# Advanced Topics in Survival Analysis 2

## Transcript

For full resource, see: <https://www.ncrm.ac.uk/resources/online/all/?id=20860>

Hello. I'm Dr Oliver Perra and this is the second part of my resources on advanced topics in survival analysis.

 In this second presentation, I will first talk about how to include predictors of event occurrence when we are running survival analysis applied to events that we are measuring using a continuous time metric, then I'm going to introduce how we can model the effects of predictors on event occurrence and then I will introduce the Cox regression model.

 Before I start, I wanted to summarise some of the issues I covered during the first presentation. So, I said that we can overcome the problems in measuring time into infinitesimally small units by breaking time into small intervals that approach but do not reach zero. In this way we can then estimate functions for different intervals of interest, different time intervals of interest, and here I'm using the lower case letter j to indicate a generic time interval of interest, whereas lower case i will indicate a generic individual in a data set in a study.

 So, with this strategy then we can calculate the survival function of event occurrence, sorry, the survival function that indicates the probability of surviving past interval j, the hazard function that represents a rate of event probability within an interval j and then the cumulative hazard function which accumulates the hazard from start of study to interval j. And I mentioned the Nelson-Aalen method as well as the negative log of the survivor function.

 I will keep using the lung cancer data set that is available with the survival R library to provide an example of how we can include predictors in survival analysis when events are measured, when event occurrence is measured using a continuous time metric. So, the lung cancer data set reports the day of death of a number of patients. So, the event of interest is the death of patients and the time metric is days.

 In the introduction to survival analysis for the National Centre for Research Methods, I had provided examples where predictors were included in analysis of events measured in a discrete time metric. But the principles that we can use are analogous, so if you're interested, you can refer back to those resources I had prepared for the National Centre for Research Methods.

 Here I will consider first a categorical predictor “sex” which then I dummy coded as male. So, male equal one represent males, male equal zero represent females. And as usual with these resources I have prepared, you will also find the scripts I have used to create these graphs and calculate the statistics and you can run them on the examples to replicate what I have done.

 So, we can start by exploring the Kaplan-Meier survival function, the cumulative hazard function using the Nelson-Aalen method and the kernel-smoothed hazard function. I report all these statistics here in those graphs.

 You can also realise the kernel-smoothed hazard function does not cover the full range of event times which is usually the case in different data sets. But here you can see that basically the first thing that probably jumps out of this graph is that many patients have a lower probability of survival across most of the study period. The cumulative hazard function also suggests some acceleration around the first year anniversary, so 365 days, and also we can also note the rise in the right end of the graph of the cumulative hazard functions, but as I mentioned in the previous results, we need to be careful with these because the estimates are usually based on a much smaller risk set so they tend to be also more erratic.

 But given the complexity of these estimates, it might seem difficult to build a statistical model to investigate the association between covariates or predictors and event occurrence.

 A strategy to overcome the problems in modelling this type of data then is to build a statistical model starting from the cumulative hazard functions, and here I reported the formula that estimates the cumulative hazard function based on the negative log of the survivor function. This is also a formula I introduced in the first part of these resources.

 This estimate of the cumulative hazard function capitalises on the relationship between hazard and survival and estimates the cumulative hazard between the start of the study and the interval of interest tj as the negative log. So, the natural logarithm, the negative natural logarithm of the survival function of that interval.

 The problem with these estimates is that they have a semi bounded nature so they are bounded from below by zero, they cannot be less than zero, they cannot be below zero, and so they cannot be negative but there is no bound above. And in order to deal with this problem, we can use the log, so the natural logarithm of the negative logarithm of the survival function. And this may seem quite complicated, but the advantage of taking the log of the log of the negative log is that this is an unbounded function that can range between minus infinite and plus infinite. This is often called the log-log survival function, but you can take other names such as the log negative log survivor function.

 And again, it might seem quite convoluted, the idea to take the log of a log, but here show how the log of the negative log survivor function transforms the distribution. On the left side of the slide, you can see the negative log survivor function, but on the right side of the slide you can see the log-log survivor function and you can see that the negative values in the later, so in the log-log function only indicate cumulative hazard that is less than one. So, cumulative hazard that is less than one translate into negative values of the log-log.

