Relative risk regression models

Assume that we have a sample of *n* individuals, and let $N_i(t)$ count the observed occurrences of an event of interest for individual i as a function of (study) time *t*

We assume that the intensity process of $N_i(t)$ may be given as

$$\lambda_i(t) = \underbrace{Y_i(t)}_{\text{at risk indicator}} \cdot \underbrace{\alpha(t \mid \mathbf{x}_i)}_{\text{hazard rate (intensity)}}$$

The hazard rate for individual i takes the form

$$\alpha(t \mid \mathbf{x}_i) = \underbrace{\alpha_0(t)}_{\text{baseline hazard}} \underbrace{r(\boldsymbol{\beta}, \mathbf{x}_i(t))}_{\text{hazard ratio (relative risk)}}$$

where $\mathbf{x}_i(t) = (x_{i1}(t), x_{i2}(t), \ldots, x_{ip}(t))^T$ is a vector of (possibly) time-dependent covariates

We assume $r(\boldsymbol{\beta}, \mathbf{0}) = 1$, so the baseline hazard $\alpha_0(t)$ is the hazard for an individual with all covariates equal to zero

The common choice of relative risk function is

$$r(\boldsymbol{\beta}, \mathbf{x}_i(t)) = \exp\left(\boldsymbol{\beta}^T \mathbf{x}_i(t)\right) = \exp\left(\beta_1 x_{i1}(t) + \cdots + \beta_p x_{ip}(t)\right)$$

which gives Cox's regression model

e^{β_j} is the hazard ratio for one unit's increase in the *j*-th covariate, keeping all other covariates constant

Ordinary ML-estimation does not work for the relative risk regression models (due to the nonparametric baseline)

We estimate β by maximizing the partial likelihood

$$L(\boldsymbol{\beta}) = \prod_{T_j} \frac{r(\boldsymbol{\beta}, \mathbf{x}_{i_j}(T_j))}{\sum_{l \in \mathscr{R}_j} r(\boldsymbol{\beta}, \mathbf{x}_l(T_j))}$$

where $T_1 < T_2 < \cdots$ are the observed event times

and $\mathscr{R}_j = \{l \mid Y_l(T_j) = 1\}$ is the risk set at T_j

The maximum partial likelihood estimator $\widehat{\boldsymbol{\beta}}$ enjoys "the usual properties" of ML-estimators

Thus $\widehat{\boldsymbol{\beta}}$ is approximately multivariate normally distributed around the true value of $\boldsymbol{\beta}$ with a covariance matrix that may be estimated by $\mathbf{I}(\widehat{\boldsymbol{\beta}})^{-1}$, where

$$\mathbf{I}(\boldsymbol{\beta}) = \left\{ -\frac{\partial^2}{\partial \beta_h \partial \beta_j} \log L(\boldsymbol{\beta}) \right\}$$

is the observed information matrix

To test the null hypothesis $H_0: \beta_j = 0$, we may use the Wald test statistic

$$Z = \frac{\hat{\beta}_j}{se\left(\hat{\beta}_j\right)}$$

which is approximately standard normally distributed under the null hypothesis

To obtain a confidence interval for the hazard ratio e^{β_j} we transform the limits of the standard confidence interval for β_j to get the 95% confidence interval :

$$\exp\left\{\hat{\beta}_j \pm 1.96\, se(\hat{\beta}_j)\right\}$$

We may use the likelihood ratio statistic to test a composite null hypothesis, where on want to test the hypothesis that *r* of the regression coefficients are zero (or equivalently, after a reparameterization, that there are *r* linear restrictions among the regression coefficients)

If $\boldsymbol{\beta}^*$ is the maximum partial likelihood estimator under the null hypothesis, the likelihood ratio test statistic takes the form

$$\chi^2_{LR} = 2\left\{\log L(\widehat{\boldsymbol{\beta}}) - \log L(\boldsymbol{\beta}^*)\right\}$$

and it is approximately chi-squared distributed with *r* df under the null hypothesis

We may estimate the cumulative baseline hazard

$$A_0(t) = \int_0^t \alpha_0(u)du$$

by the Breslow estimator

$$\widehat{A}_0(t) = \sum_{T_j \le t} \frac{1}{\sum_{l \in \mathscr{R}_j} r(\widehat{\boldsymbol{\beta}}, \mathbf{x}_l(T_j))}$$

This is a generalization of the Nelson-Aalen estimator

Stratified models

So far we have assumed a common baseline hazard for all individuals, i.e.

