Introduction to Survival Analysis

Survival analysis models factors that influence the time to an event. Ordinary least squares regression methods fall short because the time to event is typically not normally distributed, and the model cannot handle censoring, very common in survival data, without modification.

What is Survival Analysis?

- Survival analysis is generally defined as a set of methods for analyzing data where the outcome variable is the time until the occurrence of an event of interest. The event can be death, occurrence of a disease, marriage, divorce, etc.
- Survival analysis is long established within actuarial science and medical research but seems infrequently used in general data science projects.
- Time-to-event analyses are very widely applicable to all sorts of real-world behaviours - not just studies of lifespan in actuarial or medical science.

Survival Analysis

In this series of blogposts:

- We'll introduce survival analysis as a vital technique in any statisticians toolkit
- We'll demonstrate a general approach to undertaking a data science project: accessing, cleaning and storing data, and using a range of open-source analytical tools that enable rapid iteration, data exploration, modelling and visualisation
- We'll use a real-world dataset and seek to both match and improve upon some existing analysis already undertaken and released by a third-party.

Survival Analysis

By the end you should have a better understanding of the theory, some tools and techniques, and hopefully gain some ideas about how survival analysis can be applied to all manner of event-based processes that are often crucial to business operations.
So firstly...

Survival Analysis

What is Survival Analysis? Wikipedia defines survival analysis as: a branch of statistics that deals with analysis of time duration until one or more events happen, such as death in biological organisms and failure in mechanical systems.

Survival Analysis

Actuarial Application

- We might, for example, expect an actuary to try to predict what proportion of the general population will survive past a certain age.
- The actuary might want to furthermore know the rate of change of survival with time (the hazard function), and the characteristics of individuals which most influence their survival.

Survival Analysis

More generally, we can use survival analysis to model the expected time-to-event for a wide variety of situations:

- **User shopping behaviour** e.g. the elapsed time between a user subscribing to an online shopping service and ordering their first product
- **Crop yields and harvesting** e.g. the duration between seeding a field and the majority of the crops being ready for harvest
- **Radioactive halflife** e.g. the time until half the atoms in a luminous blob2 of tritium have decayed into hydrogen and helium-3
- **Hardware failure rates** e.g. the expected lifetime for a piece of machinery before component failure.
- **Customer subscription persistence** e.g. the expected time for a customer to remain subscribed to a cellphone service before churning

Survival Analysis

Illustrating the basics

- Imagine we have a fleet of haulage trucks and we're particularly interested in the elapsed time between purchase and first maintenance event (aka repairs aka servicing). We could use this analysis to:
- Identify which manufacturers and models of trucks require the least repair and favour buying those again in future
- Identify contributing factors leading to trucks needing earlier repairs and try to mitigate
- Anticipate likely spikes of activity for fleet repair during the year and ensure funds are available in advance

Survival Analysis

Sketching a survival curve We observe:
The relative duration of our study is measured from time of purchase until the end of the second year (24 months). This is a relative time and may begin at a different time for each truck
The survival function is a measure of how many trucks are serviced at each point in time. It drops quickly and then flattens out; it crosses the 50% line at 10 months, meaning that by 10 months we can expect 50% of all the trucks in the fleet to have had their first service.
About 36% of trucks remain unserviced at the end of the first 24 months; conversely, about 64% will have a service event during their first 24 months

Survival Analysis

Fundamental Measurements: The survival function S(t)
The survival function, $S(t)$, of an individual is the probability that they survive until at least time tt.

$$S(t) = Pr(T > t)$$

where t is a time of interest and T is the time of event.
The survival curve is non-increasing (the event may not reoccur for an individual) and is limited within [0,1]. Note that the event might not happen within our period of study and we call this right-censoring. This happens in the above example where for 36% of trucks, all we know is that their first service happens some time after 24 months.