 But the main point here, if you look at the graph on the right, is that, like all log transformations, the result of the transformation changes the magnitude of the difference between values. So, the distance that we observe, the small distances that we observe between values in the graph before we apply the log transformation are expanded in the log transform distribution on the right, whereas the distances between large values that we observed on the left side graph become more compressed in the right side graph. So, when we applied log transformations, basically we are emphasising the distance between values that were previously closer and we are reducing the distances between values that were before larger before the transformation.

 And this is particularly useful because it allows to highlight between group differences in the early stages of the studies which tend to be more precise because of the risks that being larger. And on the other hand, this log transformation diminishes the between group differences in the later values, which tend to be more erratic. So, it does have some advantages to use the log of the log because now we have some estimates or a representation of estimates that are easier to interpret.

 So, now we need to develop a model that can represent a relationship between the log cumulative hazard function or the log-log and the predictors of interest. So, here represented a log-log of estimated for males and females. So, how can we model this relationship between sex of the participants and the log-log estimate?

 Similarly to the discrete time example I presented in the introductory resources to survival analysis for the National Centre for Research Methods, the strategy is to express the log cumulative hazard function as the sum of a baseline function and the weighted linear combination of predictors, in this case, male. So, this is basically the Cox model that is based on a seminal paper written by Cox in 1972, so here represented as the Cox model. Here capital H0 indicates the cumulative hazard function when all predictors in the model are equal to zero. In this example, it will be the cumulative hazard function for females, so when males is zero then we are looking at the cumulative hazard function for females. In the bottom of the slide then I partitioned the model, the Cox model, for females and males, which illustrates, and once we estimate a low cumulative hazard function of males, the log commutative hazard function of, sorry, when we estimate the log cumulative hazard function of females, the log cumulative hazard function of males will be based on that of females, but it will just shift vertically and the size and the direction of this shift, this vertical shift, in the function of males will be determined by the value of the β1 coefficient that I put in bold at the bottom of this slide.

 So, this β1 coefficient represents the vertical shift in the estimates of males compared to the baseline that we have, that we estimate.

 So, this is the starting point, but before we move on, I also wanted to highlight something else. Given the Cox model, we can take the anti-log of the terms of both sides. So, here I represented a formula of the Cox model with the log of the cumulative hazard function, but if we take the anti-log of both terms on both sides of the equal sign, we obtain this other formula, this equation where the Cox model basically is being expressed as in terms of the cumulative hazard function and we can see that the cumulative hazard function estimated by the Cox model is equal to the baseline hazard function, so capital H0 for interval tj multiplied by the ℇ number exponentiated by the coefficient that we had estimated for males.

 So, this should highlight an important characteristic of the model that when we are considering the model in terms of the log cumulative hazard function, we are assuming a linear relationship between the estimated outcome and the predictors. So, the effect of male will be constant and its magnitude is represented by the β coefficient. But when we express the model as a cumulative hazard function, the fact of male is not linear since it involves exponentiating the ℇ number and multiplication. So, the result is that when we consider the hypothetical or estimated cumulative hazard function for males and females postulated by the model, these cumulative hazard functions will not be equidistant, whereas when we are considering the log-log of the cumulative hazard function, the distance between males and females will be constant over time, but you will not be constant when we are considering the hazard functions rather than the commutative hazard functions because, as you can see in the bottom of the slide, then the relationship between the baseline and the estimated coefficient for males is multiplicative, it's not additive.

 But even then I need to highlight that the Cox model assumes a constant effect of the predictor over time. Even if the cumulative hazard function derived from the model is not equidistant. And this is because first of all the model is based and build up really from considering the low cumulative hazard function you see on the top of this slide. So, even when we transform the estimates from log communitive hazard to cumulative hazard, the model has not changed and is still based on assuming a constant effect of the predictor as I showed on the top left of this slide.

 But the other reason to say that the predictor’s effect is constant over time lies in the fact that when we consider the cumulative hazard function, the relative distance in the functions of different groups will be identical over time and I demonstrate this here. If we take the ratio between the cumulative hazard function of males and that of females, the two H0 tj, so the two cumulative, the two baseline cumulative hazard functions in the division cancel each other. So, the relative distance between the cumulative hazard function of males and females will be constant and equal to the exponentiation of the β parameter.