$$\alpha(t \mid \mathbf{x}_i) = \alpha_0(t)\, r(\boldsymbol{\beta}, \mathbf{x}_i(t))$$

When this is not a realistic assumption, one may adopt a stratified version of the model

Then the study popolation is grouped into *k* strata, and for an individual in stratum *s* we assume that the hazard takes the form:

$$\alpha(t \mid \mathbf{x}_i, \text{stratum } s) = \alpha_{s0}(t)\, r(\boldsymbol{\beta}, \mathbf{x}_i(t))$$

Note that the effects of the covariates are assumed to be the same accross strata, while the baseline hazard may vary between strata

We now estimate β by maximizing the partial likelihood

$$L(\boldsymbol{\beta}) = \prod_{s=1}^{k} \prod_{T_{sj}} \frac{r(\boldsymbol{\beta}, \mathbf{x}_{i_j}(T_{sj}))}{\sum_{l \in \mathscr{R}_{sj}} r(\boldsymbol{\beta}, \mathbf{x}_l(T_{sj}))}$$

where $T_{s1} < T_{s2} < \cdots$ are the observed event times in stratum *s* and $\mathscr{R}_{sj}$ is the risk set in this stratum at time T_{sj}

The maximum partial likelihood estimator $\widehat{\boldsymbol{\beta}}$ enjoys similar properties as for the situation without stratification and statistical test may be performed as before

We may estimate the stratum-specific cumulative baseline hazards

$$A_{s0}(t)=\int_0^t \alpha_{s0}(u)du$$

by the Breslow estimators

$$\widehat{A}_{s0}(t)=\sum_{T_{sj}\le t}\frac{1}{\sum_{l\in\mathscr{R}_{sj}} r(\widehat{\boldsymbol{\beta}},\mathbf{x}_l(T_{sj}))}$$

As before these provide the basis for estimating cumulative hazards and survival functions for given values of fixed covariates (or given paths of external time-varying covariates)

Assumptions for Cox regression

We consider a Cox regression model with fixed covariates:

$$\alpha(t\,|\,\mathbf{x})=\alpha_0(t)\exp(\boldsymbol{\beta}^T\mathbf{x})$$

Note that the model assumes:

1) Log-linearity:

$$\log\{\alpha(t\,|\,\mathbf{x})\}=\log\{\alpha_0(t)\}+\boldsymbol{\beta}^T\mathbf{x}$$

2) Proportional hazards:

$$\frac{\alpha(t\,|\,\mathbf{x}_2)}{\alpha(t\,|\,\mathbf{x}_1)}=\exp\{\boldsymbol{\beta}^T(\mathbf{x}_2-\mathbf{x}_1)\}\qquad \text{(independent of time)}$$

We will indicate how these assumptions may be checked (this material is not in the ABG-book, cf page 134)

Check of log-linearity

We will check log-linearity for a numeric covariate, say covariate 1, assuming that log-linearity is ok for the remaining covariates

A simple way to do this is as follows:

- First we make a categorical variable by grouping into *k* groups according to the value of covariate 1
- Then we fit a model with separate effects for each group:

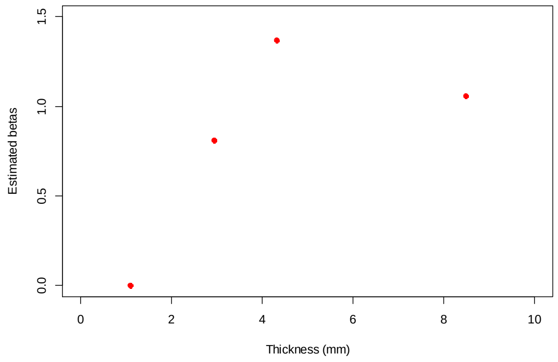
$$\alpha(t\,|\,\mathbf{x}_2,\text{group } g)=\alpha_0(t)\;e^{\beta_{1g}}\;\exp(\boldsymbol{\beta}_2^T\mathbf{x}_2)$$

$$\beta_{11}=0\quad\text{and}\quad\beta_{1g}\quad\text{are the effects og groups } g=2,\dots,k$$

$$\mathbf{x}_2=(x_2,x_3,\dots,x_p)^T\quad\text{and}\quad\boldsymbol{\beta}_2=(\beta_2,\beta_3,\dots,\beta_p)^T$$

- Then we plot the estimates $\hat{\beta}_{1g}$ versus a representative value x_{1g} for each group (e.g. midpoint or group mean) and see if we get a (fairly) straight line relationship

