
Modelldiagnostik

Dr. Dirk Hasenclever
IMISE, Leipzig

Hasenclever@IMISE.uni-Leipzig.de

Aufgaben der Modelldiagnostik

- Hat man ein Modell an einen Datensatz gefittet, hat man nur das beste Modell **der betrachteten Modellfamilie** gefunden.
- Es bleibt zu prüfen, ob
 - 📁 Der beste Fit überhaupt gut ist (**Globalfit**)
 - 📁 **Modellannahmen** (z.B. Proportional Hazard) verletzt sind.
 - 📁 Wie die **funktionale Form des Einflusses stetiger Kovariater** richtig angesetzt wurde.
 - 📁 Ob es **Outlier** oder besonders **einflußreiche Beobachtungen** gibt?

Globalfit

- Der Globalfit wird typischerweise durch die Analyse geeigneter **Residuen** beurteilt:
- **Lineare Regression:**

$$res_i := y_i - \tilde{y}_i \quad \text{mit} \quad \tilde{y}_i = \mu + \beta^T z_i$$

- **Logistische Regression:**

$\hat{p}_i - \tilde{p}_i$ mit $\hat{p}_i$ Häufigkeit in Subgruppe i und

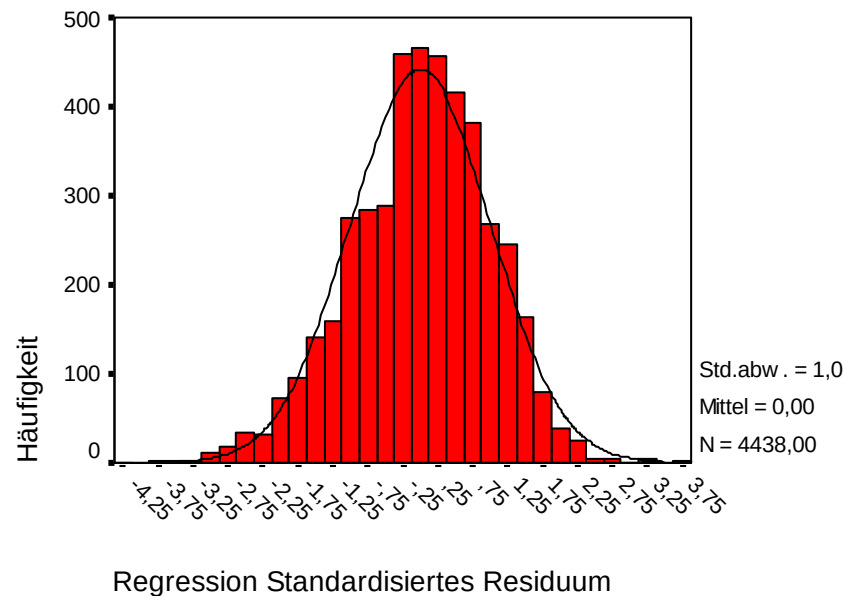
$$\tilde{p}_i = \frac{\exp(\alpha + \beta^T z_i)}{1 + \exp(\alpha + \beta^T z_i)} \text{ der Modellschätzer für } i$$

