TESTING MODEL ASSUMTPIONS AND CHOOSING PREDICTORS

A) How to assess Cox model fit

0) Premises

Cox model is the most common form of survival analysis when you want o include multiple variables in the same model. How do we know if it is good?

1) 3 methods to check it

The residual is a key tool to assess various aspects of model fit. A residual is calculated for each observation and gives some measure of the difference between the actual value from the data set and what the model predicts for that patient. Depending on the type of regression and mathematical basis we will have different types of residuals.

In particular for the cox model there will be mainly 3:

- Schoenfeld residual = whatever the shape of the hazard function, however the patient's risk of death changed over time, we can ignore all that if this assumption holds. In this course we are looking at death after hospitalization for heart failure. Suppose we want to compare the hazard for females and males; the model will give us a hazard ratio for males compared with females. This ratio is all we need if the 2 hazard functions are parallel or "proportional". And we can test this using Schoenfeld residuals: if these residuals for gender correlate with the follow-up time since hospital admission. If they do not correlate with time, then the assumption is valid.
- Martingale residual = they allow us to test whether your continuous predictor such as age has a linear relation with the outcome or whether we need to add more terms to the model such as age squared. Martingales have a mean of 0, values near 1 represents patients who died earlier than predicted, large negative values represent patients who died later than predicted. The resulting plot should give you a nice straight line if the assumption is valid.
- Deviance residual = it is used to individuate influential points, namely data points that are unusual enough to have a big influence on the model's coefficients, that is the model's hazard rations in a cox model. Just like in linear regression one unusual point can dramatically affect the size of one or more hazard ratios.

As with other types of regression the cox model has some useful kinds of residuals to test modern fit and model assumptions. For cox these are:

- The Schoenfeld to test for the proportional hazard assumption
- The martingale to test for linear relations for continuous predictors
- The deviance to test for patients with unusual data that have a big effect of the hazard ratios.

B) Cox proportional hazards assumption

0) Premises

Assuming things that are not true can get you into trouble and the same holds with cox regression, however we can test cox assumptions.

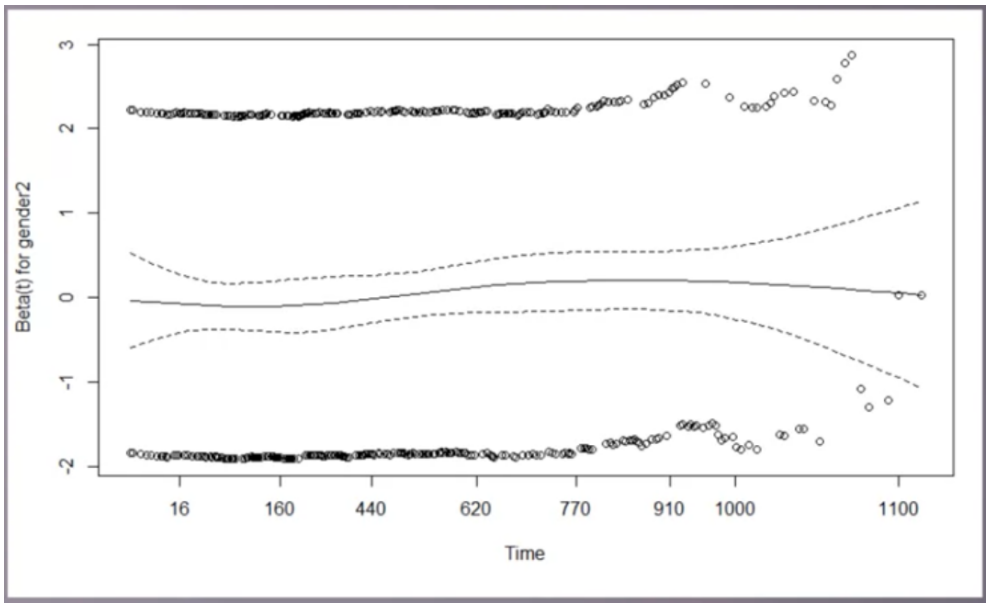
How can we test the most famous assumption of the cox model, namely proportional hazards -> the assumption that hazards are proportional for a given predictor, say gender, means that when you plot the hazards for females and that of males, the 2 lines are largely parallel. This means that for any point

in time we can multiply the hazard for females and obtain the hazard for males. If this relation holds then the hazard ratio for females compared to males is a good one number summary of the relation.

1) How can we test this?
- Graphically
- With a p-value = always useful

But always look also at the data. Schoenfeld residuals is the one used to check for the proportional hazards assumption in the cox model. In R coxph -> store it in an R object -> to this object apply the cox.zph function and store the result. Print the result to get the p value and plot the result to get the plot. With the p value we are looking to get a higher one and we can use the usual 0.05 cutoff: if the p value is > than 0.05 there is no strong evidence for a relation btw the residuals and time and this is a good thing: we hope that the model has correctly handled the relation btw gender and time so that there's nothing left off that relation for the residuals to spot. That is what we want to see.

