Kurzskript zur Vorlesung Nichtlineare und nichtparametrische Methoden, SS2019

Antje Jahn

Hochschule Darmstadt, Fachbereich MN

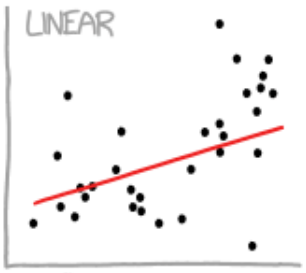
Dieses Skript erhebt weder Anspruch auf Vollständigkeit noch auf Fehlerfreiheit. Es enthält nur eine grobe Zusammenfassung ausgewählter Vorlesungsinhalte.

Sollten Sie Fehler oder Unklarheiten entdecken, so bin ich für eine Rückmeldung dankbar!

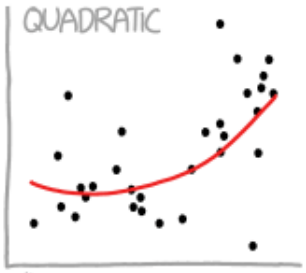
Das Skript ist nur für die Teilnehmer der Vorlesung gedacht. Es darf nicht weiter gegeben oder kopiert werden!

Inhaltsverzeichnis

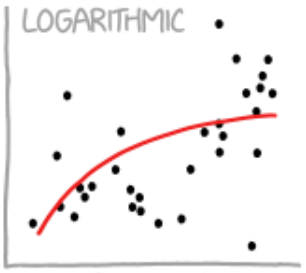
CURVE-FITTING METHODS
AND THE MESSAGES THEY SEND



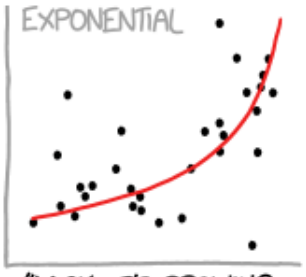
"HEY, I DID A REGRESSION."



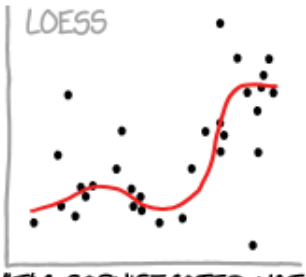
"I WANTED A CURVED LINE, SO I MADE ONE WITH MATH."



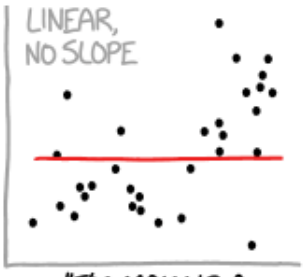
"LOOK, IT'S TAPERING OFF!"



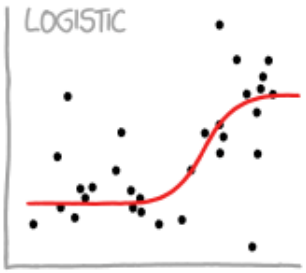
"LOOK, IT'S GROWING UNCONTROLLABLY!"



"I'M SOPHISTICATED, NOT LIKE THOSE BUMBLING POLYNOMIAL PEOPLE."

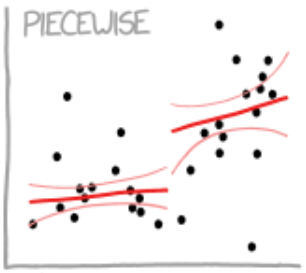


"I'M MAKING A SCATTER PLOT BUT I DON'T WANT TO."

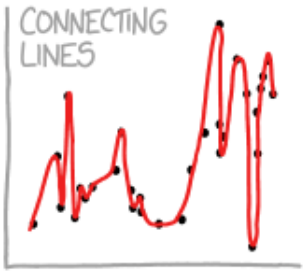


"I NEED TO CONNECT THESE TWO LINES, BUT MY FIRST IDEA DIDN'T HAVE ENOUGH MATH."

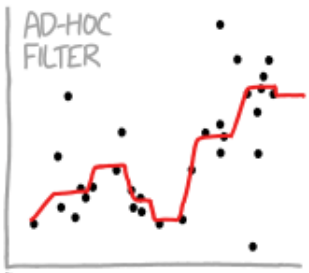
"LISTEN, SCIENCE IS HARD. BUT I'M A SERIOUS PERSON DOING MY BEST."



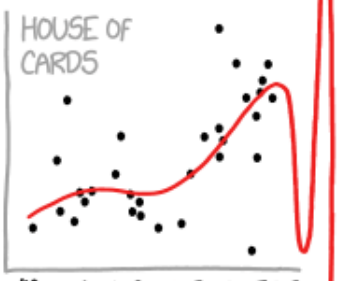
"I HAVE A THEORY, AND THIS IS THE ONLY DATA I COULD FIND."



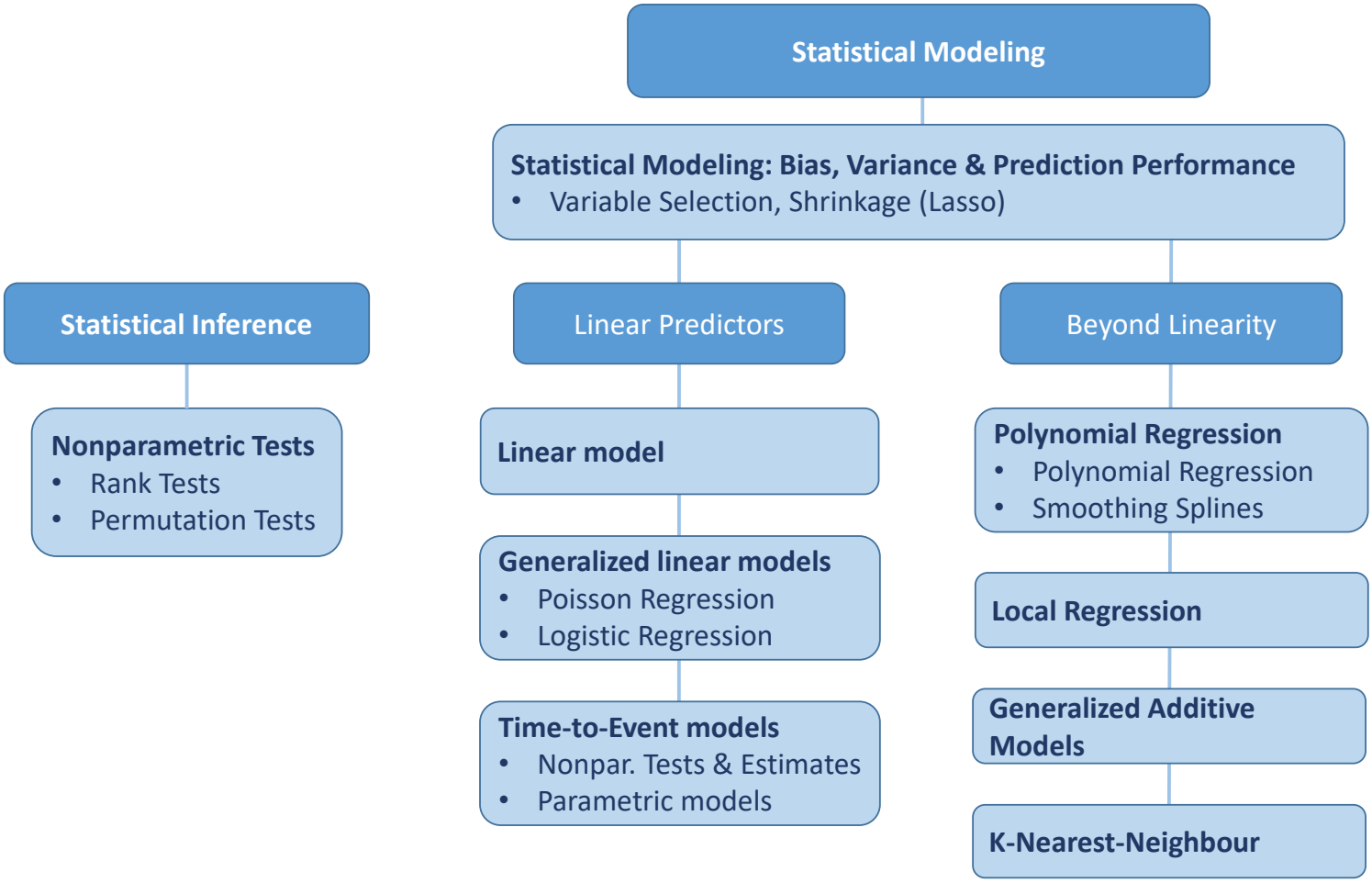
"I CLICKED 'SMOOTH LINES' IN EXCEL."



