

Chapter 17: Causal Survival Analysis

Contents

1	17.1 Hazards and Risks	2
1.1	Survival, Risk, and Hazard	2
1.2	Risk vs. Hazard	3
2	17.2 From Hazards to Risks	4
2.1	The Event Indicator D_k	4
2.2	From Hazards to Survival	4
2.3	Nonparametric and Parametric Estimation	5
3	17.3 Why Censoring Matters	6
3.1	The Censoring Indicator	6
3.2	Why Naive Estimation Fails	7
3.3	Correct Estimation Under As-If Randomly Assigned Censoring	7
4	17.4 IP Weighting of Marginal Structural Models	7
4.1	Counterfactual Survivals	7
4.2	Estimation via IP Weighting	8
5	17.5 The Parametric G-Formula	8
5.1	Identification via Standardization	8
5.2	Estimation	9
6	17.6 G-Estimation of Structural Nested Models	9
6.1	Structural Nested AFT Models	9
6.2	Selection Bias from Administrative Censoring	10
7	Summary	11
8	References	12

In previous chapters we have been concerned with causal questions about treatment effects on outcomes occurring at a particular time point. For example, we 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.

This chapter is based on Hernán and Robins (2020, chap. 17).

Key challenge: Survival outcomes involve time, and individuals may be censored before experiencing the event. The estimation of hazards is often a useful intermediate step for the estimation of survivals

and risks. This chapter shows how IP weighting and the parametric g-formula extend naturally to survival outcomes.

1 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.

Definition 1.1 (Administrative End of Follow-Up and Administrative Censoring). Most follow-up studies have a date after which there is no information on any individuals: the **administrative end of follow-up**. Individuals who do not develop the event of interest before the administrative end of follow-up have their survival time **administratively censored**.

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. The administrative censoring time is the same—120 months—for all individuals. Of the 1629 individuals, only 318 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. 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.

In addition to administrative censoring, survival analyses may also need to handle non-administrative types of censoring, such as loss to follow-up and competing events (see Fine Point 17.1). In previous chapters we 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. In this chapter, we focus on administrative censoring and defer non-administrative censoring to Part III.

1.1 Survival, Risk, and Hazard

The **survival probability** $\Pr[T > k]$, or simply the survival at month k , is the proportion of individuals who survived through time k . If we calculate the survivals at each month until the administrative end of follow-up $k_{\text{end}} = 120$ 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 with subsequent values of $k = 1, 2, \dots, k_{\text{end}}$.

Alternatively, we can define the **risk**, or **cumulative incidence**, at time k as one minus the survival:

$$\Pr[T \leq k] = 1 - \Pr[T > k]$$

The cumulative incidence curve starts at $\Pr[T \leq 0] = 0$ and increases monotonically during the follow-up.

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**:

$$\Pr[T = k \mid T > k - 1]$$

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, we will refer to the discrete-time hazard as, simply, the hazard.

In our example, if we assume (provisionally—until Section 17.4—that quitters ($A = 1$) and non-quitters ($A = 0$) are marginally exchangeable, the unadjusted survival curves at 120 months were 76.2% among quitters and 82.0% among non-quitters. Equivalently, the 120-month risk was 23.8% among quitters and 18.0% among non-quitters. These figures are unadjusted for confounding and thus do not have a causal interpretation; see Section 17.4 for the adjusted estimates.

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.

1.2 Risk vs. 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. The risk will stay flat or increase as k increases.
- The denominator of the **hazard**—the number of individuals alive at k —varies over time and its numerator includes only recent events. The hazard may increase or decrease over time.

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.

Definition 1.2 (Competing Event). 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: individuals who die from other causes (say, cancer) cannot ever develop stroke.

i Fine Point 17.1: Competing Events

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. In addition, the censoring may introduce selection bias under the null, which would require adjustment 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 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.

None of the above strategies provides a satisfactory solution to the problem of competing events. For a detailed description of approaches to handle competing events and their challenges, see

2 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 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.

Definition 2.1 (Person-Time Data Format). 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 1 at $k = 1$, and so on until the follow-up of person 1 ends. 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.

2.1 The Event Indicator D_k

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 $T > k$.

Using the time-varying outcome variable D_k , we can define:

Definition 2.2 (Survival, Risk, and Hazard).

- **Survival** at k : $\Pr[D_k = 0] = \Pr[T > k]$
- **Risk** at k : $\Pr[D_k = 1] = \Pr[T \leq k]$
- **Hazard** at k : $\Pr[D_k = 1 \mid D_{k-1} = 0]$

For $k = 1$ the hazard is equal to the risk because everybody is, by definition, alive at $k = 0$.

