

Chapter 17

CAUSAL SURVIVAL ANALYSIS

In previous chapters we have been concerned with causal questions about the treatment effects on outcomes occurring at a particular time point. For example, we have estimated the effect of smoking cessation on weight gain measured in the year 1982. Many causal questions, however, are concerned with treatment effects on the time until the occurrence of an event of interest. For example, we may want to estimate the causal effect of smoking cessation on the time until death, whenever death occurs. This is an example of a *survival analysis*.

The use of the word “survival” does not imply that the event of interest must be death. The term “survival analysis”, or the equivalent term “failure time analysis”, is applied to any analyses about time to an event, where the event may be death, marriage, incarceration, cancer, flu infection, etc. Survival analyses require some special considerations and techniques because the failure time of many individuals may occur after the study has ended and is therefore unknown. This chapter outlines basic techniques for survival analysis in the simplified setting of time-fixed treatments.

17.1 Hazards and risks

Suppose we want to estimate the average causal effect of smoking cessation A (1: yes, 0: no) on the time to death T with time measured from the start of follow-up. This is an example of a *survival analysis*: the outcome is time to an event of interest that can occur at any time after the start of follow-up. In most follow-up studies, the event of interest is not observed to happen for all, or even the majority of, individuals in the study. This is so because most follow-up studies have a date after which there is no information on any individuals: the *administrative end of follow-up*.

After the administrative end of follow-up, no additional data can be used. Individuals who do not develop the event of interest before the administrative end of follow-up have their survival time administratively censored, i.e., we know that they survived beyond the administrative end of follow-up, but we do not know for how much longer. For example, let us say that we conduct the above survival analysis among the 1629 cigarette smokers from the NHEFS who were aged 25-74 years at baseline and who were alive through 1982. For all individuals, the start of follow-up is January 1, 1983 and the administrative end of follow-up is December 31, 1992. We define the administrative censoring time to be the difference between the date of administrative end of follow-up and date at which follow-up begins. In our example, this time is the same—120 months—for all individuals because the start of follow-up and the administrative end of follow-up are the same for everybody. Of the 1629 individuals, only 318 individuals died before the end of 1992, so the survival time of the remaining 1311 individuals is administratively censored.

Administrative censoring is a problem intrinsic to survival analyses—studies of smoking cessation and death will rarely, if ever, follow a cohort of individuals until extinction—but administrative censoring is not the only type of censoring that may occur in survival analyses. Like any other causal analyses, survival

In a study with staggered entry (i.e., with a variable start of follow-up date) different individuals will have different administrative censoring times, even when the administrative end of follow-up date is common to all.

Fine Point 17.1

Competing events As described in Section 8.5, a competing event is an event (typically, death) that prevents the event of interest (e.g., stroke) from happening: individual who die from other causes (say, cancer) cannot ever develop stroke. In survival analyses, the key decision is whether to consider competing events a form of non-administrative censoring.

- If the competing event is considered a censoring event, then the analysis is effectively an attempt to simulate a population in which death from other causes is somehow either abolished or rendered independent of the risk factors for stroke. The resulting effect estimate is hard to interpret and may not correspond to a meaningful estimand (see Chapter 8). In addition, the censoring may introduce selection bias under the null, which would require adjustment (by, say, IP weighting) using data on the measured risk factors for the event of interest.
- If the competing event is not considered a censoring event, then the analysis effectively sets the time to event to be infinite. That is, dead individuals are considered to have probability zero of developing stroke between their death and the administrative end of follow-up. The estimate of the effect of treatment on stroke is hard to interpret because a non-null estimate may arise from a direct effect of treatment on death, which would prevent the occurrence of stroke.

An alternative to the handling of competing events is to create a composite event that includes both the competing event and the event of interest (e.g., death and stroke) and conduct a survival analysis for the composite event. This approach effectively eliminates the competing events, but fundamentally changes the causal question. Again, the resulting effect estimate is hard to interpret because a non-null estimate may arise from either an effect of treatment on stroke or on death. Another alternative is to restrict the inference to the principal stratum of individuals who would not die regardless of the treatment level they received. This approach targets a sort of local average effect, as defined in Chapter 16, which makes both interpretation and valid estimation especially challenging.

None of the above strategies provides a satisfactory solution to the problem of competing events. Indeed the presence of competing events raises logical questions about the meaning of the causal estimand that cannot be bypassed by statistical techniques. For a detailed description of approaches to handle competing events and their challenges, see the discussion by Young et al. (2019). More recently, Stensrud et al. (2020, 2021) proposed an approach based on separable effects, a concept discussed in Technical Point 23.3.

analysis may also need to handle non-administrative types of censoring, such as loss to follow-up (e.g., dropout from the study) and competing events (see Fine Point 17.1). In previous chapters we have discussed how to adjust for the selection bias introduced by non-administrative censoring via standardization or IP weighting. The same approaches can be applied to survival analyses. Therefore, in this chapter, we will focus on administrative censoring. We defer a more detailed consideration of non-administrative censoring to Part III of the book because non-administrative censoring is generally a time-varying process, whereas the time of administrative censoring is fixed at baseline.

In our example, the month of death T can take values subsequent from 1 (January 1983) to 120 (December 1992). T is known for 102 treated ($A = 1$) and 216 untreated ($A = 0$) individuals who died during the follow-up, and is administratively censored (that is, all we know is that it is greater than 120 months) for the remaining 1311 individuals. Therefore we cannot compute the mean survival $\hat{E}[T]$ as we did in previous chapters with the outcome of interest. Rather, in survival analysis we need to use other measures that can accommodate administrative censoring. Some common measures are the survival probability, the risk, and the hazard. Let us define these quantities, which are functions of the survival time T .

The *survival probability* $\Pr[T > k]$, or simply the survival at month k , is

For simplicity, we assume that anyone without confirmed death survived the follow-up period. In reality, some individuals may have died but confirmation (by, say, a death certificate or a proxy interview) was not feasible. Also for simplicity, we will ignore the problem described in Fine Point 12.1.

Other effect measures that can be derived from survival curves are years of life lost and the restricted mean survival time.

