

## 4 Residuen-Analyse

### 4.1 Problemstellung

a **Modellannahmen:**  $E_i \sim \mathcal{N}\langle 0, \sigma^2 \rangle$

(a)  $\mathcal{E}\langle E_i \rangle = 0$  : Linearität, Additivität.

(b) gleiche Varianz  $\text{var}\langle E_i \rangle = \sigma^2$ ,

(c) normalverteilt.

(d)  $E_i$  unabhängig,

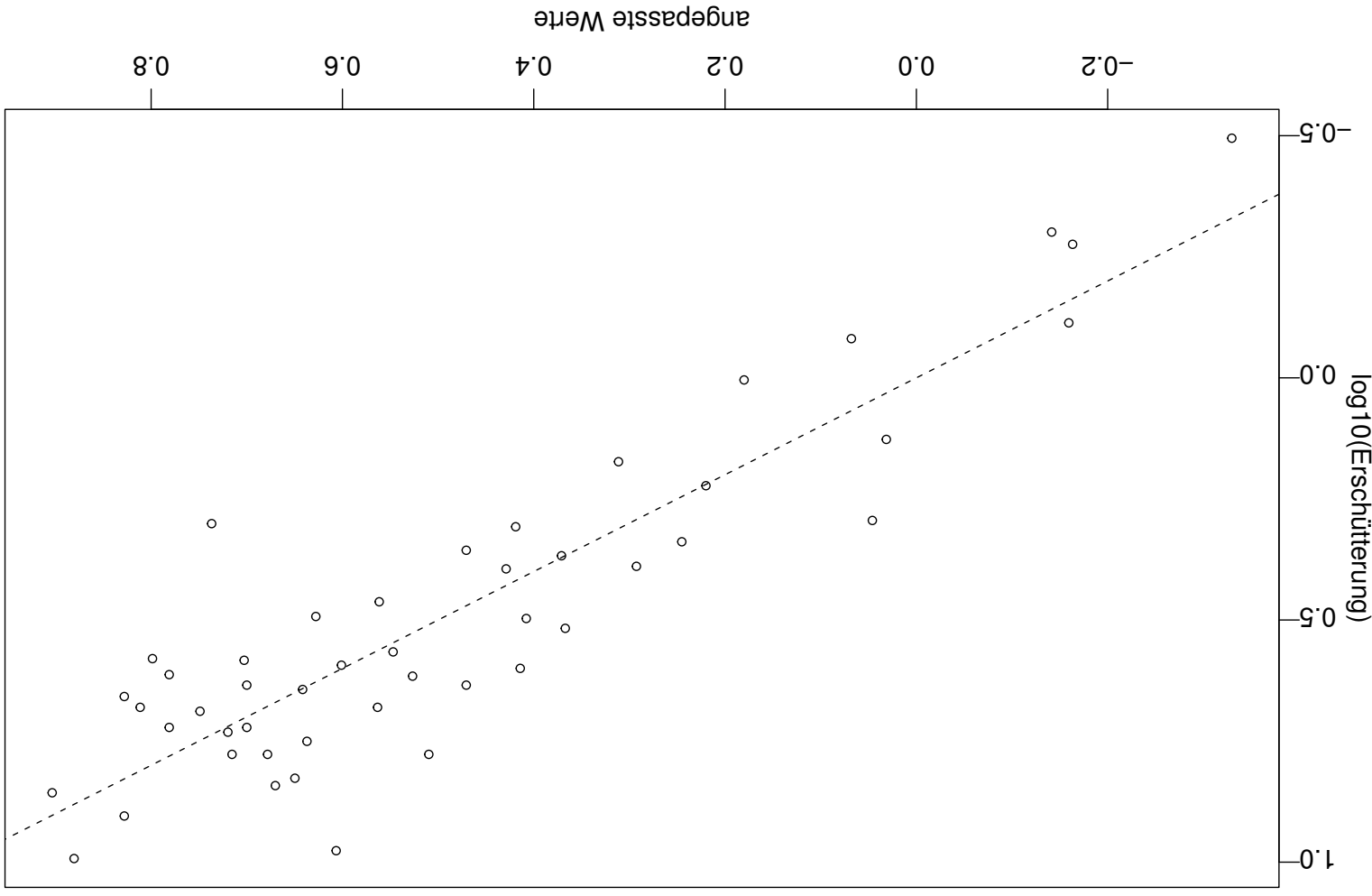
- b Voraussetzungen überprüfen!  
nicht Rechtfertigung, sondern Chance:  
**besseres Modell, explorative Datenanalyse.**
- c (Auch für **Varianzanalyse** und andere Regressionsmodelle.)  
d Verbesserungen:
- Variable **transformieren**,
  - **zusätzliche Terme**, beispielsweise Wechselwirkungen, ins Modell aufnehmen,
  - Beobachtungen gewichten,
  - allgemeinere Modelle und Methoden verwenden.

e Graf. Darstellungen (ev. Tests) = Diagnose-Untersuchung  
Abweichungen = Symptome

