# Introduction to Survival Analysis – part3

## Transcript

Video: <https://www.youtube.com/watch?v=GH2AWmWhaRs>

Full resource: <https://www.ncrm.ac.uk/resources/online/all/?id=20850>

Hello and welcome. I'm Oliver Perra and this is the third part of my introduction to Survival Analysis.

I'll just summarise what I have discussed in the two previous presentations. I have highlighted that Survival Analysis is a statistical method that is used to answer questions about whether and when an event of interest takes place. I've also described the life tables which are useful ways to summarise event occurrence over time, and then I introduced hazard function which is the conditional probability of an individual experiencing the event within that interval of interest given that the individual didn't experience the event before. And then I introduced the survival function, which is the probability of surviving past time interval for a randomly selected individual.

In this presentation, I will first illustrate how we can explore differences in event occurrence between different values of covariates or predictors. I will then discuss issues and assumptions that allow us to start building models to represent the effects of covariates and I will introduce the logit hazard function which is the basic block for building more sophisticated models in Survival Analysis.

Once again, I highlight that I will only present examples that apply to discrete time event occurrence, so events that are measured in coarse time units like years, and I do this because again, I believe that these examples provide a strong foundation for understanding more sophisticated Survival Analysis methods like Cox regression models.

I'm going to use the same data set from the previous presentation. This is a modified version of the Capaldi and others study of sexual intercourse initiation among adolescent males. The time matrix here was grade years from Grade 7 to Grade 12. One covariate in the study was parental transition before Grade 7, so adolescents reported whether they had experienced one or more changes in parental figures before Grade 7 or not. So, the variable is a proxy for lack of stability in parental relationships.

And here I showed the life tables for the two groups. The 72 adolescents who had not experienced parental transition before Grade 7 and 108 that did experience these transitions. In the material online you can find the R scripts that I've used to create these tables and the rest of outputs that I will show in this presentation.

A quick look at the tables can tell us that the risk of the event seems higher in the second group and conversely the probability of surviving is lower in this group, in the group that experienced parental transitions.

And here I have bloated the hazard function and the survival function for the two groups. When we are considering hazard function, there are two issues we need to check. The first is that shape of the hazard function for the different groups. And in this way we can check if these functions speak at the same time or rather show different patterns where the peaks happen at different times or different groups. Here, despite some differences, the shapes of the two lines appear generally similar.

The second consideration should focus on the relative level of the two lines. Do they indicate that one group is consistently at higher risk of experiencing the event? And furthermore, does the magnitude of this difference vary over time? Here the two lines indicate that adolescents male who had been exposed to parental transitions, consistently showed a higher risk of initiating sexual intercourse across Grade 7 to 12. But does this difference vary over time? Although it may appear that the differences in risk are higher in later years, overall, the variation, the level of hazard between groups across time does not seem to be considerable.

Now remember that the hazard function tells us about the risk of the event taking place within specific time periods. The survival function instead tell us about cumulative effects over time. So, for this reason, when we compare survival functions of different groups, we can get an idea about the cumulative effects of a predictor over time. Not here that the order of the two curves is reversed when we compare hazard and survival functions. Since those with parental transitions are more at risk when we consider the hazard function, they are also less likely to survive when we look at the survival function that is the less likely to avoid the target event.

So, in this case, and we can see that the cumulative effect of parental transition is quite pronounced, by Grade 12 approximately half of the adolescents with no parental transitions have had sexual intercourse, whereas roughly three-quarters of those who had been exposed to parental transitions before Grade 7 have had sexual intercourse. Note also that since the survival function of the group with no parental transition did not reach the midpoint where 15% have a probability to have experienced the event, it is not possible to calculate the exact median lifetime for this group. So, therefore I calculated that first that's our lifetime here. So, the point where one-third of the participants have experienced the event in the two groups. So, the differences in the summary statistics again highlight the group differences. A third of adolescents who had parental transitions initiate intercourse more than a year before a third of those who had not parental transitions, had not experienced parental transitions.

So, as I probably illustrated, inspections of the life tables and the plots of hazard and survival functions can be very useful in getting some insight into between group differences. And incidentally, if we wanted to get some intuition about differences in risk associated with continuous covariates, we could categorise the continuous covariables into some meaningful groups and calculate the hazard and survival functions of salient subgroups. For example, when we have a continuous covariate, we could look at the survivor and hazard function of those that scored called one standard deviation above and one standard deviation below in this, the mean of this continuous variable.