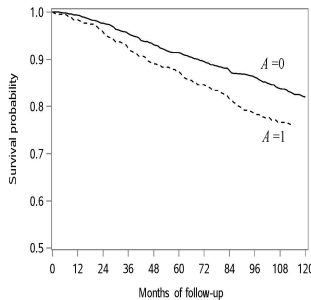


Figure 17.1

CODE: Program 17.1

A conventional statistical test to compare survival curves (the log-rank test) yielded a P-value=0.005.

the proportion of individuals who survived through time k . If we calculate the survivals at each month until the administrative end of follow-up $k_{end} = 120$ and plot them along a horizontal time axis, we obtain the *survival curve*. The survival curve starts at $\Pr[T > 0] = 1$ for $k = 0$ and then decreases monotonically—that is, it does not increase—with subsequent values of $k = 1, 2, \dots, k_{end}$. Alternatively, we can define *risk*, or cumulative incidence, at time k as one minus the survival $1 - \Pr[T > k] = \Pr[T \leq k]$. The cumulative incidence curve starts at $\Pr[T \leq 0] = 0$ and increases monotonically during the follow-up.

In survival analyses, a natural approach to quantify the treatment effect is to compare the survival or risk under each treatment level at some or all times k . Of course, in our smoking cessation example, a contrast of these curves may not have a causal interpretation because the treated and the untreated are probably not exchangeable. However, suppose for a second (actually, until Section 17.4) that quitters ($A = 1$) and non-quitters ($A = 0$) are marginally exchangeable. Then we can construct the survival curves shown in Figure 17.1 and compare $\Pr[T > k|A = 1]$ versus $\Pr[T > k|A = 0]$ for all times k . For example, the survival at 120 months was 76.2% among quitters and 82.0% among non-quitters. Alternatively, we could contrast the risks rather than the survivals. For example, the 120-month risk was 23.8% among quitters and 18.0% among non-quitters.

At any time k , we can also calculate the proportion of individuals who develop the event among those who had not developed it before k . This is the *hazard* $\Pr[T = k|T > k - 1]$. Technically, this is the discrete time hazard, i.e., the hazard in a study in which time is measured in discrete intervals—as opposed to measured continuously. Because in real-world studies, time is indeed measured in discrete intervals (years, months, weeks, days...) rather than in a truly continuous fashion, here we will refer to the discrete time hazard as, simply, the hazard.

The risk and the hazard are different measures. The denominator of the risk—the number of individuals at baseline—is constant across times k and its numerator—all events between baseline and k —is cumulative. That is, the risk will stay flat or increase as k increases. On the other hand, the denominator of the hazard—the number of individuals alive at k —varies over time t and its numerator includes only recent events—those during interval k . That is, the hazard may increase or decrease over time. In our example, the hazard at 120 months was 0% among quitters (because the last death happened at 113 months in this group) and $1/986 = 0.10\%$ among non-quitters, and the hazard curves between 0 and 120 months had roughly the shape of a letter *M*.

A frequent approach to quantify the treatment effect in survival analyses is to estimate the ratio of the hazards in the treated and the untreated, known as the *hazard ratio*. However, the hazard ratio is problematic for the reasons described in Fine Point 17.2. Therefore, the survival analyses in this book privilege survival/risk over hazard. However, that does not mean that we should completely ignore hazards. The estimation of hazards is often a useful intermediate step for the estimation of survivals and risks.

17.2 From hazards to risks

In survival analyses, there are two main ways to arrange the analytic dataset. In the first data arrangement each row of the database corresponds to one

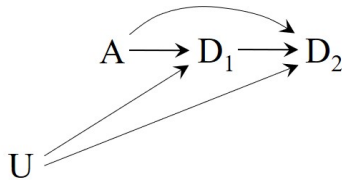


Figure 17.2

By definition, everybody had to survive month 0 in order to be included in the dataset, i.e., $D_0 = 0$ for all individuals.

person. This data format—often referred to as the “wide” format when there are time-varying treatments and confounders—is the one we have used so far in this book. In the analyses of the previous section, the dataset had 1629 rows, one per individual.

In the second data arrangement each row of the database corresponds to a person-time. That is, the first row contains the information for person 1 at $k = 0$, the second row the information for person one at $k = 1$, the third row the information for person 1 at $k = 2$, and so on until the follow-up of person one ends. The next row contains the information of person 2 at $k = 0$, etc. This person-time (or “long”) data format is the one we will use in most survival analyses in this chapter and in all analyses with time-varying treatments in Part III. In our smoking cessation example, the person-time dataset has 176,764 rows, one per person-month.

To encode survival information through k in the person-time data format, it is helpful to define a time-varying indicator of event D_k . For each person at each month k , the indicator D_k takes value 1 if $T \leq k$ and value 0 if $T > k$. The causal diagram in Figure 17.2 shows the treatment A and the event indicators D_1 and D_2 at times 1 and 2, respectively. The variable U represents the (generally unmeasured) factors that increase susceptibility to the event of interest. Note that sometimes these susceptibility factors are time-varying too. In that case, they can be depicted in the causal diagram as U_0, U_1, \dots , and so on. Part III deals with the case in which the treatment itself is time-varying.

In the person-time data format, the row for a particular individual at time k includes the indicator D_{k+1} . In our example, the first row of the person-time dataset, for individual one at $k = 0$, includes the indicator D_1 , which is 1 if the individual died during month 1 and 0 otherwise; the second row, for individual one at $k = 1$, includes the indicator D_2 , which is 1 if the individual died during month 2 and 0 otherwise; and so on. The last row in the dataset for each individual is either her first row with $D_{k+1} = 1$ or the row corresponding to month 119.

Using the time-varying outcome variable D_k , we can define survival at k as $\Pr[D_k = 0]$, which is equal to $\Pr[T > k]$, and risk at k as $\Pr[D_k = 1]$, which is equal to $\Pr[T \leq k]$. The hazard at k is defined as $\Pr[D_k = 1 | D_{k-1} = 0]$. For $k = 1$ the hazard is equal to the risk because everybody is, by definition, alive at $k = 0$.

The survival probability at k is the product of the conditional probabilities of having survived each interval between 0 and k . For example, the survival at $k = 2$, $\Pr[D_2 = 0]$, is equal to survival probability at $k = 1$, $\Pr[D_1 = 0]$, times the survival probability at $k = 2$ conditional on having survived through $k = 1$, $\Pr[D_2 = 0 | D_1 = 0]$. More generally, the survival at k is

$$\Pr[D_k = 0] = \prod_{m=1}^k \Pr[D_m = 0 | D_{m-1} = 0]$$

That is, the survival at k equals the product of one minus the hazard at all previous times. If we know the hazards through k we can easily compute the survival at k (or the risk at k , which is just one minus the survival).

