

Formelsammlung zur Vorlesung Lebensdaueranalyse

erstellt von Susanne Konrath

1 Einführung

1.1 Hazardrate und Survivalfunktion

- $T \geq 0$ stetig mit

Verteilungsfunktion $F(t) = P(T \leq t),$

Dichte $f(t) = F'(t)$

- Survivalfunktion

$$S(t) = P(T \geq t) = 1 - F(t)$$

- Hazardrate

$$\lambda(t) = \lim_{\Delta t \rightarrow 0} \frac{1}{\Delta t} P(t \leq T < t + \Delta t | T \geq t)$$

- Kumulative Hazardrate

$$\Lambda(t) = \int_0^t \lambda(u) du$$

- Zusammenhänge

$$\lambda(t) = \frac{f(t)}{1 - F(t)} = \frac{f(t)}{S(t)}$$

$$S(t) = \exp(-\Lambda(t)) = \exp\left(-\int_0^t \lambda(u) du\right)$$

- Exponential-Verteilung $\text{Ex}(\lambda)$:

$$\lambda(t) = \lambda$$

- Weibull-Verteilung $\text{We}(\lambda, \alpha)$:

$$\lambda(t) = \alpha \lambda^\alpha t^{\alpha-1}$$

- Log-logistische Verteilung $\text{LL}(\lambda, \alpha)$:

$$\lambda(t) = \frac{\alpha \lambda t^{\alpha-1}}{1 + \lambda t^\alpha}$$

- Diskrete Verweildauer T_d :

- Zeitachse zerlegt in Intervalle $[a_0, a_1), \dots, [a_{k-1}, a_k), \dots, [a_q, a_{q+1} = \infty)$

$$T_d = t_k \Leftrightarrow T \in [a_{k-1}, a_k)$$

- Diskrete Hazardrate/Hazardfunktion

$$\lambda_k = P(T_d = t_k | T_d \geq t_{k-1}) = P(T \in [a_{k-1}, a_k) | T \geq a_{k-1})$$

1.1.1 (Rechts-) Zensierung

- Modell 1 (Typ I-Zensierung)

Für jedes Individuum i , $i = 1, \dots, n$ ist fester (deterministischer) Beobachtungszeitraum $C_i = c_i$ vorgegeben. Beobachtbar ist $t_i = \min(T_i, c_i)$.

- Modell 2 (Typ II-Zensierung)

Untersuchung beendet, wenn eine vorab festgelegte Zahl von Lebensdauern T_i unzensiert beobachtet wurde.

- Modell 3 (Random Censoring)

C_i i.i.d. Zensierungszeiten, unabhängig von T_i .

- Modell 4 (Nicht-informative Zensierung)

$$P(t \leq T_i < t + \Delta t | T_i \geq t, x_i) = P(t \leq T_i < t + \Delta t | \min(T_i, C_i) \geq t, x_i)$$

1.1.2 Zählprozess-Formulierung

(t_i, δ_i) wird durch $(N_i(t), Y_i(t))$ ersetzt, mit

$$\begin{aligned} N_i(t) &= \text{Anzahl beobachteter Ereignisse in } [0, t] \text{ für Individuum } i \\ Y_i(t) &= \begin{cases} 1, & i \text{ ist zur Zeit } t \text{ unter Beobachtung und unter Risiko} \\ 0, & \text{sonst} \end{cases} \end{aligned}$$

2 Schätzung von Hazardraten und Survivalfunktionen für homogene Populationen

2.1 Nonparametrische Schätzung von Survivalfunktionen und Hazardraten

2.1.1 Sterbetafelmethode

- Zeitachse zerlegt in $q + 1$ Intervalle $[a_{k-1}, a_k)$, $k = 1, \dots, q + 1$ wobei $a_0 = 0$ und $a_{q+1} = \infty$
- Diskrete Hazardrate des k -ten Intervalls

$$\lambda_k = P(T \in [a_{k-1}, a_k) | T \geq a_{k-1})$$

- $p_k = 1 - \lambda_k = P(T \geq a_k | T \geq a_{k-1})$

$$P_k = P(T \geq a_k) = p_k \cdot \dots \cdot p_1$$

- Die erhobenen Daten sind:

n : Gesamtzahl der Individuen

d_k : Anzahl der Fälle, für die im k -ten Intervall ein Ereignis eintritt

w_k : Anzahl der Zensierungen im k -ten Intervall

R_k : "Risikomenge", d.h. die Fälle die zu Beginn des k -ten Intervalls noch unter Risiko stehen

