THE SIMPLE COX MODEL

A) Intro to Cox Model or Cox proportional hazards model

0) Premises

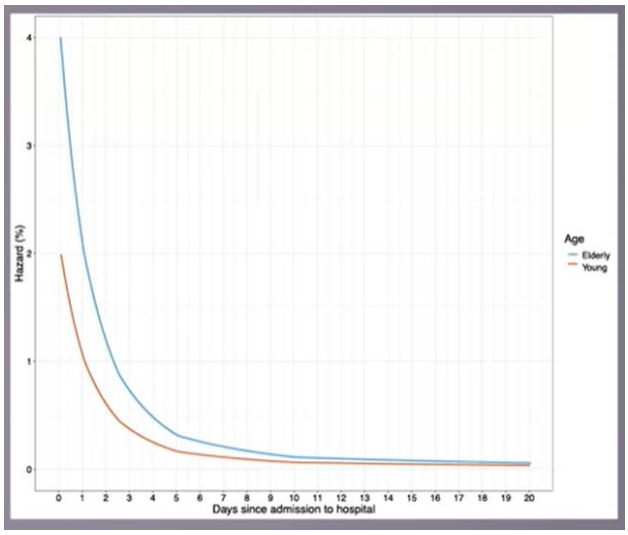
Why is it so used and famous and what is it? The km plot and log rank test are great for exploring the relation btw one predictor and mortality over time but they can only manage one predictor. Cox's approach can handle multiple predictors and in particular it is a type or regression.

It has a couple of distinguish features.

1) Proportional hazards

It means that the hazards are assumed by the model to be proportional. What are hazards? In statistics a hazard is the risk of death at a given moment in time, in general instead is the risk of having the outcome of interest. This hazard can change over time. Let's suppose we have a set of ppl with cancer and we give half of them chemio and the other half, surgery. Their risk of death or whatever or hazard is unlikely to be the same or constant over time. The way their hazard changes over time is called he HAZARD FUNCTION or HAZARD RATE.

What about proportionality?



In blue the hazard of the younger patients while in red that of the elderly.

How can we summarize these two plots and how much are the very elderly more likely to die than the younger ppl. When the 2 lines are nice like here we can summarize the relation with a number; when we say proportional for example we might say that for every time point we can multiply the hazard for the young patients by 2 and get the hazard of the elderly: one will be the some multiple of the other one and it corresponds to the HAZARD RATIO. It is actually similar to the ODDS RATIO in logistic regression.

2) Hazard function and Risk Set

I) Hazard function: The Cox model is formulated around the concept of hazards. The hazard function h(t) is the probability of the event happening at time t, given that it has not yet happened. h(t) is the probability of dying at time t having survived up to time t.

Link to hazard function explanation: http://data.princeton.edu/wws509/notes/c7s1.html

II) Risk set: just like the risk of dying changes over time, so the number of patients that are subjected to that risk change over time as ppl die or drop out. The risk set at time t is defined as the set of patients at the time t that are at risk of experiencing the event.

III) Survival analysis consists of a family of methods and one way that they differ is in their handling of drop outs and other issues when they define the risk set (we saw it when we applied the km method). In survival analysis we are also interested in the difference between survival curves of different groups of patients, we saw the log rank which gives a p-value for comparing the survival curves between different groups of patients with a km plot. The p value tells us nothing about the size of the difference btw the survival curves. This is done by dividing one hazard by another to give a hazard ratio. For ex. by dividing the hazard for females by the hazard for males, gives us a hazard ratio for females compared with males. It tells us how much more likely female patients will die than male patients.

B) Cox model on R -> look at code

C) Missing values

It is very important to understand why we have these missing data: issues of availability or collection methods. And also we might have the values but they are not as informative as they seem, example:

- Surveys on lifestyle = gender, employment, etc. when we get to the voices of exercise, alcohol and smoking, it is common for ppl to try to make them look better on these kind of questions. These data are unreliable. In surveys there is also the possibility to answer to some questions with "Prefer not to say" -> do they count for missing values?
  Now let's suppose that the survey ask for the diet too -> after 5 pages you get bored so we will have a few info or not so reliable ones -> what impression of the population the analysis will have.
- Many countries have databases for ppl who suffer from chronic diseases. For diabetes for ex ppl will have data also on their retinas but also other specific features. For those who can't have these diabetes health checks we will have again a problem of missing data and they are typically those in the poorest conditions.

