Survival Analysis in Python Statsmodels

March 9, 2020

1 Introduction

Survival analysis is used to analyze data in which the primary variable of interest is a time duration. For example, the time duration could be a person's lifespan, the time that a person survives after being diagnosed with a serious disease, the time from the diagnosis of a disease until the person recovers, or the duration of time that a piece of machinery remains in good working order.

Duration data can be used to answer questions such as:

- What is the mean duration for a population?
- What is the 75th percentile of the durations in a population?
- When comparing two populations, which one has the shorter expected or median duration?
- What are the unique associations between independent variables and a duration outcome?

Here are some key concepts in duration analysis:

- **Time origin**: The durations of interest correspond to time intervals that begin and end when something specific happens. It is important to be very explicit about what defines the time origin (time zero) from which the duration is calculated. For example, when looking at the survival of people with a disease, the time origin could be the date of diagnosis, but when looking at human lifespans ("all cause mortality")

it might make more sense to define the time origin to be the date of birth.

- **Event**: This refers to whatever must happen to conclude the time interval of interest. It may be death, or some other type of "failure", or it may be something more positive, like recovery or cure from a disease. Most survival analysis is based on the idea that every subject will eventually experience the event, although we may not observe everyone long enough to see the event happen.
- **Survival time distribution**: This is a marginal distribution defining the proportion of the population that has experienced the event on or before time T. Usually it is expressed as the *complementary survival function* (e.g. the proportion of people who have not yet died as of time T).
- **Censoring**: Censoring occurs when we do not observe when a subject experiences the event of interest, but we do have some partial information about that time. The most common form of censoring is *right censoring*, in which we observe a time T such that we know the event did not occur prior to time T. Other forms of censoring are *interval censoring* and *left censoring*.
- **Risk set**: This is the set of units (e.g. people) in a sample at a given time who may possibly experience the event at that time. It is usually the set of people who have not already experienced the event and who have not been censored (but the risk set may be only a subset of these people when using "entry times").
- **Hazard**: This is the probability of experiencing the event in the next time unit, given that it has not already occurred (technically, this is the discrete time definition of the hazard, the continuous time definition involves rates but follows the same logic).

2 Marginal survival function and hazard estimation

If there is no censoring, the marginal survival function can be estimated using the complement of the empirical cumulative distribution function of the data.

If there is censoring, the standard method for estimating the survival function is the *product-limit estimator* also known as the *Kaplan Meier estimator*.

The idea behind the Kaplan-Meier estimate is not difficult to understand. Group the data by the distinct times $t_1 < t_2 < \cdots$ ("times" here can be either event times or censoring times), and let R_t denote the risk set size at time t, and let d_t indicate the number of events at time t (if there are no ties, d_t will always be equal to either 0 or 1). The probability of the event occurring at time t (given that it has not occurred already) is estimated to be d_t/R_t. The probability of the event not occurring at time t is therefore estimated to be $1 - d_t/R_t$. The probability of making it to time t without experiencing the event is therefore estimated to be

$$(1 - d_1/R_1) \cdot (1 - d_2/R_2) \cdot ... \cdot (1 - d_t/R_t).$$

A consequence of this definition is that the estimated survival function obtained using the product limit method is a step function with steps at the event times.

A closely related calculation called the Nelson-Aalen estimator estimates the marginal hazard function.

3 Regression analysis for duration data

Regression analysis is used to understand how multiple factors of interest are related to an "outcome" of interest. If this outcome variable is a duration, we are doing "survival regression".

By direct analogy with linear regression, we might seek to model the expected survival time as a function of covariates. If there is no censoring, we could, for example, use least squares regression to relate the survival time T, or a transformation of it (e.g. $\log(T)$) to a linear function of the covariates. While this is sometimes done, it is more common to approach regression for duration data by modeling the hazard rather than modeling the duration itself.

The hazard is the conditional probability of experiencing the event of interest at time T, given that it has not yet occurred. For example, in a medical study this may be the probability of a subject dying at time T given that

the subject was still alive just before time T (in continuous time we would substitute "rate" for "probability" but we ignore this distinction here).