In this course the outcome of interest is death following a patient's first admission for heart failure. This is what we gent when we test the assumption for gender in our heart failure data set.



This graph shows the residuals for gender plotted against time since admission for heart failure. Those are the small circles along the lines where y=2 and y=-2. The middle line is what we are interested in: it is a linear regression line through these residuals and the dotted lines by the side of it gives the confidence interval for the regression line. This line is pretty flat and centered on zero, and the p value is 0,275. This is testing for a non-zero slope for the linear regression line through the residuals. There is no strong evidence that the slope in this plot is not 0. There is also no apparent relation because the p value only looks for a linear relation so it can miss the non linear ones. This is why we need the plot too. The relation does not seem to change with time since hospital admission.

If we have several variables in the model the cox.zph function will repeat the above for each predictor variable.

We will see later what will happen if this assumption is not met -> things will get a little complicated.

2) Checking the proportionality assumption (guarda anche al codice R)

If the assumption is met then the hazard lines will be roughly parallel to each other; note that this is true iff they are plotted on the log scale, namely if we take the natural logarithm of the hazards or plot them on axes on the log scale. We will use the cox.zph function.

C) What to do if the proportionality assumption is not met

0) Premises

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

1) What does non-proportional hazards mean?

a) If the relation btw males and females regarding their risk of death changes over time, it could mean, for instance, that males have a higher risk pf 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 btw gender and time -> the model is short of a coefficient.
   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) we should do the statistical tests, namely get the p values, but also do the plots. Some kind of non-proportional relationships and other assumption violations can't be detected just from a p value.
   We want now to include the interaction term and test whether it is statistically significant. For mathematical reasons we can't just include the follow-up time itself as part of the interaction; we need first to transform it. The easiest way do to this in R is via the "tt" function (short for "time transform") -> see R code.

b) Code:
   The ouput agrees with the earlier approach and says that the interaction btw gender and transformed time is not statistically significant, namely there is no apparent violation of the proportionality assumption -> good news. The p value from this approach (circa 0,5) is not the same as that from the earlier one because the methods are different, though it is always nice when they give the same message!

c) If the assumption is violated, then
   i) one option is to include this interaction.
   ii) 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 2 TIME PERIODS. We can fit one model when things are fine, namely 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.
   iii) There is also a third simple way to deal with the problem: stratify the analysis by the variable causing the problems. If it is gender, for instance, then just fit separate models for males and females. The cons of this approach is that it is no longer possible to compare the effect of each gender on mortality.

   iv) Feedback on practice quiz

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

logistic regression the difference btw the model's predicted values and the actual values from the data are the residuals. Cox regression also generates residuals, and Schoenfeld are one type mentioned previously.

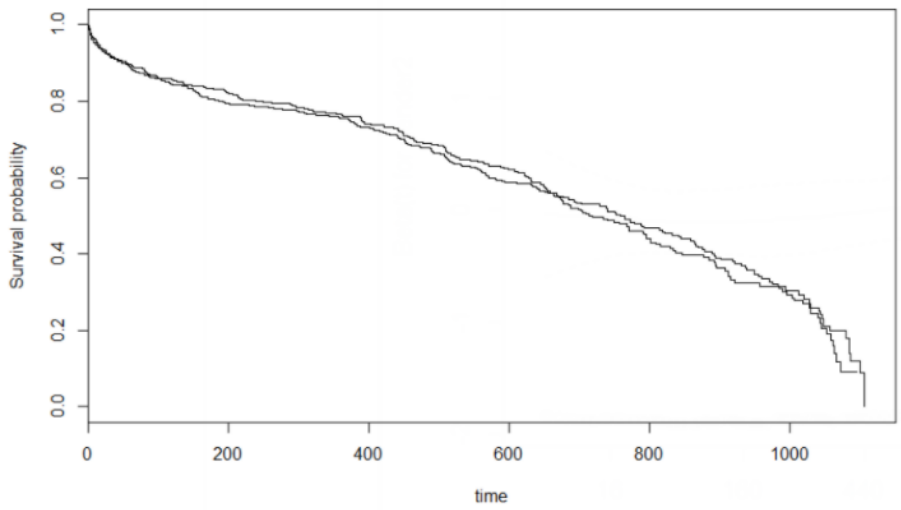
The relatively high p value from the test confirms what our eyes suggest from looking at the plot. The line is pretty flat, meaning that the effect of gender changes little during the follow-up period. That's good news.

A prettier version of the above plot can be obtained via the ggcoxzph() function, which produces Schoenfeld residuals against the transformed time for each covariate.

