Rough notes: Survival analysis in R for public health

<https://www.coursera.org/learn/survival-analysis-r-public-health/supplement/cVjMc/life-tables>

Cox doesn’t care about the distribution of survival times or what the hazard function looks like. This is why it’s called “semi-parametric”: it has some parameters – those of the predictors – but it has no parameters to describe the hazard function for patients with a value of zero for the predictors (i.e., patients with age zero and all the reference categories for the categorical variables). (For completeness, the simple proportion alive at a given time point and the Kaplan-Meier estimate are examples of non-parametric survival analysis.)

However, making assumptions about the **shape**of the hazard function – adding parameters to the model to describe the shape, making the model “fully parametric” – can lead to better prediction. More accurate prediction of a patient’s survival time or risk of death within a given timeframe is vital for enabling the patient and his or her doctor and clinical team to make decisions regarding treatment. Risk models can put patients into, for example, low-, medium- or high-risk in a process called risk stratification, and high-risk patients can be offered different treatment plans from low-risk ones. It may be that you can do better than the Cox model in terms of risk prediction for a given data set and patient outcome. The Cox model was developed to look at the effect of covariates on the hazard function rather than to estimate survival times. A fully parametric can help here, especially if the Cox model assumptions are violated (more on the Cox assumptions later in the course).

There are **several such fully parametric models** such as Weibull, exponential, log-normal, and log-logistic models, where hazard function has to be specified. The **Weibull**distribution is used widely in medicine because of its flexibility: its hazard function can be increasing, decreasing, or constant over time. A special case of it is an exponential distribution, which is simple because it has only one parameter. This is because the hazard function is constant when the survival time is exponentially distributed. If you want a hazard that increases and then decreases over time, try either the log-logistic or the log-normal.

There are further extensions to the basic survival analysis approach, such as allowing for the fact that the values of some predictors change over time (Cox can deal with this) and handling multiple events (patient outcomes) in the same model. This is useful for disease recurrence, for example. Also, like so many statistical methods, survival analysis can be run in a **Bayesian** framework. Bayesian analysis is in general more complicated but very powerful. In essence, it involves mixing your data and your prior beliefs about what is related to what and deriving probabilities that something is true. What I’m teaching on this course and throughout this series of courses within the specialisation is called classical or “frequentist” statistics. In the classical framework, there’s no formal use of prior knowledge in the underlying maths. The answer you get is completely driven by the data. There’s a philosophical as well as a mathematical difference between the two, and much has been written about it – it’s a huge subject, way too big to go into here. It’s often claimed that there are two rival camps, but (happily) it’s also often claimed that there are no such camps and that many people, including me, use both methods (which often give similar results anyway in practice).

Cox proportional hazards assumptions

The assumption that hazards are proportional for a given predictor can be tested graphically and with a p-value

Cox.zph function to get p-value, and plot object

Technically speaking, the function cox.zph() correlates for each predictor the corresponding set of scaled Schoenfeld residuals with time, to test for independence between residuals and time.

There are three types of residual in cox hazards regression model

Cox.zph uses Schoenfeld residual which is equivalent to normal residual as in regression.  The maths behind the calculation of the residuals is different, mostly because of the censoring, but we don’t need to worry about that.

Deviance residuals are transformations of martingale residuals and help you look for outliers or influential data points. You can either examine the influence of each data point on the coefficients or plot the distribution of the residuals against the covariate

It’s also possible to check outliers by visualizing the deviance residuals, which are normalized transformations of the martingale residual and should be roughly symmetrically distributed about zero with a standard deviation of 1

Another issue is whether any continuous variables that you assume to have a linear relation with the outcome actually do have a linear relation. If you fit age as a single term in the model, then that’s what you’re assuming. The martingale residual is used to test this assumption

Suppose your test for proportional hazards gives you a clear suggestion that the assumption isn't met. What should you do?

To answer this, you need to think about what having non-proportional hazards means. If the relation between males and females regarding their risk of death changes over time, it could mean, for instance, that males have a higher risk of death early on during the follow-up period but at some point, the relation changes so that females have a higher risk of death. One way of putting this is that there is a statistical **interaction**between gender and time. The model is short of a coefficient. If you add a coefficient for this interaction, which allows for the difference in risk by gender to change over time, then the problem would be solved.