Now we need to move from these intuitions to more formal models. The hazard function is the main estimate of interest here since it represents the risk of event occurrence within time periods. But to model this outcome, we need to consider models that describe the shape of the whole hazard function over time, and that may be difficult because the hazard function may have different peaks and flows. This task is also complicated by the bounded nature of those estimates. The hazard function is a conditional probability, so its values should lie within zero and one, which makes the task of modelling the outcome difficult. To deal with this issue, Cox recommended transforming the probabilities into alts and log alts.

And here I illustrate those transformations applied to us through the hazard function. So, the logit is the natural logarithm of the odds. If you have round logistic regressions you may already be familiar with the key characteristics of logits. The key advantage of using logits though, is that those values are not bounded. The logit of the midpoint in a probability distribution, that is the logit of 0.50, is zero. So, logit values below zeros correspond to probabilities less than 0.50, so less than 50%. Why the logit of zero is an infinitesimally small value, logite values above zero correspond to probability values above 50%, while the logit of one is an infinitesimally large value.

The logit transformation is also useful in turning skewed distribution into more symmetrical ones. And one way of understanding how the logit transformation works is to remember that those are particularly useful in representing differences in magnitude of values. And in fact, if we look at the logit transformation in the bottom plot on the right, we can see that the distance between the two functions in the first years are larger in the logit scale compared to the hazard values and are smaller in later years when the hazard values are larger.

So, these changes should actually highlight another key advantage of using logit transformations. The distance between the two functions over time, it's easier to compare. The logits make the difference between functions, the gap between functions more stable and this facilitates modelling. So, the logit transformation, it's very useful in dealing with the hazard functions and modelling.

So now we have a better metric for our outcome of interest, the logits of the hazard function. And we can think about what models can describe these functions more adequately. And we could, for example, hypothesise a linear model which here I represented using dashed lines. We could think of other linear models, for example, the quadratic model. Whichever model we choose, the task of starting modelling is facilitated if we follow a set of assumptions. One is that there is an estimated logit hazard function for each value of the predictor. In this case, we will estimate two hazard functions because we have a predictor that only takes two values. When we work with continuous predictors, we estimate hazard functions for each value of the continuous perimeter.

The task of modelling is also facilitated if we assume that the functions for different levels of predictors have the same shape. So, for example, here we assume that both functions follow a linear trend in this example. And finally, we assume that the distance between logits of the two functions is the same in each time interval. For example, here we see that the two hypothetical linear trends are parallel, the distance between the two values is the same at each time point. But if we were hypothesising different curves, we will still find it easier if we assume that the distance between the functions for the different levels of the covariate are constant across time. And I will illustrate this in the following examples.

The advantage of these assumptions in model building should be evident. For example, if we assume the shapes of hazard functions are the same for different levels of the covariate, our models are simpler and mostly forward. But as all assumptions, this may be unattainable and I'll advise in some cases. When you delve into Survival Analysis beyond this introductory presentation I am working on, you will learn to test these assumptions and relax them when they are clearly inadequate.

So, in this example of data, I'm using a linear model may appear relatively appropriate, but in many cases logits of hazard display trends that are difficult to reconcile to common models like a linear or a quadratic one. For this reason, we may start analysis by specifying a general model that has no constraints on the shape of the logit hazard function. So, a general model may be closer to the data although a general unconstrained model is not parsimonious. It has a course in terms of parsimony, so it's mainly not ideal in some circumstances. But I will use general specification of a model without constraints, so without a predefined shape. And how can we build this general specification model? The idea is to firstly specify a baseline function. This is the logit hazard function when the values of the predictors we are considering or the covariates we are considering are equal to zero. So, in this case the baseline function corresponds to the logit hazard of adolescents with no parental transitions, they didn't experience parental transitions.

A general unconstrained curve for the baseline or line actually, for the baseline model can then be specified by assuming there are different intercepts that represent the expected values of the logit hazard in every time interval for those in the baseline group.

So how does that work in practice? And to better understand it, it's important to consider how the data should be organised when we are looking at Survival Analysis.

And the best way to organise the data for Survival Analysis is the person-period data set. If you have experience with multi-level models, this is very similar to the transformation from a wide data set where every row corresponds to a participant to a long data set where every row corresponds to a participant’s measurement occasion and measurement occasions are nested within participants.

