$ in (6.25) if $\beta_j^a < 0$. It follows that the negative of the value given by (6.26) is also a valid solution for the minimum detectable effect.
- The use of the factor $1-\rho_j^2$ to account for covariate adjustment carries over from linear to Cox models. However, there is no analog to the reduction in residual variance that can result from including covariates in linear models, so that the adjustment to these calculations using $1-\rho_j^2$ is less likely to be conservative.
- The `stpower cox` command does incorporate the factor $1-\rho_j^2$ to account for covariate adjustment, via the `r2` option. In using sample size calculators that do not allow for this adjustment, unadjusted estimates of n or d should be inflated by $1/(1-\rho_j^2)$; similarly the minimum detectable effect estimate should be inflated by $\sqrt{1/(1-\rho_j^2)}$. To calculate power in such calculators, use $n\psi(1-\rho_j^2)$ in place of $n\psi$ as an input.
- In Sect. 4.8, we showed how the standard error $\mathrm{SE}(\hat{\beta}_j)$ plays a central role in sample size, power, and minimum detectable effect calculations for regression problems. $\mathrm{SE}(\hat{\beta}_j)$ is a large-sample approximation in Cox models, and more exact small-sample computations using the t-distribution do not carry over from the linear model. Simulations of power may be a more reliable guide when the calculated or available sample size is small.
- The alternative calculations (4.15)–(4.17) presented in Sect. 4.8, which use an estimate $\tilde{\mathrm{SE}}(\hat{\beta}_j)$ based on published results for an appropriately adjusted model using $\tilde{n}$ observations, carry over directly. There we showed that

$$n = \frac{(z_{1-\alpha/2} + z_\gamma)^2 \tilde{n} [\tilde{\mathrm{SE}}(\hat{\beta}_j)]^2}{(\beta_j^a)^2}. \tag{6.31}$$

  Similarly, power in a new sample of size n is given by

$$\gamma = 1 - \Phi\left[z_{1-\alpha/2} - |\beta_j^a| / [\sqrt{\tilde{n}/n}\tilde{\mathrm{SE}}(\hat{\beta}_j)]\right]. \tag{6.32}$$

  Finally, the minimum detectable effect in a new sample of size n can be obtained as

$$\pm\beta_j^a = (z_{1-\alpha/2} + z_\gamma)\sqrt{\tilde{n}/n}\tilde{\mathrm{SE}}(\hat{\beta}_j). \tag{6.33}$$

  In implementing these calculations, care must be taken to obtain the SE of the regression coefficient β_j, not the SE of the hazard ratio e^{β_j}. This can be computed from a 95% CI for the hazard ratio as $\tilde{\mathrm{SE}}(\hat{\beta}_j) = \log(UL/LL)/3.92$, where UL and LL are the upper and lower confidence bounds. We must also ensure that X_j is measured on the same scale as in the published results.

To illustrate these calculations, we first calculate the sample size providing 80% power in a two-sided test with α of 5% to detect an effect of bilirubin levels on survival, adjusting for the effects of hepatomegaly, edema, and spiders, as suggested by the analysis shown previously in Table 6.12.

Table 6.25 Sample size calculation for effect of bilirubin on mortality risk

```
. stpower cox, failprob(.15) hratio(1.15) sd(4.5) r2(0.2025)

Estimated sample size for Cox PH regression
Wald test, log-hazard metric
Ho: [b1, b2, ..., bp] = [0, b2, ..., bp]

Input parameters:
         alpha =    0.0500  (two sided)
            b1 =    0.1398
            sd =    4.5000
         power =    0.8000
     Pr(event) =    0.1500
            R2 =    0.2025

Estimated number of events and sample size:
             E =        25
             N =       166

. display (invnormal(.975)+invnormal(0.8))^2/((log(1.15)*4.5)^2*0.15*
(1-.2025)) 165.87573
```

The new study will have a shorter 2-year follow-up, as compared to the average 5.5 year follow-up in the DPCA Trial, with an estimated 15% cumulative mortality. Based on the DPCA results, we estimate that $\sigma_{x_j} \approx 4.5$ mg/dL and that $\rho_j \approx 0.45$ (so $\rho_j^2 = 0.2025$), indicating substantial variance inflation. We hypothesize that the hazard ratio per mg/dL increase in bilirubin level will be 1.15 (so $\beta_j^a = \log 1.15$). Table 6.25 shows results using the `stpower cox` command in Stata as well as a calculation using (6.24). The two estimates are essentially identical; a quick calculation using $\psi = 0.15$ shows that the expected number of events based on (6.24) is 25.