$n_k = |R_k|$: Anzahl der Objekte unter Risiko zu Beginn des k -ten Intervalls

- $\hat{\lambda}_k = \frac{d_k}{n_k - \frac{w_k}{2}}$

- Schätzung für die Survivorfunktion $S(a_k) = P(T \geq a_k)$:

$$\hat{P}_k = \hat{p}_k \cdot \dots \cdot \hat{p}_1$$

- Schätzung

$$\hat{P}(T \in [a_{k-1}, a_k)) = \hat{P}_{k-1} - \hat{P}_k$$

- Schätzung für Dichte im Intervall $[a_{k-1}, a_k)$

$$\hat{f}_k = \frac{\hat{P}_{k-1} - \hat{P}_k}{h_k} = \frac{\hat{P}_{k-1} \cdot \hat{\lambda}_k}{h_k}, \quad h_k = a_k - a_{k-1}$$

2.1.2 Kaplan–Meier (Product-Limit) Schätzer

- $t_{(1)} < \dots < t_{(k)}$ sind die geordneten Werte der Stichprobe.

$$\hat{S}(t) = \begin{cases} 1, & t < t_{(1)} \\ \prod_{t_{(k)} \leq t} \left(1 - \frac{d_k}{n_k}\right), & t \geq t_{(1)} \end{cases}$$

- Geschätzte Varianz (Greenwood–Formel)

$$\widehat{Var}\{\hat{S}(t)\} = \hat{S}(t)^2 \cdot \sum_{k:t_{(k)} \leq t} \frac{d_k}{n_k(n_k - d_k)}$$

- Punktweise Konfidenzintervalle

$$\hat{S}(t) \pm z_{1-\frac{\alpha}{2}} \cdot \sqrt{\widehat{Var}\{\hat{S}(t)\}}$$

2.1.3 Nelson–Aalen–Schätzer für kumulierte Hazardrate

$$\hat{\Lambda}(t) = \begin{cases} 0, & t < t_{(1)} \\ \sum_{t_{(k)} \leq t} \frac{d_k}{n_k}, & t \geq t_{(1)} \end{cases}$$

- Geschätzte Varianz

$$\hat{\sigma}_{\Lambda}^2(t) = \sum_{t_{(k)} \leq t} \frac{d_k}{n_k^2}$$

- Punktweise Konfidenzintervalle

$$\hat{\Lambda}(t) \pm z_{1-\frac{\alpha}{2}} \hat{\sigma}_{\Lambda}(t)$$

- "Breslow"–Schätzer als Alternative zum Kaplan–Meier–Schätzer

$$\hat{S}_B(t) = \exp(-\hat{\Lambda}(t))$$

2.1.4 Schätzung der Hazardrate $\lambda(t)$

- Glätte $\Delta\{\hat{\Lambda}(t_{(k)})\} = \hat{\Lambda}(t_{(k)}) - \hat{\Lambda}(t_{(k-1)})$, $k = 1, \dots, m$ („Pseudodaten“) mit einer Kernschätzung

- Ramlau–Hansen–Schätzer

$$\hat{\lambda}(t) = \frac{1}{h} \sum_{k=1}^m K\left(\frac{t - t_{(k)}}{h}\right) \Delta\{\hat{\Lambda}(t_{(k)})\}$$

K : Kern wie beim (Kern–)Dichteschätzen

2.1.5 Tests zum Vergleich von Hazardraten (und Survivalfunktionen)

- Zwei-Stichproben-Fall

$$H_0 : \lambda_1(t) \equiv \lambda_2(t)$$

$$H_1 : \lambda_1(t) \not\equiv \lambda_2(t)$$

- Der log-Rang (log-rank)-Test

Poolen alle Ereigniszeiten (egal ob zu Gruppe 1 oder Gruppe 2) zugehörig zu

$$t_{(1)} < t_{(2)} < \dots < t_{(r)}$$

d_{1j} : Anzahl von Ereignissen aus Gruppe 1 zum Zeitpunkt $t_{(j)}$

d_{2j} : Anzahl von Ereignissen aus Gruppe 2 zum Zeitpunkt $t_{(j)}$

n_{1j} : Anzahl von Individuen unter Risiko aus Gruppe 1 kurz vor $t_{(j)}$

n_{2j} : Anzahl von Individuen unter Risiko aus Gruppe 2 kurz vor $t_{(j)}$

$d_j = d_{1j} + d_{2j}$: Anzahl aller Ereignisse zum Zeitpunkt $t_{(j)}$

$n_j = n_{1j} + n_{2j}$: Anzahl aller Individuen unter Risiko kurz vor $t_{(j)}$

- d_{1j} hypergeometrisch verteilt mit Parametern n_{1j} , d_j und n_j

$$E(d_{1j}) = e_{1j} = n_{1j} \cdot \frac{d_j}{n_j}$$

$$Var(d_{1j}) = v_{1j} = \frac{n_{1j} \cdot n_{2j} \cdot d_j (n_j - d_j)}{n_j^2 (n_j - 1)}$$