"I HAD AN IDEA FOR HOW TO CLEAN UP THE DATA. WHAT DO YOU THINK?"



"AS YOU CAN SEE, THIS MODEL SMOOTHLY FITS THE– *WAIT NO NO DON'T EXTEND IT AAAAAA!!*"


Statistical Modeling
Statistical Modeling: Bias, Variance & Prediction Performance
• Variable Selection, Shrinkage (Lasso)
Statistical Inference
Nonparametric Tests
• Rank Tests
• Permutation Tests
Linear Predictors
Linear model
Generalized linear models
• Poisson Regression
• Logistic Regression
Time-to-Event models
• Nonpar. Tests & Estimates
• Parametric models
Beyond Linearity
Polynomial Regression
• Polynomial Regression
• Smoothing Splines
Local Regression
Generalized Additive Models
K-Nearest-Neighbour

1 Begriffsbestimmungen

1.1 Nichtparametrische und parametrische Verfahren

In parametrischen statistischen Verfahren wird zugrundegelegt, dass die betrachteten Zufallsvariablen aus einer Familie von Wahrscheinlichkeitsverteilungen stammt. Ein oder mehrere Parameter spezifizieren dabei die Verteilung. Die Verfahren haben häufig das Ziel

- diese Parameter zu schätzen und damit Aussagen über die Verteilung von Zufallsvariablen oder den Zusammenhang zwischen Zufallsvariablen zu treffen
- und/oder Hypothesen über diese Parameter zu testen.

Beispiele: t-Test, ML-Schätzer, lineare Regression

Nichtparametrische Verfahren legen dagegen keine Verteilungsfamilie zugrunde. Beispiele:

- Statistische Tests bei kleinen Stichproben, wenn eine Approximation der Testverteilung mit dem zentralen Grenzwertsatz nicht möglich ist (–> Rang- und Permutationstests)
- Modellschätzer, die keine eine Annahme über die Form des Zusammenhangs zwischen Y und $X_1, \dots X_n$ zugrundelegen (–> K-Nächste-Nachbarn)

1.2 Statistisches Modell

Wir benutzen die folgenden Bezeichnungen:

- Abhängige Variable Y (Outcome / Response)
- Unabhängige / Erklärende Variablen $X = (X_1 \dots X_p)^t$ (Kovariaten)
- n = Größe der Stichprobe / Anzahl an Beobachtungen
- p = Anzahl an erklärenden Variablen / Kovariaten
- Stichprobe der Größe n: (y_i, x_i^t), $i = 1 \dots n$
- $x_i = (x_{i1}, \dots x_{ip})^t$

Ein statistisches Modell

$$Y = f(X) + \epsilon$$

beschreibt den Zusammenhang zwischen X und Y. Wir unterscheiden zwischen dem wahren Modell, welches den wahren Zusammenhang beschreibt (true model / data generating model) und dem Analyse-Modell, welches wir bei der statistischen Analyse zugrundelegen (analysis

model). Ein aus Daten geschätztes Modell bezeichnen wir mit $\hat{f}$. Sofern nicht genauer spezifiziert, bezeichnet "Modell" das Analyse-Modell.

Definitionen des statistischen Modells in der Literatur:

- A set of probability distributions on the sample space (Cox & Hinkley, 1974)
- A simplification or approximation of the reality (Burnham & Anderson, 2002)
- A model represents, often in considerably idealized form, the data generating process (Wikipedia)
- Statistical models are simple mathematical rules derived from empirical data describing the association between an outcome and several explanatory variables (Dunkler et al, 2014)

Additive und Lineare Modelle

Additives Modell:

$$Y = b_0 + f_1(X_1) + \cdots + f_p(X_p) + \epsilon$$

oder

$$E(Y|x_1, \ldots x_p) = b_0 + f_1(x_1) + \cdots + f_p(x_p)$$

Lineares Modelle:

$$Y = b_0 + b_1X_1 + \cdots + b_pX_p + \epsilon$$

oder

$$E(Y|x_1, \ldots x_p) = b_0 + b_1x_1 + \cdots + b_px_p$$

Dabei heisst $b_0 + b_1X_1 + \cdots + b_pX_p$ der lineare Prädiktor.

Ziele der statistischen Modellierung

- Überprüfen von Hypothesen (statistische Tests)
- Erklären der Assoziation zwischen dem Outcome und den Kovariaten (erklärendes Modell): Fokus liegt auf der Interpretation von Regressionskoeffizienten, Konfidenzintervallen und p-Werten, hoher Anpassungsgüte des Modells, sinnvoller Variablenselektion

- Vorhersage von Y auf Basis von f und X für neue Beobachtungen (Vorhersagemodell): Fokus liegt auf einem möglichst geringen Vorhersagefehler (generalization error = Vorhersagefehler auf unabhängigen neuen Daten), Interpretierbarkeit von Regressionskoeffizienten etc. weniger wichtig (bis hin zu black-box-Algorithmen), Variablenselektion auf Basis des Vorhersagefehlers

1.3 Bias und Varianz von Modellschätzern

Mögliche Fehlerquellen bei der statistischen Modellierung:

- Wichtige Kovariablen werden nicht modelliert
- Kovariablen ohne Einfluss auf Y (noise) werden modelliert
- Der funktionale Zusammenhang (f_i) wird misspezifiziert
- Das statistische Modell (f) wird misspezifiziert

2 Nichtparametrische Tests und Rangtests

Bei parametrischen Tests (z.B. Gauss-Test, t-Test...) ist für jedes θ aus dem Parameterraum die Verteilung der Zufallsvariablen bekannt. Die Nullhypothese ist ein Punkt oder ein Intervall des Parametrraums, d.h. auch unter der Nullhypothese (einfach Nullhypothesen) oder dem Rand der Nullhypothese (zusammengesetzte Nullhypothesen) ist die Verteilung der Zufallsvariablen und damit die Verteilung der Teststatistik bekannt. Der Ablehnbereich kann daher über die Quantile dieser Verteilung bestimmt werden. Bei nichtparametrischen Tests wird dagegen nicht vorausgesetzt, dass die Verteilung der Zufallsvariablen über einen Parameter definiert und damit unter der Nullhypothese bekannt ist. Wir betrachten hier zwei Typen von nichtparametrischen Tests:

- Rangtests: Die Teststatistik basiert nur auf den Rängen der beobachteten Werte, nicht auf den Werten selbst
- Permutationstests: Die beobachteten Werte werden als fest, nicht zufällig betrachtet. Die Zuteilung der Versuchsbedingungen (z.B. die Gruppenzuteilung) wird als Zufallsvektor behandelt.
- Monte-Carlo-Permutationstests

