Regression Models for Survival Data

- We restrict attention to proportional hazards models:
  - Parametric models: The Weibull-Model
  - Semiparametric models: The Cox-Model
- Inference via Maximum Likelihood
  - Likelihood function for survival data

The Weibull-Distribution

- Survival time T is a positive random variable with density function

$$f(t;\mu,\alpha) = \frac{\alpha}{\mu}\left(\frac{t}{\mu}\right)^{\alpha-1} \exp\left(-\left(\frac{t}{\mu}\right)^{\alpha}\right)$$

where $\mu > 0$ and $\alpha > 0$.
- Special case $\alpha = 1$: exponential distribution with mean μ
- General property:

$$\mathrm{E}(T) \quad = \quad \mu \cdot \Gamma(1 + 1/\alpha)$$

Survivor and Hazard-function of the Weibull distribution

- The distribution function of the Weibull distribution is

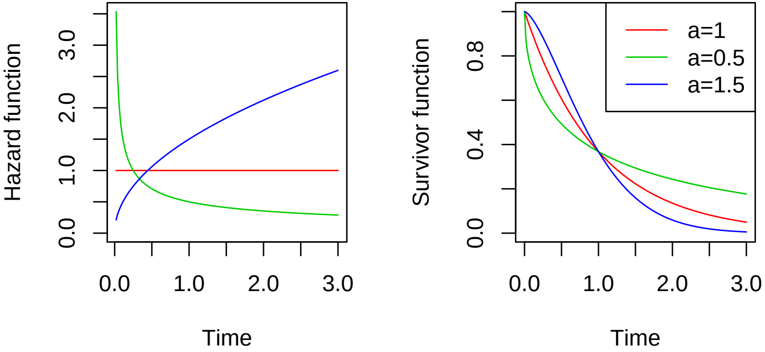
$$F(t;\mu,\alpha) = \mathrm{P}(T \le t;\mu,\alpha) = 1 - \exp\left(-\left(\frac{t}{\mu}\right)^{\alpha}\right)$$

  The survivor function is simply $S(t) = 1 - F(t)$

- From $h(t) = f(t)/S(t)$ it follows that

$$h(t;\mu,\alpha) = \frac{\alpha}{\mu}\left(\frac{t}{\mu}\right)^{\alpha-1}$$

Typical Weibull Hazard Functions and Corresponding Survivor Functions



The Likelihood Function for Survival Data

- Independent observations (t_i, δ_i), $i = 1, \dots, n$ with
  - survival time t_i
  - censoring indicator

$$\delta_i = \begin{cases} 1 & \text{if } i\text{-th observation is not censored} \\ 0 & \text{if } i\text{-th observation is censored} \end{cases}$$

- Let θ denote the unknown parameters. The likelihood function is

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{n} \left\{ f(t_i)^{\delta_i} (S(t_i))^{1-\delta_i} \right\} \\ &= \prod_{i=1}^{n} \left\{ \left(\frac{f(t_i)}{S(t_i)} \right)^{\delta_i} S(t_i) \right\} \\ &= \prod_{i=1}^{n} \left\{ h(t_i)^{\delta_i} S(t_i) \right\} \end{aligned}$$

The Weibull Proportional Hazards Model

- Reparametrize the Weibull model using $\lambda = \mu^{-\alpha}$, then

$$h(t) = \lambda \alpha t^{\alpha - 1} \text{ and } S(t) = \exp(-\lambda t^{\alpha}).$$

- Now incorporate covariates $\mathbf{x}_i$ in the hazard function:

$$h_i(t; \mathbf{x}_i) = \exp(\mathbf{x}_i^T \boldsymbol{\beta}) \cdot \lambda \alpha t^{\alpha - 1}$$

- The model assumes that individuals i and j with covariates $\mathbf{x}_i$ and $\mathbf{x}_j$ have proportional hazard functions:

$$\frac{h_i(t; \mathbf{x}_i)}{h_j(t; \mathbf{x}_j)} = \frac{\exp(\mathbf{x}_i^T \boldsymbol{\beta})}{\exp(\mathbf{x}_j^T \boldsymbol{\beta})} = \exp((\mathbf{x}_i - \mathbf{x}_j)^T \boldsymbol{\beta})$$

- The quantities $\exp(\beta_i)$ can be interpreted as hazard ratios.