In survival regression, we view the hazard as a function that is determined by the covariates. For example, the hazard may be determined by age and gender. A very popular form of hazard regression models the conditional hazards (for each covariate specification) as a collection of parallel functions, specifically

$$h(t, x) = b(t) \cdot \exp(c_1 x_1 + \cdots + c_p x_p),$$

where $b(t)$ is the "baseline hazard function", the scalars $c_0, \ldots, c_p$ are unknown regression coefficients, and the x_j are the observed covariates for one subject. This model can also be written in log form

$$\log h(t, x) = \log b(t) + c\prime x$$

where $c = [c_1 \ldots, c_p]$, $x = [x_1, ..., x_p]$. Thus, the log hazard is modeled as a time varying intercept plus a linear predictor that is not time varying (there are generalizations in which the linear predictor is also time varying).

This regression model is called proportional hazards regression or the "Cox model". A key feature of this model is that it is possible to estimate the coefficients c_j using a partial likelihood that does not involve the baseline hazard function. This makes the procedure "semi-parametric".

The key point to remember about interpreting this model is that a coefficient c_j, for a covariate, say age, has the property that the hazard (e.g. of dying) changes multiplicatively by a factor of $\exp(c_j)$ for each unit increase in the covariate's value.

Note that an intercept (either implicit or explicit) is never included in any model of this type, since it can be absorbed into the baseline hazard function $b(t)$.

4 More advanced topics in proportional hazards regression

1. Proportional hazards regression models allow the data to be stratified. Stratifiation is a partitioning of the data into groups. In a hazard regression, each stratum has its own baseline hazard function, i.e. $b(t)$ above is replaced with $b_{s(i)}(t)$, where $s(i)$ is the stratum for observation i. The consequence of doing this is that the coefficients c_j are only based on comparisons among people in the same group. This means that the results are unaffected by confounding factors that are stable within groups. It is common to use stratification as a proxy for confounders that are difficult to measure. For example, in social research the geographic location of a person's residence may be used to define strata.

2. In many settings, we do not observe every subject from their time origin. If we begin monitoring a subject at a time t, then they could not have had the event before that time. Thus, the subject should be removed from the risk set for all times prior to t. This can be accomplished by specifying t as an *entry time*.

3. Survival regression can use weights to project results from a sample to a population that differs from the population the data were sampled from.

5 More on censoring

A large part of survival analysis is concerned with appropriately handling censoring (if there is no censoring, it is generally possible to analyze the log durations using standard methods that are not specialized for survival analysis). Censoring can be a subtle topic. All survival methods have limitations on the type of censoring they can handle, and it is not always easy or even possible to determine in a given setting whether a survival method can accommodate the type of censoring that is present.

To make things more concrete, we usually imagine that every subject has both an event time T and a censoring time C. That is, every subject would eventually experience the event (if there were no censoring), and would even-

tually be censored (if the event did not happen). We observe $\min(T, C)$, and an indicator of whether the event occured (i.e. that $\min(T, C)$ is equal to T). This is sometimes said to be a *counterfactual* way of looking at the analysis, since only one of T or C is observable in the real world.

The key requirement for most surival methods is that we have "independent censoring", meaning that T and C are statistically independent quantities. Since we never observe both T and C for the same person, it is usually not possible to directly assess whether T and C are dependent. Knowledge about the data collection process can sometimes be used to assess whether it is plausible that independent censoring holds.

For example, one type of censoring that is quite common is "administrative censoring". This occurs when a study has a fixed data collection window, say a thee year interval from January 1, 2012 to January 1, 2015. Suppose that people are randomly recruited into the study, and if the event has not occured by January 1, 2015, the subject is considered to be censored. Subjects who are recruited into the study later are more likely to be censored. As long as the reruitment date is not dependent with the true survival time, T and C are independent. We can imagine a setting when administrative censoring may induce dependence, e.g. if the subjects recruited later in the study were healthier than those recruited earlier. But in many situations, this can be excluded as a likely circumstance based on knowledge of how the study was conducted.

On the other hand, in some cases there is strong reason to believe that subjects are more likely to be censored as they grow sicker, which may mean that T and C are positively dependent. For example, if we have medical study in which the data come from insurance records, as people get sicker they are more likely to become unable to work, and may have to quit their job (leading to them being censored). Similarly, as people age they may retire or become eligible for Medicare, leading to age-dependent censoring. Since age is likely correlated with survival time, this could also induce dependent censoring.

In survival regression, we usually only need T and C to be independent given the covariates. Thus, one strategy for dependent censoring is to identify covariates such that T and C become independent after conditioning on the covariates. For example, age or a measure of overall health may be sufficient to substantially reduce the dependence between T and C.