Residuen Lineare Regression I

Prüfe Normalverteilung der Residuen

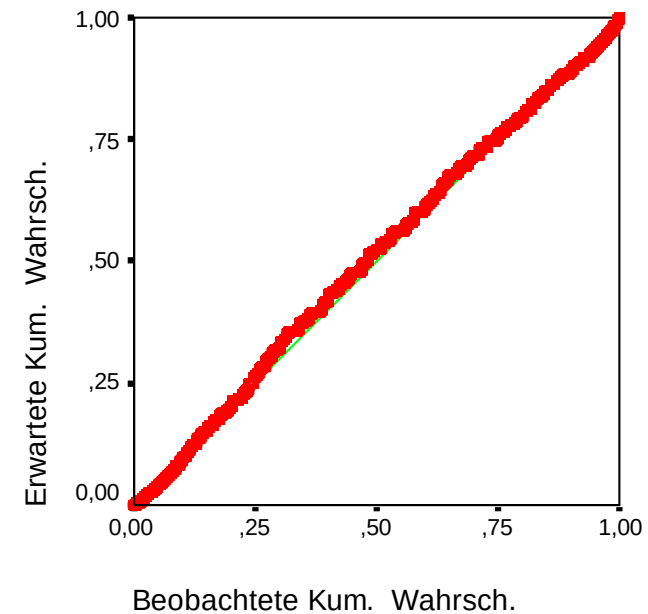
Histogramm

Abhängige Variable: Hämoglobin



P-P-Diagramm von Standardisiertes

Abhängige Variable: Hämoglobin



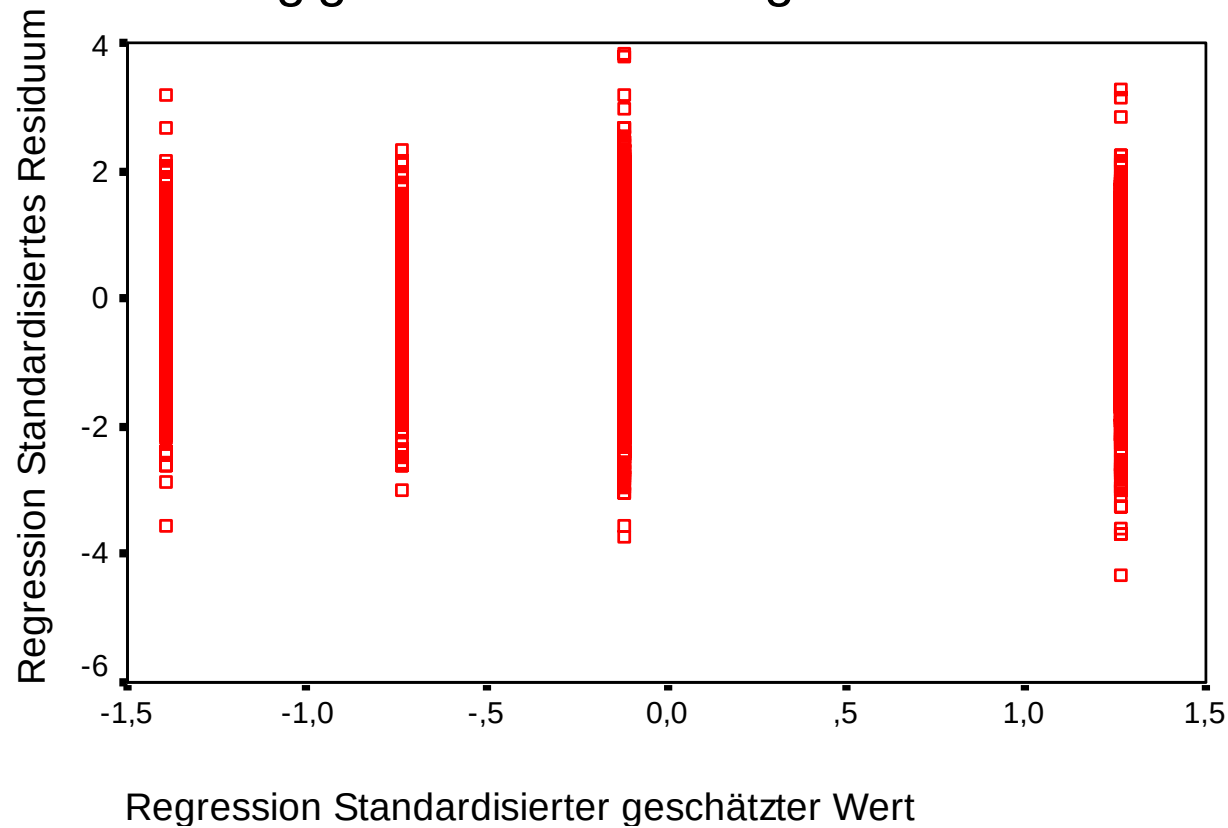
Residuen Lineare Regression II

Streudiagramm

Abhängige Variable: Hämoglobin

Prüfe

Varianzhomogenität



Transformation von Dichten

Sei X eine Zufallsvariable mit Dichte $f_x(x)dx$ und Verteilungsfunktion $F_x(x)$.

Sei h eine streng monotone stetig differenzierbare Transformation mit Umkehrabbildung h^{-1}

Definiere die Zufallsvariable $Y: = h(X)$.

Bestimme die Dichte $f_y(y) dy$: [Transformationsformel für Integrale!]

$$F_y(y) = F_x(h^{-1}(y)) = \int_{h^{-1}(-\infty)}^{h^{-1}(y)} f_x(w)dw = \int_{-\infty}^y f_x(h^{-1}(u)) \frac{d(h^{-1})}{du}(u)du$$


Also

$$f_y(y) dy = f_x(h^{-1}(y)) \cdot \frac{d(h^{-1})}{dy}(y)dy$$

Transformation von Dichten (Bsp. Lognormalverteilung)

Beispiel :

Sei $X \sim N(\mu, \sigma^2)$

Definiere Zufallsvariable $Y := \exp(X)$

Y heißt lognormalverteilt, da $\ln(Y)$ normalverteilt ist.

Bestimme die Dichte der Lognormalverteilung :

$$F_Y(y) = F_X(\ln(y)) = \int_{\ln(0)}^{\ln(y)} f_X(w) dw = \int_{-\infty}^y f_X(\ln(u)) \frac{d(\ln(u))}{du} du = \int_{-\infty}^y f_X(\ln(u)) \frac{1}{u} du$$

!

Also

$$f_Y(y) dy = \frac{1}{y} \cdot \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln(y) - \mu)^2}{2\sigma^2}\right) dy$$

Transformation durch Verteilungsfunktion

P-Werte sind unter H_0 gleichverteilt auf $[0,1]$

Sei T Zufallsvariable mit Dichte f_T und Verteilungsfunktion F_T

Definiere Zufallsvariable $P := F_T(T)$

Bestimme die Dichte von P :

$$\begin{aligned} F_P(p) &= F_T(F_T^{-1}(p)) = \int_{F_T^{-1}(0)}^{F_T^{-1}(p)} f_T(w) dw = \int_0^p f_T(F_T^{-1}(u)) \cdot \frac{dF_T^{-1}(u)}{du} du \\ &= \int_0^p f_T(F_T^{-1}(u)) \frac{1}{\left(\frac{dF_T(F_T^{-1}(u))}{du} \right)} du = \int_0^p f_T(F_T^{-1}(u)) \frac{1}{f_T(F_T^{-1}(u))} du = \int_0^p du \end{aligned}$$

Ableitung der
Umkehrfunktion



Also $f_P(p) = 1$ und $P \sim \text{Uniform}(0,1)$

Cox-Snell-Residuen I

- $X=S(T) \sim \text{uniform}[0,1]$
(vergleich früherer Beweis)

- Die Kumulierte Hazardfunktion

$$\boxed{Y=H_S(X) = -\ln(X) = -\ln(S(T))}$$

ist standard-exponential verteilt nach der Transformationsformel:

$$f_y(y) dy = f_x(h^{-1}(y)) \cdot \left| \frac{d(h^{-1})}{dy} \right| (y) dy = 1 \cdot \left| \frac{d(\exp(-y))}{dy} \right| (y) dy = \exp(-y) dy$$

- Für die Standard-Exponentialverteilung S_{exp} gilt allgemein: :

$$S_{\text{exp}}(h) = \exp(-h) \text{ und } H_{S_{\text{exp}}}(h) = -\ln(S_{\text{exp}}(h)) = h$$

Cox-Snell-Residuen II

Sei $\hat{S}_{\text{cox}}(t \mid \hat{\beta}, z)$ die Modellbasierte geschätzte Überlebenskurve für einen Patienten mit Kovariaten z zum Zeitpunkt t .

Wenn keine zensierten Werte vorliegen, sollten bei guter Modellanpassung die Residuen :

$$\text{res}_j := -\ln(\hat{S}_{\text{cox}}(t_j \mid \hat{\beta}, z_j)) \quad j \text{ über alle Patienten}$$

Standard - Exponential verteilt sein.