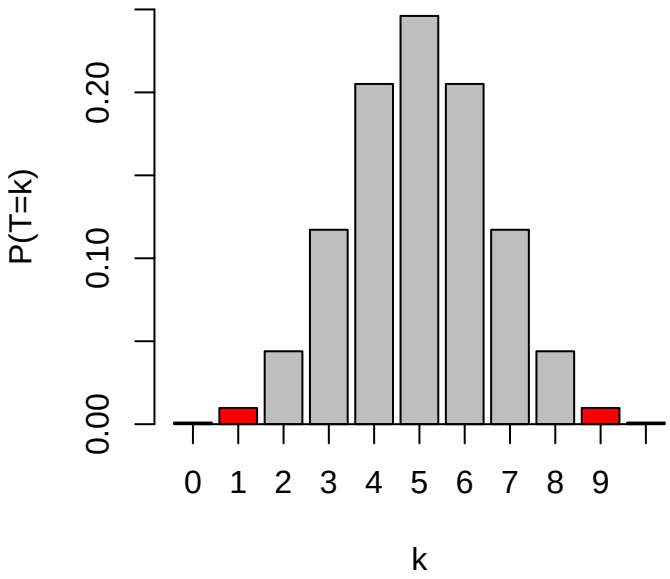
2.1 Einstichprobenproblem / verbundene Stichproben

2.1.1 Vorzeichen-Test

1. $X_1, \ldots, X_n$ unabhängig identisch verteilte Zufallsvariablen; X_i habe stetige Verteilung ; x_{med} sei der Median von X, d.h. $P(X \leq x_{med}) = P(X \geq x_{med}) = 0.5$
2. $H_0 : \{x_{med} = \delta_0\}$ vs $H_1 : \{x_{med} \neq \delta_0\}$
3. Festlegung des Signifikanzniveaus α
4. Unter H_0 gilt
$$P(X < \delta_0) = P(X < x_{med}) = 0.5$$
Das heißt
$$Y_i := \begin{cases} 1, & X_i < \delta_0 \\ 0, & X_i \geq \delta_0 \end{cases}$$
ist ein Bernoulli-Experiment mit $Y_i \underset{H_0}{\sim} B(1, 0.5)$ und
$$T := \sum_{i=1}^{n} Y_i \underset{H_0}{\sim} B(n, 0.5)$$
5. Finde $c_{\alpha/2}$ als den größten Wert $\in \{0, 1, \ldots, n\}$ für den die Verteilungsfunktion der $B(n, 0.5)$-Verteilung $\leq \frac{\alpha}{2}$ ist. Definiere dann als Ablehnbereich
$$C = \{a; a \leq c_{\frac{\alpha}{2}} \text{ oder } n - \alpha \leq c_{\frac{\alpha}{2}}\}$$

Bemerkungen:

- Beobachtungen $x_i = \delta_0$ können theoretisch nicht auftreten, da stetige Verteilung. In der Praxis werden diese Beobachtungen weggelassen, so dass sich der Parameter n (Stichprobengröße) entsprechend reduziert
- Der Test wird in der Praxis auch für ordinalskalierte Daten angewandt.
- Für $n \geq 30$ gilt unter H_0 approximativ $B(n, 0.5) \sim N(0.5 \cdot n, 0.25 \cdot n)$, d.h. alternativ
$$T := \frac{\sum_{i=1}^{n} Y_i - 0.5n}{\sqrt{0.25n}} \underset{H_0}{\sim} N(0, 1)$$
mit Ablehnbereich $C =]-\infty; -z_{1-\alpha/2}[\cup]z_{1-\alpha/2}; \infty[$.



2.1.2 Wilcoxon-Vorzeichen-Rang-Test

Anwendungsbereich:

- Alternative zum Vorzeichen-Test bei symmetrischen Verteilungen
- Anwendung bei ordinalskalierten Daten, z.B. Antwortskalen

1. $X_1, \ldots, X_n$ unabhängig identisch verteilt, X_i stetig und symmetrisch; x_{med} der Median von X
2. Die Nullhypothese soll geprüft werden, ob die zentrale Lage der Verteilung bei 0 liegt, d.h.
$$H_0 : \{x_{med} = \delta_0\} \quad vs \quad H_1 : \{x_{med} \neq \delta_0\}$$
3. Festlegung des Signifikanzniveaus α
4. Wahl der Teststatistik: Definiere
$$D_i := X_i - \delta_0$$
Bilde Ränge der $|D_i|$ und definiere
$$W^+ := \sum_{i;D_i>0} rang(|D_i|) \quad W^- := \sum_{i;D_i<0} rang(|D_i|)$$

Unter H_0 erwarten wir wegen der Symmetrie der Verteilung $w^+ \approx w^-$. Da weiter $w^+ + w^- = \frac{n(n+1)}{2}$ ist unter $H_0 \quad E(W^+) = \frac{n(n+1)}{4}$.

5. Der Ablehnbereich für $T = min(W^+, W^-)$ in Abhängigkeit von n ist tabelliert.

Bemerkungen:

- Beobachtungen $x_i = \delta_0$ können theoretisch nicht auftreten, da stetige Verteilung. In der Praxis werden diese Beobachtungen weggelassen, so dass sich der Parameter n (Stichprobengröße) entsprechend reduziert
- Treten Beobachtungen $X_i = X_j$ auf (Bindungen) so wird allen der durchschnittliche Rang zugewiesen. Dies sollte bei Stetigkeitsannahme nicht zu häufig vorkommen. Ansonsten gibt es korrigierte Teststatistiken (Korrektur der Varianzschätzung)
- Der Test wird in der Praxis auch für ordinalskalierte Variablen mit symmetrischer Verteilung angewandt.
- Für $n \geq 30$ gilt unter H_0 approximativ

$$W^+ \underset{H_0}{\sim} \mathcal{N}\left(\frac{n(n+1)}{4}, \frac{n(n+1)(2n+1)}{24}\right)$$

  d.h. alternativ

$$T := \frac{W^+ - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} \underset{H_0}{\sim} \mathcal{N}(0,1)$$

  mit Ablehnbereich $C =]-\infty; -z_{1-\alpha/2}[\cup]z_{1-\alpha/2}; \infty[$.

- Der Test wird - wie der verbundene t-Test - häufig auch für verbundene Stichproben angewandt, d.h. für die Differenzen $X_i - Y_i$
- Ohne die Symmetrie der Verteilung kann der Test antikonservativ werden

2.1.3 Permutationstest für verbundene Stichproben

$(X_i, Y_i)_{i=1\ldots n}$ unabhängig identisch verteilte Zufallsvariablen und $D_i = X_i - Y_i$ symmetrisch um θ. Die Nullhypothese $\{\theta = 0\}$ soll getestet werden. Möglich wäre ein Wilcoxon-Vorzeichen-Rang-Test. Alternative ist ein Permutationstest, der auch unter Bindungen valide bleibt. Unter H_0 gilt $(X_i - Y_i) \sim (Y_i - X_i)$ (exchangeability unter H_0) und damit mit $D_i = X_i - Y_i$: Es sei $M \subset \{1, \ldots n\}$ beliebig.

$$T := \frac{1}{n}\sum_{i=1}^{n} D_i = \frac{1}{n}\sum_{i=1}^{n} X_i - Y_i \underset{H_0}{\sim} T_M = \frac{1}{n}\Big(\sum_{i\in M} X_i - Y_i + \sum_{i\notin M} Y_i - X_i\Big)$$

und damit

$$P(T \geq t) = \frac{|\{M \subset \{1 \dots n\}, T_M \geq t\}|}{|\{M \subset \{1 \dots n\}\}|} = \frac{|\{M \subset \{1 \dots n\}, T_M \geq t\}|}{2^n}$$

