# Advanced Topics in Survival Analysis 3

## 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 third part of my resources on advanced topics in survival analysis.

In this third presentation, I will provide discussion on how the Cox regression model is a semi-parametric model and then we'll introduce a useful statistic, the risk scores, and also I will illustrate methods that can help recover a baseline function from the estimates of the Cox regression model. And then I will mention some ways to relax some of the model assumptions, particularly an example of time varying effects of predictors.

To summarise what I've covered in the previous presentation, I introduced the Cox regression model and I highlighted that the estimates are based on the log of the negative log survivor function, or as the log of the cumulative hazard function, but mathematical properties mean that the model can be expressed in terms of the estimated cumulative hazard and estimated raw hazard which provides basically a way to estimate and say something about the hazard functions starting by modelling the cumulative hazard functions.

But the Cox regression model has some key assumptions, so one is that the fitted log hazard functions are available for each value of the predictors, and if there are more than one predictor, that should be also log hazard functions for each combination of values of the different predictors. And also, the fitted log hazard functions have an identical shape and the distance between log hazard functions is identical at every possible interval.

So, I start here by providing some statistics taken from the lung cancer patient example that is available, sorry, the lung cancer patient data set that is available with the R library survival and you can see there how I first calculated with statistics if you look at the scripts that are available with the resources I have prepared, so you can follow the examples on I've used and replicate them using R.

And as you can see here, the only statistic that we get from the Cox regression model is a coefficient for the predictor sex which I dummy coded as male. So, I will highlight here that the Cox regression model is semi-parametric, but what it means is that the model makes fewer parametric assumptions on the phenomena of interest than fully parametric models. And in fact, a key characteristic of the Cox model, which I mentioned in the previous presentation, is that the model does not make assumptions on the baseline hazard function or the log hazard function. All that the model does is estimating the predictor’s effect on the log baseline function. And in fact, here you can see that the results of the lung cancer data set do not report a baseline. They just report a coefficient for males. There is no estimation of a baseline function and all we obtained from the model is the rate of hazard changes for males compared to females.

This semi-parametric quality or characteristic of the Cox regression model actually makes it more probably very attractive because in many ways we don't have to worry about specifying a shape for the baseline function, we don't have to make assumptions about the baseline function.

But it also has some downsides and the downside is that we do not obtain predicted hazard values. We do obtain hazard ratio, so we can estimate the rate of change in hazard associated with the predictors, but we do not have a direct estimation of the baseline hazard function itself.

This may not be such a huge problem in reality because, as I illustrated in the first of these presentations, it is very difficult to properly estimate the hazard function when we are measuring time continuously. When we are measuring the occurrence of events using a continuous time metric, the hazard function tends to be very erratic, so we do not easily can represent those as our functions. But we might still be interested in at least estimating the baseline cumulative hazard function to make sense of the data or to have some insights into the data. And statisticians have developed some ways in which we can recover a baseline function. But before I illustrate this, I will introduce another way to summarise the results of a Cox model which is based on individual risk scores, and this is really a precursor to understand and how we can recover a baseline function.

Individual risk scores allow to summarise the effects of several predictors in a model, in a Cox regression model, so they are attractive, especially when we are looking at complex models with multiple predictors. In these cases, it may be difficult to appreciate the role of different predictors and risk scores provide some help, as I will illustrate.

The purpose of the risk scores is to compare the risk of event occurrence for a participant in the study compared to the risk of an ideal baseline individual that might not even exist. The baseline individual is basically a participant that displays value zero in all the predictors.

Here I'm going to use the same data set I've used before, so this is the lung cancer patients’ data set from the survival library in R and the event of our of interest is the death of patients measured in day.

So, here I ran a model where the predictors were sex and the patient’s age at the start of the study. I centred age at the median value of the sample, so baseline individual in this example will be there for a female patient of median age because I had centred age at the median by the sample media. As usual, you can see the scripts I've used to create and calculate these examples with R and those scripts are provided with these resources.

So, here the risk score will try to assess the risk of death for patients that have different sexes and whose age varies around the median. The risk score can be calculated by considering the transformation of the Cox model equation. So, here I put the Cox model expressed as an hazard function. But if we consider individuals that have non-zero values in the predictors, the estimated hazard function of this individual is equal to the exponentiation of the sum of the β coefficients multiplied by the predictors, by the value of the predictors.

For a baseline individual instead, since the predictors all have value zero, the estimated hazard function is just equal to the baseline hazard function. All the predictors are equal to zero so the exponentiation of zero is equal one and therefore the risk score of a baseline participant is equal to the baseline hazard function.

But if we want to compare the hazard of an individual with non-zero values in the predictors against a baseline individual, we need to divide the risk scores and I represented this division in the right, lower right part of the slide. But since the baseline hazard function is both on the numerator and the denominator of this division, they cancel each other out. So, basically this means that the risk score of an individual with non-zero values in the predictors is equal to the exponentiation of the sum of the product of the β parameter with the respective value of the predictor. So, it's equal to this expression you see at the bottom right of this slide.