The idea here is that every participant contributes different periods to the data set, different time intervals. For example, Participant 1 contributes only three periods, that is there are observations of this participant up until the point when he reported an event, initiated the event remember, his initiation of sexual intercourse.

Participant 2 contributes six time periods up until the point where the study stopped. No event was ever recorded during this these periods and so there is no event being recorded for this participant and the participant is censored.

Participant 17 instead only contributed one period since he experienced the event within this initial period of the study.

Once we have structured the data set in this way, we can create a set of dummies for each time period. These dummies are set to one in the time period they represent and zero in all the other periods. So, they're basically indexing different time periods. So, in this way, once we estimate intercepts alpha for each time period, so if we build a model where we have intercepts and we assume that there are different intercepts for each time period, these intercepts only switched on, so to speak, during that specific time period. That is by multiplying the dummy time period indices with the alpha intercept, the intercept is only relevant within its specific time period.

So, I'll try to represent here that when we are considering Period 7, for example, Year 7, the only parameter alpha, the only intercept that remain when we multiplicate the dummies for different periods by the alphas, is the alpha, the intercept for Period 7 that is -2.40.

And this hopefully will be clearer when I show you the formula for the model and how it can be graphed. So, we can specify a line that does not follow a preset shape, such as a linear or quadratic shape, but a line that rather can be jack and irregular and thus be closer to the data observed. And to understand how this happens, I also put the formal definition of the baseline model where I specified that the logit of the hazard function for individual i in time period j is the sum of the product of the dummies for each time period and the relative intercept alpha values.

So, say when the time period is seven, we take this equation, substitute the P indexes with the values in the row corresponding to Period 7 and also substitute the alphas with the values in the row for Period 7. And we will obtain that the logit hazard for Period 7 is equal to its specific intercept -2.40.

So, effectively each time period then has its own period specific intercept and therefore we obtain a jack line like the one displayed here. So, every time period basically has different intercepts which can then raise or lower the value of the estimated logit of hazard functions.

I have specified a baseline model that is a model when covariates are set to zero. But we also have covariance in the model and, for example, we have a parental transition, PT covariate. This was recorded before Grade 7, so it will not change afterwards. It's a time invariant covariance because its value is set before the start of the study.

But the person-period data set makes it easy also to include time varying covariates. Here I just created a fictional covariate X and assumed this varied between zero and three across time periods. A similar time varying covariate in a study like this may be, for example, frequency of alcohol use in the previous year which research suggest is a significant predictor of whether and when adolescents have intercourse for the first time.

But how do we include those covariates in the estimation of logit hazards? Here I put the formal definition of the model, so the first part is the is the baseline model I have just described. The second part of the equation shows that we assume different slopes for each covariate, and note that the value of the time invariant covariate PT, parental transition, only changes from one individual i to another. Instead, the value of the time varying covariate X will change across individuals i and time periods j. So, it's a time varying covariance. And the slopes represent expected changes in event occurrence associated with one unit increase in that covariate while controlling for the effect of the other covariates in the model, just like other regression models. So, this formulation is very akin to other regression models basically.

However, it's also worthwhile reflecting on what this equation is doing conceptually. In Survival Analysis we are interested in answering questions about the relationship between the timing of an event and covariates or predictors. So, time is conceptually an outcome when we are talking about Survival Analysis. However, this model that I have illustrated this equation is basically making a time index it by periods into a predictor. So, paradoxically, the conceptual outcome of interest has become the predictor. But this is possible because this formulation of the model, this equation, is changing the question from what is the association between timing of an event and predictors?

To a slightly different question, what is the association between the risk of event occurrence in each time period and predictors? So, by answering the second question about risk of event occurrence in each period and predictors, we can ask the first question about associations of predictors with timing of events, with events timing.

So here I'm going to consider only one covariate, parental transition. And we can see that when this covariate PT, parent transition, is zero, indicating adolescents that did not experience parental transitions, we obtain the baseline function that I described before. But when the covariate is one for cases that did experience parental transitions before the start of the study, we are basically adding the slope of the covariate to the baseline model in each time period. So basically, representing it in the plot, this means that the logits of the hazard function of cases that experience parental transitions will just shift in each time period by the degree that corresponds to the covariate slope.

So, in this case we can see that the line of those that had experienced parental transitions shifts constantly upward represented increased risk for those who had experienced parental transitions before the start of the study.