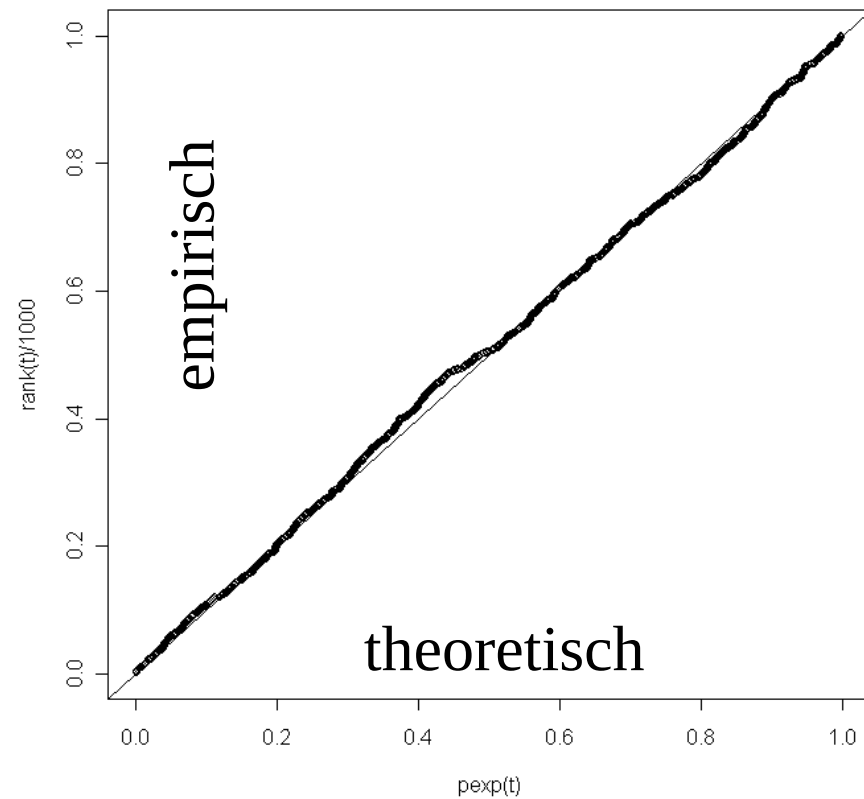
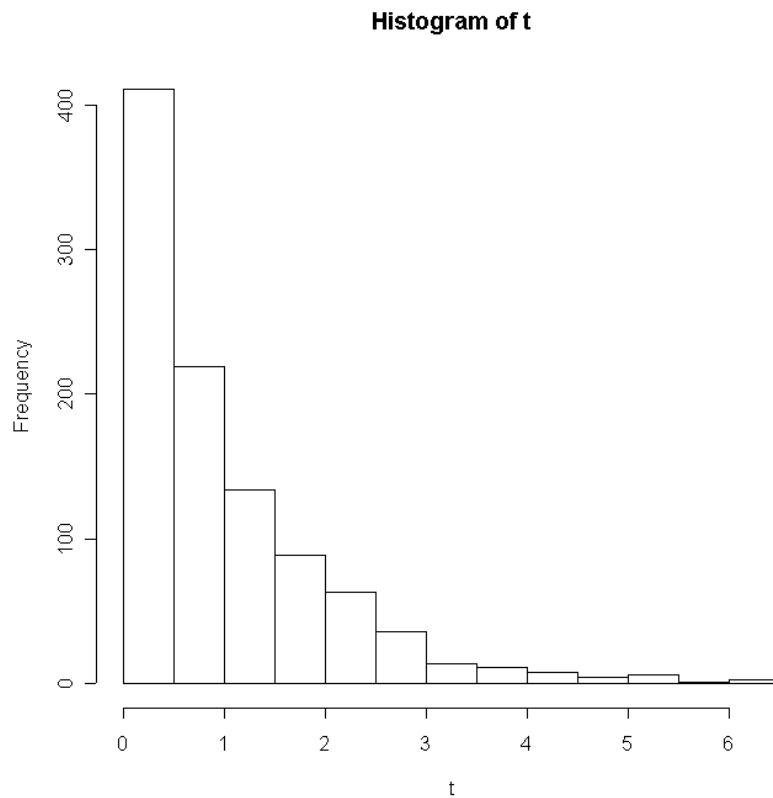
Cox-Snell-Residuen III

Übereinstimmung mit einer theoretischen Verteilung kann allgemein mittels eines q - q Plots überprüft werden.

Dabei werden die Quantile der Werte der empirischen Verteilung gegen die theoretischen Quantile zu denselben Werten aufgetragen.

Die Punkte sollten nahe der Winkelhalbierenden liegen.