2.2 Zweistichprobenproblem / unverbundene Stichproben

2.2.1 Wilcoxon-Rangsummen-Test (Mann-Whitney-U-Test)

Anwendungsbereich:

- Vergleich von zwei unabhängigen Stichproben bzgl. der zentralen Lage einer stetigen Variablen
- Alternative zum Zwei-Stichproben-t-Test bei kleinen Fallzahlen und Abweichung von der Normalverteilungsannahme

1. $X_1, \dots, X_n, Y_1 \dots Y_m$ unabhängige Zufallsvariablen, $X_i, i = 1 \dots n$ stetig mit Verteilungsfunktion F_X und $Y_i, i = 1 \dots m$ stetig mit Verteilungsfunktion F_Y. Weiter gelte das Lokations- (Shift-) Modell, d.h. $F_X(x) = F_Y(x + \theta)$ für ein $\theta \in \mathbb{R}$.
2. Die Nullhypothese soll geprüft werden, ob die beiden Verteilungen übereinstimmen, d.h.

   $$H_0 : \{\theta = 0\} = \{F_X(x) = F_Y(x) \forall x \in \mathbb{R}\} \quad vs \quad \{H_1 : \theta \neq 0\}$$

   Bzw. einseitig:

   $$H_0 : \{\theta \leq 0\} \quad vs \quad \{H_1 : \theta > 0\}$$

3. Festlegung des Signifikanzniveaus α
4. Wahl der Teststatistik: Bilde Ränge $rg(X_i)$ der Beobachtungen $X_1, \dots X_n, Y_1, \dots Y_m$ und definiere als Teststatistik

   $$T := \sum_{i=1}^{n} rg(X_i)$$

   Unter H_0 kann die (diskrete) Verteilung von T kombinatorisch hergeleitet werden und die Ablehnregionen sind in Abhängigkeit von n und m tabelliert.

Bemerkungen:

- Umgang mit Bindungen: Bei $x_i = x_j$ werden die Ränge zufällig verteilt (keine Konsequenz für die Teststatistik). Bei $x_i = y_j$ werden Durschnittsränge gebildet
- Für $n \geq 30$ gilt unter H_0 approximativ

  $$T \underset{H_0}{\sim} \mathcal{N}\left(\frac{n(n+m+1)}{2}, \frac{nm(n+m+1)}{12}\right)$$

d.h. alternativ

$$\frac{T - \frac{n(n+m+1)}{2}}{\sqrt{\frac{nm(n+m+1)}{12}}} \underset{H_0}{\sim} \mathcal{N}(0,1)$$

mit Ablehnbereich $C =]-\infty; -z_{1-\alpha/2}[\cup]z_{1-\alpha/2}; \infty[$.

2.2.2 Permutationstest

Unter denselben Annahmen wie beim Wilcoxon-Rangsummen-Test, aber ohne die Voraussetzung der Stetigkeit (d.h. Bindungen sind kein Problem), kann der folgende Permutationstest durchgeführt werden:

$$Z := (X_1, \dots X_n, Y_1, \dots Y_m) = (Z_1, \dots Z_{n+m})$$

Unter H_0 (exchangeability under H_0) gilt

$$Z \sim Z_\pi \quad \text{für alle Permutationen mit } Z_\pi = Z_{\pi(1)}, \dots Z_{\pi(n+m)}$$

un damit

$$T = T(Z) := \frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{m}\sum_{i=1}^{m} Y_i \underset{H_0}{\sim} \frac{1}{n}\sum_{i=1}^{n} Z_{\pi(i)} - \frac{1}{m}\sum_{i=n+1}^{n+m} Z_{\pi(i)} = T(Z_\pi) =: T_\pi$$

und

$$p = P(T \geq t) = \frac{|\{\pi; T_\pi \geq t\}|}{|\{\pi\}|}$$

3 Statistische Modellierung: Vorhersagefehler

Es liegt ein wahres datengenerierendes statistisches Modell

$$Y = f(X) + \epsilon, \qquad E(\epsilon) = 0, Var(\epsilon) = \sigma^2, \epsilon \text{ und X sind unabhängig}$$

zugrunde. Ein statistisches Verfahren schätzt f aus einer Stichprobe und liefert als Ergebnis $\hat{f}$. Für eine neue Beobachtung mit $X = x$ kann dann y als $\hat{y} = \hat{f}(x)$ geschätzt werden. Wie gut ist nun das statistische Verfahren zur Schätzung von f bzw. wie groß ist der zu erwartende Vorhersagefehler?

Dazu benötigen wir zunächst eine Verlustfunktion, die den Fehler zwischen y und $\hat{y}$ beschreibt.

Verlustfunktion

Als Verlustfunktion bezeichnen wir eine Funktion $L : \mathbb{R}^2 \to \mathbb{R}$, die den Fehler zwischen einem beobachteten Wert y und einem geschätzten Wert $\hat{y} = \hat{f}(x)$ definiert, z.B.

$$L(y, \hat{f}(x)) = \begin{cases} (y - \hat{f}(x))^2 & \text{quadratische Verlustfunktion / quadratischer Fehler} \\ |y - \hat{f}(x)| & \text{absolute Verlustfunktion / absoluter Fehler} \\ -2 * logL(y|\hat{\theta}(x)) & \text{Log-Likelihood-Funktion} \end{cases}$$

mit L der Likelihood von y unter aus dem Modell geschätzten Prädiktor $\hat{\theta}(x)$.

Bemerkungen:

- Diese Verlustfunktionen eignen sich gut für stetige Outcomes, in Klassifikationsverfahren (diskrete Outcomes) werden häufig andere Verlustfunktionen genutzt.
- Wir schreiben $L(y, \hat{f}(x))$ als Wert der Verlustfunktion für eine Realisierung (x, y) von (X, Y). Wir schreiben $L(Y, \hat{f}(X))$, wenn X und Y Zufallsvariablen bezeichnen und damit auch L eine Zufallsvariable ist. Diese beschreibt den (zufälligen) Vorhersagefehler für eine (zufällige) Trainingsstichprobe, die $\hat{f}$ bestimmt, und eine (zufällige) Teststichprobe (X, Y). Der erwartete Vorhersagefehler des statistischen Verfahrens, $E(L(Y, \hat{f}(X))$, kann als Gütekriterium des Verfahrens und zum Vergleich verschiedener statistischer Modelle dienen.

Fehlerdefinitionen

Der Trainingsfehler beschreibt den mittleren Fehler auf den Trainingsdaten, d.h. den Daten auf denen das Modell geschätzt wurde. Dieser unterschätzt i.d.R. den Vorhersagefehler

$$\text{Training Error: } \bar{err} := \frac{1}{n}\sum_{i=1}^{n} L(y_i, \hat{f}(x_i)) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2 = MSE = n \times RSS$$

Der Testfehler (generalization error) beschreibt den erwarteten Vorhersagefehler auf neuen unabhängigen Daten (X, Y)

$$\text{Test Error: } Err := E_{(X,Y)}(L(Y, \hat{f}(X))|traindata)) = E_{(X,Y)}((Y - \hat{f}(X))^2|traindata))$$

Der erwartete Testfehler beschreibt den erwarteten Vorhersagefehler, wenn auch die Modellanpassung $\hat{f}$ als zufällig angenommen wird, d.h. betrachtet nicht allein das konkrete aus unseren Daten geschätzte Modell.

$$\text{Expected Test Error: } E_{train}(Err) = E_{traindata}E_{(X,Y)}(L(Y, \hat{f}(X)|traindata))$$