The hazard at k , $\Pr[D_k = 1 | D_{k-1} = 0]$, can be estimated nonparametrically by dividing the number of cases during the interval k by the number of individuals alive at the end of interval $k - 1$. If we substitute this estimate into the above formula the resulting nonparametric estimate of the survival $\Pr[D_k = 0]$ at k is referred to as the Kaplan-Meier, or product-limit, estima-

Fine Point 17.2

The hazards of hazard ratios When using the hazard ratio as a measure of causal effect, two important properties of the hazard ratio need to be taken into account.

First, because the hazards vary over time, the hazard ratio generally does too. That is, the ratio at time k may differ from that at time $k + 1$. However, many published survival analyses report a single hazard ratio, which is usually the consequence of fitting a Cox proportional hazards model that assumes a constant hazard ratio by ignoring interactions with time. The reported hazard ratio is a weighted average of the k -specific hazard ratios, which makes it hard to interpret. If the risk is rare and censoring only occurs at a common administrative censoring time k_{end} , then the weight of the hazard ratio at time k is proportional to the total number of events among untreated individuals that occur at k . (Technically, the weights are equal to the conditional density at k of T given $A = 0$ and $T < k_{end}$.) Because it is a weighted average, the reported hazard ratio may be 1 even if the survival curves are not identical. In contrast to “the” hazard ratio, ratios and differences of survival probabilities and risks are defined with respect to a fixed time period, e.g., the 5-year survival difference, the 120-month risk ratio.

Second, even if we presented the time-specific hazard ratios, their causal interpretation is not straightforward. Suppose treatment kills all high-risk individuals by time k and has no effects on others. Then the hazard ratio at time $k + 1$ compares the treated and the untreated individuals who survived through k . In the treated group, the survivors are all low-risk individuals (because the high-risk ones have already been killed by treatment); in the untreated group, the survivors are a mixture of high-risk and low-risk individuals (because treatment did not weed out the former). As a result the hazard ratio at $k + 1$ will be less than 1 even though treatment is not beneficial for any individual.

This apparent paradox is an example of selection bias due to conditioning on a post-treatment variable (i.e., being alive at k) which is affected by treatment. For example, the hazard ratio at time 2 is the probability $\Pr[D_2 = 1 | D_1 = 0, A]$ of the event at time 2 among those who survived time 1. As depicted in the causal diagram of Figure 17.3, the conditioning on the collider D_1 will generally open the path $A \rightarrow D_1 \leftarrow U \rightarrow D_2$ and therefore induce an association between treatment A and event D_2 among those with $D_1 = 0$. This built-in selection bias of hazard ratios does not happen if the survival curves are the same in the treated and the untreated, i.e., if there are no arrows from A into the indicators for the event. Hernán (2010) described an example of this problem.

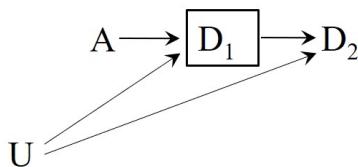


Figure 17.3

tor. Figure 17.1 was constructed using the Kaplan-Meier estimator, which is an excellent estimator of the survival curve, provided the total number of failures over the follow up period is reasonably large. Typically, the number of cases during each interval is low (or even zero) and thus the nonparametric estimates of the hazard $\Pr[D_k = 1 | D_{k-1} = 0]$ at k will be very unstable. If our interest is in estimation of the hazard at a particular k , smoothing via a parametric model may be required (see Chapter 11 and Fine Point 17.3).

An easy way to parametrically estimate the hazards is to fit a logistic regression model for $\Pr[D_{k+1} = 1 | D_k = 0]$ that, at each k , is restricted to individuals who survived through k . The fit of this model is straightforward when using the person-time data format. In our example, we can estimate the hazards in the treated and the untreated by fitting the logistic model

$$\text{logit } \Pr[D_{k+1} = 1 | D_k = 0, A] = \theta_{0,k} + \theta_1 A + \theta_2 A \times k + \theta_3 A \times k^2$$

Although each person occurs in multiple rows of the person-time data structure, the standard error of the parameter estimates outputted by a routine logistic regression program will be correct if the hazards model is correct.

where $\theta_{0,k}$ is a time-varying intercept that can be estimated by some flexible function of time such as $\theta_{0,k} = \theta_0 + \theta_4 k + \theta_5 k^2$. The flexible time-varying intercept allows for a time-varying hazard and the product terms between treatment A and time ($\theta_2 A \times k + \theta_3 A \times k^2$) allow the hazard ratio to vary over time. See Technical Point 17.1 for details on how a logistic model approximates a hazards model. Functions other than the logit (e.g., the probit) can also be used to model dichotomous outcomes and therefore to estimate hazards.

We then compute estimates of the survival $\Pr[D_{k+1} = 0 | A = a]$ by multiplying the estimates of one minus the estimates of $\Pr[D_{k+1} = 1 | D_k = 0, A = a]$

Fine Point 17.3

Models for survival analysis. Methods for survival analysis need to accommodate the expected censoring of failure times due to administrative end of follow-up.

Nonparametric approaches to survival analysis, like constructing Kaplan-Meier curves, make no assumptions about the distribution of the unobserved failure times due to administrative censoring. On the other hand, parametric models for survival analysis assume a particular statistical distribution (e.g., exponential, Weibull) for the failure times or hazards. The logistic model described in the main text to estimate hazards is an example of a parametric model.

Other models for survival analysis, like the Cox proportional hazards model and the accelerated failure time (AFT) model, do not assume a particular distribution for the failure times or hazards. In particular, these models are agnostic about the shape of the hazard when all covariates in the model have value zero—often referred to as the baseline hazard. These models, however, impose a priori restrictions on the relation between the baseline hazard and the hazard under other combinations of covariate values. As a result, these methods are referred to as *semiparametric* methods.

See the book by Hosmer, Lemeshow, and May (2008) for a review of applied survival analysis. More formal descriptions can be found in the books by Fleming and Harrington (2005) and Kalbfleisch and Prentice (2002).

CODE: Program 17.2

provided by the logistic model, separately for the treated and the untreated. Figure 17.4 shows the survival curves obtained after parametric estimation of the hazards. These curves are a smooth version of those in Figure 17.1.