Cox-Snell-Residuen III



Stichprobe N=1000 aus Standardexponentialverteilung

q-q Plot

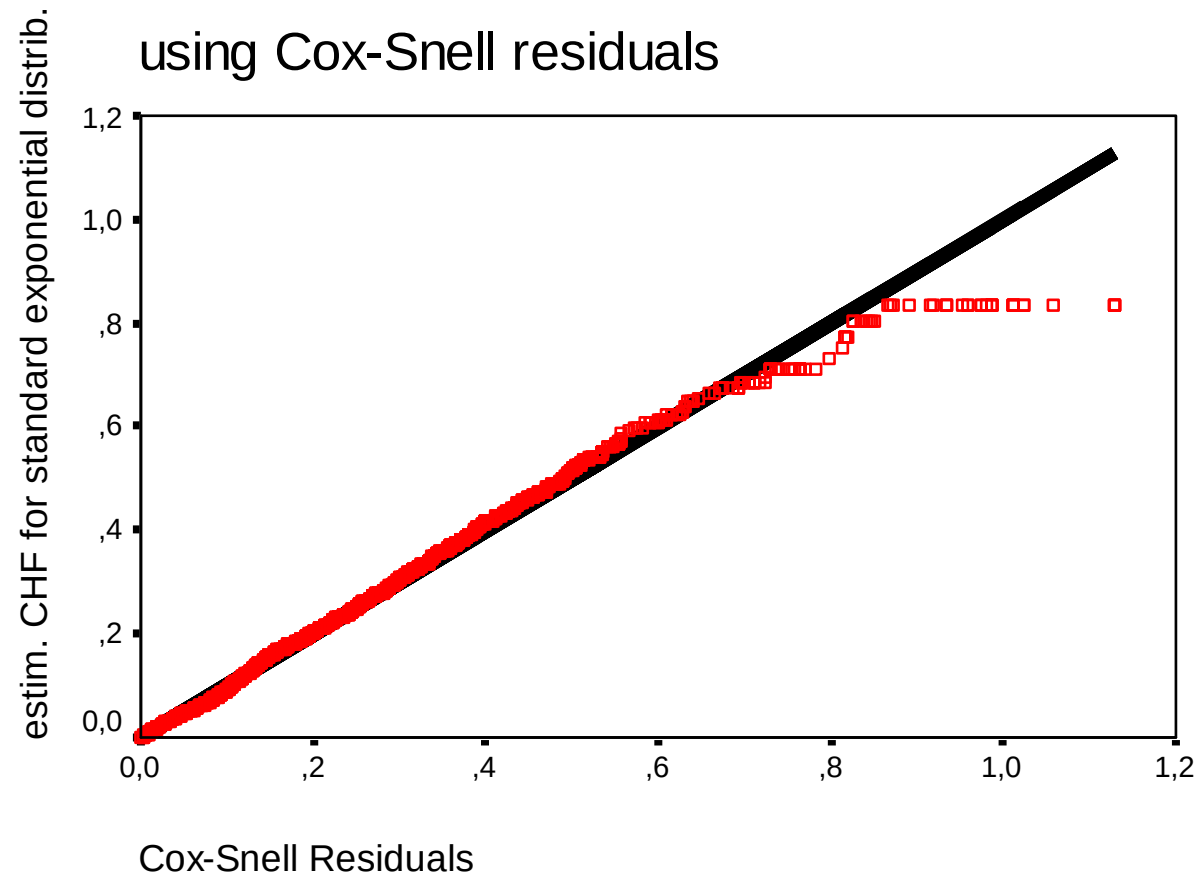
Cox-Snell-Residuen IV

- Bei zensierten Daten sollten die Residuen eine zensierte Stichprobe aus der Standardexponentialverteilung ergeben.
- Die empirischen Quantile für den q-q Plot müssen dann über z.B. Kaplan Meier Schätzer mit res_j als Zeit geschätzt werden.
- „Daten“ also $(\text{res}_j, \boxtimes_j)$ j über alle Patienten

Beispiel Cox-Snell Residuen

Global fit Assessment of IPS

using Cox-Snell residuals



Diagnostik: Proportional Hazard Annahme

Zur Überprüfung der Proportional Hazard Annahme für diskrete Gruppen :

$$\text{Falls } S_1(t) = S_0(t)^R \Leftrightarrow$$

$$\ln(S_1(t)) = R \cdot \ln(S_0(t)) \Leftrightarrow$$

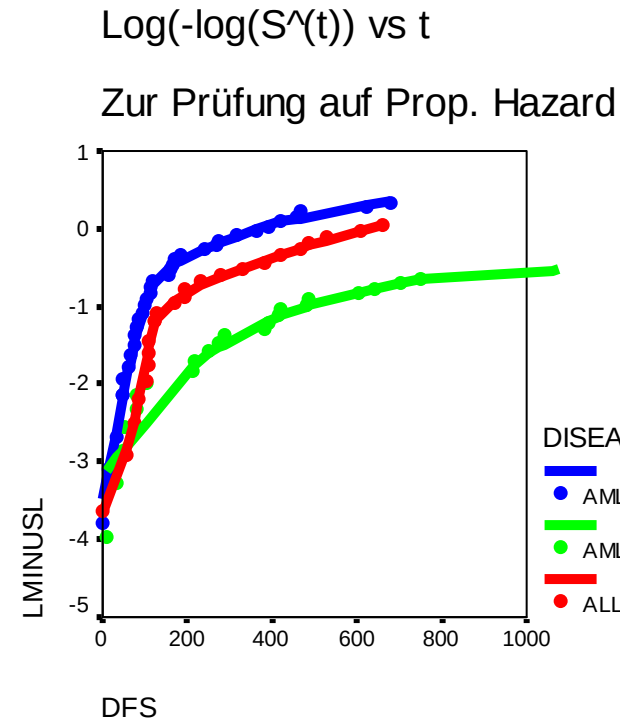
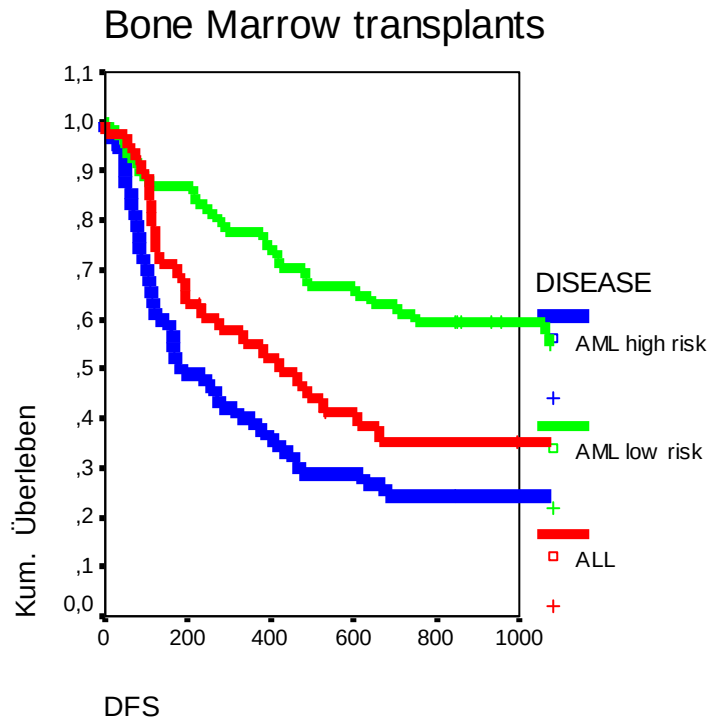
$$\ln(-\ln(S_1(t))) = \ln(-\ln(S_0(t))) + \ln(R)$$

Also plote $\ln(-\ln(\hat{S}_i(t)))$ $i = \text{über diskrete Gruppen gegen } t$.