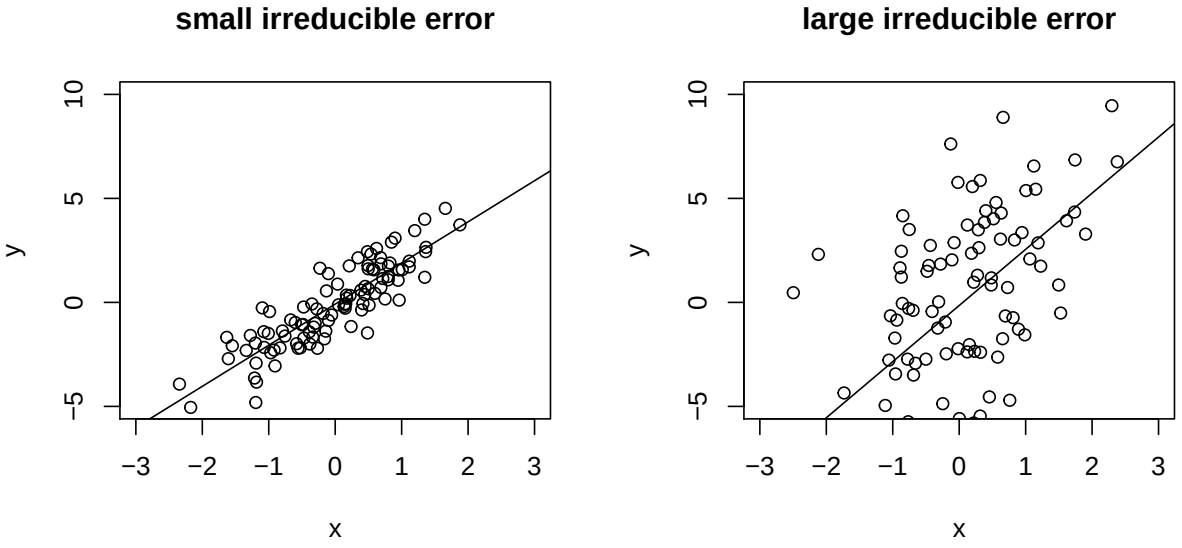
Bias-Varianz-Zerlegung

Wir legen im folgenden immer die quadratische Verlustfunktion zugrunde. Es gilt

$$\begin{aligned} \text{Expected Test Error in x} &:= E(L(Y, \hat{f}(x))|X = x) \\ &= \sigma^2 + Bias^2(\hat{f}(x)) + Var(\hat{f}(x)) \\ &= \text{Irreducible Error} + Bias^2(\hat{f}(x)) + \text{Var}(\hat{f}(x)) \end{aligned}$$

Bemerkungen

- Ist der irreducible error groß im Verhältnis zu f(X), wird kein Prognosemodell eine präzise Vorhersage liefern
- In einem guten Vorhersagemodell sind Bias($\hat{f}$) und Var($\hat{f}$) klein
- Oft kann eine Reduktion der Varianz auf Kosten einer Erhöhung des quadratischen Bias erreicht werden
- Underfitting: $\hat{f}$ hat einen (zu) großen Bias
- Overfitting: $\hat{f}$ hat eine (zu) große Bias
- Dieser Bias-Varianz-Trade-Off wird z.B. in regularisierten Regressionsmodellen (shrinkage, variable selection...) über Hyper-Parameter (λ) optimiert



Modellwahl auf Basis von training, test und expected test error

- Validierungsstichprobe
- Kreuzvalidierung
- Informationskriterien

$$AIC := -2LL + 2p$$

$$BIC := -2LL + p \cdot log(n)$$

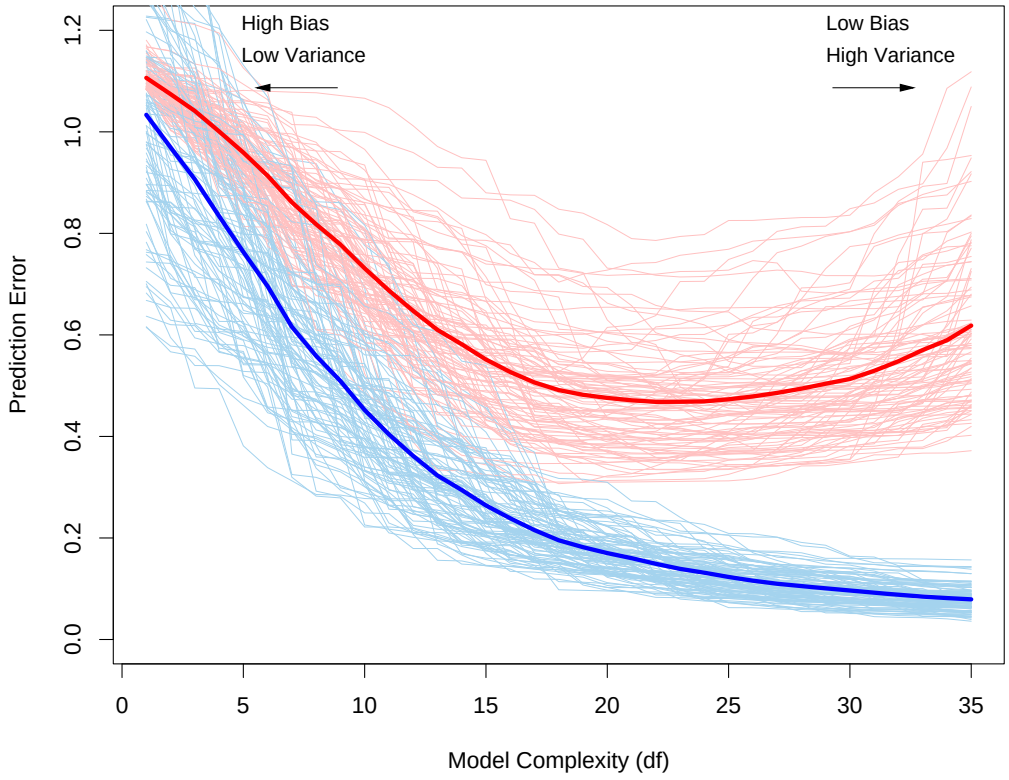
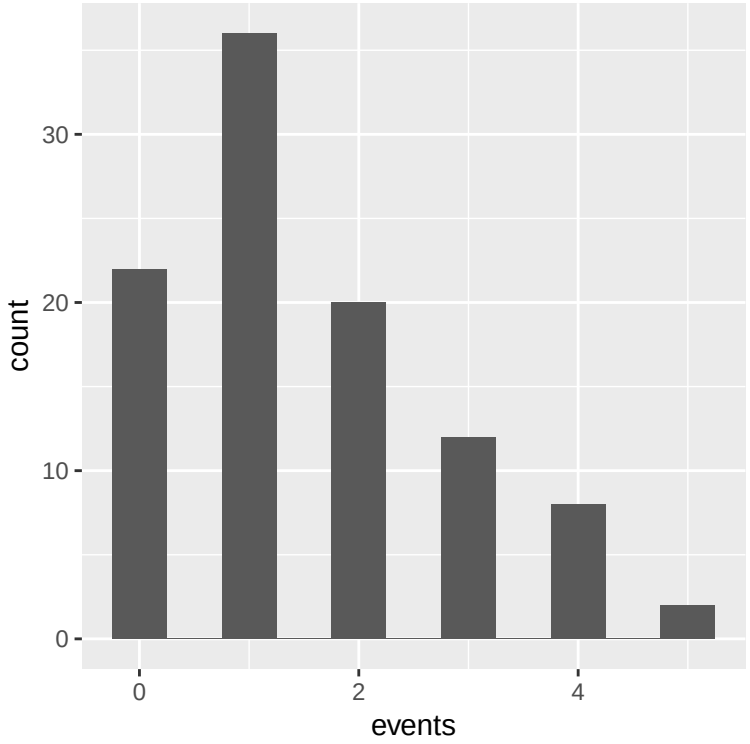