Melanoma data: Checking log-linearity by grouping tumor thickness in a model with sex and ulceration as the other covariates

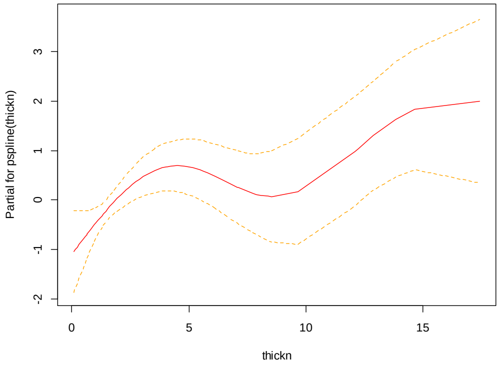


A more advanced method is to fit a penalized smoothing spline $s(x_1)$ for the effect of covariate 1 :

$$\alpha(t\,|\,\mathbf{x}) = \alpha_0(t)\exp\{s(x_1) + \boldsymbol{\beta}_2^T\mathbf{x}_2\}$$

and see if the spline estimate becomes fairly linear

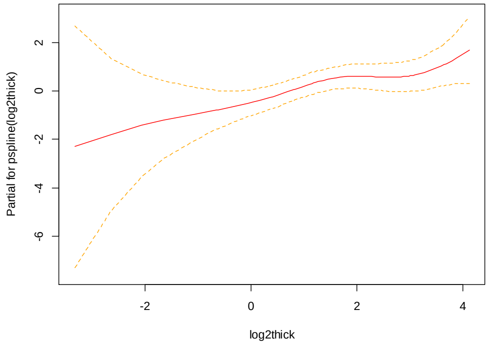
Melanoma data: Checking log-linearity by using a spline for tumor thickness in a model with sex and ulceration as the other covariates



When the effect of a numeric covariate is not log-linear, we may transform the covariate or use a grouped version of it (for "advanced" use we may work with splines)

For the melanoma data, the plots indicate that we may use log-thickness as covariate (and then log2 is a good choice)

Melanoma data: Checking log-linearity by using a spline for log2 of tumor thickness in a model with sex and ulceration as the other covariates



Graphical check of proportional hazards

We will check proportionality for one covariate, say covariate 1, assuming that poportionality is ok for the remaining covariates

If proportionality is ok for covariate 1, we have:

$$\alpha(t\,|\,\mathbf{x}) = \alpha_0(t)\,e^{\beta_1 x_1}\exp(\boldsymbol{\beta}_2^T\mathbf{x}_2)$$

where $\mathbf{x}_2 = (x_2, x_3, \ldots, x_p)^T$ and $\boldsymbol{\beta}_2 = (\beta_2, \beta_3, \ldots, \beta_p)^T$

To check the assumption, we make *k* strata based on the value of x_1 (after grouping a numerical covariate):

$$\alpha(t\,|\,\mathbf{x}_2, \text{stratum } s) = \alpha_{s0}(t)\exp(\boldsymbol{\beta}_2^T\mathbf{x}_2)$$

If proportionality is ok, we have

$$\alpha_{s0}(t) = \alpha_0(t)\,e^{\beta_1 x_1(\text{stratum } s)}$$

Thus if proportionality is ok, we have the following relation between the cumulative hazards

$$A_{s0}(t) = A_0(t)\,e^{\beta_1 x_1(\text{stratum } s)}$$

which implies

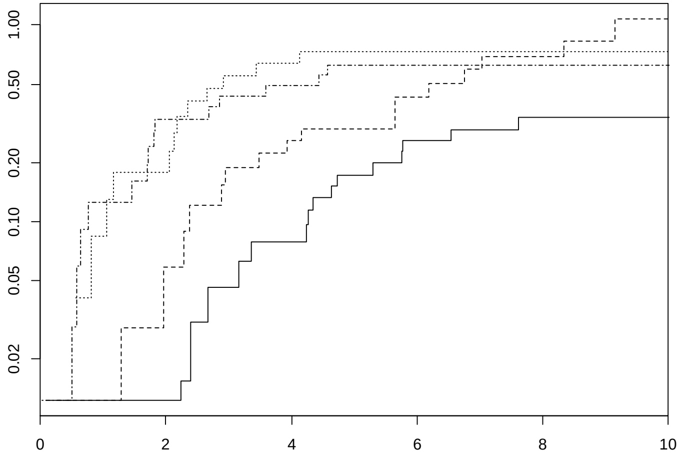
$$\log A_{s0}(t) = \log\{A_0(t)\} + \beta_1\, x_1(\text{stratum } s)$$