And here from the equations I also provide a practical example using the lung cancer data set. Again, you can follow this if you look at the scripts I have prepared with these resources. So, the results of the Cox regression model are in the table on top. So, we see that we have coefficients for male and age and then we have in the panel in the table on the bottom of this slide, the three cases, three participants in the study with their age, their sex and the age centred. So, if we substitute those values of age centred and male to the equation I've written here, we can easily calculate the risk scores. And here, for example, we can see that the participant number one being male and 11 years older compared to the median age displays an increased risk score, whereas participant number seven being female and younger than the median age displays a lower risk score.

So, it's important to remember that risk scores are measured in relative not absolute terms. So, they measure the relative level of hazard of a participant in comparison to the baseline of that function in comparison to a baseline participant.

But where these risk scores are particularly useful is in providing some intuition about the mechanisms at play in determining risk, particularly in complex model. It is rare that few predictors have large effects that dominate over others and because of that it is difficult often to assess the role of predictors from the parameters of the model alone. By looking at the risk scores, we can get some ideas and some intuition about how the combination of predictors and predictor values together with the coefficients affect participants’ level of risk. So, they can be useful in providing some more insights into the mechanism that determine risk.

But the risk scores are also important in and play a key role in recovering a baseline function. Remember that these risk scores basically compare the level of risk of a participant to the risk of a baseline individual, therefore the risk scores provide an alternative matrix for assessing the side of the risk set. And the idea is basically to use this property to estimate what is at the beginning of every time interval the total risk scores for participants that are still in the risk set. And in this way it is possible to estimate the reduction in remaining risk across intervals. And using this totals, using the total amount of remaining risk in different intervals, it is then possible to compute the baseline conditional probabilities, basically.

There are two ways basically to estimate the baseline functions based on these properties or the risk scores, and these were different ways only differ in approximations. So, the most used method is the one that is dubbed the product limit method. Some software will report it as the product limit method. And using this method it is then possible to estimate a baseline function as well as an average function whereby the values of the predictors are set, fixed at the sample average. In practice this means centring all the variables, even all the predictors, even the dichotomous ones. But this does not have relevance for the model parameter, so it is acceptable to do so and therefore recover a function of an average individual and a sample. And again, you can find scripts with these resources that worked out these methods, so you can follow them.

Here, the example I used is based again on the lung cancer data and in this example I introduced another predictor, the ECOG performance score given to the patient by a physician. I dichotomised this score to indicate whether the patient is bedridden most of the time.

And here are the results of the Cox regression model I report in the table. So, you can see the coefficients and the hazard ratios associated with those different predictors.

But using these methods then I can recover baseline function for, so the survival function and the cumulative hazard function of baseline participants. So, a participant that has a zero value on males, so a female has zero score in the ECOG assessment and has age that is equal to the median centred age.

And I represented in this graph the baseline function and I compared this function against an average individual, so an individual that has the sample average of male, proportion of male, the average proportion of the dichotomised ECOG score, and in this case I decided to centre, to use centred age anyway, but I could have chosen a different average or a different value for centred age as well.

The point really is that it's possible then to compare the functions of a baseline individual against the function of individuals that have different values in their predictors.

But these methods then also give the possibility of estimating the functions of prototypical individuals in the sample, so individuals that display meaningful and relevant combination of predictor values. So, we can use the properties of risk scores to estimate the functions, and in this example here I reported, for example, the survival function of individuals that display different combinations of values in the predictors. And again, you can find the R scripts I have used to create this graph where I report the survival function of a baseline individual that has values of zero in all the predictors and then, for example, a bedridden male, a male that is not bedridden, or an older male, for example. I've changed the age of the individuals represented in the yellow line to be 12 years over the median, for example.

The point is that you can create different combinations of prototypical individuals and plot their functions in order to get some insight regarding the role of the different predictors in affecting the risk, the hazard and the survival rate in the studies, of individuals in the studies.

Here I have just used the same method to also calculate the cumulative hazard function by those groups, those prototypical groups, and you can see that, for example, bedridden male has a higher cumulative risk over time and it might be interesting to also look at the changes in the rate of the cumulative hazard function for bedridden males and inspecting those type of graphs can be very valuable in giving better intuition and clearer understanding of what the model is suggesting, what the model is showing.

So, I illustrated that the Cox regression model is semi-parametric. As I said, it doesn't concern itself with modelling or assuming a shape for the baseline function. There are different types of models that are more traditionally based on parametric assumptions on the baseline function, but the Cox regression model doesn't need that and also statisticians have demonstrated that the Cox regression model is very efficient, even when the baseline function has a shape that can be easily represented or represented according to some known shape. So, even in those cases where it might be possible to safely assume a shape for the baseline function, the Cox regression model performs almost as effectively, but the Cox regression model performs a lot more effectively and efficiently when we are unable to make assumptions on the baseline function. So, it has a lot of flexibility and that is also why it has become so popular.