The validity of this procedure requires no misspecification of the hazards model. In our example, this assumption seems plausible because we obtained essentially the same survival estimates as in the previous section when we estimated the survival in a fully nonparametric way. A 95% confidence interval around the survival estimates can be easily constructed via bootstrapping of the individuals in the population.

17.3 Why censoring matters

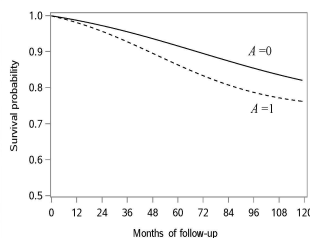


Figure 17.4

The only source of censoring in our study is a common administrative censoring time $k_{end} = 120$ that is identical for all individuals. In this simple setting the procedure described in the previous section to estimate the survival is overkill. One can simply estimate the survival probabilities $\Pr[D_{k+1} = 0|A = a]$ by the fraction of individuals who received treatment a and survived to $k + 1$, or by fitting separate logistic models for $\Pr[D_{k+1} = 0|A]$ at each time, for $k = 0, 1, \dots, k_{end}$.

Now suppose that individuals start the follow-up at different dates—there is staggered entry into the study—but the administrative end of follow-up date is common to all. Because the administrative censoring time is the difference between the administrative end of follow-up and the time of start of follow-up, different individuals will have different administrative censoring times. In this setting it is helpful to define a time-varying indicator C_k for censoring by time k . For each person at each month k , the indicator C_k takes value 0 if the administrative end of follow-up is greater than k and 1 otherwise. In the person-time data format, the row for a particular individual at time k includes the indicator C_{k+1} . We did not include this variable in our dataset because $C_{k+1} = 0$ for all individuals at all times k before 120 months. In the general case with random (i.e., individual-specific) administrative censoring, the indicator C_{k+1} will transition from 0 to 1 at different times k for different

Technical Point 17.1

Approximating the hazard ratio via a logistic model. The (discrete-time) hazard ratio $\frac{\Pr[D_{k+1}=1|D_k=0, A=1]}{\Pr[D_{k+1}=1|D_k=0, A=0]}$ is $\exp(\alpha_1)$ at all times $k+1$ in the hazards model $\Pr[D_{k+1}=1|D_k=0, A] = \Pr[D_{k+1}=1|D_k=0, A=0] \times \exp(\alpha_1 A)$. If we take logs on both sides of the equation, we obtain $\log \Pr[D_{k+1}=1|D_k=0, A] = \alpha_{0,k} + \alpha_1 A$ where $\alpha_{0,k} = \log \Pr[D_{k+1}=1|D_k=0, A=0]$.

Suppose the hazard at $k+1$ is small, i.e., $\Pr[D_{k+1}=1|D_k=0, A] \approx 0$. Then one minus the hazard at $k+1$ is close to one, and the hazard is approximately equal to the odds: $\Pr[D_{k+1}=1|D_k=0, A] \approx \frac{\Pr[D_{k+1}=1|D_k=0, A]}{\Pr[D_{k+1}=0|D_k=0, A]}$. We then have

$$\log \frac{\Pr[D_{k+1}=1|D_k=0, A]}{\Pr[D_{k+1}=0|D_k=0, A]} = \text{logit } \Pr[D_{k+1}=1|D_k=0, A] \approx \alpha_{0,k} + \alpha_1 A$$

That is, if the hazard is close to zero at $k+1$, we can approximate the log hazard ratio α_1 by θ_1 in a logistic model $\text{logit } \Pr[D_{k+1}=1|D_k=0, A] = \theta_{0,k} + \theta_1 A$ like the one we used in the main text (Thompson 1977). As a rule of thumb, the approximation is often considered to be accurate enough when $\Pr[D_{k+1}=1|D_k=0, A] < 0.1$ for all k .

This rare event condition can almost always be guaranteed to hold: we just need to define a time unit k that is short enough for $\Pr[D_{k+1}=1|D_k=0, A] < 0.1$. For example, if D_k stands for lung cancer, k may be measured in years; if D_k stands for infection with the common cold virus, k may be measured in days. The shorter the time unit, the more rows in the person-time dataset used to fit the logistic model.

people.

Our goal is to estimate the survival curve that would have been observed if nobody had been censored before k_{end} , where k_{end} is the maximum administrative censoring time in the study. That is, our goal is to estimate the survival $\Pr[D_k=0|A=a]$ that would have been observed if the value of the time-varying indicators D_k were known even after censoring. Technically, we can also refer to this quantity as $\Pr[D_k^{\bar{c}=\bar{0}}=0|A=a]$ where $\bar{c} = (c_1, c_2 \dots c_{k_{end}})$. As discussed in Chapter 12, the use of the superscript $\bar{c} = \bar{0}$ makes explicit the quantity that we have in mind. We sometimes choose to omit the superscript $\bar{c} = \bar{0}$ when no confusion can arise. For simplicity, suppose that the time of start of follow-up was as if randomly assigned to each individual, which would be the case if there were no secular trends in any variable. Then the administrative censoring time, and therefore the indicator \bar{C} , is independent of both treatment and death time.

We cannot validly estimate this survival $\Pr[D_k=0|A=a]$ at time k by simply computing the fraction of individuals who received treatment level a and survived and were not censored through k . This fraction is a valid estimator of the joint probability $\Pr[C_{k+1}=0, D_{k+1}=0|A=a]$, which is not what we want. To see why, consider a study with $k_{end}=2$ and in which the following happens:

- $\Pr[C_1=0] = 1$, i.e., nobody is censored by $k=1$
- $\Pr[D_1=0|C_0=0] = 0.9$, i.e., 90% of individuals survive through $k=1$
- $\Pr[C_2=0|D_1=0, C_1=0] = 0.5$, i.e., a random half of survivors is censored by $k=2$
- $\Pr[D_2=0|C_2=0, D_1=0, C_1=0] = 0.9$, i.e., 90% of the remaining individuals survive through $k=2$

The fraction of uncensored survivors at $k=2$ is $1 \times 0.9 \times 0.5 \times 0.9 = 0.405$. However, if nobody had been censored, i.e., if $\Pr[C_2=0|D_1=0, C_1=0] =$

1, the survival would have been $1 \times 0.9 \times 1 \times 0.9 = 0.81$. This example motivates how correct estimation of the survivals $\Pr[D_k = 0|A = a]$ requires the procedures described in the previous section. Specifically, under (as if) randomly assigned censoring, the survival $\Pr[D_k = 0|A = a]$ at k is

$$\prod_{m=1}^k \Pr[D_m = 0|D_{m-1} = 0, C_m = 0, A = a] \text{ for } k < k_{end}$$

The estimation procedure is the same as described above except that we either use a nonparametric estimate of, or fit a logistic model for, the cause-specific hazard $\Pr[D_{k+1} = 1|D_k = 0, C_{k+1} = 0, A = a]$.