FIGURE 7.1. *Behavior of test sample and training sample error as the model complexity is varied. The light blue curves show the training error* $\overline{\text{err}}$*, while the light red curves show the conditional test error* $\text{Err}_{\mathcal{T}}$ *for* 100 *training sets of size* 50 *each, as the model complexity is increased. The solid curves show the expected test error* Err *and the expected training error* $\text{E}[\overline{\text{err}}]$*.*

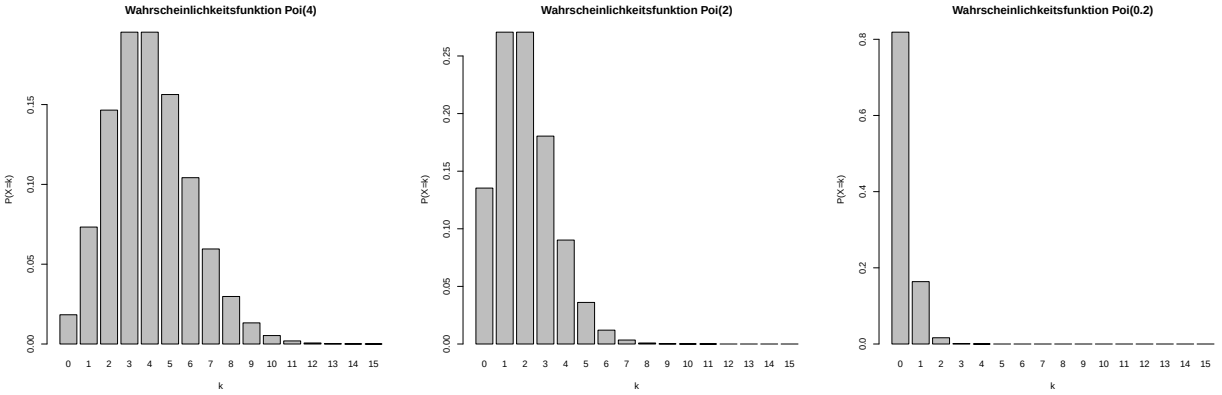
4 Verallgemeinerte lineare Modelle (GLM)

4.1 Poisson-Regression

Beispiel Rekurrenzen nach Ersttumortherapie:



Wahrscheinchkeitsfunktion der Poisson-Verteilung:



Ziel: Modellierung von Kovariateneffekten auf $E(Y)$, d.h. Schätzung von $E(Y|x)$

Definition 4.1. Poisson-Regressionsmodell

Es seien $Y_1 \dots Y_n$ (bedingt) unabhängige Zufallsvariablen mit Wertebereich $\mathbb{N}$. $X_1 \dots X_n$ seien p-dimensionale Zufallsvektoren (oder deterministisch). Dann ist durch

$$Y_i|X_i = x_i \sim Poi(\lambda_i)$$

mit $\lambda_i = exp(b^T x_i) \quad bzw. \quad log(\lambda_i) = b^T x_i$ ein Poisson-Regressions-Modell definiert.

Interpretation der Regressionskoeffizienten:

$$E(Y_i|x_i) = exp(b^T x_i)$$

ML-Schätzung

Wir nehmen im folgenden an, dass x_i schon eine führende 1 und der Koeffizientenvektor b entsprechend auch den Intercept b_0 enthält

$$\begin{aligned} L(b) &= \prod_{i=1}^{n} P(Y_i = y_i|b,x) \prod_{i=1}^{n} e^{-exp(b^T x_i)} \frac{exp(b^T x_i)^{y_i}}{y_i!} \\ l(b) &= \sum_{i=1}^{n} y_i b^T x_i - exp(b^T x_i) - log(y_i!) \\ \Delta l &= \sum_{i=1}^{n} x_i (y_i - exp(b^T x_i)) \\ &= s(b) \quad \text{Score-Funktion} \end{aligned}$$

$s(b) = 0$ ist ein nichtlineares Gleichungssystem zu dessen Lösung numerische Verfahren, z.B. Newton-Raphson, herangezogen werden. Hierzu wird die Hesse-Matrix von l an der Stelle des ML-Schätzers, H, benötigt, die in diesem Kontext auch "beobachtete Informationsmatrix" heisst.

Eigenschaften der Schätzer

Asymptotisch gilt

$$\hat{b} \sim N(b, F^{-1}(\hat{b})) \quad mit\ F^{-1}(\hat{b}) = \hat{Cov}(\hat{b})$$

F ist die Fisher-Matrix $= E(-H(\hat{b})) = Cov(s(\hat{b}))$

Statistische Tests und Konfidenzintervalle

Wald-Test:

$$H_0 = \{b_j = 0\} \; vs \; H_1 = \{b_j \neq 0\}$$

$$T = \frac{\hat{b}_j}{\sqrt{F^{-1}(\hat{b})_{jj}}}$$

oder $1-\alpha$-CI für b_j

$$\hat{b}_j \pm z_{\alpha/2}\sqrt{F^{-1}(\hat{b})_{jj}}$$

Likelihood-Ratio-Test

$$T = -2(l_{M1}(\hat{b}_{M1}) - l_{M2}(\hat{b}_{M2})) \underset{H_0}{\sim} \chi^2_1$$

wobei $M_1, \hat{b}_{M1}$ und $M_2, \hat{b}_{M2}$ die Modelle und Schätzer jeweils ohne und mit Kovariable X_j

Anpassungsgüte

AIC:

$$AIC = -2l(\hat{b}) + 2(p+1)$$

Prüfung der Modellannahmen

Pearson-Residuen: Unter der modellierten Poisson-Verteilung gilt $E(Y_i|x_i) = Var(Y_i|x_i) = \exp(b^T x_i)$. Daher sollten die Pearson-Residuen

$$\frac{y_i - \exp(\hat{b}^T x_i)}{\sqrt{\exp(\hat{b}^T x_i)}} = \frac{y_i - \hat{\eta}_i}{\sqrt{\hat{\eta}_i}}$$

keine Struktur aufweisen (mean=0, var=1).

4.2 Poisson-Regression mit offset

Angenommen, die Follow-Up-Zeit der Beobachtungen ist unterschiedlich und t_i bezeichnet die Follow-Up-Länge von Individuum i. Definiere

$$Y_i = \text{Anzahl Ereignisse in Zeitintervall } [0, t_i]$$

Dann wird ein Poisson-Regressionsmodell defniert durch

$$Y_i|x_i, t_i \sim Poi(t_i\eta_i) = Poi(t_i \exp(b^T x_i)) = Poi(\exp(log(t_i) + b^T x_i))$$

(denn die erwartetet Ereigniszahl sollte sinnvollerweise proportional zur Beobachtungslänge ansteigen. $log(t_i)$ heisst offset-Variable und muss dem Statistikprogramm übergeben werden.

