6. Data exploration, model choice and model checking

September 1 – 25, 2015

Data exploration, model choice and model checking

- Exploratory data analysis
- Model choice
- Goodness-of-fit and model checking
- Model expansion through stratification

(1) Exploratory analysis

- It is about interrogating your data!
- Kaplan-Meier plots of survival function
  - stratified by different grouping variables (e.g. treatment vs. no treatment)
- Nelson–Aalen plots of cumulative hazards
  - in particular, when competing risk or multi-state models
- Preliminary checking of
  - parametric assumptions
  - proportionality assumptions

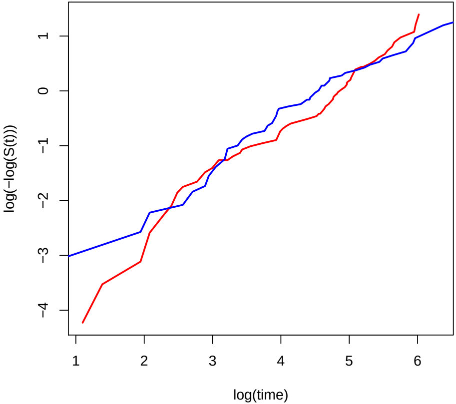
Exploring parametric assumptions

- Compare the non-parametric estimate of the cumulative hazard against its theoretical form under the asssumed parametric model
- For example, consider two groups (strata) under the Weibull regression model: $\log(-\log(\hat{S}(t)))$ should be a linear function of $\log(t)$ in both groups (with dummy covariates $Z = 0$ or $Z = 1$):

$$\begin{aligned}\log(-\log(S(t; Z, \theta)) &= \log(\Lambda(t; Z, \theta)) \\ &= \log((t/\alpha)^{\gamma} \exp(\beta Z)) \\ &= \gamma \log(t) - \gamma \log(\alpha) + \beta Z\end{aligned}$$

Example

- The log-log plots for the veteran data, stratified by 'treatment' (red = standard treatment; blue = experimental treatment). Ref. exercise 4 in practical 2.



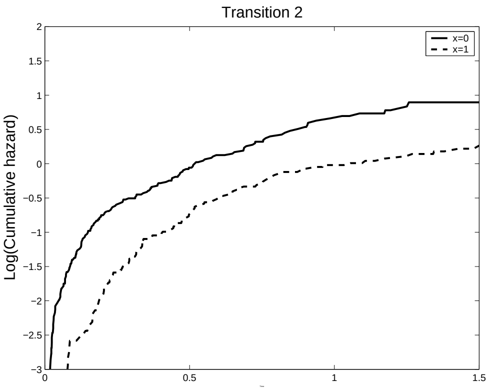
N.B. Under the Weibull model, the slopes should approximate γ and the distance of the two curves should approximate β (the log relative rate)

Exploring proportionality

- In general, compare non-parametric estimates of the cumulative hazard across different strata (of a categorical variable Z)
- If the proportional hazards model is appropriate, the curves for the different groups (strata) should be parallel and their (vertical) distance correspond to log relative rates
- For example, for two groups ($Z = 1$ or $Z = 0$):

$$\begin{aligned}\log(-\log(S(t; Z, \theta)) &= \log(\Lambda(t; Z, \theta)) \\ &= \log(\Lambda_0(t)\exp(\beta Z)) \\ &= \log(\Lambda_0(t)) + \beta Z\end{aligned}$$

Exploring proportionality: example



(2) Model choice

- Nested models can be compared by the likelihood ratio test:

$$-2\log\left(\frac{\sup_{\theta_1} L(\theta_1, \theta_2)}{\sup_{\theta} L(\theta)}\right) \sim \chi^2_q,$$

where L is the likelihood of the larger model with
a $(p+q)$-dimensional parameter vector $\theta = (\theta_1, \theta_2)$, and
θ_2 has dimension q (so q is the difference in the number of parameters)

Example

- The Weibull model reduces to the exponential model by choosing the shape parameter (γ) as 1
- The two models are thus nested. The difference of deviances has a χ^2_1 distribution:

$$-2\log\left(\frac{\sup_{\alpha} L(\alpha,1)}{\sup_{\alpha,\gamma} L(\alpha,\gamma)}\right) \sim \chi^2_1,$$

where L is the Weibull likelihood

- In R, nested models can (sometimes) be compared with the *anova* command from the output objects $c1$ and $c2$: *anova(c1,c2)*.
- See exercise 2 in practical 2.

Prediction or explanation?

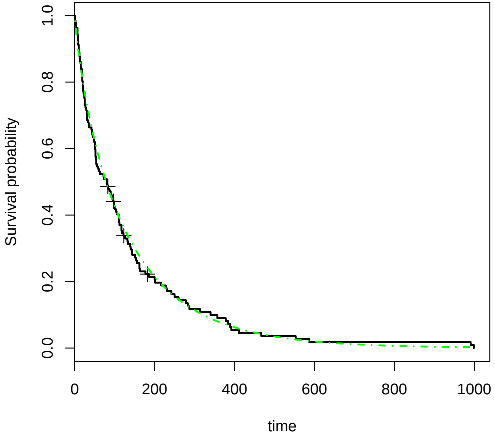
- If the ultimate aim of the analysis is prediction, rather than explanation, different information criteria for model selection may be used
- Akaike's Information Criterion (AIC)
  - AIC = -2*(log-likelihood) + 2*(number of parameters)
  - penalises models with too many parameters
  - the smaller, the better the model's predictive ability
  - command *extractAIC*(object) in R