Weibull-Regression in R

Function `survreg` in library `survival`:

```
> library(survival)
> m1 <- survreg(Surv(time, d) ~ cenc0, data = pbc, dist = "weibull")
> print(summary(m1))
```

```
Call:
survreg(formula = Surv(time, d) ~ cenc0, data = pbc, dist = "weibull")
            Value Std. Error     z        p
(Intercept)  8.08     0.1116 72.41 0.00e+00
cenc0       -1.12     0.2157 -5.21 1.90e-07
Log(scale)  -0.12     0.0864 -1.39 1.65e-01

Scale= 0.887

Weibull distribution
Loglik(model)= -848.1   Loglik(intercept only)= -859.2
Chisq= 22.23 on 1 degrees of freedom, p= 2.4e-06
Number of Newton-Raphson Iterations: 5
n= 184
```

Interpretation of Parameters

- Attention: A different parametrization is used with intercept ν, scale parameter σ and covariate effects γ_j
- Relationship to original parametrization:

$$\begin{aligned} \beta_j &= -\gamma_j/\sigma \\ \alpha &= \sigma^{-1} \\ \mu &= \exp(\nu) \end{aligned}$$

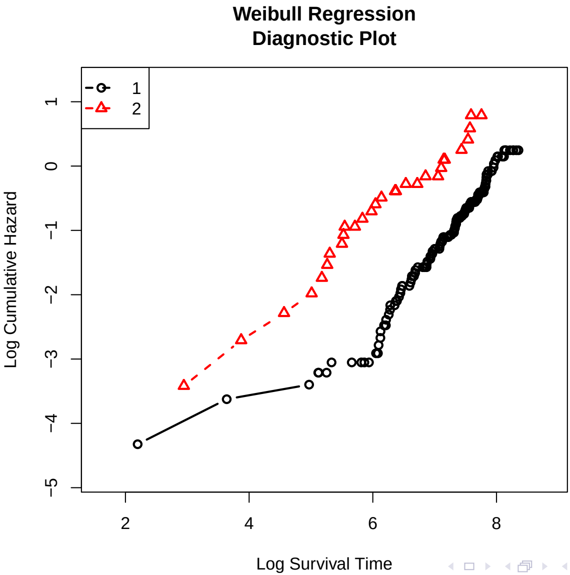
- Standard error via Delta-Rule
- In the example we obtain $\hat{\mu} = 3243.19$, $\hat{\sigma} = 0.89$ and $\hat{\beta} = 1.27$, so the estimated hazard ratio is $\exp(\hat{\beta}) = 3.55$.

A Convenience Function

```
> library(biostatZH)
> m1b <- WeibullReg(Surv(time, d) ~ cenc0, data = pbc)
> print(m1b)
```

```
$formula
Surv(time, d) ~ cenc0

$coef
           Estimate           SE
lambda 0.0001099469 8.274432e-05
alpha  1.1275557227 9.741921e-02
cenc0  1.2671884994 2.407996e-01

$HR
            HR       LB       UB
cenc0 3.550855 2.214950 5.692485

$ETR
            ETR        LB        UB
cenc0 0.3250304 0.2129504 0.4961002

$summary

Call:
survreg(formula = formula, data = data, dist = "weibull")
            Value Std. Error     z        p
(Intercept)  8.08     0.1116 72.41 0.00e+00
cenc0       -1.12     0.2157 -5.21 1.90e-07
Log(scale)  -0.12     0.0864 -1.39 1.65e-01

Scale= 0.887

Weibull distribution
```

Checking Proportional Hazards



The Event Time Ratio

- The p-th quantile of a Weibull-distributed random variable is

$$t_p = \mu(-\log p)^{\frac{1}{\alpha}}.$$

- The log event time ratio (ETR) for an individual with covariates $\mathbf{x}_i$ relative to an individual with covariates $\mathbf{x}_j$ is $\gamma^T(\mathbf{x}_i - \mathbf{x}_j)$.
- In the case of just two treatments, the log event time ratio is simply given by γ.
- Note: the ETR is not just the reciprocal of the hazard ratio HR, since HR also depends on the scale parameter.

A more general model

```
> library(biostatZH)
> m2 <- WeibullReg(Surv(time, d) ~ cenc0 + treat, data = pbc)
> print(m2)
$formula
Surv(time, d) ~ cenc0 + treat

$coef
           Estimate           SE
lambda 0.0001001886 7.670051e-05
alpha  1.1302715611 9.774858e-02
cenc0  1.2664547760 2.406032e-01
treat  0.1478776910 2.043361e-01

$HR
            HR        LB       UB
cenc0 3.548251 2.2141779 5.686121
treat 1.159371 0.7767678 1.730429

$ETR
            ETR        LB        UB
cenc0 0.3261209 0.2137739 0.4975107
treat 0.8773636 0.6156946 1.2502414

$summary
```

Interpretation

- For fixed Cholestase group, the hazard rate in the treatment group is 1.16 (95% CI: (0.78,1.73))
- For fixed Cholestase group, the event time ratio in the treatment group is 0.88 (95% CI: (0.62,1.25))