verschiedene Abweichungen → Syndrom → Diagnose → Therapie

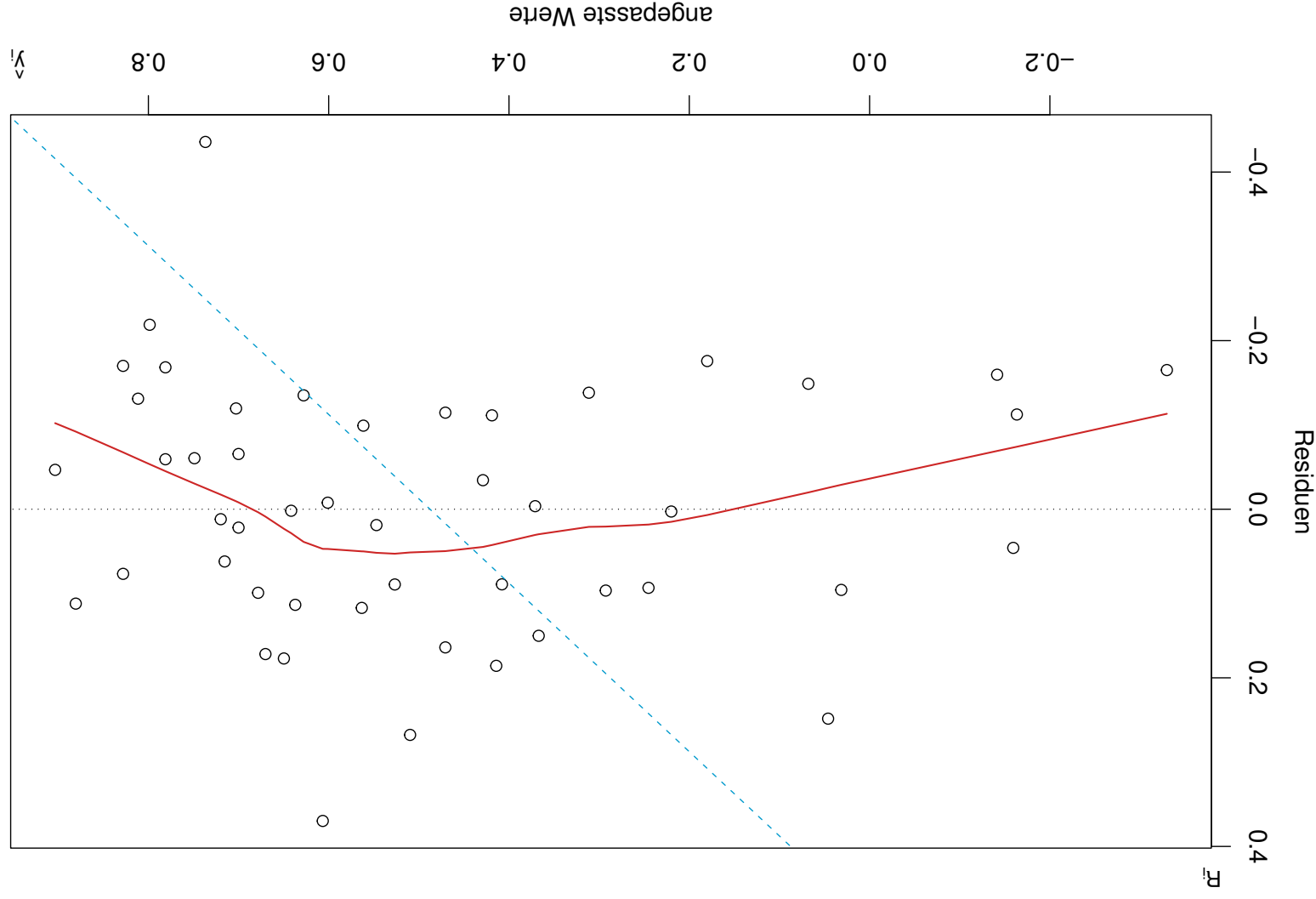
## 4.2 Residuen und angepasste Werte

- a
- Einfache Regression
  - Streudiagramm betrachten!
  - Multiple Regression
  - mehrere  $X^{(j)}$
  - angepasste Werte



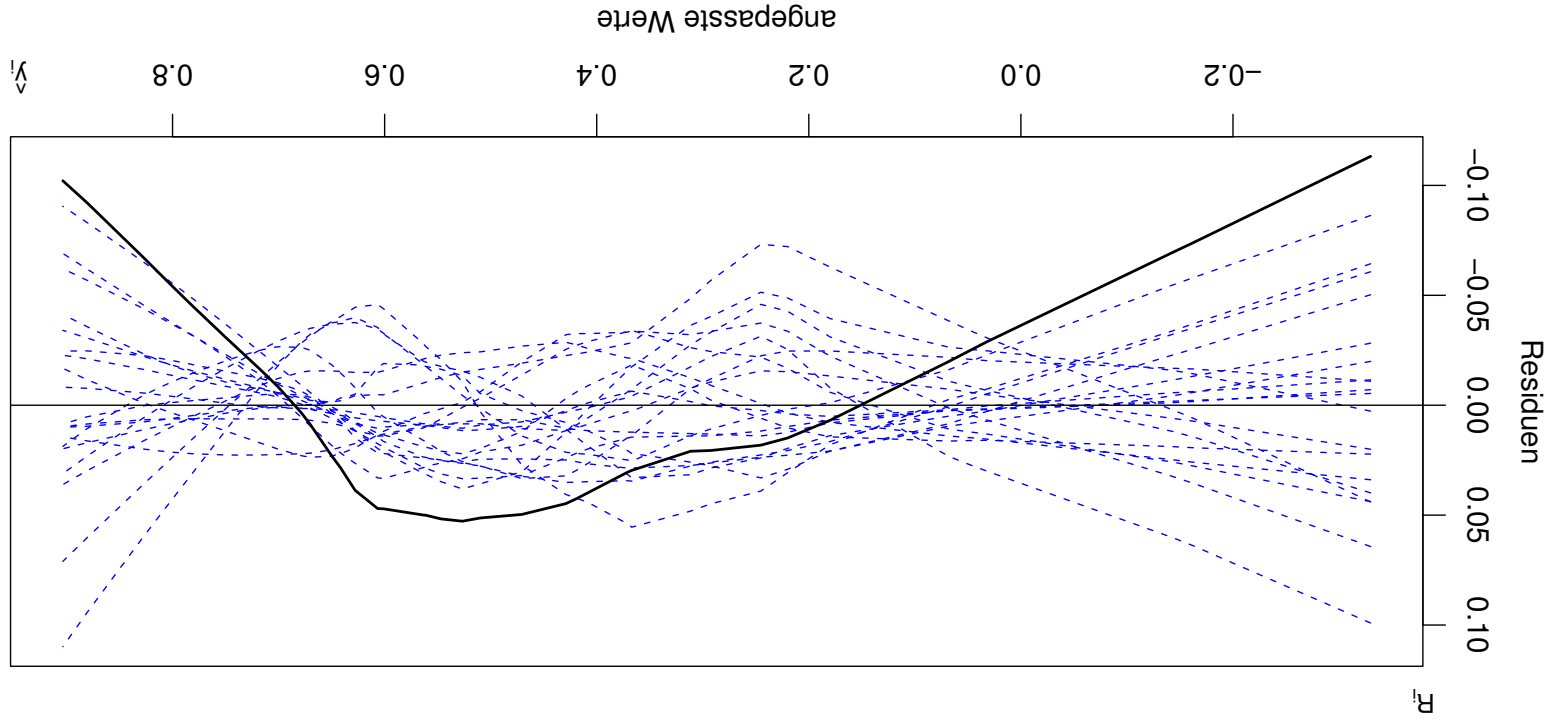
- b Welche Abweichungen von Voraussetzungen könnte man hier sehen?
- (a) **Regressionsfunktion: Verlauf** der Punkte.  
Typische Abweichung: Krümmung des mittleren Verlaufs.
- c (b) **Gleiche Varianzen: Streubreite** der Punkte um die Gerade.  
Typische Abweichung: Punkte laufen gegen rechts auseinander.
- d (c) **Verteilung der Fehler: Streuen Punkte** **symmetrisch** um die Gerade?  
Ausreißer?
- e **Wie beurteilen?**
- Abweichung im Bereich des Zufalls?
  - Abweichung gefährlich? Antwort abhängig vom Zweck der Studie!

f Variante des Diagramms  $Y$  gegen  $\hat{y}$  zeigt Abweichungen genauer:  
**Tukey-Anscombe-Diagramm**: Residuen gegen angepasste Werte