- Teststatistik

$$U_L = \sum_{j=1}^r (d_{1j} - e_{1j})$$

$$Var(U_L) = \sum_{j=1}^r v_{1j} = V_L$$

- Log-Rang-Test- oder Mantel-Haenszel-Test-Statistik

$$W_L = \frac{U_L^2}{V_L} \stackrel{a}{\sim} \chi_1^2$$

- Alternative: Wilcoxon-Test

$$U_W = \sum_{j=1}^r (d_{1j} - e_{1j}) n_j$$

$$Var(U_W) = V_W = \sum_{j=1}^r v_{1j} \cdot n_j^2$$

- Wilcoxon-Test-Statistik

$$W_W = \frac{U_W^2}{V_W} \stackrel{a}{\sim} \chi_1^2$$

2.2 Likelihood-Schätzung für parametrische Hazardmodelle

2.2.1 Likelihood für rechtszensierte Daten

- Daten (t_i, δ_i)

$$L(\theta) = \prod_{i=1}^n f(t_i)^{\delta_i} S(t_i)^{1-\delta_i} = \prod_{i=1}^n \lambda(t_i)^{\delta_i} \exp \left\{ - \int_0^{t_i} \lambda(s) ds \right\}$$

2.2.2 Weitere Zensierungsmechanismen

- Beiträge zur Likelihood:

Exakt beobachtete Lebensdauer	$f(t)$
Rechtszensierte Lebensdauer	$S(C_r)$
Linkszensierte Lebensdauer	$1 - S(C_l)$
Intervallzensierte Lebensdauer	$S(L) - S(R)$
Linkstrunkierte Beobachtungen	$\frac{f(t)}{S(Y)}$
Rechtstrunkierte Beobachtungen	$\frac{f(t)}{1-S(Y)}$

3 Regressionsmodelle für Survivaldaten

3.1 Parametrische Transformationsmodelle

3.1.1 Modelle

- $Y = \log(T) = x'\beta + \sigma\varepsilon$
- $T = \exp(x'\beta + \sigma\varepsilon)$
- Fehlerverteilung für ε ist von bekanntem Typ:

Verteilung von ε bzw. $\log(T)$	Verteilung von T
$\varepsilon \sim N(0, 1)$ Normal-Verteilung	log-normal Verteilung
Extremwert-Verteilung	Weibull- (Exponential-) Verteilung
logistische Verteilung	log-logistische Verteilung
log-Gamma Verteilung	Gamma-Verteilung (2 Parameter)
log-generalisierte Gamma-Verteilung	generalisierte Gamma-Verteilung (3 Parameter)

3.2 Das Proportional-Hazards-Modell von Cox

3.2.1 Modell

- $\lambda(t; x_i) = \lambda_0(t) \exp(x'_i\beta)$
- $\lambda_0(t)$ Baseline Hazardrate
- $x'\beta$ enthält keine Konstante (Intercept)
- Proportionalität der Hazardraten

$$\frac{\lambda(t; x_1)}{\lambda(t; x_2)} = \frac{\lambda_0(t) \exp(x'_1\beta)}{\lambda_0(t) \exp(x'_2\beta)} = \exp((x_1 - x_2)'\beta)$$

3.2.2 Partial-Likelihood-Inferenz für das Cox-Modell

- $t_{(1)} < \dots < t_{(k)}$ bezeichnen die Ereigniszeitpunkte der unzensierten Beobachtungen.
- Partielle Likelihood

$$PL(\beta) = \prod_{i=1}^k \frac{\exp(x'_{(i)}\beta)}{\sum_{j \in R(t_{(i)})} \exp(x'_j\beta)}$$

- Maximum Partial–Likelihood–Schätzer

$$\hat{\beta} = \operatorname{argmax}\{\log PL(\beta)\}$$

- (partielle) Score–Funktion

$$s(\hat{\beta}) = \frac{\partial \log PL(\beta)}{\partial \beta}$$

- $\hat{\beta} \stackrel{a}{\sim} N(\beta, \hat{V}(\hat{\beta}))$

$$V(\hat{\beta}) = \text{Inverse der "Informations–Matrix"} - \frac{\partial s(\hat{\beta})}{\partial \beta'}$$