Extensions

- There are alternative parametric models within the more general framework of accelerated failure time models.
- In `survreg` available:
  - `dist="gaussian"`
  - `dist="logistic"`
  - `dist="lognormal"`
  - `proportional odds model: dist="loglogistic"`
- Details in D. Collett: *Modelling Survival Data in Medical Research*, Chapter 6, 2nd Edition, Chapman & Hall

The Cox-Model

- Similar to the Weibull model, the semiparametric Cox-Model assumes proportional hazards

$$h_i(t; \mathbf{x}_i) = \exp(\mathbf{x}_i^T \boldsymbol{\beta}) \cdot h_0(t)$$

  but lets the baseline hazard function $h_0(t)$ completely unspecified.
- A conditional likelihood approach allows to estimate the coefficients $\boldsymbol{\beta}$, no matter which form $h_0(t)$ has.

The Likelihood in the Cox-Model

- Idea: Consider all non-censored events with (ordered) time points $t_{(1)}, \dots, t_{(r)}$.
- Let $R(t_{(j)})$ denote the set of individuals which are under risk at time $t_{(j)}$.
- The conditional probability, that the j-th individual dies at time $t_{(j)}$ is then

$$\frac{h_j(t_{(j)}; \mathbf{x}_{(j)})}{\sum_{i \in R(t_{(j)})} h_i(t_{(j)}; \mathbf{x}_i)} = \frac{\exp(\mathbf{x}_{(j)}^T \beta)}{\sum_{i \in R(t_{(j)})} \exp(\mathbf{x}_i^T \beta)}$$

- Identical to the conditional likelihood contribution in matched case control studies!
- The likelihood function is the product over $j = 1, \dots, r \rightarrow$ numerical optimization

Cox-Regression in R

Function coxph in library survival:

```
> m3 <- coxph(Surv(time, d) ~ cenc0, data = pbc)
> print(summary(m3))

Call:
coxph(formula = Surv(time, d) ~ cenc0, data = pbc)

  n= 184

        coef exp(coef) se(coef)    z Pr(>|z|)
cenc0 1.3231    3.7550   0.2455 5.39 7.04e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

      exp(coef) exp(-coef) lower .95 upper .95
cenc0     3.755     0.2663     2.321     6.075

Rsquare= 0.12   (max possible= 0.991 )
Likelihood ratio test= 23.51  on 1 df,   p=1.243e-06
Wald test            = 29.05  on 1 df,   p=7.037e-08
Score (logrank) test = 33.36  on 1 df,   p=7.657e-09
```

Ties

- For identical survival times (ties) the likelihood contribution is more complex, just as in matched case-control studies with more than one case per stratum.
- A fast approximate method is the default method for ties in `coxph`, but an exact method is also available.

```
Call:
coxph(formula = Surv(time, d) ~ cenc0, data = pbc, method = "exact")

  n= 184

        coef exp(coef) se(coef)     z Pr(>|z|)
cenc0 1.3232    3.7555   0.2456 5.388 7.11e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

      exp(coef) exp(-coef) lower .95 upper .95
cenc0     3.756     0.2663     2.321     6.077

Rsquare= 0.12   (max possible= 0.991 )
Likelihood ratio test= 23.5  on 1 df,   p=1.249e-06
Wald test            = 29.03  on 1 df,   p=7.114e-08
Score (logrank) test = 33.33  on 1 df,   p=7.767e-09
```

A More General Model

```
Call:
coxph(formula = Surv(time, d) ~ cenc0 + treat, data = pbc, method = "exact")

  n= 184

        coef exp(coef) se(coef)     z Pr(>|z|)
cenc0 1.3246    3.7606   0.2454 5.397 6.79e-08 ***
treat 0.1625    1.1765   0.2056 0.790    0.429
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

      exp(coef) exp(-coef) lower .95 upper .95
cenc0     3.761     0.2659    2.3245     6.084
treat     1.176     0.8500    0.7862     1.760

Rsquare= 0.123   (max possible= 0.991 )
Likelihood ratio test= 24.13  on 2 df,   p=5.771e-06
Wald test            = 29.69  on 2 df,   p=3.573e-07
Score (logrank) test = 33.97  on 2 df,   p=4.195e-08
```

Final Comments

- An estimation of the baseline hazard function $h_0(t)$ is possible after estimating the regression coefficients β, but typically not of main interest. Alternatively the cumulative baseline hazard function or the baseline survivor function can be obtained.
- There are several techniques (graphical, tests, residual analysis) to check the proportional hazards assumption.
- Time-varying covariates can also be considered.