1 The problem of survival analysis

Survival analysis concerns analyzing the time to the occurrence of an event. For instance, we have a dataset in which the times are 1, 5, 9, 20, and 22. Perhaps those measurements are made in seconds, perhaps in days, but that does not matter. Perhaps the event is the time until a generator's bearings seize, the time until a cancer patient dies, or the time until a person finds employment, but that does not matter, either.

For now, we will just abstract the underlying data-generating process and say that we have some times—1, 5, 9, 20, and 22—until an event occurs. We might also have some covariates (additional variables) that we wish to use to "explain" these times. So, pretend that we have the following (completely made up) dataset:

```
time      x
   1      3
   5      2
   9      4
  20      9
  22     -4
```

Now what is to keep us from simply analyzing these data using ordinary least-squares (OLS) linear regression? Why not simply fit the model

$$\texttt{time}_j = \beta_0 + \beta_1 x_j + \epsilon_j, \qquad \epsilon_j \sim \mathrm{N}(0, \sigma^2)$$

for $j = 1, \ldots, 5$, or, alternatively,

$$\ln(\texttt{time}_j) = \beta_0 + \beta_1 x_j + \epsilon_j, \qquad \epsilon_j \sim \mathrm{N}(0, \sigma^2)$$

That is easy enough to do in Stata by typing

```
. regress time x
```

or

```
. generate lntime = ln(time)
. regress lntime x
```

These days, researchers would seldom analyze survival times in this manner, but why not? Before you answer too dismissively, we warn you that we can think of instances for which this would be a perfectly reasonable model to use.

1.1 Parametric modeling

The problem with using OLS to analyze survival data lies with the assumed distribution of the residuals, ϵ_j. In linear regression, the residuals are assumed to be distributed normally; that is, time conditional on x_j is assumed to follow a normal distribution:

$$\texttt{time}_j \sim \mathrm{N}(\beta_0 + \beta_1 x_j, \sigma^2), \qquad j = 1, \ldots, 5$$

The assumed normality of time to an event is unreasonable for many events. It is unreasonable, for instance, if we are thinking about an event with an instantaneous risk of occurring that is constant over time. Then the distribution of time would follow an exponential distribution. It is also unreasonable if we are analyzing survival times following a particularly serious surgical procedure. Then the distribution might have two modes: many patients die shortly after the surgery, but if they survive, the disease might be expected to return. One other problem is that a time to failure is always positive, while theoretically, the normal distribution is supported on the entire real line. Realistically, however, this fact alone is not enough to render the normal distribution useless in this context, because σ^2 may be chosen (or estimated) to make the probability of a negative failure time virtually zero.

At its core, survival analysis concerns nothing more than making a substitution for the normality assumption characterized by OLS with something more appropriate for the problem at hand.

Perhaps, if you were already familiar with survival analysis, when we asked, "why not linear regression?" you offered the excuse of right-censoring—that in real data we often do not observe subjects long enough for all of them to fail. In our data there was no censoring, but in reality, censoring is just a nuisance. We can fix linear regression easily enough to deal with right-censoring. It goes under the name censored-normal regression, and Stata's `intreg` command can fit such models; see [R] **intreg**. The real problem with linear regression in survival applications is with the assumed normality.

Being unfamiliar with survival analysis, you might be tempted to use linear regression in the face of nonnormality. Linear regression is known, after all, to be remarkably robust to deviations from normality, so why not just use it anyway? The problem is that the distributions for time to an event might be dissimilar from the normal—they are almost certainly nonsymmetric, they might be bimodal, and linear regression is not robust to these violations.

Substituting a more reasonable distributional assumption for ϵ_j leads to parametric survival analysis.

1.2 Semiparametric modeling

That results of analyses are being determined by the assumptions and not the data is always a source of concern, and this leads to a search for methods that do not require assumptions about the distribution of failure times. That, at first blush, seems hopeless. With survival data, the key insight into removing the distributional assumption is that, because events occur at given times, these events may be ordered and the analysis may be performed exclusively using the ordering of the survival times. Consider our dataset:

```
time      x
   1      3
   5      2
   9      4
  20      9
  22     -4
```

Examine the failure that occurred at time 1. Let's ask the following question: what is the probability of failure after exposure to the risk of failure for 1 unit of time? At this point, observation 1 has failed and the others have not. This reduces the problem to a problem of binary-outcome analysis,

```
time      x      outcome
   1      3            1
   5      2            0
   9      4            0
  20      9            0
  22     -4            0
```

and it would be perfectly reasonable for us to analyze failure at `time` = 1 using, say, logistic regression

$$\begin{aligned} &= \Pr(\text{failure after exposure for 1 unit of time}) \\ &= \Pr(\texttt{outcome}_j = 1) \\ &= \frac{1}{1+\exp(-\beta_0 - x_j\beta_x)} \end{aligned}$$

for $j = 1, \ldots, 5$. This is easy enough to do:

```
. logistic outcome x
```

Do not make too much of our choice of logistic regression—choose the analysis method you like. Use probit. Make a table. Whatever technique you choose, you could do all your survival analysis using this analyze-the-first-failure method. To do so would be inefficient but would have the advantage that you would be making no assumptions about the distribution of failure times. Of course, you would have to give up on being able to make predictions conditional on x, but perhaps being able to predict whether failure occurs at `time` = 1 would be sufficient.