To check proportionality, we may fit a stratified Cox model and plot

$$\log \hat{A}_{s0}(t) \quad \text{versus time } t \text{ for each stratum } s$$

The plots should be (fairly) parallel if proportionality is ok

Melanoma data: Checking proportionality of tumor thickness in a model with sex and ulceration as the other covariates



Test of proportional hazards

One way to obtain a formal test for proportional hazard is to fit a model of the form

$$\alpha(t\,|\,\mathbf{x}) =$$

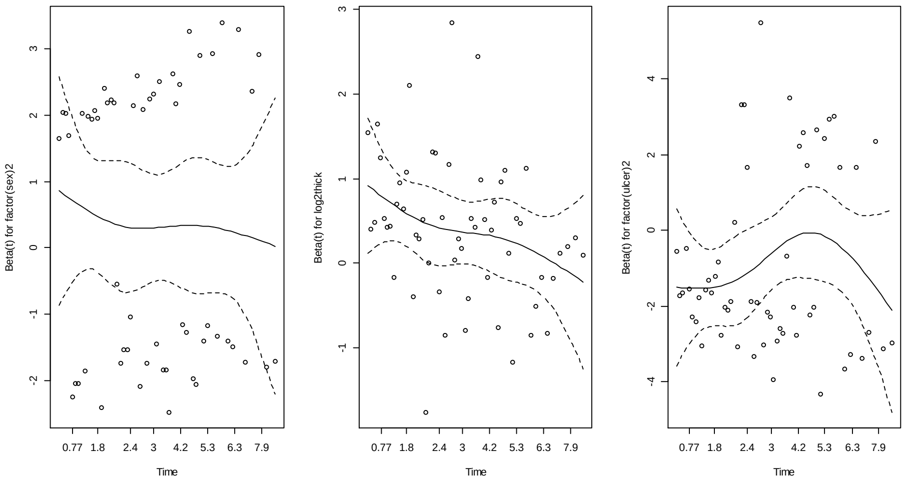
$$\alpha_0(t)\exp\left\{\beta_{11}x_{i1} + \beta_{12}x_{i1}g(t) + \cdots + \beta_{p1}x_{ip} + \beta_{p2}x_{ip}g(t)\right\}$$

for a known function $g(t)$, e.g. $g(t) = \log t$

We then test the null hypotesis that one or all of $\beta_{j2} = 0$

	chisq	p
factor(sex)2	0.230	0.6312
log2thick	4.022	0.0449
factor(ulcer)2	0.956	0.3283
GLOBAL	8.773	0.0325

Melanoma data: Plots that indicate possible time dependent effects of the covariates



Graphical check of global model fit

One way of preforming a global model check is by means of the grouped martingale residual processes described in section 4.1.3 in the ABG-book (which is not part of the curriculum)

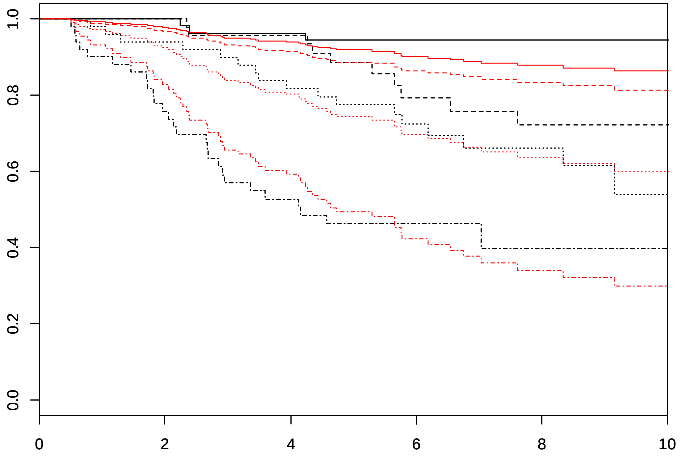
Here we describe another graphical check for survival data that may easily be implemented in R

The check is performed by grouping the individuals into *k* groups according to the prognostic index

$$\hat{\eta}_i = \hat{\boldsymbol{\beta}}^T\mathbf{x}_i$$

and then compare the Kaplan-Meier estimates for these groups with the fitted survival curves obtained from the fitted Cox model (computed for the mean value of the prognostic index in each group)

Illustration for melanoma data with sex, log tumor thickness and ulceration as covariates:



0.0
0.2
0.4
0.6
0.8
1.0
0
2
4
6
8
10