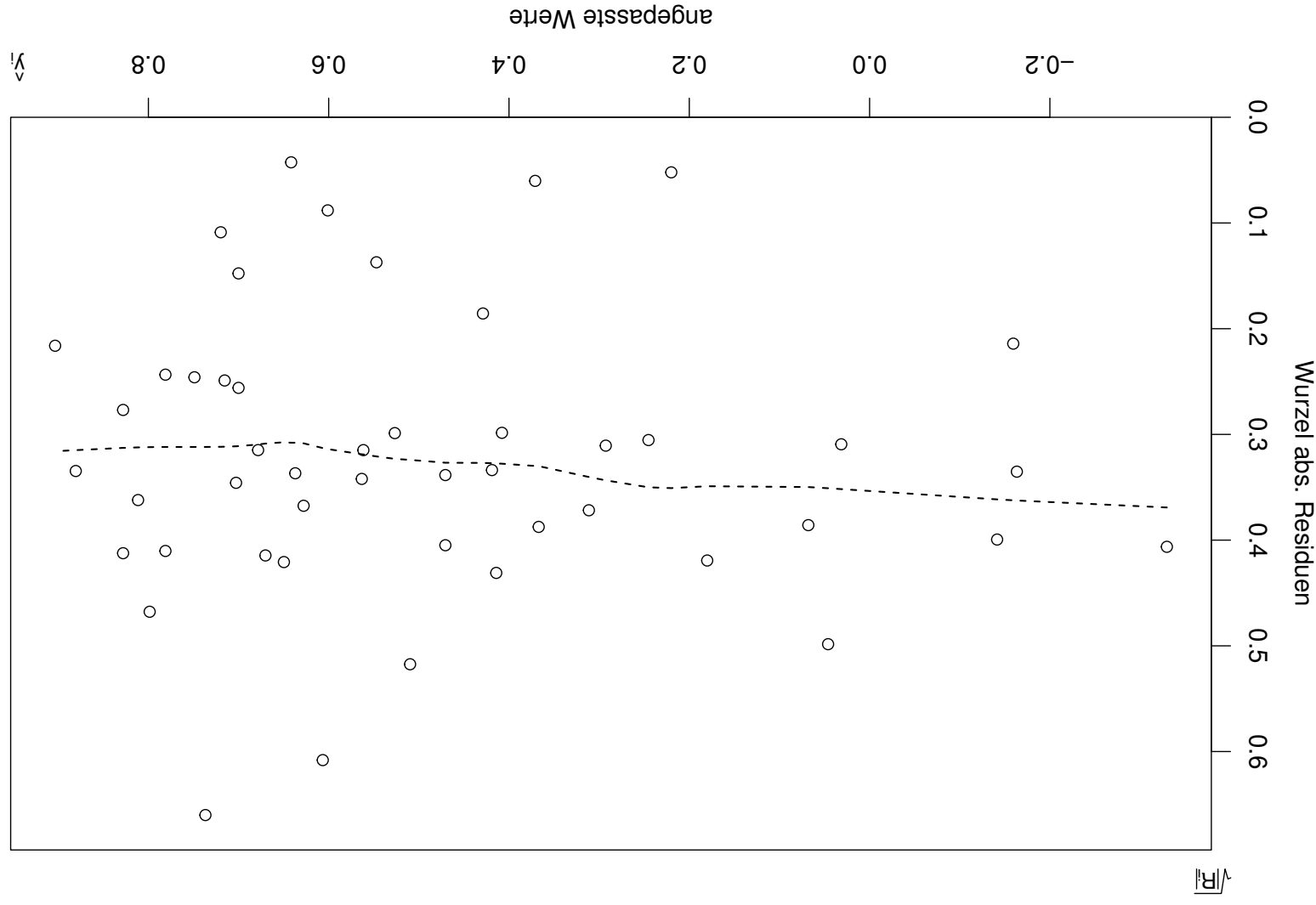


Referenzlinie für konstante  $Y$ -Werte.

- g (a) **Regressionfunktion:**  $\mathcal{E}\langle E_i \rangle = 0$ . Mittelwert der  $R_i$  über „Fenster“ der  $\hat{y}_i$
- gleitendes Mittel → Glättung „lowess“, **robust**
  - Nichtparametrische Regression.
  - h Abweichung zufällig? → Simulation
  - 19 zusätzliche Kurven. Ist die beobachtete „die extremste“?
  - „grafischer Test“

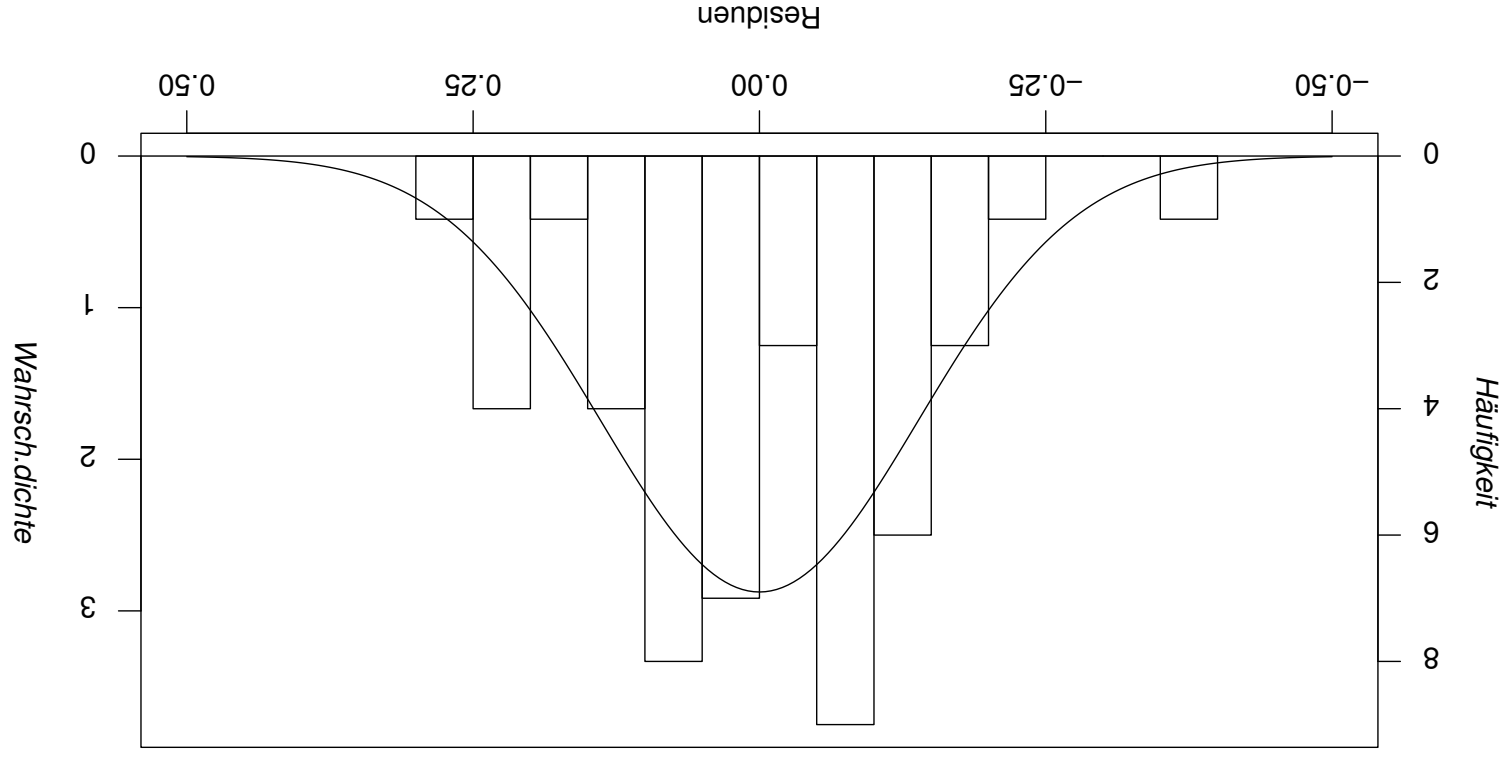


! (b) Gleiche Varianzen:  $\sqrt{|R_i|}$  besser lowess für  $\sqrt{|R_i|}$  gegen  $\hat{y}_i$ .