Unter der Proportional Hazard Annahme sollten Kurven ~ gleichen Abstand haben.

Verwende einfach Kaplan - Meier oder Überlebenskurven vermittelt COX - Modell stratifiziert nach der Gruppenvariable. Statifikation bedeutet, dass nach Gruppenvariable separat der Baseline geschätzt wird, der Einfluss der Kovariaten aber als quantitativ gleich angenommen wird.

Diagnostik: Nicht proportionaler Hazard



Low risk AML hat abweichende Hazardfunktion, da im log(-log) Plot nicht im gleichen vertikalen Abstand.

Martingal Residuum

Bestimme die funktionale Form des Einflusses einer stetigen Kovariate :

linear,

quadratisch,

logarithmisch,

logistisch,

binäre Sprungfunktion...???

....

Definiere Martingal - Residuum mittels der Differenz von
Zensurindikator δ_i und dem COX - Snell Residuum res_i :

$$M_i := \delta_i - res_i$$

Martingal Residuum II

Eigenschaft der Martingal - Residuen :

Beschreibe $h(t | Z^*, z) = h_0(t) \cdot \exp(\beta^{*T} Z^*) \cdot \exp(f(z))$ das wahre Modell.

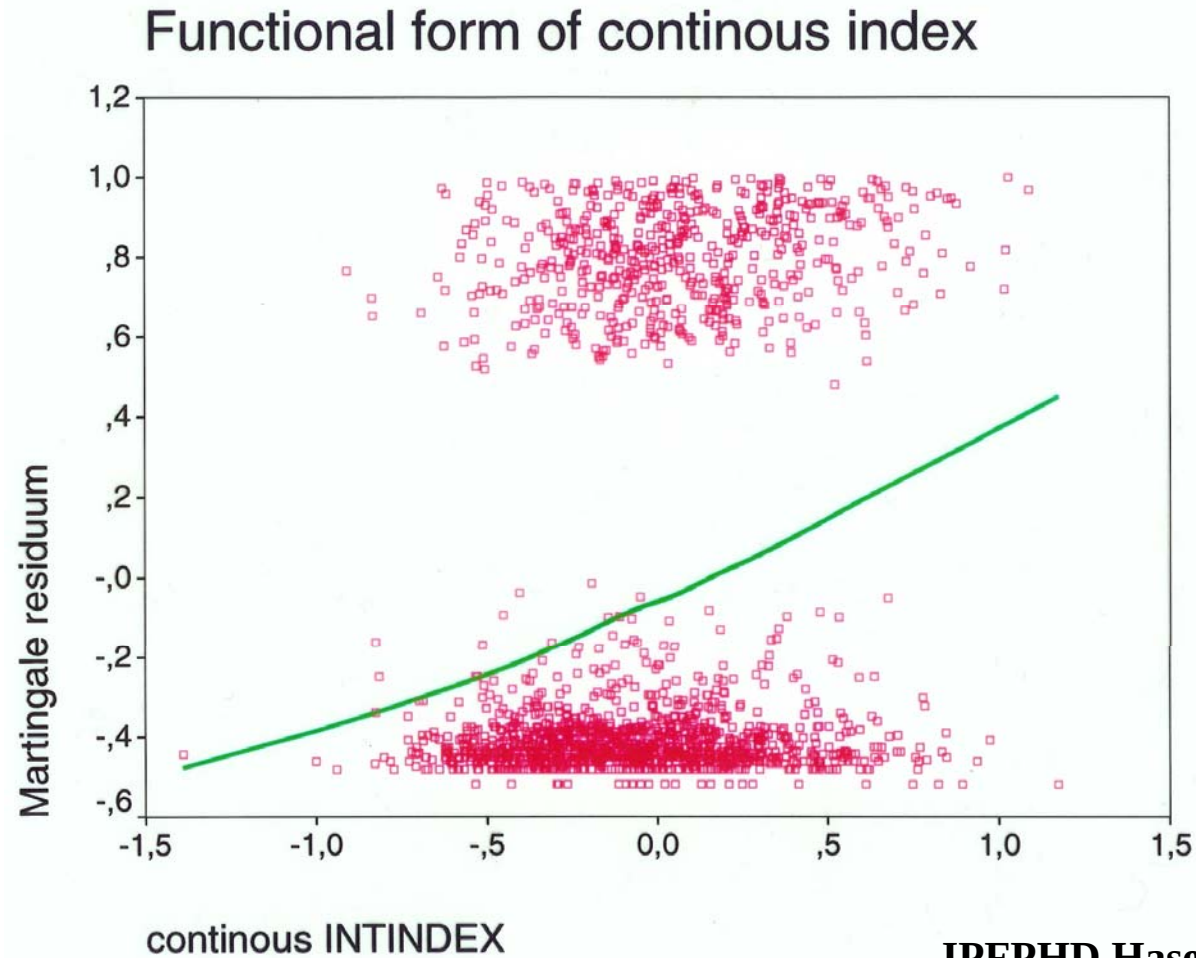
Bezeichne z die interessierende stetige Kovariate, Z^* die anderen Kovariaten und $f(z)$ die korrekte funktionale Form des Einflusses von z :