So missing values are an annoying fact of life with medical and public health data and they take on different forms. The data that we have can give a distorted picture of reality if the missing data are not missing at random. Various tricks exist to deal with this problem but to choose the right one we must first understand its course.

D) The prevention and handling of the missing data (link: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3668100)

0) Premises

Gives tips on how to plan good data collection.

Missing data can reduce the statistical power of a study, can produce biased estimates, reduce the representativeness of the samples and lastly complicate the analysis of the study, leading to invalid conclusions.

Def.: Missing data or values is defined as the data value that is not stored for a variable in the observation of interest.

Being this a common issue many studies have focused on handling the missing data, on the consecutive problems and the methods to avoid or minimize such in medical research.

1) Types of missing data (typically 3):
    - Missing completely at random (MCAR) = defined as when the probability is not related to either a specific value or to the set of observed responses. If data are missing by design, because of an equipment failure or because the samples are lost in transit, such data are regarded as being MCAR. The statistical advantage is that the analysis remains unbiased.
    - Missing at random (MAR) = is a more realistic assumption. Data are MAR when the probability of missing responses depends on the set of observed responses and not related to the specific missing values which is expected to be obtained. MAR does not mean that the missing data can be ignored. If a dropout variable is MAR, we may expect that the probability of a dropout of the variable in each case is conditionally independent of the variable, which is obtained currently and expected to be obtained in the future, given the history of the obtained variable prior to that case.
    - Missing not at random (MNAR) = The cases of MNAR data are problematic; the only way to obtain an unbiased estimate of the parameters in such a case is to model the missing data. The model may be incorporated into a more complex one for estimating the missing values.

2) Techniques for handling the missing data
   The best possible method of handling the missing data is to prevent the problem by well-planning the study and collecting the data carefully. In clinical research:
    i) The study design should limit the collection od data to those who are participating in the study. This can be achieved by:
        a. minimizing the number of follow-up visits,
        b. collecting only the essential information at each visit,
        c. developing user friendly forms and surveys.
    ii) Before the beginning of the clinical research develop a detailed documentation, protocol, standard language and collection and storage of the data and procedure to follow for screening the participants.
    iii) Before the beginning of the clinical research a training should be conducted to instruct the personnel.
    iv) Perform a small pilot study in order to identify the possible issues.
    v) Set a priori targets for the unacceptable level of missing data; monitor and report the data collection di conseguenza.
    vi) The study investigators should identify and engage the participants at greatest risk.
    vii) Finally, if a patient decides to drop out from the follow up, record the reason for subsequent analysis.

One technique for handling the missing data is to use robust data analysis techniques. With robust methods we indicate all those techniques where there is confidence that mild to moderate violations

of the assumptions will produce little to no bias or distortion in the conclusions drawn on the population.

I) Listwise or case deletion

The most common approach and it involves simply omitting those cases with missing data and analyze the remaining data. If the assumption of the MCAR is satisfied, this technique produces unbiased estimates and conservative results, on the contrary it will produce bias.
Appropriate when the sample is large and the MCAR assumption is satisfied.

II) Pairwise deletion

It eliminates info only when the particular data-point needed to test a particular assumption is missing. It uses all the info observed. Main issues:

- The parameters of the model will stand on different sets of data with different statistics, such as the sample size and standard errors;
- It can produce an intercorrelation matrix that is not positive definite.

It is less biased for the MCAR or MAR data and the appropriate mechanisms are included as covariates. If there are many missing observations the analysis will be deficient.

III) Mean substitution

The mean variable of a variable is used in place of the missing data value for that same variable. With missing values that are not strictly random the mean substitution may lead to inconsistent bias and it add no new information but only increases the sample size and leads to an underestimate of the errors.

IV) Regression imputation

Process of replacing the missing data with estimated values. It preserves all cases by replacing the missing data with a probable value estimated by the available info. As in the mean substitution no novelty is added, while the sample size has been increased and the standard error is reduced.

V) Last observation carried forward

Many studies are performed with the longitudinal or time-series approach in which the subjects are repeatedly measured over a series of time-points. This method replaces every missing value with the last observed value from the same subject. However, it makes a strong assumption and therefore produces a biased estimate of the treatment effect and underestimated the variability of the estimated result.