When we have a predictor with few categories, we can also use Kaplan-Meier plot as an informal visual check. If the predictor satisfies the proportional hazard assumption, then the graph of the survival function versus the survival time should yield parallel curves. This method does not work well for continuous predictors or categorical ones with many levels because the graph becomes too "cluttered".

I) KM plot for gender

```
km_fit <- survfit(Surv(fu_time, death) ~ gender)
autoplot(km_fit) or in alternative
plot(km_fit, xlab = "time", ylab = "Survival probability") # label the axes
```



The lines give the survival probability for each gender at each time point. The two lines are pretty parallel over time throughout the follow-up period, though in this case the fact that they are almost on top of one another makes it easy to judge.

II) Using the other types of residuals in Cox regression

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

a) Specifying the argument type="dfbeta" plots the estimated changes in the regression coefficients on deleting each observation (patient) in turn.

```
res.cox <- coxph(Surv(fu_time, death) ~ age)
```

```
ggcoxdiagnostics(res.cox, type = "dfbeta",
                      linear.predictions = FALSE, ggtheme = theme_bw())
```

b) Checking for outliers:

It is possible to check for outliers by visualizing the deviance residuals, which are normalized transformations of the martingale residual and should be roughly symmetrically distributed around zero with standard deviation of 1.

We recall that in the normal distribution, 5% of the observations are more than 1,96 standard deviations from the mean. So if the std is 1, then only 5% of the obs should be bigger than 1,96 or more negative than -1,96. If you have more than that proportion, then your model doesn't fit the data as well as it should and some observations are a problem. This is just the same issue as with the other types of 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.

- Positive values correspond to individuals that "died too soon" compared with expected survival times.
- Negative values correspond to individual that "lived too long" compared with expected survival times.
- Very large or small values are outliers, which are poorly

Code:

```
res.cox <- coxph(Surv(fu_time, death) ~ age)
ggcoxdiagnostics(res.cox, type = "deviance",
              linear.predictions = FALSE, ggtheme = theme_bw())
```

c) Do the continuous variables that we assume to have a linear relation with the outcome actually have a linear relation?

If for example we fit age as a single term in the model, then that's what you are assuming. The martingale residual is used to test this assumption via the ggcoxfunctional() function.

Martingale residuals may present any value between minus infinity and 1, and have a mean of zero:

- Martingale residuals close to 1 represent individuals that "died too soon",
- Large negative values correspond to individuals that "lived too long".

The plots should give a nice straight line if the assumption is valid.

D) How to choose predictors for a regression model

0) Premises

This was covered in the course of Logistic Regression for Public Health, where it was explained that some common ways of choosing predictors for a multiple regression model:

- Forwards and stepwise selection were simply too awful to contemplate,
- Backwards elimination does sometimes work ok.
- It is always good to use a priori knowledge but it is not always sufficient, particularly when we have a lot of possible variables.

Less often we will 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 this 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.

1) How to apply backwards elimination:
   i) Fit the model containing all your chosen predictors, either all you're a priori ones or all your available ones
   ii) Store all the coefficients from that model
   iii) Remove in one go all predictors whose p value is above the pre-defined threshold, typically the usual 0,05 or we could 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
   iv) Compare the coefficients for the remaining predictors with their coefficients from the original model.
2) Checks to make when using backwards elimination

If the coefficients haven't changed much from the original model, then we have our final model.

We can now check the residual and other model assumptions.

If, however, we have a predictor whose coefficient has changed noticeably then we need to find the variable(s) that we have removed that are correlated with this affected predictor. We can do this via trial and error, namely adding back in, one of the removed variables at a time until the affected predictor's coefficient is back to its original value and keep the removed variable in the model.

For example, say that blood pressure (HR=1.40 , p=0,002) was retained but cholesterol was removed because it was not statistically significant (HR=1.05, p=0,155). Then we removed cholesterol from the original model and the HR for blood pressure goes to 1.50. We consider this a big enough change to worry about and therefore we add cholesterol back in, restoring the original HR value for blood pressure.

3) How big is "big enough to worry about"?

It is typically arbitrary. Anything less than 0.05. It depends on how the results are going to be used, for example 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 ppl can be invited for screening for some disease based on their estimated risk of developing that disease, using the coef of 1.30 instead of 1.50 can greatly affect the number of ppl invited. In the latter such a difference won't be so important.

4) 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 big.

E) Practice in R: Running a Multiple Cox model

0) Premises

In this exercise we have to try the following methods:

- A priori knowledge only, using ingo from a report (link: https://www.ncbi.nlm.nih.gov/books/NBK513479/) to the funder of a grant

- Inclusion of all available predictors
- Backwards elimination following the inclusion of all available predictors
- For the last of these, check that the proportionality assumption is valid for each variable remaining in the model after the elimination process.