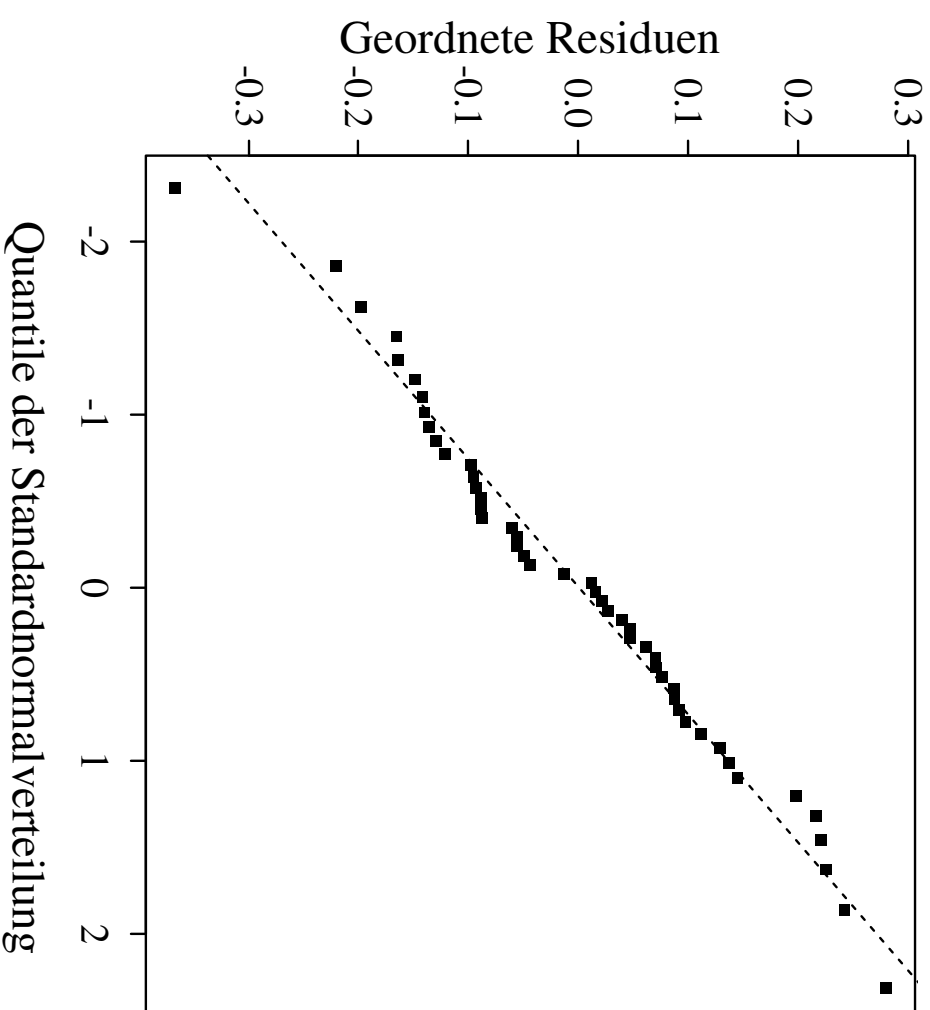
## 4.3 Verteilung der Fehler

- a (c) Normalverteilung? Histogramm der  $H_i$  resp. Residuen  $R_i$  ! Die Zielgröße  $Y$  muss nicht normalverteilt sein !