The `stpower cox` command can also be used to calculate minimum detectable effects. In the DPCA trial, suppose an ancillary study is being considered to evaluate the independent association of mortality with a novel risk marker, to be measured using stored baseline specimens. There were 125 deaths among 312 participants, so $\psi = 125/312 = 0.40$; equivalently, $d = n\psi = 125$. We hypothesize that the new marker will be highly correlated with available prognostic measures ($\rho_j \approx 0.5$), yet hope that it will provide additional predictive information. Initial testing suggests that the SD of the new marker is approximately 1.5 mg/dL. We hypothesize that higher levels of the marker will be associated with lower risk. What hazard ratio per mg/dL increase in the new marker will be detectable with 80% power in a two-sided test with α of 5%?

Table 6.26 shows that `stpower cox` and (6.26) give essentially the same result: for the DPCA sample to provide 80% power to reject $\beta_j = 0$, the mortality hazard must be independently reduced by approximately 18% for each mg/dL increase in the novel marker.

Table 6.26 Minimum detectable effect of a novel marker

```
. stpower cox, n(312) failprob(.40) sd(1.5) r2(0.25) power(.8) hr

Estimated hazard ratio for Cox PH regression
Wald test, hazard metric
Ho: [b1, b2, ..., bp] = [0, b2, ..., bp]

Input parameters:
         alpha =    0.0500  (two sided)
            sd =    1.5000
             N =       312
         power =    0.8000
     Pr(event) =    0.4000
            R2 =    0.2500

Estimated number of events and hazard ratio:
             E =       125
        hratio =    0.8244

. display exp(-(invnormal(.975)+invnormal(0.8))/(1.5*sqrt(125*(1-0.25))))
.82456636
```

6.8 Summary

Survival data exhibit novel features including right-censoring, interval censoring, truncation, and competing risks. The Cox proportional hazards model is suited to the special features of survival data and summarizes the effects of covariates through hazard ratios. The Cox model has much in common with other regression models; in particular, issues of confounding, mediation, and interaction are dealt with in similar ways. Specialized techniques are required to calculate predicted survival and to examine the assumption of proportional hazards. The Cox model can be ended to handle TDCs and stratification. Competing risks arise when other events may preclude observing the event of interest. Extensions to the proportional hazards model for competing risks data can be based on the cause-specific hazard function (which models the effect of covariates directly on the event of interest) or can be based on the Fine–Gray model (which allows for the effect of covariates which occur through competing events). The two approaches provide complementary perspectives on the effect of covariates in the presence of competing risks.

6.9 Further Notes and References

The Cox model has proven popular because it is computationally feasible and flexible. Alternatives include the *accelerated failure time model* (Wei 1992) or the *additive hazards model* (Aalen 1989). These models are less popular and statistical techniques for them are less well developed. By contrast, there are extensively developed techniques for parametric survival regression (implemented in Stata with the `streg` package). Parametric models require us to make assumptions about the form

of the baseline hazard function and have proved less popular because the parametric assumptions sacrifice robustness without substantial efficiency gains. Useful references include Chap. 5 of Marubini and Valsecchi (1995) and Chap. 12 of Klein and Moeschberger (1997).

Some more complex survival data settings are beyond the scope of chapter. For instance, there may be more than a single event per subject, yielding clustered or hierarchical survival data. See Wei and Glidden (1997) for an overview of possible approaches, including analogs of the *marginal* and *random effects* models described for repeated continuous and binary outcomes in Chap. 7. The are both available options in Stata `stcox` command—the marginal by using the `vce(cluster)` option and random effects by using the `shared` option.

Stata provides extensive capabilities for fitting and assessing Cox models. For instance, more flexible model for time-varying hazards than those discussed in Sect. 6.4.2.6 could be developed by treating time as continuous (using the `tvc`) option in conjunction with splines. A complete suite of parametric survival analysis methods are also provided. The flexible `stset` command handles complex patterns of censoring and truncation.

Applied book-length treatments on survival analysis are available by Miller et al. (1981) and Marubini and Valsecchi (1995). These two texts strike a nice balance in their completeness and orientation toward biomedical applications. The texts by Klein and Moeschberger (1997) and Therneau and Grambsch (2000) are very complete in their coverage of tools for survival analysis in general and the Cox model in particular Chap. 3 of Klein and Moeschberger (1997) provides a complete discussion on left-truncation, interval censoring, and general censoring patterns.

Sometimes time-to-event data can be more effectively handled using an alternative framework. In particular, consider cohort studies in which interval-censored outcomes are ascertained at each follow-up visit. One alternative is to use the continuation ratio model, referenced in Chap. 5, for time to the first such event. This can be seen as a discrete-time survival model, where the time scale is measured in visits (or intervals). Where appropriate, another, often more powerful, alternative is to use a logistic model for repeated binary measures, covered in Chap. 7. A closely related issue is the handling of Finally, some time-to-event data has no censored values. In that situation, techniques covered in Chap. 8 can provide a useful regression framework for dealing with the skewness and heteroscedasticity such data are likely to exhibit.