Survival Analysis

The hazard function $\lambda(t)$ The hazard function $\lambda(t)$ is a related measure, telling us the probability that the event TT occurs in the next instant $t + \delta t$, given that the individual has reached timestep t:

$$\lambda(t) = \lim_{\delta_t \to 0} Pr(t \leq T < t + \delta_t | T > t)\delta_t$$

With some maths we can work back to the Survival function:

$$S(t) = exp(\int \lambda(u)du)$$

The hazard function $\lambda(t)$ is non-parametric, so we can fit a pattern of events that is not necessarily monotonic.

Survival Analysis

Other measurements and considerations

- The cumulative hazard function is an alternative representation of the time-to-event behaviour is the cumulative hazard function $\Lambda(t)$, which is essentially the summing of the hazard function over time, and is used by some models for its greater stability. We can show:

$$\Lambda(t) = \int_0^t \lambda(u)du = -logS(t)$$

- The simple relation of $\Lambda(t)$ to the survival function $S(t)$ is a nice property, exploited in particular by the Cox Proportional Hazards and Aalen Additive models, which we'll demonstrate later.

Survival Analysis

Stating it slightly differently, we can relate the attributes of the individuals to their survival curve:
This powerful approach is known as **Survival Regression**. In the trucks example, we might want to know the relative impact of engine size, hours of service, geographical regions driven etc upon the time from first purchase to first service.

Survival Analysis

The half-life

- We saw the half-life of truck repair illustrated above (the orange arrows).
- Most people are familiar with this measure and it's exactly as it says on the tin:
  select a group of individuals - our fleet of trucks - and measure how long it takes for the event of interest to occur. once the event of interest - the first service repair - occurs for half of the population, that period is aka the half-life. note that we can't state exactly which truck or exactly when, just work on the aggregate values.
- There's nothing particularly special about the half-life, and we might be interested in the time taken for e.g. 25% of the trucks to come in for first service, or 75% or 90% etc.

Censoring

- One important concept in survival analysis is censoring.
- The survival times of some individuals might not be fully observed due to different reasons.
- In life sciences, this might happen when the survival study (e.g., the clinical trial) stops before the full survival times of all individuals can be observed, or a person drops out of a study, or for long-term studies, when the patient is lost to follow up.
- In the industrial context, not all components might have failed before the end of the reliability study.
- In such cases, the individual survives beyond the time of the study, and the exact survival time is unknown. This is called right censoring.

Survival Analysis

- During a survival study either the individual is observed to fail at time T, or the observation on that individual ceases at time c.
- Then the observation is **min(T,c)** and an indicator variable Ic shows if the individual is censored or not.
- The calculations for hazard and survivor functions must be adjusted to account for censoring.
- Statistics and Machine Learning Toolbox functions such as ecdf, ksdensity, coxphfit, mle account for censoring.

Survival Analysis

Censoring

- Censoring is a type of missing data problem common in survival analysis.
- Other popular comparison methods, such as linear regression and t-tests do not properly accommodate for censoring.
- This makes survival analysis attractive for data from randomized clinical studies.

Survival Analysis

Types of Censoring

In an ideal scenario, both the birth and death rates of a patient is known, which means the lifetime is known.

- **Right censoring** occurs when the 'death' is unknown, but it is after some known date. e.g. The 'death' occurs after the end of the study, or there was no follow-up with the patient.
- **Left censoring** occurs when the lifetime is known to be less than a certain duration. e.g. Unknown time of initial infection exposure when first meeting with a patient.

Survival Analysis

Censoring and truncation

- We're measuring time-to-event in the real world and so there's practical constraints on the period of study and how to treat individuals that fall outside that period. Censoring is when the event of interest (repair, first sale, etc) occurs outside the study period, and truncation is due to the study design.
- The following discussion continues our hypothetical truck maintenance study:
- We imagine that our data source is a set of database extracts taken at the start of 2014 for a 3-year period from mid-2010 to mid-2013, and this is all of the data we could extract from the company database

Survival Analysis

It may be that we have very little data about service-intervals longer than 24 months, so despite the study period covering 36 months, when we calculate survival curves we decide to only look at the first 24 months of a trucks life
All trucks remain in daily operation through end-2014, none are sold or scrapped.