Often we are not ready to assume that censoring is as if randomly assigned. When there is staggered entry, an individual's time of administrative censoring depends on the calendar time at study entry (later entries have shorter values of k_{end}) and calendar time may itself be associated with the outcome. Therefore, the above procedure will need to be adjusted for baseline calendar time. In addition, there may be other baseline prognostic factors that are unequally distributed between the treated ($A = 1$) and the untreated ($A = 0$), which also requires adjustment. The next sections extend the above procedure to incorporate adjustment for baseline confounders via g-methods. In Part III we extend the procedure to settings with time-varying treatments and confounders.

17.4 IP weighting of marginal structural models

When the treated and the untreated are not exchangeable, a direct contrast of their survival curves cannot be endowed with a causal interpretation. In our smoking cessation example, we estimated that the 120-month survival was lower in quitters than in non-quitters (76.2% versus 82.0%), but that does not necessarily imply that smoking cessation increases mortality. Older people are more likely to quit smoking and also more likely to die. This confounding by age makes smoking cessation look bad because the proportion of older people is greater among quitters than among non-quitters.

Let us define $D_k^{a, \bar{c}=\bar{0}}$ as a counterfactual time-varying indicator for death at k under treatment level a and no censoring. For simplicity of notation, we will write $D_k^{a, \bar{c}=\bar{0}}$ as D_k^a when, as in this chapter, it is clear that the goal is estimating the survival in the absence of censoring. For additional simplicity, in the remainder of this chapter we omit $C_k = 0$ from the conditioning event of the hazard at k , $\Pr[D_{k+1} = 0|D_k = 0, L = l, A]$. That is, we write all expressions as if all individuals had a common administrative censoring time, like in our smoking cessation example.

Suppose we want to compare the counterfactual survivals $\Pr[D_{k+1}^{a=1} = 0]$ and $\Pr[D_{k+1}^{a=0} = 0]$ that would have been observed if everybody had received treatment ($a = 1$) and no treatment ($a = 0$), respectively. That is, the causal contrast of interest is

$$\Pr[D_{k+1}^{a=1} = 0] \text{ vs. } \Pr[D_{k+1}^{a=0} = 0] \text{ for } k = 0, 2, \dots, k_{end} - 1$$

Because of confounding, this contrast may not be validly estimated by the contrast of the survivals $\Pr[D_{k+1} = 0|A = 1]$ and $\Pr[D_{k+1} = 0|A = 0]$ that we described in the previous sections. Rather, a valid estimation of the quantities $\Pr[D_{k+1}^a = 0]$ for $a = 1$ and $a = 0$ typically requires adjustment for

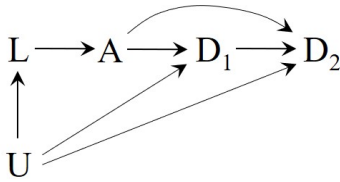


FIGURE 17.5

CODE: Program 17.3

confounders, which can be achieved through several methods. This section focuses on IP weighting. Let us assume that the treated ($A = 1$) and the untreated ($A = 0$) are exchangeable within levels of the L variables, as represented in the causal diagram of Figure 17.5. Like in Chapters 12 to 15, L includes the variables sex, age, race, education, intensity and duration of smoking, physical activity in daily life, recreational exercise, and weight. We also assume positivity and consistency. The estimation of IP weighted survival curves has two steps.

First, we estimate the stabilized IP weight SW^A for each individual in the study population. The procedure is exactly the same as the one described in Chapter 12. We fit a logistic model for the conditional probability $\Pr[A = 1|L]$ of treatment (i.e., smoking cessation) given the variables in L . The denominator of the estimated SW^A is $\widehat{\Pr}[A = 1|L]$ for treated individuals and $(1 - \widehat{\Pr}[A = 1|L])$ for untreated individuals, where $\widehat{\Pr}[A = 1|L]$ is the predicted value from the logistic model. The numerator of the estimated weight SW^A is $\widehat{\Pr}[A = 1]$ for the treated and $(1 - \widehat{\Pr}[A = 1])$ for the untreated, where $\widehat{\Pr}[A = 1]$ can be estimated nonparametrically or as the predicted value from a logistic model for the marginal probability $\Pr[A = 1]$ of treatment. See Chapter 11 for details on predicted values.

The application of the estimated weights SW^A creates a pseudo-population in which the variables in L are independent from the treatment A , which eliminates confounding by those variables. In our example, the weights had mean 1 (as expected) and ranged from 0.33 to 4.21.

Second, using the person-time data format, we fit a hazards model like the one described above except that individuals are weighted by their estimated SW^A . Technically, this IP weighted logistic model estimates the parameters of the marginal structural logistic model

$$\text{logit } \Pr[D_{k+1}^a = 0 | D_k^a = 0] = \beta_{0,k} + \beta_{1,a} + \beta_{2,a} \times k + \beta_{3,a} \times k^2$$

That is, the IP weighted model estimates the time-varying hazards that would have been observed if all individuals in the study population had been treated ($a = 1$) and the time-varying hazards if they had been untreated ($a = 0$).

The estimates of $\Pr[D_{k+1}^a = 0 | D_k^a = 0]$ from the IP weighted hazards models can then be multiplied over time (see previous section) to obtain an estimate of the survival $\Pr[D_{k+1}^a = 0]$ that would have been observed under treatment $a = 1$ and under no treatment $a = 0$. The resulting curves are shown in Figure 17.6.