4.3 Verallgemeinerte lineare Modelle

ALM	GLM
linearer Prädiktor ($\eta_i = b^T x_i$)	linearer Prädiktor ($\eta_i = b^T x_i$)
$\eta_i = E(Y_i \vert x_i)$	$\eta_i = g(E(Y_i \vert x_i))$ mit Link-Fkt. g^*
$Y \vert x$ normalverteilt	Verteilung von $Y \vert x$ in Exponentialfamilie
$Y \vert x_i$ konstante Varianz	-

* g invertierbar; Wertebereich von g nicht beschränkt; g^{-1}:= Response-Funktion

5 Modelle und nichtparametrische Methoden für Ereigniszeitdaten

Beobachtet werden unabh. identisch verteilte Realisisierungen einer Zufallsvariable T, die die Zeit bis zum Eintreten eines bestimmtesn Ereignis beschreibt. Beispiele

- T = Zeit von Diagnose bis Tod (Überlebenszeit)
- T = Zeit von Diagnose bis Genesung
- T = Zeit bis zum Ausfall einer technischen Komponente

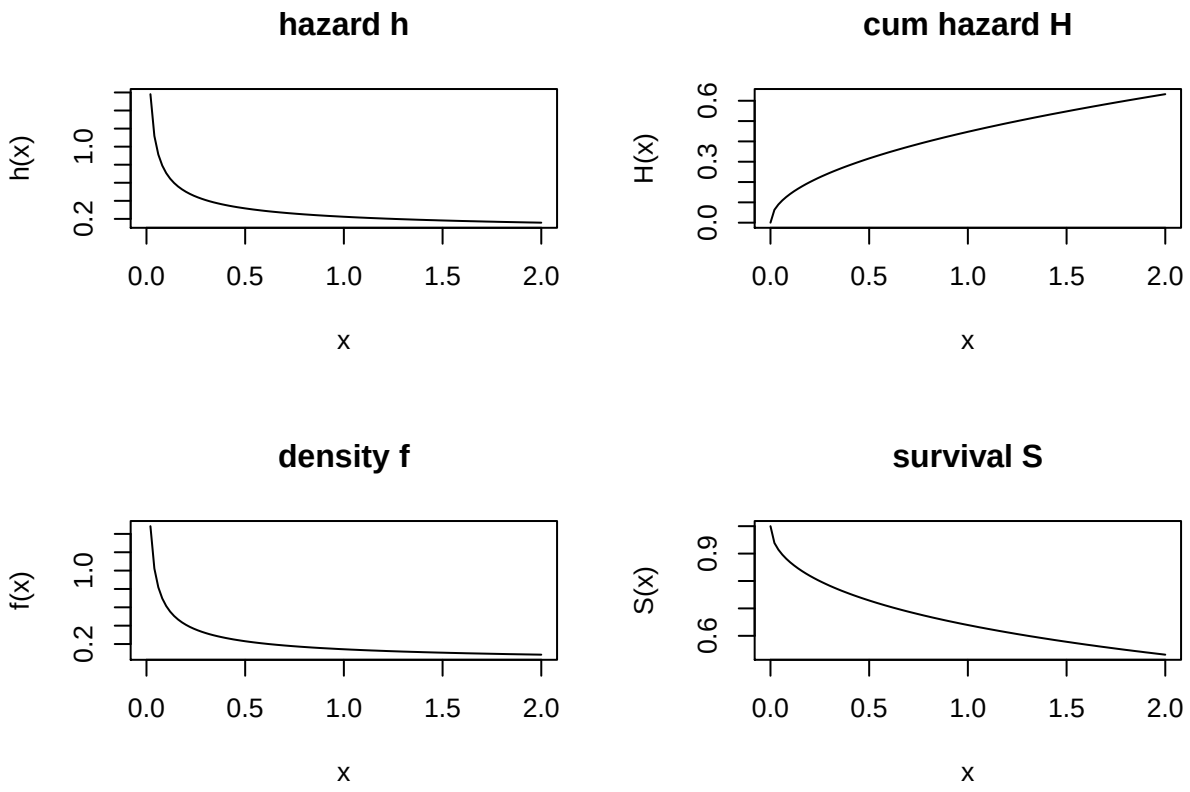
Wir nehmen im folgenden immer an, dass T eine stetige Zufallsvariable ist mit Dichte f.

Dichte, Überlebens- und Hazardfunktion

$$
\begin{aligned}
F(t) &:= P(T \leq t) = \int_0^t f(x)dx \quad \text{Verteilungsfunktion / cumulative incidence function} \\
S(t) &:= 1 - F(t) = P(T > t) \quad \text{Überlebensfunktion / survivor function} \\
h(t) &:= lim_{\Delta \downarrow 0} \frac{P(t \leq T < t + \Delta | T \geq t)}{\Delta} \quad \text{hazard rate / instantaneous event rate} \\
H(t) &:= \int_0^t h(x)dx \quad \text{cumulative hazard function}
\end{aligned}
$$

Dabei kann jede dieser Funktionen aus einer der anderen Funktionen berechnet werden, d.h. eine Funktion spezifiziert die Verteilung eindeutig.

Beispiel:



Es gilt:

$$\begin{aligned} h(t) &= \frac{f(t)}{S(t)} \\ H(t) &= -log(S(t)) \\ S(t) &= exp(-H(t)) \end{aligned}$$

Ziele der Überlebenszeitanalysen

- Schätzer $\hat{S}(t)$ oder Vorhersage $\hat{S}(t|x)$
- Einfluss von Kovariablen auf $S(t)$
- Identifikation von statistisch signifikanten Unterschieden in den Überlebenswahrscheinlichkeiten zwischen zwei oder mehr Gruppen

Rechtszensierung

- Ereigniszeiten t_i, die wir nicht beobachten, heissen zensiert
- Wissen wir nur, dass $t_i > c_i$ für ein $c_i > 0$, so heisst t_i rechts-zensierte Beobachtung
- Wir nehmen im folgenden immer an, dass T_i und die Zensierungszeit C_i unabhängig sind (ggf. bedingt auf die Kovariaten)

5.1 Nichtparametrische Schätzer von S(t)

- $(T_i, C_i)_{i=1\ldots n}$ sind iid Zufallsvariablen
- $t_1, \ldots t_n)$ sind die beobachteten Ereigniszeiten, d.h. die Realisierungen von $min(T_i, C_i)$
- r der n beobachteten Ereigniszeiten sind unzensiert, n-r sind zensiert
- $t_{(1)} < \cdots < t_{(r)}$ sind die geordenten nicht-zensierten Ereigniszeiten
- n_j ist die Anzahl an Individuen, die unmittelbar vor $t_{(j)}$ noch unter Risiko stehen, d.h. weder zensiert noch verstorben vor $t_{(j)}$ sind, $j = 1 \ldots r$
- d_j ist die Anzahl an Individuen, die z,Zp. $t_{(j)}$ versterben

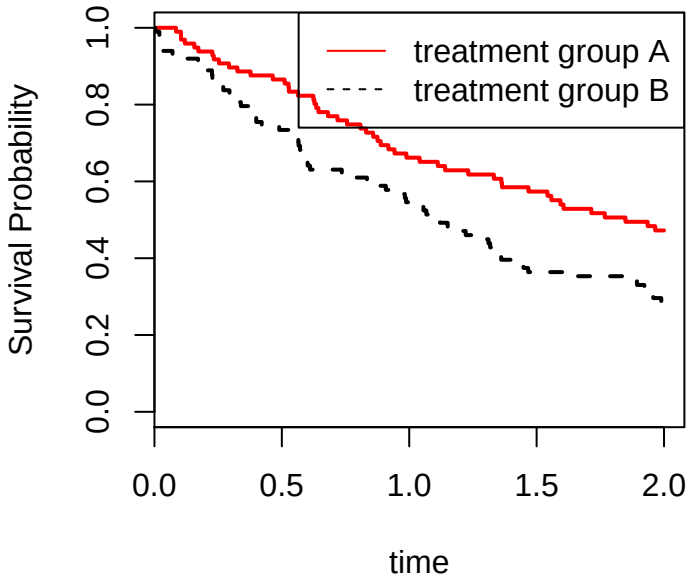
Kaplan-Meier-Schätzer:

$$\hat{S}(t) \quad = \quad \prod_{i=1}^{k(t)} \frac{n_i - d_i}{n_i} \qquad \text{mit } k(t) = max\{j; t_{(j)} \leq t\}$$

Bemerkung: Ohne Zensierungen ist $n_j - d_j = n_{j+1}$ und damit

$$\hat{S}(t_{(j)}) = 1 - \hat{F}(t)$$

mit $\hat{F}$ die empirische Verteilungsfunktion.