So, this is consistent with the goal of creating models where the logit hazard function of different covariate values have the same shape and the distance between these two functions is the same across time period is constant across time periods.

So, now we can estimate the parameters of a similar model, the model we have specified here, using maximum likelihood methods. And indeed, we can estimate these parameters starting from the person-period data set. And just regressed the event indicator on the time period indicators and the covariates using logistic regression. So, we can just apply, just apply logistic regression on the variables that are represented in the roles of the person-period data set.

One objection to using logistic regression in this way may highlight that time periods j are nested within individuals, so they are not really independent variables. However, this objection is addressed by considering that the hazard functions represent conditional probabilities of event occurrence in a specific time period. So, these probabilities are conditional on the individual not experiencing the event in the previous time periods. This means basically that the records in the person-period data set arose in the person-period data set are conditionally independent.

And so, using logistic regressions, we can estimate the parameters and here in the table I also report the estimations of these parameters and you can find the script that I've used to estimate these parameters. That is included with the resources of these presentations.

Here I represented the estimates of the logit model in the top panel on the right, and these estimates are in the logit scale and both logit functions for the baseline group who didn't experience parental transitions and the other group who experienced parental transitions. Both logit functions have the same shape across time and the distance between the two estimates is the same in every time period. This is where the assumptions that I applied on the model and they are confined in the plot. But when we transform this estimate in other scales, we see that the picture changes and here in particular look at the larger panel on the bottom left. This boundary represents a yield of the event for the two groups.

And we can see that the difference between the values of the odds changes over time periods. However, if we take the odds of each time period for the different levels of the covariate and calculate the ratios of these odds, we will obtain the same number, in this example 0.447 which means the relative distance of the odds is constant, or as the odds of the outcome for those who experienced parental transitions are proportional to the odds of the baseline group, those that did not experience a parental transition.

This characteristic of the odds model follows from constraining the logit estimates to have this constant distance. Another way of a similar constraint is therefore called the proportional odds model. Here I also reported the hazard estimates calculated by transforming the logit estimates. These are represented in the panel in the bottom right of the slide. So, these are the fitted on the estimated hazard functions for the two groups. And although the three lines appear very similar to those of the estimated odds for the two groups, the hazard estimates are not proportional. They do not have the same relative distance as the odds do. In this example it is not evident that they are different but they are and there isn't an assumption that the proportion of the hazard estimates, it's a constant across time periods.

This leads me to highlight the importance of exercising some caution when interpreting the results of Survival Analysis and particularly when we consider results in different scales. For example, if we consider the estimated hazard functions from the model I have just fitted and brought them in a graph like this, we won't notice an increasing gap between the event probability across the two groups. And we may be tempted to assume that this increasing gap may indicate an interaction between the covariate parental transition and time. That is, we might be tempted to assume that the effect of parental transition on the risk of initiating the course changes over time. However, a similar conclusion will be entirely wrong.

And in fact, we need to keep in mind that the model that I specified and reported here again in its formal definition, only included a linear effect of parental transition on the logits of the hazard function. The actual model therefore is not the one represented by the estimated hazard functions after transformation, the one reported on the right panel, but is the one represented by the logits of the hazard function, the one described by the figure on the left.

The fact that a constant effect of a covariate on the logged scale may look like something different when the estimates are transformed on the hazard function scale, as in this example, that means that when we look at the hazard transform hazard functions, we might be misled. But in fact, we should not draw a conclusion from observing the hazard functions plots when the actual model was actually expressed in the logit scale.

So, the point is it is important to align the readers and to keep in mind that what is being implied by the actual model and keep in check warranted interpretations based on reporting the model parameters in different scales.

And with this I conclude my presentation. And to summarise, I have shown how it's possible to explore covariate effects by looking at live tables and plots of survival and hazard function for different levels of echo variate for different values of echo variate.

I then showed how we can start modelling in Survival Analysis using the logit of the hazard function. And then I showed how we can start obtaining maximum likelihood estimation through logistic regressions when and after we have set our data in a person-period data set.

And again, I will highlight those models I just illustrated have stringent assumptions which should be tested and can be relaxed. And if you delve into Survival Analysis more, you will learn about ways to test these assumptions and relax them and build more flexible models.

So, I hope this was helpful, but look at the material hive also created for these resources that show examples and exercises using R.

So, thank you very much again for your attention and remember to check the webpage of the National Centre for Research Methods for new resources and courses.

Thanks, thank you very much, bye.

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