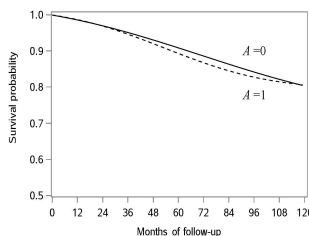


Figure 17.6

In our example, the 120-month survival estimates were 80.7% under smoking cessation and 80.5% under no smoking cessation; difference 0.2% (95% confidence interval from -4.1% to 3.7% based on 500 bootstrap samples). Though the survival curve under treatment was lower than the curve under no treatment for most of the follow-up, the maximum difference never exceeded -1.4% with a 95% confidence interval from -3.4% to 0.7% . That is, after adjustment for the covariates L via IP weighting, we found little evidence of an effect of smoking cessation on mortality at any time during the follow-up. The validity of this procedure requires no misspecification of both the treatment model and the marginal hazards model.

17.5 The parametric g-formula

In the previous section we estimated the survival curve under treatment and under no treatment in the entire study population via IP weighting. To do so, we adjusted for L and assumed exchangeability, positivity, and consistency. Another method to estimate the marginal survival curves under those assumptions is standardization based on parametric models, i.e., the parametric g-formula.

The survival $\Pr[D_{k+1}^a = 0]$ at $k+1$ under treatment level a is the weighted average of the survival conditional probabilities at $k+1$ within levels of the covariates L and treatment level $A = a$, with the proportion of individuals in each level l of L as the weights. That is, under exchangeability, positivity, and consistency, $\Pr[D_{k+1}^a = 0]$ equals the standardized survival

$$\sum_l \Pr[D_{k+1} = 0 | L = l, A = a] \Pr[L = l].$$

For a formal proof, see Section 2.3.

Therefore, the estimation of the parametric g-formula has two steps. First, we need to estimate the conditional survivals $\Pr[D_{k+1} = 0 | L = l, A = a]$ using our administratively censored data. Second, we need to compute their weighted average over all values l of the covariates L . We describe each of these two steps in our smoking cessation example.

For the first step we fit a parametric hazards model like the one described in Section 17.2, except that the variables in L are included as covariates. If the model is correctly specified, it validly estimates the time-varying hazards $\Pr[D_{k+1} = 1 | D_k = 0, L, A]$ within levels of treatment A and covariates L . The product of one minus the conditional hazards

$$\prod_{m=0}^k \Pr[D_{m+1} = 0 | D_m = 0, L = l, A = a]$$

is equal to the conditional survival $\Pr[D_{k+1} = 0 | L = l, A = a]$. Because of conditional exchangeability given L , the conditional survival for a particular set of covariate values $L = l$ and $A = a$ can be causally interpreted as the survival that would have been observed if everybody with that set of covariates had received treatment value a . That is,

$$\Pr[D_{k+1} = 0 | L = l, A = a] = \Pr[D_{k+1}^a = 0 | L = l]$$

Therefore the conditional hazards can be used to estimate the survival curve under treatment ($a = 1$) and no treatment ($a = 0$) within each combination of values l of L . For example, we can use this model to estimate the survival curves under treatment and no treatment for white men aged 61, with college education, low levels of exercise, etc. However, our goal is estimating the marginal, not the conditional, survival curves under treatment and under no treatment.

For the second step we compute the weighted average of the conditional survival across all values l of the covariates L , i.e., we standardize the survival to the confounder distribution. To do so, we use the method described in Section 13.3 to standardize means: standardization by averaging after expansion of dataset, outcome modeling, and prediction. This method can be used even when some of the variables in L are continuous so that the sum over values l is formally an integral. The resulting curves are shown in Figure 17.7.

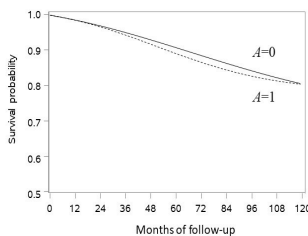


Figure 17.7

In Chapter 12 we referred to models conditional on all the covariates L as faux marginal structural models.

CODE: Program 17.4

The procedure is analogous to the one described in Chapter 13.

In our example, the survival curve under treatment was lower than the curve under no treatment during the entire follow-up, but the maximum difference never exceeded -2.0% (95% confidence interval from -5.6% to 1.8%). The 120-month survival estimates were 80.4% under smoking cessation and 80.6% under no smoking cessation; difference 0.2% (95% confidence interval from -4.6% to 4.1%). That is, after adjustment for the covariates L via standardization, we found little evidence of an effect of smoking cessation on mortality at any time during the follow-up. Note that the survival curves estimated via IP weighting (previous section) and the parametric g-formula (this section) are similar but not identical because they rely on different parametric assumptions: the IP weighted estimates require no misspecification of a model for treatment and a model for the unconditional hazards; the parametric g-formula estimates require no misspecification of a model for the conditional hazards.

17.6 G-estimation of structural nested models

The previous sections describe causal contrasts that compare survivals, or risks, under different levels of treatment A . The survival was computed from hazards estimated by logistic regression models. This approach is feasible when the analytic method is IP weighting of marginal structural models or the parametric g-formula, but not when the method is g-estimation of structural nested models. As explained in Chapter 14, structural nested models are models for conditional causal contrasts (e.g., the difference or ratio of covariate-specific means under different treatment levels), not for the components of those contrasts (e.g., each of the means under different treatment levels). Therefore we cannot estimate survivals or hazards using a structural nested model.

We can, however, consider a structural nested log-linear model to model the ratio of cumulative incidences (i.e., risks) under different treatment levels. *Structural nested cumulative failure time models* do precisely that (see Technical Point 17.2), but they are best used when failure is a rare event because log-linear models do not naturally impose an upper limit of 1 on the risk. For non-rare failures, we can instead consider a structural nested log-linear model to model the ratio of cumulative survival probabilities (i.e., $1 - \text{risk}$) under different treatment levels. *Structural nested cumulative survival time models* do precisely that (see Technical Point 17.2), but they are best used when survival is rare because log-linear models do not naturally impose an upper limit of 1 on the survival. A more general option is to use a structural nested model that models the ratio of survival times under different treatment options. That is, an accelerated failure time (AFT) model.

Let T_i^a be the counterfactual time of survival for individual i under treatment level a . The effect of treatment A on individual i 's survival time can be measured by the ratio $T_i^{a=1}/T_i^{a=0}$ of her counterfactual survival times under treatment and under no treatment. If the survival time ratio is greater than 1, then treatment is beneficial because it increases the survival time; if the ratio is less than 1, then treatment is harmful; if the ratio is 1, then treatment has no effect. Suppose, temporarily, that the effect of treatment is the same for every individual in the population.