Dann gilt für die Martingal - Residua des COX - Modells mit Z^* , aber ohne z :

$$E[M_j | z_j] \approx [f(z_j) - a] \cdot b$$

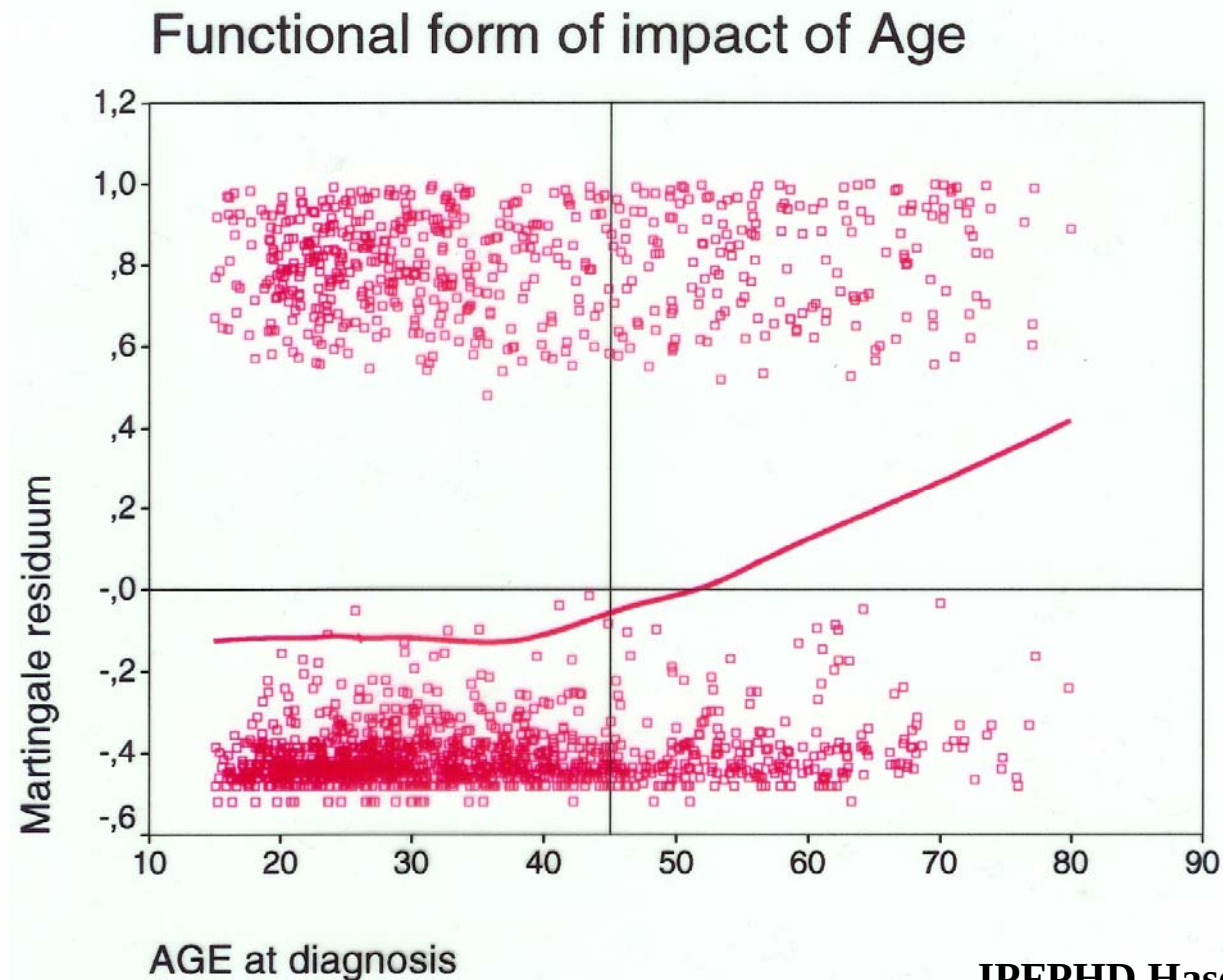
mit Konstanten a, b , die insbesondere nicht von z abhängen.

Funktionale Form: Linear



IPFPD Hasenclever 1998

Funktionale Form: „Rampe“



IPFPHD Hasenclever 1998

Glätten von Punktwolken I

Problem : Gegeben Punktwolke

- möglicherweise mit diversen Ausreißern -

Wie kann man den Zusammenhang
durch eine glatte Kurve beschreiben?