1) Objective 1

Model patient, primary care and hospital factors associated with re-admission and mortality for patients with heart failure and COPD. Combine the data set we have used up until now with primary care (practice level) and hospital level data on resources and performance.

2) First task

Review the relevant section of the report – the predictors of one year mortality for heart failure (Objective 1) – and pick the significant patient-level predictors from the model. In the report logistic regression rather than survival analysis was uses, as they were only interested in the fact of death within the year of follow up rather than the specific timing, but we can assume that the set of predictors is also likely to apply to our survival analysis model.

3) Next task

Look at the set of predictors and compare them with what we have in our data set. We might not have them all. We will notice that age was fitted in groups in the report, but we can try and use it also as a continuous variable and thereby assume a linear relation btw age and the hazard of mortality.

For the index length of hospital stay (LOS) and the number of previous missed appointments, we can just fit them as continuous variables too.

4) Fit the model

With what we have and see whether we can apply backwards elimination to reduce the number of predictors. Lastly we can check whether any of the coefficients (HRs) for the statistically significant predictors were affected by removing the non-significant ones. If they were, then decide whether we need to add them back in.

5) NOTES

The exercise raised various important issues that you may want to discuss with your peers. All these issues are extremely common in analysing public health data in general, and some are of course specific to survival analysis. Among the issues you could discuss are:

- Choosing the variables from the report to include in the model
- How to program them in R
- How to apply backwards elimination, e.g. p-value cut-offs
- How to choose the final set of predictors
- Any issues regarding testing the proportionality assumption
- Why your final set of predictors differs from those in the report

F) Results of the exercise on model selection and backwards elimination

1) A priori list of predictors

First task: read the relevant section of the report to find what predictors to study. This was table 6, "Odds ratios with 95% Cis for possible patient, trust and primary care predictors of mortality within 1 year of admission for HF and COPD patients". Ignoring the COPD results and the non-patient predictors the a priori list of predictors should comprise:

- Age
- Sex
- Ethnicity
- IHD (ischaemic heart disease)
- valvular disease
- PVD (peripheral vascular disease)
- prior stroke
- COPD
- pneumonia
- hypertension (had an inverse relation with mortality)
- renal disease
- cancer
- cancer with metastases
- mental health disorders
- cognitive impairment (senility and dementia combined)
- LOS
- Number of missed prior outpatient appointments

2) Preparing the data for Cox regression

It is necessary to read the documentation and summarize the distribution of the variables we haven't yet used. Most of them are simply binary, cognitive impairment needs to be created from a combination of 2 existing flags.

Now, how can we include each variable in the model. For binary variables we will generally want the zero to be the reference category. For the variables that can be considered continuous we should plot their relation with the outcome and decide whether we can just assume a linear relation or whether something more complicated is needed. Here assumption that linear is fine.

3) Interpretation of the R code:

First thig we look at are the standard errors, which are all comfortably low: not less than 0.1 and not big enough to worry about.

Then we look at the p values to see which predictors we could drop to make the final table.

Then backwards elimination.

Compare the 2 sets of coefficients (see table); the table uses hazard rations namely the numbers from the exp(coef) bit:

Predictor	Full Model HR	Reduced Model HR
Age	1.06	1.06
Gender	0.81	0.76
Valvular disease	1.21	1.27
Pneumonia	1.45	1.57
Metastatic cancer	8.98	12.20
Cognitive impairment	1.39	1.43

Descriptive statistics btw cognitive impairment and death shows that only 79 patients had cognitive impairment of whom 59 died. If we were doing logistic regression, we could convert this into an odds of death. The impaired patients are 3 times more likely to die than they are to survive. For those without such impairment, the risk of death is 47% and the odds are 433/488=0.89, which is pretty high. When we have other predictors the odds ratio will change, and because the odds for those without cognitive impairment are so high, it doesn't take much for the odds ratio for those with the impairment to change a lot.

Health, odds ratios can get very high for only small changes in the underlying probability. It's the same principle for hazard ratios from survival analysis. A change in the model (the removal of various covariates) that has little effect on HRs that are close to 1 – for example for age and gender – has a big effect on HRs that are far from 1, in this case much larger than 1. In summary, the change from an HR of 8.8 to 12.2 is not surprising and not important in this case.

4) Testing the proportionality assumption on the remaining variables (still R code)

We are left with six variables to test -> nice high p values, so all is well on that front.

5) Why this final model differs from that in the report?

Even though similar data and variables were used in the report and to create our model we still got different results, why? Various reasons:

- The data were similar but not the same
- The original set of variables was similar but not the same = the original set of variables in the project was larger than what we used.
- The statistical methods and outcome were different = the report used logistic regression with an outcome of mortality within one year of the admission, whereas in your survival analysis, the follow-up period was often much longer than a year. This will have an impact on the observed relations between each predictor and the outcome.
- The size of your data set was much smaller than in the report