Nelson-Aalen-Schätzer:

$$\hat{S}_{NA}(t) = \prod_{i=1}^{k(t)} \exp(-\frac{d_j}{n_j})$$

Schätzer von H(t)

$$\hat{H}(t) = -log(\hat{S}(t)) = -log\Big(\prod_{j=1}^{k} \frac{n_j - d_j}{n_j}\Big) = -\sum_{j=1}^{k} log\Big(\frac{n_j - d_j}{n_j}\Big)$$

Schätzer der medianen Überlebenszeit

Für die mediane Überlebenszeit $t_{(50)}$ gilt $S(t_{(50)}) = 0.5$

$$\hat{t}_{(50)} = \begin{cases} min\{t_{(j)}; \hat{S}(t_{(j)}) < 0.5\}; \hat{S}(t) \neq 0.5 \forall t \\ \frac{t_{(j)}+t_{(j+1)}}{2}; \hat{S}(t_{(j)}) = 0.5 \end{cases}$$

5.2 Nichtparametrische Tests auf Gleichheit der Verteilungen

Es seien nun jeweils unabhängig identisch verteilte Ereigniszeiten $X_1, \dots X_n$ und $Y_1, \dots Y_m$ gegeben, z.B. aus zwei verschiedenen Populationen/Gruppen mit $S_i, h_i, f_i, F_i, H_i, i = 1, 2$.

Aus den Stichprobendaten dieser beiden Gruppen werden die geordneten nicht-zensierten Ereigniszeiten $t_{(1)} < t_{(2)} < ...t_{(r)}$ beobachtet.

- $n_{1j}, n_{2j}, \dots n_j$ die Anzahl an Ind. unter Risiko vor $t_{(j)}$ in den Gruppen 1, 2 und Gesamt
- $d_{1j}, d_{2j}, \dots d_j$ die Anzahl an Ereignissen zum Zp. $t_{(j)}$ in den Gruppen 1, 2 und Gesamt

Zu einem Zp. $t_{(j)}$ ergibt sich die folgende Datensituation:

Gruppe	Anzahl Events zum Zp. $t_{(j)}$	Anzahl den Zp. $t_{(j)}$ Überlebender	Anzahl unter Risiko vor $t_{(j)}$
1	d_{1j}	$n_{1j} - d_{1j}$	n_{1j}
2	d_{2j}	$n_{2j} - d_{2j}$	n_{2j}
Total	d_j	$n_j - d_j$	n_j

Logrank-Test:

Zum Testen der Nullhypothese

$$H_0 = \{S_1(t) = S_2(t) \forall t \in \mathbb{R}\}$$

kann die Logrank-Teststatistik herangezogen werden:

$$T := \frac{\sum_{j=1}^r (d_{1j} - E(d_{1j}))}{\sqrt{Var(\sum_{j=1}^r (d_{1j} - E(d_{1j})))}} = \frac{\sum_{j=1}^r (d_{1j} - n_{1j}\frac{d_j}{n_j})}{\sqrt{\sum_{j=1}^r Var(d_{1j})}} \underset{H_0}{\sim} N(0,1)$$

Wilcoxon-Test:

$$T := \frac{\sum_{j=1}^r n_j (d_{1j} - E(d_{1j}))}{\sqrt{\sum_{j=1}^r n_j^2 Var(d_{1j})}} \underset{H_0}{\sim} N(0,1)$$

Der Wilcoxon-Test gewichtet die Avweichungen $d_{1j} - E(d_{1j})$ mit der Anzahl an Personen unter Risiko. D.h. Unterschiede in $S(t)$ zu frühen Zeitpunkten t werden stärker gewichtet als die zu späteren Zeitpunkten.

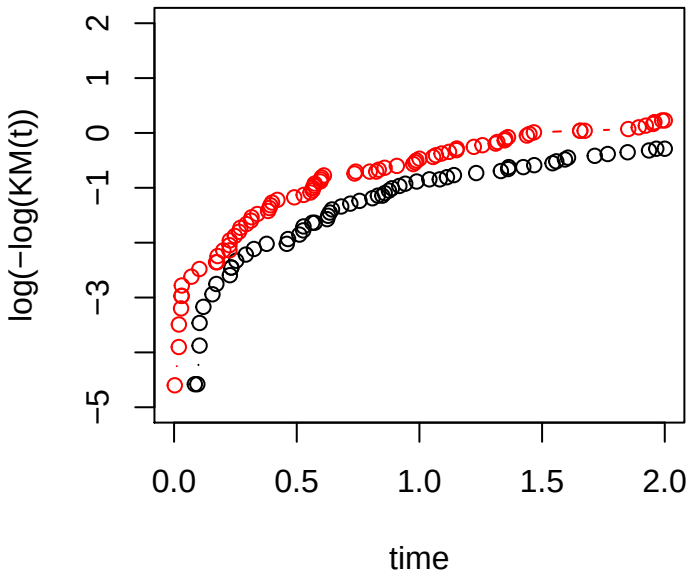
5.3 Die Proportional Hazards Annahme

Der Logrank-Test ist vorzuziehen, wenn die PH-Annahme erfüllt ist, d.h.

$$\frac{h_1(t)}{h_2(t)} \equiv \phi$$

Eine graphische Methode, um dies zu überprüfen ergibt sich aus:

$$\begin{aligned} & h_1(t) = \phi h_2(t) \\ \Leftrightarrow \quad & H_1(t) = \phi H_2(t) \\ \Leftrightarrow \quad & log(H_1(t)) = log(\phi) + log(H_2(t)) \end{aligned}$$



5.4 Parametrische Modelle

Wird die hazard-Funktion parametrisch spezifiziert (und damit auch f, S und H) können die Parameter und damit die Verteilung mit Maximum-Likelihood-Mehoden geschätzt werden.

Exponential-Modell

$$\begin{aligned} h(t) &\equiv \lambda, \quad \lambda > 0 \\ H(t) &= \lambda t \\ S(t) &= exp(-\lambda t) \\ f(t) &= \lambda exp(-\lambda t) \end{aligned}$$

Medianes Überleben:

$$S(t) = 0.5 \Leftrightarrow exp(-\lambda t) = 0.5 \Leftrightarrow t = -\frac{log(0.5)}{\lambda} = \frac{log(2)}{\lambda}$$

Weibull-Modell

$$\begin{aligned} h(t) &= \lambda\gamma t^{\gamma-1}, \quad \lambda > 0, \gamma > 0 \\ H(t) &= \lambda\gamma \int_0^t x^{\gamma-1} dx = \lambda x^{\gamma}|_0^t = \lambda t^{\gamma} \\ S(t) &= exp(-\lambda t^{\gamma}) \\ f(t) &= -(-\lambda)\gamma t^{\gamma-1} exp(-\lambda t^{\gamma}) = \lambda\gamma t^{\gamma-1} exp(-\lambda t^{\gamma}) \end{aligned}$$

Medianes Überleben:

$$S(t) = 0.5 \Leftrightarrow exp(-\lambda t^{\gamma}) = 0.5 \Leftrightarrow t^{\gamma} = \frac{log(2)}{\lambda} \Leftrightarrow t = \left(\frac{log(2)}{\lambda}\right)^{1/\gamma}$$

Proportional hazard Exponential- / Weibull-Modell

$$h(t|x) = exp(b^T x) h_0(t)$$

mit $h_0(t) = \lambda$ bzw. $h_0(t) = \lambda\gamma t^{\gamma-1}$

6 Polynomiale Regression und Regression Splines

7 Lokale Regression

8 Verallgemeinerte additive Modelle (GAM)