6.10 Problems

Problem 6.1. Divide the hazard ratio for `bilirubin` by its standard error in Table 6.4 and compare the result to the listed value of `z`. Also compute a CI for this hazard ratio by adding and subtracting 1.96 times its standard error from the hazard ratio estimate. Are the results very different from the CI listed in the output, which is based on computations on the coefficient scale?

Problem 6.2. In the ACTG 019 data, treatment `rx` is coded ZDV = 1 and placebo = 0. Define a new variable `rxplus11` which is coded ZDV = 12 and placebo = 11; this can be done using the Stata command `generate rxplus11=rx+11`. Fit a Cox model with `rxplus11` as the only predictor, then fit a second Cox model with `rx` as the only predictor. How do the two results compare?

Problem 6.3. Using the ACTG 019 data from Problem 6.2, recode treatment so it is coded ZDV = 0 and placebo = 1. How do the hazard ratios, CIs, likelihood ratio (LR), and Wald tests compare to the original coding? If any are different, how are they different?

Problem 6.4. Using the PBC dataset, calculate the hazard ratio for values of albumin = 2.5, 3.5, and 4.0, using `albumin = 3` as the reference level assuming the log-hazard is linear in albumin. The PBC dataset is available at `http://www.biostat.ucsf.edu/vgsm`.

Problem 6.5. For the PBC dataset, fit a model with cholesterol and bilirubin. Interpret the results, as you would in a paper, reporting the hazard ratios for a 100 mg/dL increase in cholesterol and a 10 mg/dL increase in bilirubin. Is the relationship between cholesterol and survival confounded by bilirubin?

Problem 6.6. Calculate a hazard ratio and CI for a 5-year increase in age by computing the fifth power of the estimated hazard ratio and its confidence limits, using the results for a 1-year increase in Table 6.9. Compare the result to a fit of the Cox model using a re-scaled version of the variable.

Problem 6.7. Using the model in Table 6.14 and taking Table 6.15 as your guide, calculate the effect of hepatomegaly among those on placebo. Then, derive and calculate the contrast required to identify the effect of hepatomegaly among those on DPCA. Given these, derive and fit the linear contrast to test for interaction. How does it compare with the test of interaction for comparing the effect of DPCA treatment across hepatomegaly that was given in Sect. 6.2.10?

Problem 6.8. For the ACTG 019 dataset, write out the Cox model allowing for an interaction between ZDV treatment `rx` and the baseline CD4 cell count `cd4`.

(a) Express the test of the null hypothesis of no interaction between CD4 and treatment in terms of the parameters of the model.
(b) Again using the parameters of the model, what is the hazard ratio for a ZDV-treated subject with x CD4 cells compared with a placebo-treated subject with x CD4 cells?
(c) Fit the model. Does there appear to be an interaction between treatment and CD4 stratum? If so, what is the interpretation?
(d) What are the hazard ratios for ZDV as compared to placebo for patients with 500, 109, and 50 CD4 cells, respectively?

Problem 6.9. We can also control for the effect of bilirubin in the PBC mortality data using stratification rather than adjustment. One way to categorize is to create approximately equal-size groups. In Stata, for example, you can categorize by

quintile of bilirubin using the command `xtile cat5=bilirubin, nq(5)`. Try fitting a Cox model for `cholesterol` stratified by `bilirubin`, stratified at 2, 3, 10, and 50 levels. What is the trade-off in increasing the number of levels? What number of levels works best? (Hint: Balance adjustment against the size of the standard error).

Problem 6.10. Using the PBC dataset, apply the methods of Sect. 6.4.2 for examining proportional hazards to the variable `hepatomegaly` and interpret the results.

6.11 Learning Objectives

(1) Define right-censoring, hazard function, proportional hazards, left-truncation, competing risks data, and TDCs.
(2) Be able to:

- Convert a predictor to a new unit scale
- Derive the hazard ratio between two groups defined by their predictor values
- Interpret hazard ratio estimates, Wald test p-values, and CIs
- Calculate and interpret the likelihood-ratio test comparing two nested Cox models
- Detect and model interaction using the Cox model
- Detect nonproportional hazards using log-minus-log and smoothed hazard ratio plots, and the Schoenfeld test
- Use stratification to control for a covariate with nonproportional effects

(3) Understand:

- When to use survival techniques
- Why the semi-parametric form of the Cox model is desirable
- Why the Cox model is "multiplicative"
- How the stratified Cox model relaxes the proportional hazard assumption
- How to address confounding, mediation, and interaction using a Cox model
- The difference between modeling cause-specific hazards and cumulative incidence functions for competing risk data
- Recognize settings which are beyond the scope of this chapter, including interval and dependent censoring, and repeated-events data