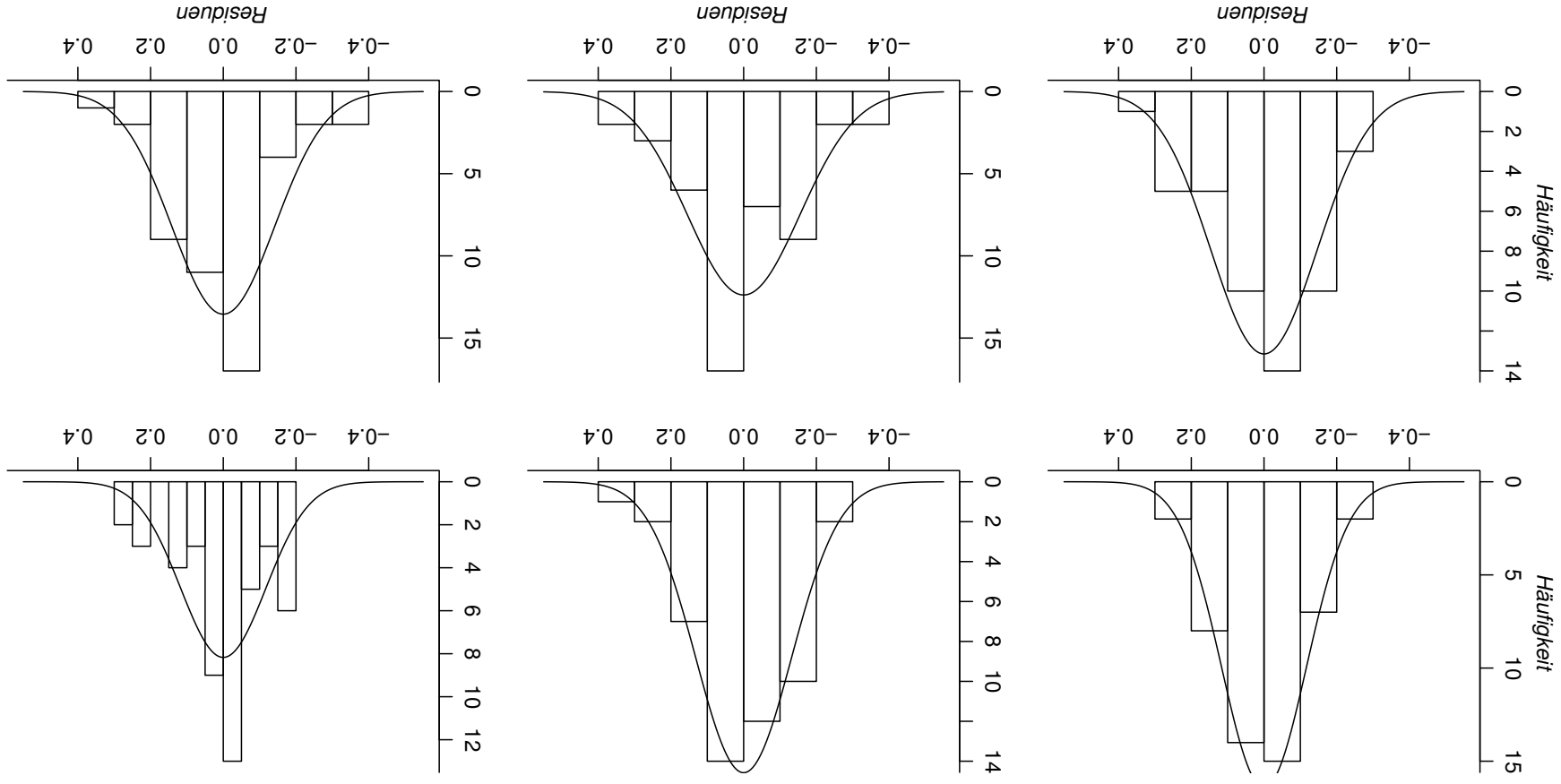
4.3

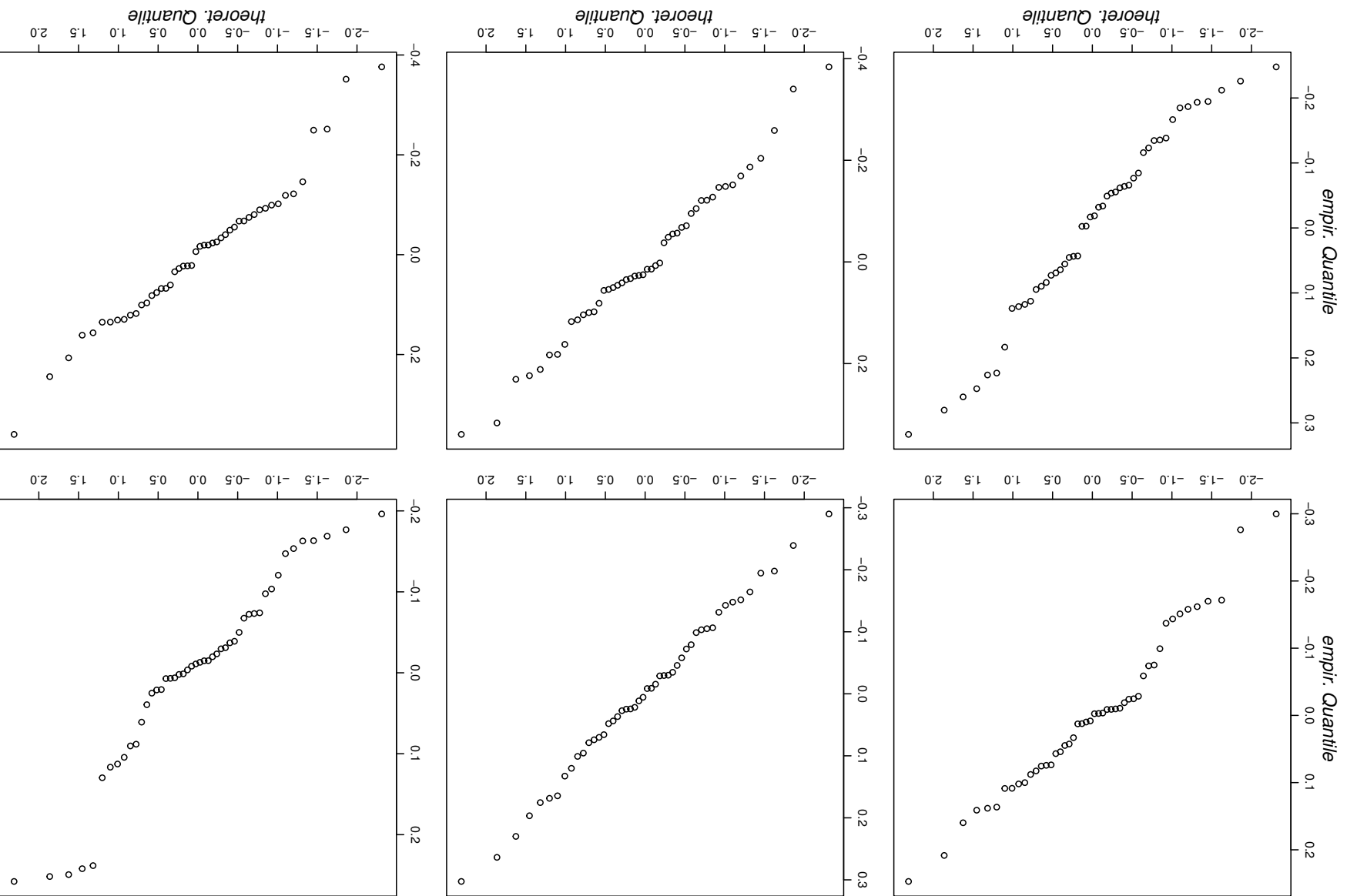
b Raffinierter: Quantil-Quantil-Diagramm (QQ-Plot, normal plot)



4.3

- c Abweichungen zufällig? **Anpassungstest** (goodness of fit test)
- d Oder simulieren!





e Verteilung der Zufallsfehler? – Zufallsfehler  $E_i \neq$  Residuen  $R_i$   
 $R_i = Y_i - \hat{y}_i$  beides zufällig.  $\hat{y}_i$  hängt von  $Y_i$ , also von  $E_i$  ab.

$$f \quad R_i \sim \mathcal{N}(0, \sigma^2(1 - H_i))$$

$H_i$  leverage, Hebelarm

$$\bullet \quad Y_i \rightarrow Y_i + \Delta y_i \quad \rightarrow \quad \hat{y}_i \rightarrow \hat{y}_i + H_i \Delta y_i \quad \text{Hebelwirkung}$$

$H_i$  misst den „Abstand“ zwischen  $\bar{x}_i$  und  $\bar{x}$ .  
 einfache R.:  $H_i = (1/n) + (x_i - \bar{x})^2 / \text{SSQ}_{(X)}$ .

multiple R.:  $H_i = (1/n) + d(x_i, \bar{x})^2$ .  $d$ : Mahalanobis-Distanz.

$$\bullet \quad 0 \leq H_i \leq 1, \quad \text{ave } \langle H_i \rangle = p/n.$$

Residuen **standardisieren**, damit sie alle die gleiche Verteilung haben:  $\varepsilon$

$$\tilde{R}_i = R_i / \left( \hat{\sigma} \sqrt{1 - H_{ii}} \right)$$

Verwende stand. Residuen zur Überprüfung der Verteilung!

Meistens ist der Unterschied der Varianzen  $\text{var}\langle R_i \rangle$  klein, deshalb genügen unstandardisierte Residuen auch.