2.2 From Hazards to Survival

The survival probability at k is the product of the conditional probabilities of having survived each interval between 0 and k .

Theorem 2.1 (Survival as a Product of Hazards). *The survival at k is:*

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

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).

Example: The survival at $k = 2$, $\Pr[D_2 = 0]$, is equal to the survival probability at $k = 1$, $\Pr[D_1 = 0]$, times the survival probability at $k = 2$ conditional on having survived through $k = 1$, $\Pr[D_2 = 0 \mid D_1 = 0]$.

2.3 Nonparametric and Parametric Estimation

The hazard at k , $\Pr[D_k = 1 \mid D_{k-1} = 0]$, 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 $k - 1$.

Definition 2.3 (Kaplan-Meier (Product-Limit) Estimator). The resulting nonparametric estimate of the survival $\Pr[D_k = 0]$ at k is referred to as the **Kaplan-Meier**, or **product-limit**, estimator.

An easy way to parametrically estimate the hazards is to fit a **logistic regression model** for $\Pr[D_{k+1} = 1 \mid D_k = 0]$ restricted to individuals who survived through k . In our example, we can estimate the hazards in the treated and untreated by fitting:

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

where $\theta_{0,k} = \theta_0 + \theta_4 k + \theta_5 k^2$ is a flexible time-varying intercept. The product terms between treatment A and time allow the hazard ratio to vary over time (see Technical Point 17.1 for details on how a logistic model approximates a hazards model).

We then compute estimates of the survival $\Pr[D_{k+1} = 0 \mid A = a]$ by multiplying the estimates of one minus the hazard from the logistic model, separately for the treated and the untreated.

i 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. 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. The hazard ratio at time 2 is $\Pr[D_2 = 1 \mid D_1 = 0, A]$. 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.

Definition 2.4 (Nonparametric, Parametric, and Semiparametric Models for Survival Analysis). **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.

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.

i 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.

i Technical Point 17.1: Approximating the Hazard Ratio via a Logistic Model

The discrete-time hazard ratio $\frac{\Pr[D_{k+1}=1 | D_k=0, A=1]}{\Pr[D_{k+1}=1 | D_k=0, A=0]}$ equals $\exp\{(\alpha_1)\}$ at all times $k+1$ in the hazards model $\Pr[D_{k+1} = 1 | D_k = 0, A] = \Pr[D_{k+1} = 1 | D_k = 0, A = 0] \times \exp\{(\alpha_1 A)\}$. Suppose the hazard at $k+1$ is small, i.e., $\Pr[D_{k+1} = 1 | D_k = 0, A] \approx 0$. Then the hazard is approximately equal to the odds, and we have:

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

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 the logistic model. As a rule of thumb, the approximation is often considered 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 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.

3 17.3 Why Censoring Matters

The only source of censoring in our study is a common administrative censoring time $k_{\text{end}} = 120$ that is identical for all individuals. In this simple setting the procedure described in the previous section to estimate the survival is straightforward.

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.

3.1 The Censoring Indicator

In this setting it is helpful to define a time-varying **indicator for censoring** C_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.

Our goal is to estimate the survival curve that would have been observed if nobody had been censored before k_{end} . 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. We can also refer to this quantity as $\Pr[D_k^{c=0} = 0 | A = a]$, where the superscript $c = 0$ makes explicit that we want the survival under no censoring.

3.2 Why Naive Estimation Fails

We **cannot** validly estimate $\Pr[D_k = 0 \mid A = a]$ 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 $\Pr[C_{k+1} = 0, D_{k+1} = 0 \mid A = a]$, which is not what we want.

To see why, consider a study with $k_{\text{end}} = 2$ in which:

- $\Pr[C_1 = 0] = 1$: nobody is censored by $k = 1$
- $\Pr[D_1 = 0 \mid C_0 = 0] = 0.9$: 90% of individuals survive through $k = 1$
- $\Pr[C_2 = 0 \mid D_1 = 0, C_1 = 0] = 0.5$: a random half of survivors is censored by $k = 2$
- $\Pr[D_2 = 0 \mid C_2 = 0, D_1 = 0, C_1 = 0] = 0.9$: 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, the survival would have been $1 \times 0.9 \times 1 \times 0.9 = 0.81$.

3.3 Correct Estimation Under As-If Randomly Assigned Censoring

Theorem 3.1 (Identification of Survival Under As-If Randomly Assigned Censoring). *Under (as-if) randomly assigned censoring, the survival $\Pr[D_k = 0 \mid A = a]$ at k is:*

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

*This requires estimating the **cause-specific hazard** $\Pr[D_{k+1} = 1 \mid D_k = 0, C_{k+1} = 0, A = a]$.*

The estimation procedure is the same as described above except that we estimate the cause-specific hazard.