There is nothing magical about the first death time; we could instead choose to analyze the second death time, which here is `time` = 5. We could ask about the probability of failure, given exposure of 5 units of time, in which case we would exclude the first observation (which failed too early) and fit our logistic regression model using the second and subsequent observations:

```
. drop outcome
. generate outcome = cond(time==5,1,0) if time>=5
. logistic outcome x if time>=5
```

In fact, we could use this same procedure on each of the death times, separately.

Which analysis should we use? Well, the second analysis has slightly less information than the first (because we have one less observation), and the third has less than the first two (for the same reason), and so on. So we should choose the first analysis. It is, however, unfortunate that we have to choose at all. Could we somehow combine all these analyses and constrain the appropriate regression coefficients (say, the coefficient on x) to be the same? Yes, we could, and after some math, that leads to semiparametric survival analysis and, in particular, to Cox (1972) regression if a conditional logistic model is fit for each analysis. Conditional logistic models differ from ordinary logistic models for this example in that for the former we condition on the fact that we know that `outcome==1` for only one observation within each separate analysis.

However, for now we do not want to get lost in all the mathematical detail. We could have done each of the analyses using whatever binary analysis method seemed appropriate. By doing so, we could combine them all if we were sufficiently clever in doing the math, and because each of the separate analyses made no assumption about the distribution of failure times, the combined analysis also makes no such assumption.

That last statement is rather slippery, so it does not hurt to verify its truth. We have been considering the data

```
time    x
   1    3
   5    2
   9    4
  20    9
  22   -4
```

but now consider two variations on the data:

```
time     x
   1.1     3
   1.2     2
   1.3     4
  50.0     9
  50.1    -4
```

and

```
time       x
      1      3
    500      2
   1000      4
  10000      9
 100000     -4
```

These two alternatives have dramatically different distributions for time, yet they have the same temporal ordering and the same values of x. Think about performing the individual analyses on each of these datasets, and you will realize that the results you get will be the same. Time plays no role other than ordering the observations.

The methods described above go under the name semiparametric analysis; as far as time is concerned, they are nonparametric, but because we are still parameterizing the effect of x, there exists a parametric component to the analysis.

1.3 Nonparametric analysis

Semiparametric models are parametric in the sense that the effect of the covariates is still assumed to take a certain form. Earlier, by performing a separate analysis at each failure time and concerning ourselves only with the order in which the failures occurred, we made no assumption about the distribution of time to failure. We did, however, make an assumption about how each subject's observed x value determined the probability (for example, a probability determined by the logistic function) that a subject would fail.

An entirely nonparametric approach would be to do away with this assumption also and to follow the philosophy of letting the dataset speak for itself. There exists a vast body of literature on performing nonparametric regression using methods such as lowess or local polynomial regression; however, such methods do not adequately deal with censoring and other issues unique to survival data.

When no covariates exist, or when the covariates are qualitative in nature (gender, for instance), we can use nonparametric methods such as Kaplan and Meier (1958) or the method of Nelson (1972) and Aalen (1978) to estimate the probability of survival past a certain time or to compare the survival experiences for each gender. These methods account for censoring and other characteristics of survival data. There also exist methods such as the two-sample log-rank test, which can compare the survival experience across gender by using only the temporal ordering of the failure times. Nonparametric methods make assumptions about neither the distribution of the failure times nor how covariates serve to shift or otherwise change the survival experience.

1.4 Linking the three approaches

Going back to our original data, consider the individual analyses we performed to obtain the semiparametric (combined) results. The individual analyses were

Pr(failure after exposure for exactly 1 unit of time)
Pr(failure after exposure for exactly 5 units of time)
Pr(failure after exposure for exactly 9 units of time)
Pr(failure after exposure for exactly 20 units of time)
Pr(failure after exposure for exactly 22 units of time)

We could omit any of the individual analyses above, and doing so would affect only the efficiency of our estimators. It is better, though, to include them all, so why not add the following to this list:

Pr(failure after exposure for exactly 1.1 units of time)
Pr(failure after exposure for exactly 1.2 units of time)
...

That is, why not add individual analyses for all other times between the observed failure times? That would be a good idea because the more analyses we can combine, the more efficient our final results will be: the standard errors of our estimated regression parameters will be smaller. We do not do this only because we do not know how to say anything about these intervening times—how to perform these analyses—unless we make an assumption about the distribution of failure time. If we made that assumption, we could perform the intervening analyses (the infinite number of them), and then we could combine them all to get superefficient estimates. We could perform the individual analyses themselves a little differently, too, by taking into account the distributional assumptions, but that would only make our final analysis even more efficient.

That is the link between semiparametric and parametric analysis. Semiparametric analysis is simply a combination of separate binary-outcome analyses, one per failure time, while parametric analysis is a combination of several analyses at *all* possible failure times. In parametric analysis, if no failures occur over a particular interval, that is informative. In semiparametric analysis, such periods are not informative. On the one hand, semiparametric analysis is advantageous in that it does not concern itself with the intervening analyses, yet parametric analysis will be more efficient if the proper distributional assumptions are made concerning those times when no failures are observed.

When no covariates are present, we hope that semiparametric methods such as Cox regression will produce estimates of relevant quantities (such as the probability of survival past a certain time) that are identical to the nonparametric estimates, and in fact, they do. When the covariates are qualitative, parametric and semiparametric methods should yield more efficient tests and comparisons of the groups determined by the covariates than nonparametric methods, and these tests should agree. Test disagreement would indicate that some of the assumptions made by the parametric or semiparametric models are incorrect.