Theoretische Verteilung der Residuen

$$\begin{aligned} \widehat{\beta} &= (\widetilde{\mathbf{X}}_T^T \widetilde{\mathbf{X}}_T)^{-1} \widetilde{\mathbf{X}}_T^T \widetilde{\mathbf{y}} \\ \widetilde{\beta} &= \widehat{\beta} - \bar{y} \mathbf{1}_T \end{aligned}$$

$$\bar{y} \mathbf{H} :=$$

$$\bar{y} = \bar{y} - \widehat{y} = \bar{y} - \widetilde{\mathbf{X}}_T^T \widetilde{\beta}$$

$$\mathcal{E} \langle \bar{y} \rangle = \mathcal{E} \langle \bar{y} \rangle - \mathcal{E} \langle \widetilde{\mathbf{X}}_T^T \widetilde{\beta} \rangle = \widetilde{\mathbf{X}}_T^T \widetilde{\beta} - \bar{y} \mathbf{1}_T = \bar{\mathbf{0}}$$

$$\bar{y} = \bar{y} - \widehat{y} = \bar{y} - \bar{y} \mathbf{H} = (\mathbf{I} - \mathbf{H}) \bar{y}$$

$$R_i = (1 - H_{ii}) y_i - \sum_{k \neq i} H_{ik} y_k$$

$$\text{var} \langle R_i \rangle = \sigma^2 \left( (1 - H_{ii})^2 + \sum_{k \neq i} H_{ik}^2 \right) = \sigma^2 (1 - 2H_{ii} + \sum_k H_{ik}^2)$$

Man muss zeigen, dass  $\sum_k H_{ik}^2 = H_{ii}$  ist. Dann ist

$$\text{var} \langle R_i \rangle = \sigma^2 (1 - H_{ii}) \text{ bewiesen.}$$

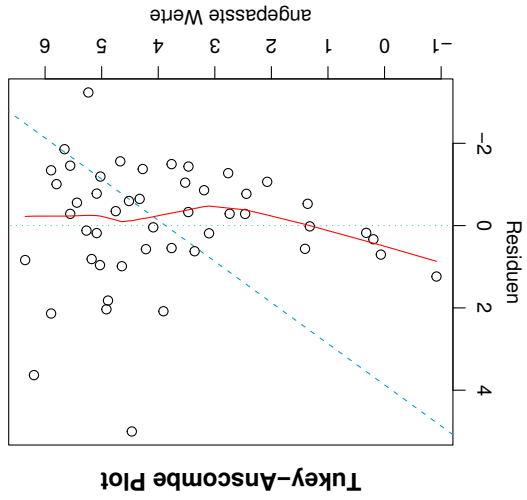
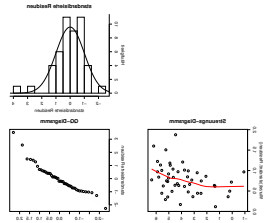
Das  $\lambda$ -Element ist also  $H_{\lambda}$

$$H = \tilde{X}^{-1} \tilde{X} \tilde{X} = \tilde{X}^{-1} \tilde{X} \tilde{X} = \tilde{X}^{-1} \tilde{X} \tilde{X}$$

Es ist  $\sum_k H_{\lambda}^k = (H H)_{\lambda}$  und

## 4.4 Zielgröße transformieren?

- a Symptome  $\rightarrow$  Syndrom  $\rightarrow$  Diagnose  $\rightarrow$  Therapie.  
Umgekehrt: Krankheit  $\rightarrow$  Syndrom  
Falsche / fehlende Transformation der Zielgröße  $\rightarrow$  ???  
Bsp. Sprengungen: Fehlende log-Transformation  $\rightarrow$  ???



b Syndrom:

- nach oben gekrümmte Glättung,
  - nach rechts trichterförmig zunehmende Streuung,
  - schiefe Verteilung der Residuen – bis auf 1 Ausreißer nach unten.
- c „Transformations-Syndrom“

### e First aid transformations

- Logarithmus-Transformation für **Konzentrationen und Beträge**
  - Wurzeltransformation für **Zählraten**
  - Arcus-Sinus-Transformation  $\tilde{y} = \arcsin \sqrt{\tilde{y}}$  für **Anteile** (Prozentzahlen/100).
- sollten für solche Daten immer angewendet werden

(wenn es keine Gegenründe gibt), auch für erklärende Variable!

### f Logarithmus-Transformation für Beträge (Mengen, Konzentrationen, ...).

$$\tilde{Y} = \log_{10}\langle Y \rangle \text{ und } \tilde{X} = \log_{10}\langle X \rangle.$$

$$\log_{10}\langle Y_i \rangle = \alpha + \beta \log_{10}\langle x_i \rangle + E_i$$

$$Y_i = 10^\alpha x_i^\beta 10^{E_i} \text{ Potenzgesetz, Fehler multiplikativ.}$$

Weitere Terme: multiplikative Wirkung!

8 Schwierigkeit:  $\log\langle 0 \rangle = -\infty$ . Abhilfe:  $\tilde{Y} = \log\langle Y + c \rangle$ . Wahl von  $c$ ?  
Bitte nicht  $c = 1$ ! Vorschlag:  $c = \text{med}\langle Y^k \rangle / s^{2.9}$  mit  $s = \text{med}\langle Y^k \rangle / q_{0.25}\langle Y^k \rangle$

h **Box-Cox-Transformation.**

$$g_\lambda(x) = \begin{cases} \frac{x^\lambda - 1}{\lambda} & \text{für } \lambda \neq 0, \\ \ln(x) & \text{für } \lambda = 0. \end{cases}$$

! Transformation ändert die Regressionsfunktion! Erlaubt?  
Hängt von der Anwendung ab!

k Kann (monotone) Transformation der Zielgröße helfen?  
Referenzlinie im TA-Diagramm betrachten!

## 4.5 Ausreisser und langschwänzige Verteilung

a **Ausreisser**. Beobachtungen, die schlecht zum Modell passen.

b Grober Fehler? → korrigieren.

Wenn nichts Spezielles war, darf man Ausreisser weglassen?

Ja: Voraussetzung nicht erfüllt →

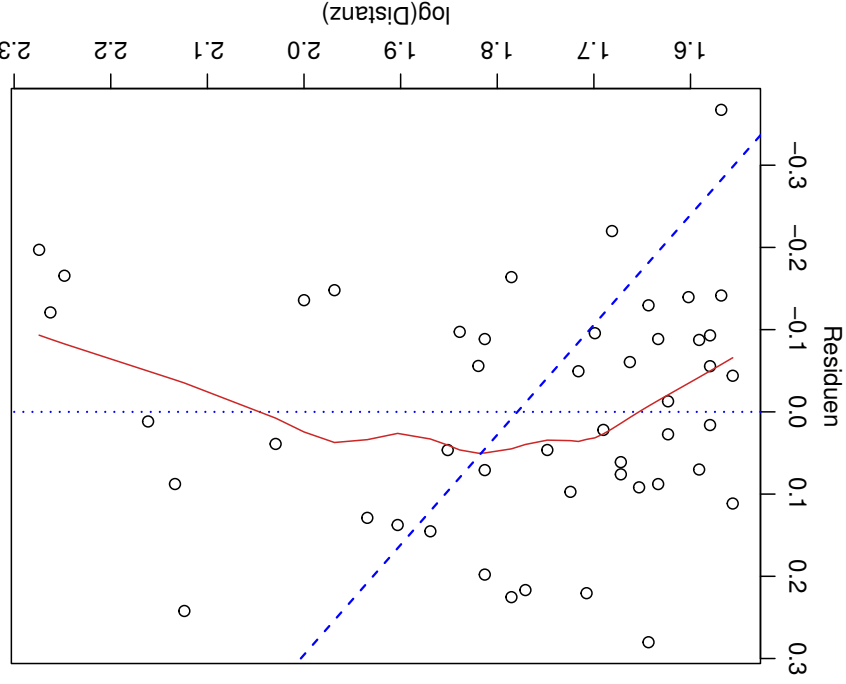
c **Langschwänzige Verteilung**. Kleinste Quadrate nicht optimal.

Max.lik. für langschw. Vert. → weniger Gewicht für extreme Beob.