Idee : $y_i = g(x_i) + \varepsilon_i \quad i = 1, \dots, n$

$g()$ glatte Funktion

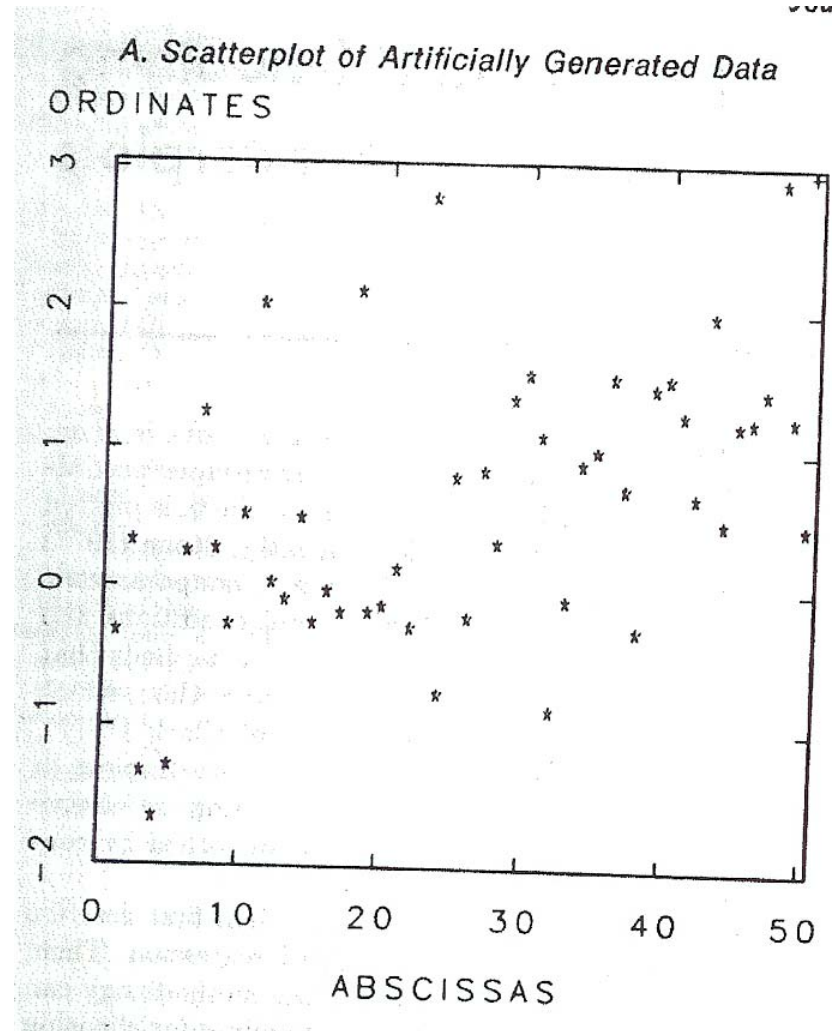
$E(\varepsilon_i) = 0$ und

gemeinsame konstante Skala

Ansatz : Da $g()$ glatt, enthalten

Nachbarnpunkte von x_i Information über $g(x_i)$

Robust locally weighted Regression
(LOWESS) Cleveland 1979



Glätten von Punktwolken II

Beschreibe Nachbarschaft durch Gewichtsfunktionen :

Die Funktion $w : \mathbb{R} \rightarrow \mathbb{R}$ heißt Gewichtsfunktion :

$$w(x) > 0 \text{ für } |x| < 1$$

$$w(x) = 0 \text{ für } |x| \geq 1$$

$$w(-x) = w(x)$$

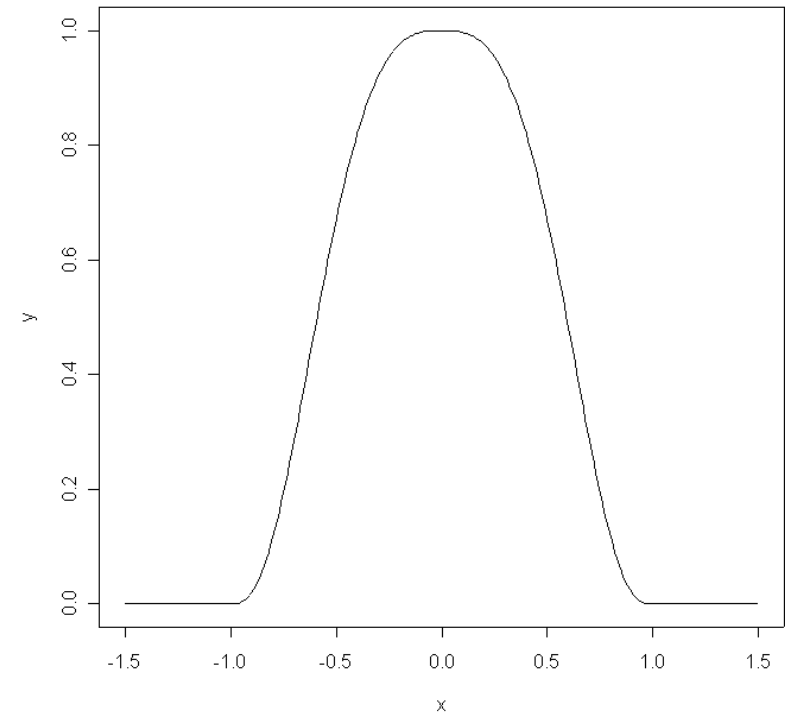
$$w(x) \text{ monoton fallend für } x \geq 0$$

z.B.
$$w(x) = \begin{cases} (1-|x|^3)^3 & |x| < 1 \\ 0 & \text{sonst} \end{cases}$$

Zu jedem x_i bezeichne $x_i^{(r)}$
den r -ten nächsten Nachbarwert.

Durch Reskalierung $\tilde{w}_i(x) = w\left(\frac{x-x_i}{|x_i - x_i^{(r)}|}\right)$ erhalten wir

zu jedem x_i eine lokale Gewichtsfunktion $\tilde{w}_i$.



Glätten von Punktwolken III

Um einen Wert für $g(x_i)$ zu gewinnen, rechne lokal gewichtete Regression (kleinste Quadrate) mit einem Polynom vom Grad d (z.B. linear $d = 1$) mit Gewichten $\tilde{w}_i$:

$$\sum_k \tilde{w}_i(x_k)(y_k - (\hat{\beta}_0 + \hat{\beta}_1 x_k))^2 \rightarrow \min$$

$\Rightarrow$ (Initialer) Schätzer für $\hat{g}(x_i) = \hat{\beta}_0 + \hat{\beta}_1 x_i$

Glätten von Punktwolken IV

Restproblem : Robustifizierung gegen Ausreißer!

Idee : Wichte offenbare Ausreißer herunter!

Residua : $e_i = y_i - \hat{g}(x_i)$

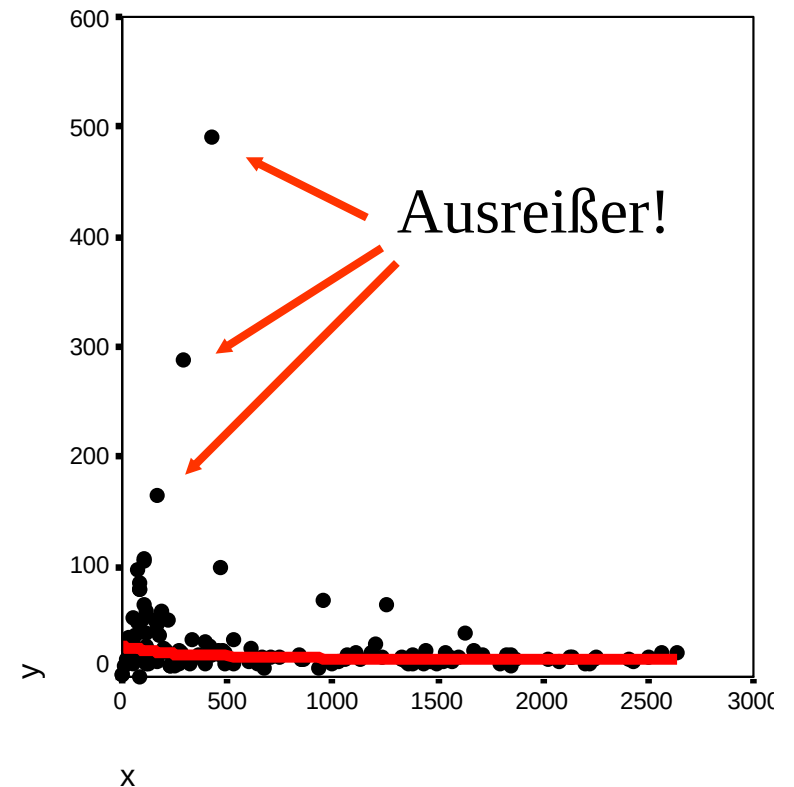
Bestimme : $s := \text{median}(e_i)$
als typische Residuumgröße.