 So, in both cases, the fact of the predictor is constant over time and here also represented two idealised examples of these functions, of these relationships in the two graphs. So, the first one is the log of the cumulative hazard function and you can see the constant distance, the constant vertical distance between the two functions for males and females, whereas in the left graph of the graph on the left side, sorry, on the right side, you can see that the relative distance between the two cumulative hazard functions is constant over time and it's equal to the exponentiation of the β coefficient.

 But another key property of the Cox model is that by capitalising on the mathematical links between the cumulative hazard and the hazard function, the model can be expressed directly in terms of hazard. So, a demonstration of this property is too complex and you can refer to some of the references I've provided, but the key point is that as the fitted log cumulative hazard function of the two groups are separated by a constant absolute amount β1, so are the fitted log hazard functions of the two groups and you can see the two simulated functions on the right panel.

 And as the predicted raw cumulative hazard functions of the two groups are separated by a constant relative amount equal to the exponentiation of β1 coefficient, so are the predicted raw hazard functions of the two groups and again, you can see the, I mean you can see this in the simulated functions on the left panel of this graph.

 So, that means that the Cox model takes an identical form regardless of whether we express the outcome as cumulative hazard or as raw hazard functions.

 And this provides another key advantage of the model, because we can interpret the parameters of the Cox model and the coefficients of the Cox model in terms of how the predictor affects the hazard function and therefore that model gives us a more intuitive and informative metric to interpret the results than the cumulative hazard function.

 So, really the bottom line is that by applying the model to cumulative hazard functions, whether they are log or raw hazard functions, cumulative hazard functions, we can also have some information about the, or at least some estimates about the raw hazard functions.

 So, the Cox model provides this very useful property that we can start by modelling cumulative hazard functions to then say something and get some results that can be applied to hazard functions.

 And here, again to drive home this point, the different equations here represent the Cox model estimated originally. So, the model estimated based on the log of the hazard of the cumulative hazard function, the capital H function represented on the first line. And then I’ve written the model in terms of the estimated cumulative hazard function, but also then I write it in terms of the estimated log hazard function, so the lower case h, and in terms of the estimated hazard function. And the model is basically the same even when we start by estimating it on the log cumulative hazard function, then we get the same coefficients and the same relationship between the different type of estimates that we are obtaining.

 And probably these properties of the model can help highlight three of the key assumptions.

 The first one is that there is a fitted log hazard function for each value of the predictor, whether it's categorical or continuous. And also, when we have more than one predictor, we assume that there are fitted log hazard functions for each combination of the predictor’s values.

 Then the other assumption is that each of these fitted log hazard functions has an identical shape. So, in the example we were using, we assume that the log hazard functions of males and females cancer patients have the same shape. It does not matter what shape this function has incidentally, but we do assume that the different levels of the predictors will be associated with fitted functions that have the same shape.

 And finally, the distance between each of these log hazard function is identical at every possible interval. Like all assumptions, this may be adequate or not, and when they are not adequate or realistic, they can be relaxed and I will mention that in the next presentation. But it's important to keep those three key assumptions in mind.

 So, I will conclude this presentation now with a summary of what I have covered, and remember also that with the resources provided, you will also find R scripts to replicate the examples and the graphs that I have displayed here.

 So, to summarise, the Cox regression model estimates the log of the negative log survivor function, that is the log of the cumulative hazard function in other words, but the mathematical relationships mean that the model can be expressed in terms of estimated cumulative hazard as well as estimated raw hazard. And that is a very important property because then we can use the model to say something that is more meaningful and easier to interpret and it can provide a very intuitive and useful way of looking at the phenomenon of interest.

 But it's also important to remember the key assumptions of this model that is a fitted log hazard function has to be there for each value of the predictor, or in case of different predictors for each, a combination of values of the predictors. The fitted log hazard functions have an identical shape and the distance between the log hazard functions is identical at every possible interval.

 So, before I close, I refer again to the webpage of the National Centre for Research Methods if you are interested in more resources about statistical and different research methods, but you will also find other material I have prepared and particularly I have prepared a set of exercises on the topics that I'm covering with these presentations and also solutions for these exercises.

 So, thank you very much for your attention.

National Centre for Research Methods (NCRM)
Social Sciences
Murray Building (Bldg 58)
University of Southampton
Southampton SO17 1BJ
United Kingdom

**Web** www.ncrm.ac.uk
**Email** info@ncrm.ac.uk
**Tel** +44 23 8059 4539
**Twitter** @NCRMUK