We could then consider the *structural nested accelerated failure time (AFT) model* $T_i^a/T_i^{a=0} = \exp(-\psi_1 a)$, where ψ_1 measures the expansion (or contraction) of each individual's survival time attributable to treatment. If $\psi_1 < 0$ then treatment increases survival time, if $\psi_1 > 0$ then treatment decreases

In fact, we may not even approximate a hazard ratio because structural nested logistic models do not generalize easily to time-varying treatments (Technical Point 14.1).

Tchetgen Tchetgen et al (2015) and Robins (1997b) described survival analysis with instrumental variables that exhibit similar problems to those described here for structural nested models.

The "nested" component is only evident when treatment is time-varying. See Chapter 21.

Technical Point 17.2

Structural nested cumulative failure time (CFT) models and cumulative survival time (CST) models. For a time-fixed treatment, a (non-nested) structural CFT model is a model for the ratio of the counterfactual risk under treatment value a divided by the counterfactual risk under treatment value 0 conditional on treatment A and covariates L . The general form of the model is

$$\frac{\Pr[D_k^a = 1|L, A]}{\Pr[D_k^{a=0} = 1|L, A]} = \exp[\gamma_k(L, A; \psi)]$$

where $\gamma_k(L, A; \psi)$ is a function of treatment and covariate history indexed by the (possibly vector-valued) parameter ψ . For consistency, $\exp[\gamma_k(L, A; \psi)]$ must be 1 when $A = 0$ because then the two treatment values being compared are identical, and when there is no effect of treatment at time m on outcome at time k . An example of such a function is $\gamma_k(L, A; \psi) = \psi A$ so $\psi = 0$ corresponds to no effect, $\psi < 0$ to beneficial effect, and $\psi > 0$ to harmful effect.

Analogously, for a time-fixed treatment, a (non-nested) structural CST model is a model for the ratio of the counterfactual survival under treatment value a divided by the counterfactual survival under treatment level 0 conditional on treatment A and covariates L . The general form of the model is

$$\frac{\Pr[D_k^a = 0|L, A]}{\Pr[D_k^{a=0} = 0|L, A]} = \exp[\gamma_k(L, A; \psi)]$$

Although CFT and CST models differ only in whether we specify a multiplicative model for $\Pr[D_k^a = 1|L, A]$ versus for $\Pr[D_k^a = 0|L, A]$, the meaning of $\gamma_k(L, A; \psi)$ differs because a multiplicative model for risk is not a multiplicative model for survival, whenever the treatment effect is non-null. When we let the time index k be continuous rather than discrete, a structural CST model is equivalent to a structural additive hazards model (Tchetgen Tchetgen et al., 2015) as any model for $\Pr[D_k^a = 0|L, A] / \Pr[D_k^{a=0} = 0|L, A]$ induces a model for the difference in the time-specific hazards of T^a and $T^{a=0}$, and vice-versa.

The use of structural CFT models requires that, for all values of the covariates L , the conditional cumulative probability of failure under all treatment values satisfies a particular type of rare failure assumption. In this “rare failure” context, the structural CFT model has an advantage over AFT models: it admits unbiased estimating equations that are differentiable in the model parameters and thus are easily solved. Page (2005) and Picciotto et al. (2012) provided further details on structural CFT and CST models. For a time-varying treatment, this class of models can be viewed as a special case of the multivariate structural nested mean model (Robins 1994). See Technical Point 14.1.

The negative sign in front of ψ preserves the usual interpretation of positive parameters indicating harm and negative parameters indicating benefit.

survival time, if $\psi_1 = 0$ then treatment does not affect survival time. More generally, the effect of treatment may depend on covariates L so a more general structural AFT would be $T_i^a / T_i^{a=0} = \exp(-\psi_1 a - \psi_2 a L_i)$, with ψ_1 and ψ_2 (a vector) constant across individuals. Rearranging the terms, the model can be written as

$$T_i^{a=0} = T_i^a \exp(\psi_1 a + \psi_2 a L_i) \quad \text{for all individuals } i$$

Following the same reasoning as in Chapter 14, consistency of counterfactuals implies the model $T_i^{a=0} = T_i \exp(\psi_1 A_i + \psi_2 A_i L_i)$, in which the counterfactual time T_i^a is replaced by the actual survival time $T_i^A = T_i$. The parameters ψ_1 and ψ_2 can be estimated by a modified g-estimation procedure (to account for administrative censoring) that we describe later in this section.

The above structural AFT is unrealistic because it is both deterministic and rank-preserving. It is deterministic because it assumes that, for each individual, the counterfactual survival time under no treatment $T^{a=0}$ can be computed without error as a function of the observed survival time T , treatment A , and covariates L . It is rank-preserving because, under this model, if

individuals i would die before individual j had they both been untreated, i.e., $T_i^{a=0} < T_j^{a=0}$, then individual i would also die before individual j had they both been treated, i.e., $T_i^{a=1} < T_j^{a=1}$.

Because of the implausibility of rank preservation, one should not generally use methods for causal inference that rely on it, as we discussed in Chapter 14. And yet again we will use a rank-preserving model here to describe g-estimation for structural AFT models because g-estimation is easier to understand for rank-preserving models, and because the g-estimation procedure is actually the same for rank-preserving and non-rank-preserving models.

Consider the simpler rank-preserving model $T_i^{a=0} = T_i \exp(\psi A_i)$ without a product term between treatment and covariates. G-estimation of the parameter ψ of this structural AFT model would be straightforward if administrative censoring did not exist, i.e., if we could observe the time of death T for all individuals. In fact, in that case the g-estimation procedure would be the same as we described in Section 14.5. The first step would be to compute candidate counterfactuals $H_i(\psi^\dagger) = T_i \exp(\psi^\dagger A_i)$ under many possible values ψ^\dagger of the causal parameter ψ . The second step would be to find the value ψ^\dagger that results in a $H_i(\psi^\dagger)$ that is independent of treatment A in a logistic model for the probability of $A = 1$ with $H_i(\psi^\dagger)$ and the confounders L as covariates. Such value ψ^\dagger would be the g-estimate of ψ .

However, this procedure cannot be implemented in the presence of administrative censoring at time K because $H_i(\psi^\dagger)$ cannot be computed for individuals with unknown T_i . One might then be tempted to restrict the g-estimation procedure to individuals with an observed survival time only, i.e., those with $T_i \leq K$. Unfortunately, that approach results in selection bias. To see why, consider the following oversimplified scenario.