- Breslow–Schätzer für die kumulative Hazardrate

$$\hat{\Lambda}_0(t) = \sum_{t_{(i)} \leq t} \frac{1}{\sum_{j \in R(t_i)} \exp(x_j' \hat{\beta})}$$

3.3 Zeitabhängige Kovariablen

- $x(t)$ vorhersagbarer stochastischer Prozess. Hinreichende Bedingung: Linkstetige Pfade

$$PL(\beta) = \prod_{i=1}^k \frac{\exp(x_{(i)}(t_{(i)})' \beta)}{\sum_{j \in R(t_{(i)})} \exp(x_j(t_{(i)})' \beta)}$$

3.4 Modellierung zeitabhängiger Effekte $\beta(t)$

- Ersetze $\beta_1 x_1$ im Prädiktor durch z.B. $\beta_1 + \beta^* t$
- $\beta_1 \cdot x_1 + \beta^* t \cdot x_1 = \beta_1 \cdot x_1 + \beta^* \underbrace{t \cdot x_1}_{\text{neue zeitabhängige Kovariable}}$
- Schätze β_1 und β^* mit PL–Ansatz

4 Modellwahl und Modelldiagnostik

4.1 Kriterien zur Modellwahl

- AIC: Akaiikes Informationskriterium

$$AIC = -2 \log PL(\hat{\beta}) + 2p \quad p: \text{Anzahl der Parameter im Modell}$$

- Devianz: Sei zunächst $\Lambda_0(t)$ fest.

$$D = 2 \sum_{i=1}^n \left[-\delta_i \log \left\{ \frac{\delta_i - \tilde{M}_i}{\delta_i} \right\} - \tilde{M}_i \right]$$

mit

$$\tilde{M}_i = \delta_i - \int_0^{t_i} \exp(x_i \hat{\beta}) d\Lambda_0(s)$$

Ersetze $\Lambda_0(t)$ durch (z.B.) den Breslow–Schätzer $\hat{\Lambda}_0(t)$.

4.2 Test auf PH–Annahme

Durch künstliche zeitabhängige Kovariable. Sei x_1 eine zeitkonstante Kovariable.

- Definiere $x_2(t) = x_1 \cdot g(t)$ $g(t)$ bekannte Funktion von t , z.B. $\ln(t)$
- Fitte Cox–Modell

$$\lambda(t | x(t)) = \lambda_0(t) \exp(\beta_1 x_1 + \beta_2 x_2(t) + \dots)$$
- Teste $H_0 : \beta_2 = 0$ gegen $H_1 : \beta_2 \neq 0$

4.3 Residuen

4.3.1 Cox–Snell–Residuen

- Falls Lebensdauer T die Survivalfunktion $S(t)$ bzw. kumulative Hazardrate $\Lambda(t)$ besitzt, dann ist die die Zufallsvariable

$$Y = -\log S(T) = \Lambda(T) = g(t)$$

exponentialverteilt mit Rate 1.

- Definition: Cox–Snell–Residuen

$$r_i = \hat{\Lambda}_0(t_i) \exp(x_i' \hat{\beta}), \quad i = 1, \dots, n$$

4.3.2 Martingal–Residuen

- Definition: Martingal–Residuen

$$M_i = \delta_i - r_i$$

- Es gilt $E(M_i) \approx 0$.

4.3.3 Devianz–Residuen

- Definition: Devianz–Residuen

$$D_i = \text{sign}(\hat{M}_i) \sqrt{2} \left[-\hat{M}_i - \delta_i \log(\delta_i - \hat{M}_i) \right]^{\frac{1}{2}}$$

5 Modifikationen und Erweiterungen von Hazardraten-Modellen

5.1 Frailty-Modelle

Frailty-Modell vom Cox-Typ

Cluster (Gruppen-) bzw. individualspezifische zufällige Effekte werden in Prädiktor einbezogen:

$$\begin{aligned} \lambda(t | x_{ij}) &= \lambda_0(t) \exp(x_{ij}^T + \gamma_i), \quad i = 1, \dots, m; j = 1, \dots, n_i \\ &= \lambda_0(t) \nu_i \exp(x_{ij}^T \beta) \end{aligned}$$

mit $\nu_i = \exp(\gamma_i)$, und z.B. γ_i i.i.d. $N(0, \tau^2)$ oder ν_i i.i.d. Gamma, mit $E(\nu_i) = 1$.

Durch den gemeinsamen clusterspezifischen Effekt γ_i bzw. ν_i sind Beobachtungen aus Cluster i positiv korreliert.