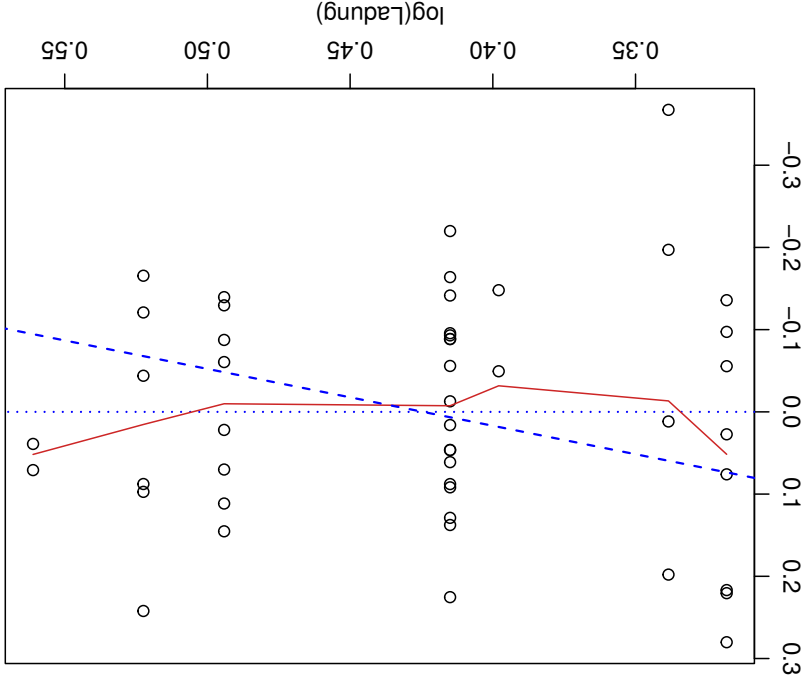
→ Block robuste Regression.

## 4.6 Residuen und erklärende Variable

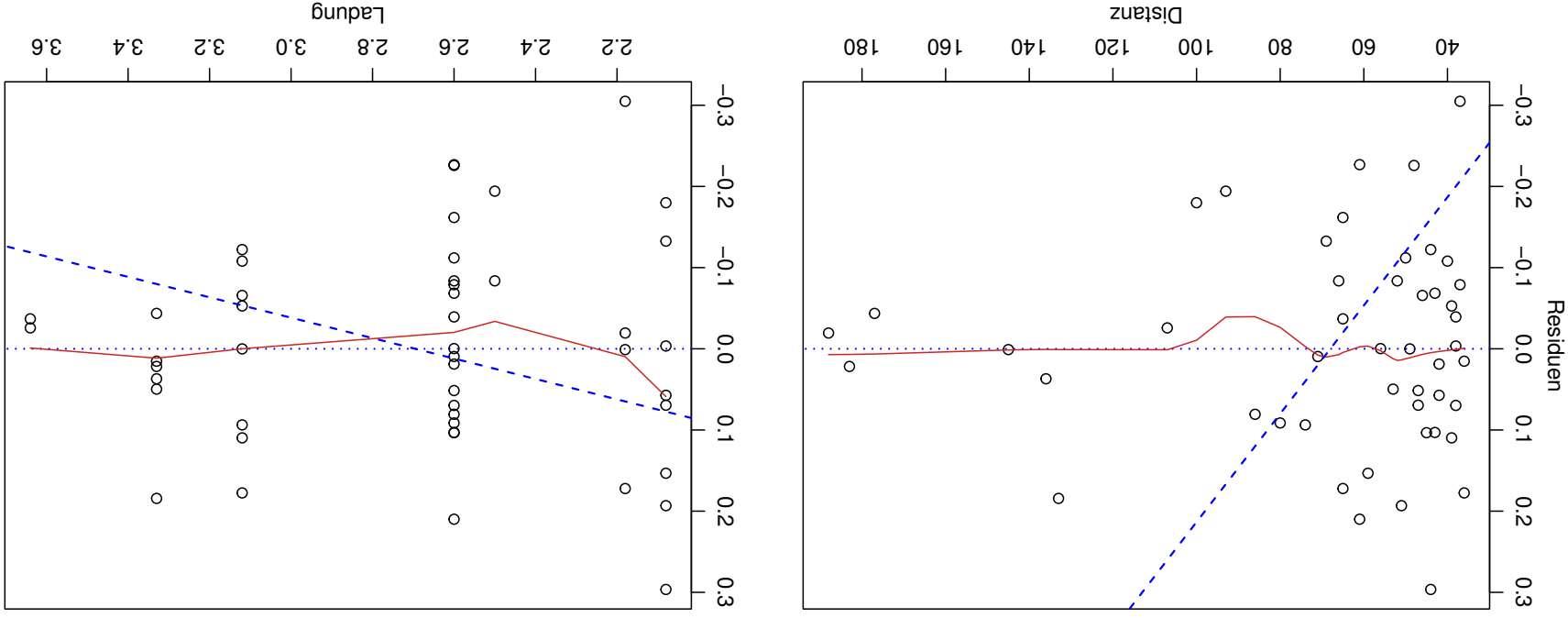
a Residuen gegen  $X^{(j)}$  auftragen!



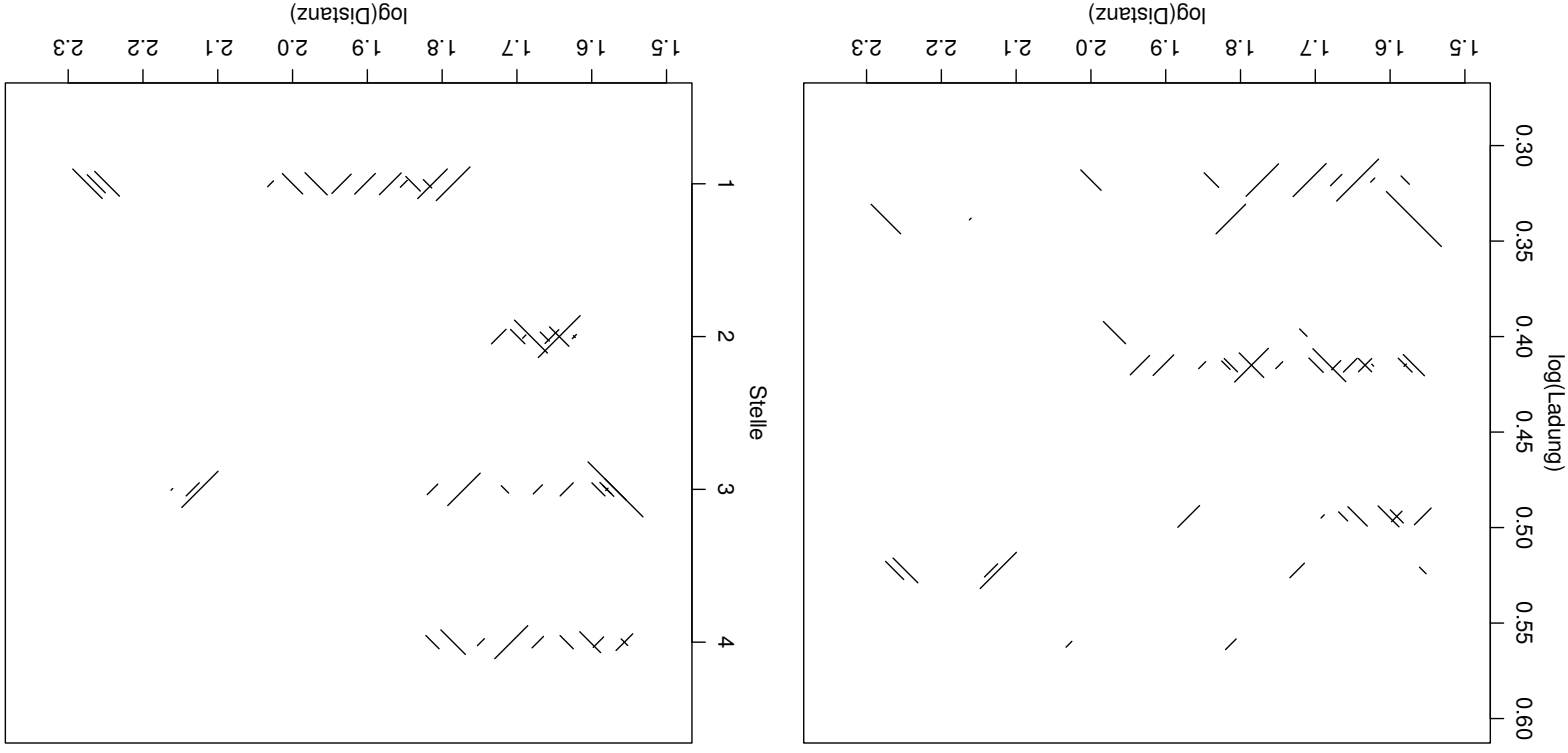
b Transformation von  $X^{(j)}$ ? Beachte Referenzlinie!