The Cox regression model is also semi-parametric in other ways because, for example, it relies basically on the ordering of event occurrence, and again some of the references I've provided with these resources discuss these issues in more detail if you are interested in them.

However, the Cox regression model does make some assumptions, so here remind them and so there are some assumptions that needs to be considered and so the model is not completely non-parametric, it is semi-parametric. And here I remind, so I put again the key assumption. The last one states the distance between the log hazard functions of the different combinations of predictor values are identical across time intervals, and this assumption may be tenable in some cases, but there are other cases where it may not be tenable and there may be cases, for example, where the facts in reducing risk of death is more pronounced in some periods. And I just wanted to say that it is possible to test this assumption of constant distance and relax it and I'll just briefly introduce an example of how to do that.

I will not go into too many details, but just to illustrate how this assumption can be relaxed to allow interactions between predictors and time, I will start from the Cox model expressed in terms of log hazard functions. And remember here that tij indicates the time interval j, for individual i in the data set, so lower case j indicates different time intervals, lower case i indicates different individuals in the sample. Lower case h0 indicates the baseline hazard function. So, the log of the hazard function for an interval j and individual i is equal to the log baseline hazard function at interval j plus the product of the individual’s values or predictor X and the β1 coefficient. So, this is the basic Cox regression.

But to this initial equation we can add another term that represents a β2 coefficient of the product or the value of predictor X and time. So, this basically represents the interaction between the values of the predictor and time, not also that in this notation I added a constant c which is subtracted from time, so c here represents a constant. This constant is just used to facilitate interpretation, it’s analogous to centring time, and we could decide to use the beginning of the study as our constant that is the constant will be times zero when the study starts so we can ignore it basically. We could choose another meaningful point in time, for example, the samples median lifetime or we can use some other meaningful time points.

In this equation also why the β1 coefficient represents how much the log hazard moves vertically for unit increase of predictor X, the β2 coefficient represents how much the log hazard will further move with each increase in time. So, it represents the moderation of the fact by time.

If the β2 is positive, it means that the log hazard function will increasingly move upward vertically with increasing time, whereas if the β2 is negative, it means the log hazard will increasingly move downward, vertically downward with increase in time units.

But another popular option in modelling interactions with time is to consider the interaction between the predictor X and the natural logarithm of time. This has some useful qualities because, as usual, logarithm transformations are useful in representing changes in the magnitude of values. So, it might be particularly useful in some instances.

However, we also need to be careful not to select values of time or the constant and the constant c that will return zero because the logarithm of zero will return negative infinity so it will present problems in estimating coefficients and so on. So, we need to make sure that we don't choose time and a constant time c that will make the values within the parenthesis time -c equal to zero if we are using a log transformation of time.

But there are other ways in which interactions between predictors and time can be expressed, for example, it may be possible to divide time piecewise into epochs that can have different durations and create dummies for the epochs and estimate coefficients for each one of these. For example, we might consider the occurrence of shareholders selling their shares and we can test if the effect of predictor on these hazard changes during epochs that correspond to before and after a series of announcements. So, we might have intervals of different duration and not equally spaced as well. And the references I provide with resources will also provide more details and worked examples if you are interested in that.

I just finally wanted to graphically represent an example of an interaction of that predictor with time and again this is a using the lung cancer data set, but you see the Cox regression model expressed as log hazard in the upper part of this slide and the graph in the upper part of this slide represents a standard Cox regression model or the predicted log hazard functions for two groups based on this standard model. So, we can see that the distance between the two functions is the same at every time interval in the study and the equation in the lower part of the slide represents the addition of an interaction term where the predictor is interacting with the logarithm of time and the graph in the lower part of the slide illustrates possible interaction effect where we can see that the distance between the two functions is not the same at every time interval in the study. So, the assumption is not maintained and we are in this example in the lower part of the slide, we are assuming and modelling an interaction between the predictor and time.

So, to summarise, the Cox regression model is a semi-parametric model. There are no assumptions concerning the baseline function and these give the model a lot of flexibility and makes it useful in many instances.

However, there are key assumptions and particularly it's important to remember the proportionality of hazards which is also a name given to the Cox regression model. Some authors object to this name to the Cox regression model because, as I illustrated, this assumption can be relaxed. So, the Cox regression model doesn't necessarily imply proportionality of hazards. And in fact, I provided an example where I've modelled an interaction between predictors and time.

I've also illustrated how to calculate individual risk scores that allow to estimate the relative hazard of individual cases compared to baseline cases, but it also allows to recover baseline functions and can be used then to model the functions of prototypical individuals in this way allowing further intuition and further insights into the results of the Cox regression model.

So, thank you very much again for your attention and I will say again that if you found this resource useful, you can also find some exercises with solutions that I have prepared as well as other references and you can find more resources on research and statistical methods on the webpage of the National Centre for Research Methods.

So, thank you very much. Goodbye.

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