## 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

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

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

(the time-dependency of the covariates is suppressed in the above notation)

Throughout we will assume that the covariate processes are predictable

This implies that:

- fixed covariates should be measured in advance (i.e. at time zero) and remain fixed throughout the study
- the values at time *t* of time-dependent covariates should be known "just before" time *t*

In most of our applications, we will restrict attention to fixed covariates

For the time-dependent covariates it is useful to distinguish between external (or exogenous) and internal (or endogenous) covariates

Examples of external time-dependent covariates are:

- Defined time-dependent covariates: the complete covariate path is given at the outset of the study (e.g. a person's age at study time *t* )
- Ancillary time-dependent covariates: the path of a stochastic process that is not influenced by the event being studied (e.g. observed level of air pollution)

(Fixed covariates are also external)

Time-dependent covariates that are not external, are called internal

One example is biochemical markers measured for the individuals during follow-up

## Relative risk regression models

Assume that the hazard rate for individual $i$ takes the form

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

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

We make no assumptions of the form of the baseline hazard

Thus the model contains a nonparametric part (the baseline hazard) and a parametric part (the relative risk function)

We say that the model is semiparametric

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

Consider two individuals, indexed 1 and 2, and assume that all components of $\mathbf{x}_1(t)$ and $\mathbf{x}_2(t)$ are equal, except the *j*-th component where $x_{2j}(t) = x_{1j}(t) + 1$

Then:

$$\frac{\alpha(t \mid \mathbf{x}_2)}{\alpha(t \mid \mathbf{x}_1)} = \frac{\alpha_0(t) \exp\left(\boldsymbol{\beta}^T \mathbf{x}_2(t)\right)}{\alpha_0(t) \exp\left(\boldsymbol{\beta}^T \mathbf{x}_1(t)\right)} = \exp\left\{\boldsymbol{\beta}^T \left(\mathbf{x}_2(t) - \mathbf{x}_1(t)\right)\right\} = e^{\beta_j}$$

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

Other possible choices of the relative risk function are:

- The additive risk function:

$$r(\boldsymbol{\beta}, \mathbf{x}_i(t)) = 1 + \boldsymbol{\beta}^T \mathbf{x}_i(t)$$

- The excess relative risk function:

$$r(\boldsymbol{\beta}, \mathbf{x}_i(t)) = \prod_{j=1}^{p} \{1 + \beta_j x_{ij}(t)\}$$

Cox regression is the only relative risk regression model implemented in R

## Partial likelihood and estimation of $\beta$

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

Instead we have to use a partial likelihood

We will se how this may be derived

The intensity process of $N_i(t)$ is given as

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

The intensity process of the aggregated counting process $N_{\bullet}(t) = \sum_{l=1}^{n} N_l(t)$ takes the form (assuming no joint events)

$$\lambda_{\bullet}(t) = \sum_{l=1}^{n} \lambda_l(t) = \sum_{l=1}^{n} Y_l(t)\, \alpha_0(t)\, r(\boldsymbol{\beta}, \mathbf{x}_l(t))$$

We consider the conditional probability of observing an event for individual $i$ at time *t* , given the past and given that an event is observed at time *t* :

$$\pi(i \mid t) = P(dN_i(t) = 1 \mid dN_{\bullet}(t) = 1, \mathcal{F}_{t-})$$

$$= \frac{P(dN_i(t) = 1 \mid \mathcal{F}_{t-})}{P(dN_{\bullet}(t) = 1 \mid \mathcal{F}_{t-})} = \frac{\lambda_i(t)}{\lambda_{\bullet}(t)}$$

$$= \frac{Y_i(t)\, r(\boldsymbol{\beta}, \mathbf{x}_i(t))}{\sum_{l=1}^{n} Y_l(t)\, r(\boldsymbol{\beta}, \mathbf{x}_l(t))}$$

Then the intensity process of $N_i(t)$ may be factorized as

$$\lambda_i(t) = \lambda_{\bullet}(t)\, \pi(i \mid t)$$

We obtain the partial likelihood by multiplying together the conditional probabilites $\pi(i \mid t)$ over all observed event times $T_1 < T_2 < \cdots$ (thereby disregarding the information on the regression coefficients contained in the aggregated process)

Then, if $i_j$ is the index of the individual who experiences an event at $T_j$ , the partial likelihood becomes

$$L(\boldsymbol{\beta}) = \prod_{T_j} \pi(i_j \mid T_j) = \prod_{T_j} \frac{Y_{i_j}(T_j)\, r(\boldsymbol{\beta}, \mathbf{x}_{i_j}(T_j))}{\sum_{l=1}^{n} Y_l(T_j)\, r(\boldsymbol{\beta}, \mathbf{x}_l(T_j))}$$

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

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

We will show (later) that 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

For general relative risk functions it may be better to use the expected information matrix. But as this coincides with the observed information matrix for Cox regression, we will not go into these details (cf. section 4.1.5)