Often we are not ready to assume that censoring is as if randomly assigned. When there is staggered entry, the calendar time at study entry may be associated with the outcome. Therefore, the above procedure will need to be adjusted for baseline confounders, which can be achieved through g-methods as described in the next sections.

4 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.

4.1 Counterfactual Survivals

Definition 4.1 (Counterfactual Survival). Let D_k^a denote a counterfactual time-varying indicator for death at k under treatment level a and no censoring. The counterfactual survival at k under treatment level a is $\Pr[D_k^a = 0]$.

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. The causal contrast of interest is:

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

Because of confounding, this contrast may not be validly estimated by the contrast of the survivals $\Pr[D_{k+1} = 0 \mid A = 1]$ and $\Pr[D_{k+1} = 0 \mid A = 0]$.

4.2 Estimation via IP Weighting

Let us assume that the treated ($A = 1$) and the untreated ($A = 0$) are exchangeable within levels of the L variables. 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.

Step 1: We estimate the stabilized IP weight SW^A for each individual in the study population, using exactly the same procedure as described in Chapter 12. The estimated weight SW^A has mean 1 (as expected); in our example the weights ranged from 0.33 to 4.21.

Step 2: Using the person-time data format, we fit a weighted hazards model:

$$\text{logit } \Pr[D_{k+1}^a = 1 \mid D_k^a = 0] = \beta_{0,k} + \beta_1 a + \beta_2(a \times k) + \beta_3(a \times k^2)$$

where individuals are weighted by their estimated SW^A . This IP weighted model estimates the parameters of the **marginal structural logistic model**, i.e., the time-varying hazards that would have been observed if all individuals in the study population had been treated ($a = 1$) and if they had been untreated ($a = 0$).

The estimates of $\Pr[D_{k+1}^a = 1 \mid D_k^a = 0]$ (the counterfactual hazards) from the IP weighted hazards model can then be converted to estimates of the survival $\Pr[D_{k+1}^a = 0] = \prod_{m=0}^k (1 - \Pr[D_{m+1}^a = 1 \mid D_m^a = 0])$ under treatment $a = 1$ and under no treatment $a = 0$.

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 correct specification of both the treatment model and the marginal hazards model.

5 17.5 The Parametric G-Formula

In the previous section we estimated the survival curve under treatment and under no treatment via IP weighting. Another method to estimate the marginal survival curves under exchangeability, positivity, and consistency is standardization based on parametric models, i.e., the **parametric g-formula**.

5.1 Identification via Standardization

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.

Theorem 5.1 (Identification of Survival Under Standardization). *Under exchangeability, positivity, and consistency, $\Pr[D_{k+1}^a = 0]$ equals the standardized survival:*

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

5.2 Estimation

The estimation of the parametric g-formula has two steps.

Step 1: Fit a parametric hazards model including the variables in L as covariates:

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

If the model is correctly specified, it validly estimates the time-varying hazards $\Pr[D_{k+1} = 1 \mid D_k = 0, L, A]$ within levels of treatment A and covariates L . The product of one minus the conditional hazards:

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

is equal to the conditional survival $\Pr[D_{k+1} = 0 \mid L = l, A = a]$.

Step 2: Compute the weighted average of the conditional survival across all values l of the covariates L —that is, standardize the survival to the confounder distribution—using the method described in Section 13.3.

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.

6 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. Therefore we cannot estimate survivals or hazards using a structural nested model.

6.1 Structural Nested AFT Models

We can, however, model the ratio of survival times under different treatment levels.

Definition 6.1 (Structural Nested Accelerated Failure Time (AFT) Model). Let T_i^a be the counterfactual time of survival for individual i under treatment level a . Suppose the effect of treatment is the same for every individual. We can then consider the **structural nested accelerated failure time (AFT) model**:

$$T_i^a / T_i^{a=0} = \exp\{(\cdot) - \psi_1 a\}$$

where ψ_1 measures the expansion (or contraction) of each individual's survival time attributable to treatment.

- $\psi_1 < 0$: treatment increases survival time
- $\psi_1 > 0$: treatment decreases survival time
- $\psi_1 = 0$: 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 $T_i^a / T_i^{a=0} = \exp\{(\psi_1 a + \psi_2 a L_i)\}$. Rearranging, the model can be written as:

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

6.2 Selection Bias from Administrative Censoring

Consider a 60-month randomized experiment. Table 1 shows the survival time under treatment and under no treatment for three types of individuals:

Table 1: Survival times under treatment and no treatment for each individual type in a 60-month experiment.

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

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 $A = 1$ (see Table 1).