VI) Maximum likelihood

Assuming that the observed data are a sample drawn from a multivariate normal distribution is relatively easy to understand. After the parameters are estimated using the available data, the missing data are estimated based on the parameters which have just been estimated.

VII) Expectation-Maximization imputation

It is a class of maximum likelihood that can be used to cerate a new data set, in which all missing values are imputed with values estimated by mle.
Steps:

i) Expectation step: estimation of parameters such as variances, covariances and means perhaps using listwise deletion.
ii) Use the estimated parameters to create a regression equation to fill in the missing data.
iii) Maximization step = uses those equations to fill in the missing data.
iv) Repeat the expectation step with the new parameters where the new regression equations are determined to "fill in" the missing data.
v) Repeat the expectation and maximization steps until the system stabilizes, when the covariance matrix for the subsequent iteration is virtually the same as that for the preceding iteration.

Important characteristic: when the new data set with no missing values is generated, a random disturbance term for each imputed value is incorporated in order to reflect the uncertainty associated with the imputation.

Cons: it can take a long time to converge and can lead to biased parameter estimated and can underestimate the standard error.

A predicted value based on the variables that are available for each case is substituted for the missing data. Because a single imputation omits the possible differences among the multiple imputations a single imputation will tend to underestimate the standard errors and thus overestimate the level of precision.

VIII) Multiple imputation

In multiple imputation the missing values are replaced with a set of plausible values which contain the natural variability and uncertainty of the right values.
Procedure:

i) Prediction of the missing data using the existing data from other variables.
ii) Replace the missing values with the predicted ones obtaining a data set called imputed data set.
iii) Iterate the procedure obtaining multiple imputed data sets.
iv) Analyze each multiple imputed data set using the standard statistical analysis procedures obtaining multiple analysis results.
v) Combine these results and run a single overall analysis result.

Pros: it restores the natural variability of the missing values incorporating the uncertainty due to the missing data. It produces valid statistical inference.

IX) Sensitivity Analysis

It is defined as the study which defines how the uncertainty in the output of a model can be allocated to the different sources of uncertainty in its inputs. When analyzing the missing data, additional assumptions on the reasons for the missing data are made.

3) Recommendations

Some amount of missing data is always expected and more attentions should be given to the missing data and how to deal with them. The best solution is to maximize the data collection when the study protocol is designed.
Single imputation and LOCF are not optimal approaches for the final analysis as they can cause bas and lead to invalid conclusions.

E) Multiple imputation in R using MICE, preceded by an intro to MCAR, MAR and MNAR. (link: http://dept.stat.lsa.umich.edu/~jerrick/courses/stat701/notes/mi.html, link: https://blog.caspio.com/what-you-need-to-know-about-data-harvesting-and-how-to-prevent-it/)

1) Methods to handle missing data

I) If the data is MCAR:
- Complete case analysis;
- Multiple imputation or any other imputation method.

II) If the data is MAR:
- Some complete cases analyses are valid under weaker assumptions than MCAR for example linear regression;
- Multiple imputation.

III) If the data is MNAR:
- Model the missing data explicitly and appropriately; jointly modeling the response and missingness.
- Generally we assume MAR whenever possible.

2) mice in R

In R we can use the mice (Multiple Imputation with Chained Equations) to perform multiple imputation and the subsequent analysis. The 3 steps of the multiple imputation are typically:
- impute the missing values with values randomly drawn from some distributions to generate m complete cases data sets.
- Perform the same analysis on each of the m data sets.
- Pool the results in some fashion.

These 3 steps are performed via the mice, with and pool functions respectively.

I) *mice* function = performs the imputation via chained equations. Chained equations are a variation of a Gibbs Sampler (an MCMC approach) that iterates between drawing estimates of missing values and estimates of parameters for distribution of the variable. Chained equations have fast convergence compared to most MCMC due to conditional independence of Yt and Yt-1.
The default number of imputations m=5.
method (argument of the mice function) = *pmm* = predictive mean matching, which has the benefits that it is semi-parametric and imputed values are restricted to observed values.

II) *with* function = it replicates using *attach*, running code, then running *detach*. When we pass a *mids* object (namely the output of a *mice()* call) to *with*, it handles the imputations.

III) *pool* function = for any method that produces an object having both *coef()* and *vcov()* methods, this function can combine the individual coefficient estimates into one "pooled" estimate, based upon Rubin's rules for pooling.