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 $e^{\beta_j}$ we transform the limits of the standard confidence interval for $\beta_j$ to get the 95% confidence interval :

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

To test the simple null hypothesis $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$ for a specified value of $\boldsymbol{\beta}_0$ (typically $\boldsymbol{\beta}_0 = \mathbf{0}$ ) we may apply the usual likelihood based tests statistics:

- The likelihood ratio test statistic:

$$\chi^2_{LR} = 2\left\{\log L(\widehat{\boldsymbol{\beta}}) - \log L(\boldsymbol{\beta}_0)\right\}$$

- The score test statistic:

$$\chi^2_{SC} = \mathbf{U}(\boldsymbol{\beta}_0)^T\, \mathbf{I}(\boldsymbol{\beta}_0)^{-1} \mathbf{U}(\boldsymbol{\beta}_0)$$

where $\mathbf{U}(\boldsymbol{\beta}) = \frac{\partial}{\partial \beta} \log L(\boldsymbol{\beta})$ is the vector of score functions

- The Wald test statistic:

$$\chi^2_W = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^T\, \mathbf{I}(\hat{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$$

All the test statistics are approximately chi-squared distributed with $p$ df under the null hypothesis

All the tests may be generalized to 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)

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

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

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

## Estimation of cumulative hazards and survival probabilities

We will estimate the cumulative baseline hazard

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

We take the aggregated counting process $N_{\bullet}(t) = \sum_{l=1}^n N_l(t)$ as our starting point

Its intensity process is given by

$$\lambda_{\bullet}(t) = \sum_{l=1}^n Y_l(t)\,\alpha_0(t)\,r(\boldsymbol{\beta},\mathbf{x}_l(t)) = \left(\sum_{l=1}^n Y_l(t)\,r(\boldsymbol{\beta},\mathbf{x}_l(t))\right)\alpha_0(t)$$

If we had known $\beta$, this would have been an example of the multiplicative intensity model

For a given value of $\beta$ , we may therefore estimate $A_0(t)$ by

$$\widehat{A}_0(t;\boldsymbol{\beta}) = \int_0^t \frac{dN_{\bullet}(u)}{\sum_{l=1}^n Y_l(u)\,r(\boldsymbol{\beta},\mathbf{x}_l(u))}$$

Since $\boldsymbol{\beta}$ is unknown, we replace it by $\widehat{\boldsymbol{\beta}}$ to obtain the Breslow estimator:

$$\widehat{A}_0(t) = \int_0^t \frac{dN_{\bullet}(u)}{\sum_{l=1}^n Y_l(u)\,r(\widehat{\boldsymbol{\beta}},\mathbf{x}_l(u))}$$

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

If all covariates are fixed, the cumulative hazard corresponding to an individual with a given covariate vector $\mathbf{x}_0$ is

$$A(t\,|\,\mathbf{x}_0) = \int_0^t \alpha(u\,|\,\mathbf{x}_0)\,du = r(\beta,\mathbf{x}_0)\,A_0(u)$$

and it may be estimated by

$$\widehat{A}(t\,|\,\mathbf{x}_0) = r(\widehat{\boldsymbol{\beta}},\mathbf{x}_0)\,\widehat{A}_0(t)$$

For a given path $\mathbf{x}_0(s);\, 0 < s \le t$ of an external time-dependent covariate, the cumulative hazard

$$A(t\,|\,\mathbf{x}_0) = \int_0^t r(\boldsymbol{\beta},\mathbf{x}_0(u))\,\alpha_0(u)\,du$$

may be estimated by

$$\widehat{A}(t\,|\,\mathbf{x}_0) = \int_0^t r(\widehat{\boldsymbol{\beta}},\mathbf{x}_0(u))\,d\widehat{A}_0(u) = \sum_{T_j \le t} \frac{r(\widehat{\boldsymbol{\beta}},\mathbf{x}_0(T_j))}{\sum_{l \in \mathscr{R}_j} r(\widehat{\boldsymbol{\beta}},\mathbf{x}_l(T_j))}$$

The corresponding survival function is given by

$$S(t \mid \mathbf{x}_0) = \mathop{\pi}_{u \leq t} \{1 - dA(u \mid \mathbf{x}_0)\}$$

and it may be estimated by

$$\widehat{S}(t \mid \mathbf{x}_0) = \mathop{\pi}_{u \leq t} \left\{1 - d\widehat{A}(u \mid \mathbf{x}_0)\right\} = \prod_{T_j \leq t} \left\{1 - \triangle\widehat{A}(T_j \mid \mathbf{x}_0)\right\}$$

Alternatively we may use (as is done in R):

$$\tilde{S}(t \mid \mathbf{x}_0) = \exp\left\{-\hat{A}(t \mid \mathbf{x}_0)\right\}$$

For practical purposes there is little difference between the two estimators

The estimators of the cumulative hazards and survival functions are approximately normal and their variances may be estimated as described in section 4.1.6 (which is not part of the curriculum)