Hence an analysis that is restricted to 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 arises when treatment has a non-null effect on survival time.

The g-estimation procedure for the AFT model must account for administrative censoring to avoid this selection bias. The structural AFT model as written 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. It is **rank-preserving** because if individual i would die before individual j had they both been untreated, then individual i would also die before individual j had they both been treated. Because of the implausibility of rank preservation, one should not generally use methods for causal inference that rely on it.

Definition 6.2 (Blipped-Down and Artificially Censored Survival Times). For a given candidate value ψ , the **blipped-down** survival time for treated individuals ($A_i = 1$) is:

$$U_i(\psi) = T_i \exp\{(\psi A_i)\}$$

This quantity equals $T_i^{a=0}$ if ψ equals the true ψ_1 . Under no unmeasured confounding, $U_i(\psi_1)$ is independent of A_i within levels of L_i .

To account for administrative censoring at time K , the **artificially censored** version of $U_i(\psi)$ is:

$$U_i^*(\psi) = \min(U_i(\psi), K \exp\{(\psi - \psi_1 A_i)\})$$

with corresponding artificial censoring indicator $\Delta_i^*(\psi) = \mathbf{1}[U_i(\psi) \leq K \exp\{(\cdot) - \psi A_i\}]$.

i Technical Point 17.3: G-Estimation of Structural Nested AFT Models

The key challenge in g-estimation of structural nested AFT models is that administrative censoring creates selection bias when treatment affects the censoring indicator (see the main text above). The solution is to create “artificially censored” versions of the observed survival times.

The g-estimator solves the estimating equation

$$\sum_i (A_i - \hat{E}[A_i | L_i]) \cdot h(U_i^*(\psi), \Delta_i^*(\psi)) = 0$$

where $h(\cdot)$ is a score function (e.g., from a rank-based or weighted log-rank statistic) and $\hat{E}[A_i | L_i]$ is the estimated propensity score. By searching over ψ , we find the value that makes treated and untreated individuals comparable in terms of their blipped-down survival times.

7 Summary

Key concepts from this chapter:

1. **Survival analysis:** Methods for time-to-event outcomes that accommodate administrative censoring
2. **Survival, risk, and hazard:** Related but distinct measures; $\Pr[D_k = 0] = \prod_{m=1}^k \Pr[D_m = 0 | D_{m-1} = 0]$
3. **Person-time data format:** Enables parametric modeling of time-varying hazards via logistic regression
4. **Why censoring matters:** Staggered entry creates individual-specific censoring times; naive estimation is biased
5. **Counterfactual survivals:** The target is $\Pr[D_{k+1}^a = 0]$ under treatment a and no censoring
6. **IP weighting of marginal structural models:** Estimates causal survival curves by weighting by SW^A
7. **Parametric g-formula:** Standardizes conditional survivals to the covariate distribution
8. **G-estimation of structural nested AFT models:** Models the ratio of survival times; requires artificial censoring to handle administrative censoring

Causal estimands for survival outcomes:

- **Causal survival difference:** $\Pr[D_k^{a=1} = 0] - \Pr[D_k^{a=0} = 0]$
- **Causal risk difference:** $\Pr[D_k^{a=1} = 1] - \Pr[D_k^{a=0} = 1]$

NHEFS smoking cessation example (120-month results):

- Unadjusted: 76.2% survival among quitters vs. 82.0% among non-quitters (confounded)
- IP weighting: 80.7% under cessation vs. 80.5% under continued smoking; difference 0.2%
- Parametric g-formula: 80.4% under cessation vs. 80.6% under continued smoking; difference -0.2%

Both adjusted methods found little evidence of an effect of smoking cessation on mortality.

Why not just use hazard ratios? (Fine Point 17.2):

- Hazard ratios vary over time; a “single” hazard ratio from Cox regression is a weighted average
- Even time-specific hazard ratios have built-in selection bias (conditioning on survival is conditioning on a collider)
- Risk and survival differences/ratios are better-defined causal estimands

Common mistakes in survival analysis:

1. Reporting only a single (time-averaged) hazard ratio from a Cox model and interpreting it causally
2. Treating competing events as independent censoring (see Fine Point 17.1)
3. Ignoring time-varying confounding (addressed in Part III)
4. Restricting the analysis to uncensored individuals when estimating AFT models, which introduces selection bias

Looking ahead: Part III extends these ideas to time-varying treatments, where survival analysis becomes even more complex. The g-formula and IP weighting remain powerful tools, but they must account for the time-varying nature of both treatment and confounders.

8 References

Hernán, Miguel A, and James M Robins. 2020. *Causal Inference: What If*. Chapman & Hall/CRC.
<https://miguelhernan.org/whatifbook>.