(3) Model checking

- Statistical procedures for model selection do not (necessarily) tell how good the model fits the actual data
- So, after fitting a parametric model, the results should be checked against the observed data
- We here give three alternatives
  - inspection of the fitted survival function or cumulative hazard against their non-parametric estimates
  - inspection of fitted residuals
  - extension of the Cox proportional hazard model

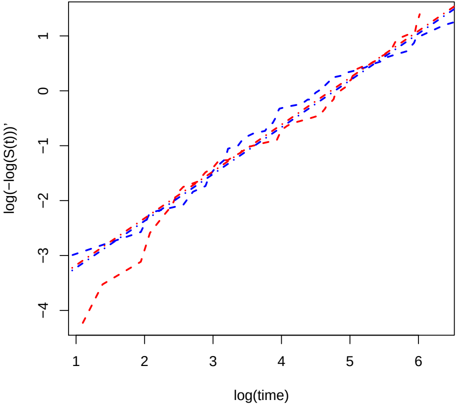
Kaplan-Meier vs. estimated survival

- Example: survival in the veteran data (KM vs the estimated survival under the Weibull model). Ref exercise 1 in practical 2.



Nelson–Aalen vs. the fitted model

- The log-log plot for the standard treatment vs. test treatment in the veteran data (non-parametric = dashed; estimated = dot-dashed)



Extension of the PH model

Allow the relative risk to vary with time. For example, consider

- $Z_1(t) = Z_1$ and $Z_2(t) = Z_1 t$. The model is

$$\lambda(t; Z_1, Z_2, \theta) \quad = \quad \lambda_0(t) \exp(\beta_1 Z_1 + \beta_2 Z_2),$$

  where β_2 measures the interaction between Z_1 and the time.
- Note that the relative risk of $Z_1 = 1$ to $Z_1 = 0$ is $\exp(\beta_1 + \beta_2 t)$, a smooth function of t.
- If $\beta_2 > 0$ then the relative risk function is increasing and $\beta_2 < 0$ then it is decreasing.

Extension of the PH model cont.

- $\beta_2 = 0$ corresponds to the proportional hazards or constant relative risk model.
- This extension can be used to test the proportionality assumption.

Unit exponentiality

- If random variable T has survival function $S(t)$, then $S(T) \sim \text{Uniform}[0, 1]$

and, equivalently,

$$\Lambda(T) = -\log(S(T)) \sim \text{Exp}(1)$$

- So, calculate "residuals":

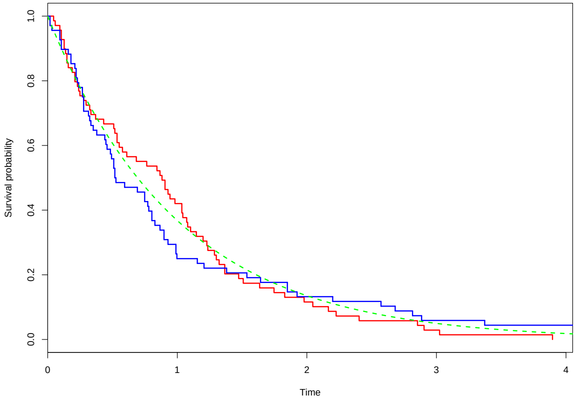
$$\hat{r}_i = \Lambda(t_i; \hat{\gamma}, \hat{\alpha}, Z_i), \;\; i = 1, ..., N$$

and check if these can be taken to arise from from the Exp(1) distribution.

- If the model is appropriate, $\hat{\Lambda}(t_i)$ are (appr.) samples from the exponential distribution with rate 1.

Example

- The Kaplan-Meier plot of the Λ-residuals in the veteran data.
- The models was Weibull regression with explanatory variable "treatment"
- The green curve shows the survival function of the Exp(1) distribution



(4) Model expansion through stratification

- So far we have assumed that the hazard rates in different subgroups (strata), defined by covariates (i.e. men and women), are *proportional* under
  - a parametric regression model
  - the Cox proportional hazards model
- Proportionality implies
  - a common *baseline* rate of failure, baseline referring to those with "baseline" values of the covariates, and
  - a multiplicative effect of covariates on the baseline
- If needed, how to expand the model through stratification?

Weibull regression with stratification

- Assuming the same shape parameter but different scale parameters, and the same effects of covariates across the strata, the stratum-specific hazard is defined as

$$\lambda_{is}(t_i; Z_i, \theta) = \alpha_s^{-1}\gamma_s(t/\alpha_s)^{\gamma_s - 1}\exp(\beta' Z_i),\ \ i = 1, ..., S$$

- A more general model with varying effects of covariates with strata:

$$\lambda_{is}(t_i; Z_i, \theta) = \alpha_s^{-1}\gamma_s(t/\alpha_s)^{\gamma_s - 1}\exp(\beta_s' Z_i),\ \ i = 1, ..., S$$

Stratified Cox analysis

- The model is now specified as

$$\lambda_{is}(t; Z_i, \theta) = \lambda_{0s}(t) \exp(\beta' Z_i)$$

N.B. In R, stratified analysis is obtained by option *strata*:

Surv(time,status)$\sim$ A + strata(B)