Trying this interaction term in the model and testing whether it is statistically significant is in fact another way of testing the proportionality assumption. If this interaction term is not statistically significant, then it follows that the assumption is valid. As is usual with any kind of regression, Cox included, you should do the statistical tests – i.e. get the p-values – but also do the plots. Some kinds of non-proportional relationships and other assumption violations can’t be detected just from a p value.

So, if the assumption is violated, then one option is to include this interaction. If the p-value is low but the hazards are proportional for most of the follow-up period, then that suggests another solution: divide the survival analysis into **two time periods**. You can fit one model when things are fine, i.e. when the assumption is valid, and another model to cover the later follow-up period when the assumption is not valid. This second model may need an interaction term, but the first one won't.

There’s also a third simple way of dealing with the problem: stratify the analysis by the variable that’s causing the problems. If it’s gender, for instance, then just fit separate models for males and females. The drawback of this approach is that it’s no longer possible to compare the effect of each gender on mortality.

**Model selection methods: how to choose your predictors**

This was covered in detail in the Logistic Regression for Public Health – similar principles apply to any type of regression, including Cox models. There, I explained some common ways of choosing predictors for a multiple regression model and that two such ways – forwards selection and stepwise selection – were simply too awful to contemplate using. A third common way, backwards elimination, does sometimes work OK. While it’s always good to make use of a priori knowledge from the literature and experts in the field, this isn’t always of sufficient help, particularly when you have a lot of possible variables. Less often you’ll have a good deal of a priori knowledge and therefore a large number of predictors that have been found to be associated with the outcome. In that situation, it can be useful to apply backwards elimination to the model with all these chosen predictors in order to reduce the size of the results table for presentation.

**How to apply backwards elimination**

Here are the steps:

1. Fit the model containing all your chosen predictors – either all your a priori ones or all your available ones (if your data set isn’t too large)
2. Store all the coefficients from that model
3. Remove in one go all predictors whose p value is above the preset threshold, typically the usual 0.05 (in a variant of this, you remove the predictor with the highest p value and refit the model, repeating steps until all the predictors have p values above the chosen threshold)
4. Compare the coefficients for the remaining predictors with their coefficients from the original model

**Checks to make when using backwards elimination**

If the coefficients haven’t changed much from the original model, then you now have your final model. You can go ahead and check the residuals and other model assumptions. If, however, you have a predictor whose coefficient has changed noticeably, then you need to find the variable(s) that you have removed that are correlated with this affected predictor. You can do this by trial and error, so add back in one of the removed variables at a time until the affected predictor’s coefficient is back to its original value. When that happens, you’ll need to keep the removed variable in the model.

For example, suppose that blood pressure was retained (original model HR=1.30, p=0.002) but cholesterol was removed because it was not statistically significant (original model HR=1.05, p=0.155). Then you removed cholesterol from the original model, and the HR for blood pressure changed from 1.30 to 1.50. You consider that a big enough change to worry about (see below for more on this). You add cholesterol back in, and the original HR for blood pressure is restored. You now need to keep both blood pressure and cholesterol in your final model. Such correlation between variables is a big reason why stepwise procedures are so unreliable.

NB: how big is “big enough to worry about” is arbitrary. Anything less than 0.05, e.g. a change from HR=1.30 to HR=1.34, is not big enough in my opinion. It’s up to you to decide this. It largely depends on how the results are going to be used, e.g. in a risk calculator for clinical decision-making, perhaps in a national screening programme, or for an epidemiological study of risk factors. In the former, where people can be invited for screening for some disease based on their estimated risk of developing that disease, using the coefficient of 1.30 instead of 1.50 can greatly affect the number of people invited. In the latter, however, such a difference won’t be of such importance, especially if all we do is take the table of hazard ratios and p values and say, “these are the predictors of the outcome, whereas these other factors are not”. The estimated size of the relation (the HR) is of secondary importance.

**Conclusion**

The question of how to choose the predictors in a regression model, be it linear, logistic, Cox or other type of regression, is a huge one when the number of possible predictors is bigger than a handful or two. As this is an introductory specialisation, I've only talked about a few approaches.