We conduct a 60-month randomized experiment to estimate the effect of a dichotomous treatment A on survival time T . Only 3 types of individuals participate in our study. Type 1 individuals are those who, in the absence of treatment, would die at 36 months ($T^{a=0} = 36$). Type 2 individuals are those who in the absence of treatment, would die at 72 months ($T^{a=0} = 72$). Type 3 individuals are those who in the absence of treatment, would die at 108 months ($T^{a=0} = 108$). That is, type 3 individuals have the best prognosis and type 1 individuals have the worst one. Because of randomization, we expect that the proportions of type 1, type 2, and type 3 individuals are the same in each of the two treatment groups $A = 1$ and $A = 0$. That is, the treated and the untreated are expected to be exchangeable.

Suppose that treatment $A = 1$ decreases the survival time compared with $A = 0$. Table 17.1 shows the survival time under treatment and under no treatment for each type of individual. Because the administrative end of follow-up is $K = 60$ months, the death of type 1 individuals will be observed whether they are randomly assigned to $A = 1$ or $A = 0$ (both survival times are less than 60), and the death of type 3 individuals will be administratively censored whether they are randomly assigned to $A = 1$ or $A = 0$ (both survival times are greater than 60). The death of type 2 individuals, however, will only be observed if they are assigned to $A = 1$. Hence an analysis that welcomes all individuals with non-administratively censored death times will have an imbalance of individual types between the treated and the untreated. Exchangeability will be broken because the $A = 1$ group will include type 1 and type 2 individuals, whereas the $A = 0$ group will include type 1 individuals only. Individuals in the $A = 0$ group will have, on average, a worse prognosis than those in the $A = 1$ group, which will make treatment look better than it really is. This selection bias (Chapter 8) arises when treatment has a non-null effect on survival time.

Robins (1997b) described non-deterministic non-rank-preserving structural nested AFT models.

Less computationally intensive approaches, known as directed search methods, for approximate searching are available in statistical software. The Nelder-Mead Simplex method is an example of a directed search method.

| | Type | | |
|-----------|------|----|-----|
| | 1 | 2 | 3 |
| $T^{a=0}$ | 36 | 72 | 108 |
| $T^{a=1}$ | 24 | 48 | 72 |

Table 17.1

 Technical Point 17.3

Artificial censoring Let $K(\psi)$ be the minimum survival time under no treatment that could possibly correspond to an individual who actually died at time K (the administrative end of follow-up). For a dichotomous treatment A , $K(\psi) = \inf \{K \exp(\psi A)\}$, which implies that $K(\psi) = K \exp(\psi \times 0) = K$ if treatment contracts the survival time (i.e., $\psi > 0$), $K(\psi) = K \exp(\psi \times 1) = K \exp(\psi)$ if treatment expands the survival time (i.e., $\psi < 0$), and $K(\psi) = K \exp(0) = K$ if treatment does not affect survival time (i.e., $\psi = 0$).

All individuals who are administratively censored (i.e., $T > K$) have $\Delta(\psi) = 0$ because there is at least one treatment level (the one they actually received) under which their survival time is greater than K , i.e., $H(\psi) \geq K(\psi)$. Some of the individuals who are not administratively censored (i.e., $T \leq K$) also have $\Delta(\psi) = 0$ and are excluded from the analysis—they are artificially censored—to avoid selection bias.

The artificial censoring indicator $\Delta(\psi)$ is a function of $H(\psi)$ and K . Under conditional exchangeability given L , all such functions, when evaluated at the true value of ψ , are conditionally independent of treatment A given the covariates L . That is, g-estimation of the AFT model parameters can be performed based on $\Delta(\psi)$ rather than $H(\psi)$. Technically, $\Delta(\psi)$ is substituted for $H(\psi)$ in the estimating equation of Technical Point 14.2. For practical estimation details, see the Appendix of Hernán et al (2005).

This exclusion of uncensored individuals from the analysis is often referred to as *artificial censoring*. See Technical Point 17.3.

CODE: Program 17.5

The point estimate of ψ is the value that corresponds to the minimum of the estimating function described in Technical Point 17.3.; the limits of the 95% confidence interval are the values that correspond to the value 3.84 (χ^2 with one degree of freedom) of the estimating function.

To avoid this selection bias, one needs to select individuals whose survival time would have been observed by the end of follow-up whether they had been treated or untreated, i.e., those with $T_i^{a=0} \leq K$ and $T_i^{a=1} \leq K$. In our example, we will have to exclude all type 2 individuals from the analysis in order to preserve exchangeability. That is, we will not only exclude administratively censored individuals with $T_i > K$, but also some uncensored individuals with known survival time $T_i \leq K$ because their survival time would have been greater than K if they had received a treatment level different from the one they actually received.

We then define an indicator $\Delta(\psi)$, which takes value 0 when an individual is excluded and 1 when she is not. The g-estimation procedure is then modified by replacing the variable $H(\psi^\dagger)$ by the indicator $\Delta(\psi^\dagger)$. See Technical Point 17.3 for details. In our example, the g-estimate $\hat{\psi}$ from the rank-preserving structural AFT model $T_i^{a=0} = T_i \exp(\psi A_i)$ was -0.047 (95% confidence interval: -0.223 to 0.333). The number $\exp(-\hat{\psi}) = 1.05$ can be interpreted as the median survival time that would have been observed if all individuals in the study had received $a = 1$ divided by the median survival time that would have been observed if all individuals in the study had received $a = 0$. This survival time ratio suggests little effect of smoking cessation A on the time to death.

As we said in Chapter 14, structural nested models, including AFT models, have rarely been used in practice. A practical obstacle for the implementation of the method is the lack of user-friendly software. An even more serious obstacle in the survival analysis setting is that the parameters of structural AFT models need to be estimated through search algorithms that are not guaranteed to find a unique solution. This problem is exacerbated for models with two or more parameters ψ . As a result, the few published applications of this method tend to use simplistic AFT models that do not allow for the treatment effect to vary across covariate values.

This state of affairs is unfortunate because subject-matter knowledge (e.g., biological mechanisms) is easier to translate into parameters of structural AFT models than into those of structural hazards models. This is especially true when using non-deterministic and non-rank preserving structural AFT models.