Wähle weitere Gewichtungsfunktion B :

$$\text{z.B.: } B(x) = \begin{cases} (1 - x^2)^2 & |x| < 1 \\ 0 & \text{sonst} \end{cases}$$

$$\delta_i = B\left(\frac{e_i}{6s}\right) \text{ quantifiziert,}$$

inwieweit der Punkt (x_i, y_i) als Ausreißer erscheint.



Glätten von Punktwolken V

Iteratives Verfahren :

Schritt 1: Lokal gewichtete Regression

$$\sum_k \delta_k \cdot \tilde{w}_i(x_k) (y_k - (\hat{\beta}_0 + \hat{\beta}_1 x_k))^2 \rightarrow \min$$

⇒ Iterativer Schätzer für $\hat{g}(x_i) = (\hat{\beta}_0 + \hat{\beta}_1 x_i)$

Schritt 2: Update der Gewichte δ_i zur Runterwichtung der Residuen

Praktisch : (*R* : *lowess* ; SPSS: LOWESS Option bei Bearbeitung von Scatterplots)

Grad des Polynoms = 1

1- 3 Iterationen genügen

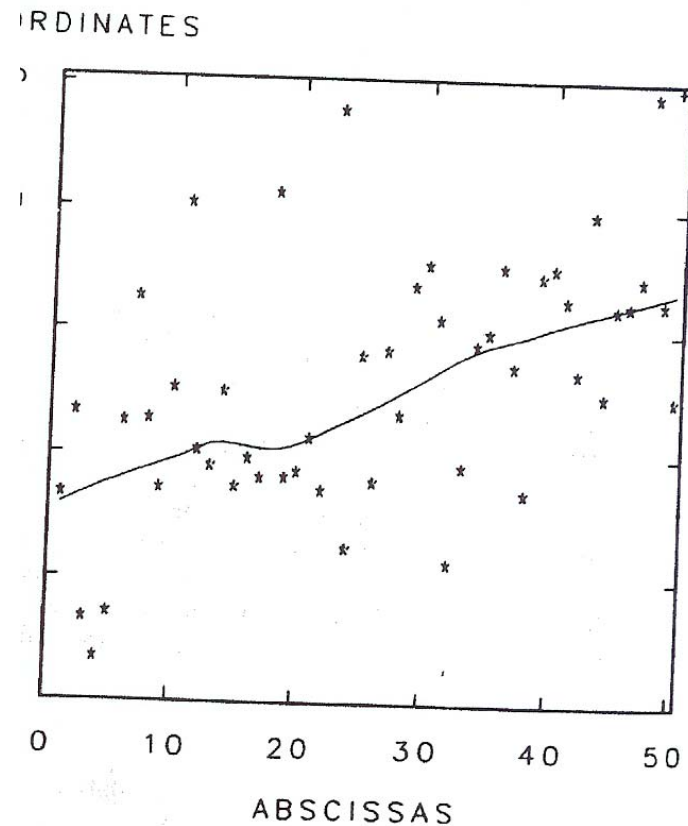
Variiere den Grad der Glättung durch % einbezogene Punkte :

Nachbarschaftsdefinition durch $r = f * n$

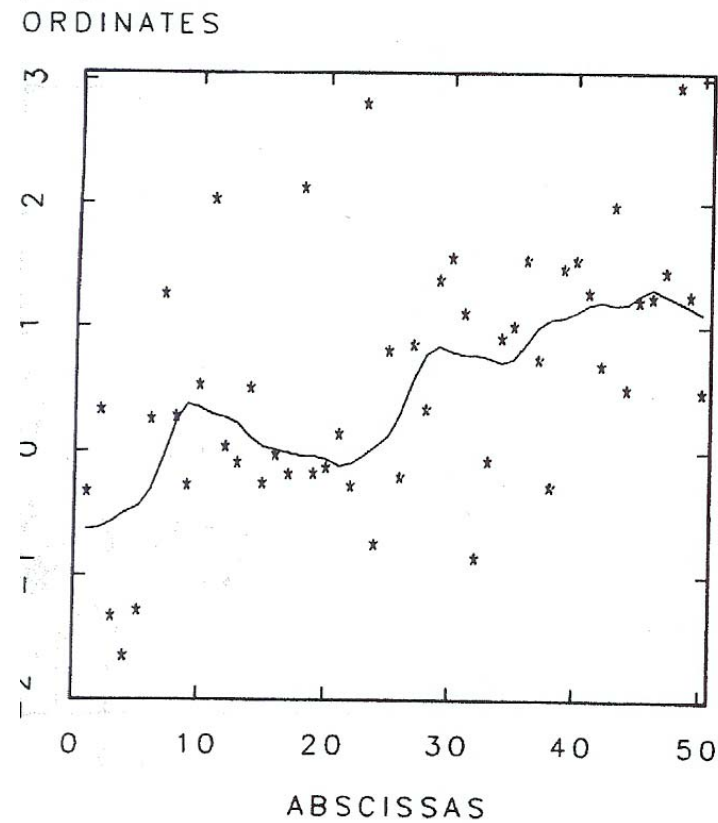
$f = 0.5$ ist oft guter Ausgangspunkt.

Glätten von Punktwolken VI

B. Scatterplot of Artificially Generated Data and Robust Smoothed Values With $f = .5$



C. Scatterplot of Artificially Generated Data and Robust Smoothed Values With $f = .2$



Unterschiedliche Glättungsgrade

Cleveland 1979