c Wenn Transformation nicht hilft: **quadratischer Term**,  
oder glatte Funktion statt linearer → Glättung, Nichtparametrische Regression



d **Sind Effekte von 2 Variablen additiv?** → Residuen gegen 2 erkl. Var. auftragen!

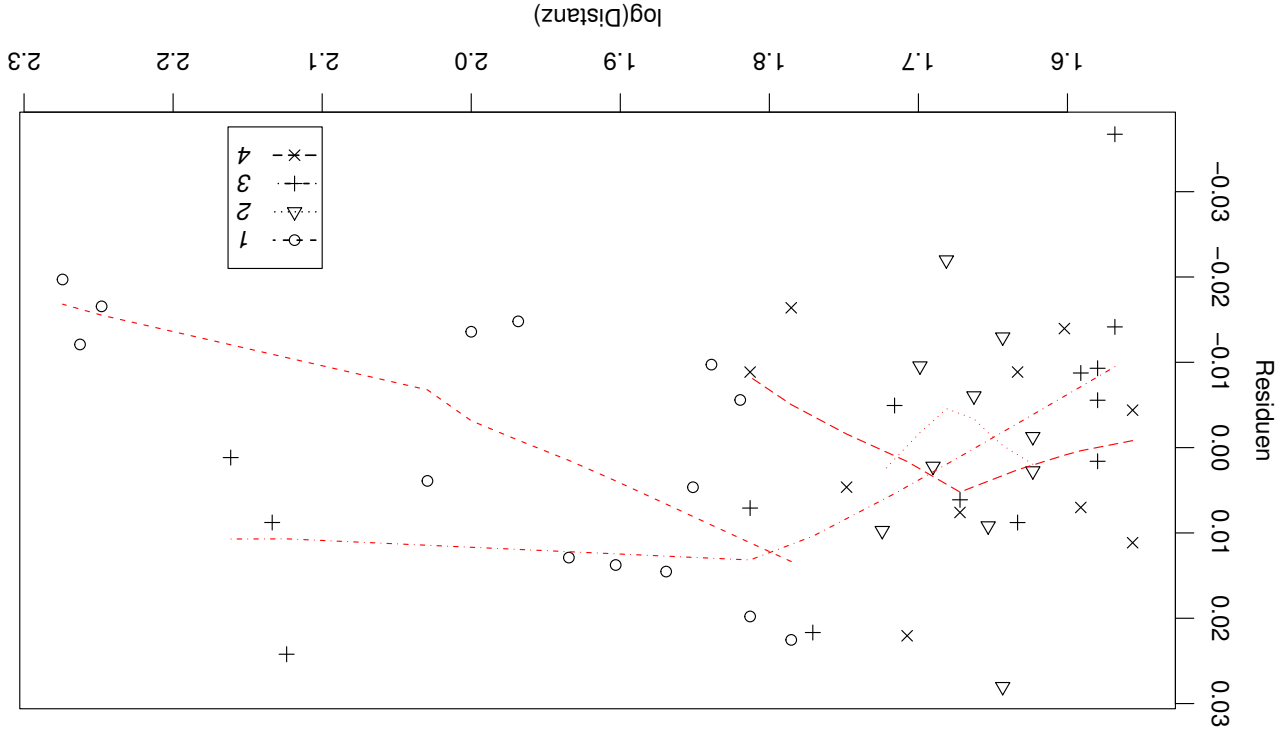


← **Wechselwirkung**  $x_{(j)}^{(k)} = x_{(j)}^{(k)} x_{(l)}^{(k)}$  für kontinuierl.  $X^{(j)}$  und  $X^{(k)}$  (oder besser  $= (x_{(j)}^{(k)} - \bar{x}_{(j)}^{(k)})(x_{(l)}^{(k)} - \bar{x}_{(l)}^{(k)})$ )

Wechselwirkung zw. kont. Var.  $X^{(j)}$  und Faktor  $X^{(k)}$ ?

= verschiedene Steigungen  $\beta_j$  für versch. Niveaus des Faktors.

e Bessere Figur für 2. Fall (Faktor):



f Varianz  $\text{var}\langle H_i \rangle$  abhängig von  $X^{(i)}$ ?  $\rightarrow$  gewichtete Regr.

## 4.7 Gewichtete lineare Regression

a **Varianzen** verschieden,  $\text{var}\langle H_i \rangle =: \sigma_i^2$ .

$\sigma_i$  bekannt. Dann: Kleines  $\sigma_i \rightarrow$  grosses Gewicht.

Formal: Max.lik.  $\rightarrow$  Gewichtete Kl. Quadrate = minimiere  $\sum_i w_i R_i^2$ ,  $w_i = 1/\sigma_i^2$

b  $\sigma_i$  unbekannt, aber  $\text{var}\langle H_i \rangle = \sigma^2 v_i \rightarrow$  Gewichte  $w_i = 1/v_i$ .

$\sigma_i$  Funktion von  $x_i^{(j)}$ ,  $\approx \sigma^2 v \langle x_i^{(j)} \rangle \rightarrow w_i = 1/v \langle x_i^{(j)} \rangle$ .

Achtung:  $\sigma_i$  Funktion von  $Y_i$ : Man darf nicht  $w_i = 1/v \langle Y_i \rangle$  nehmen  
(evtl.  $w_i = 1/v \langle \hat{y}_i \rangle$ . Nicht iterieren!)

4.7

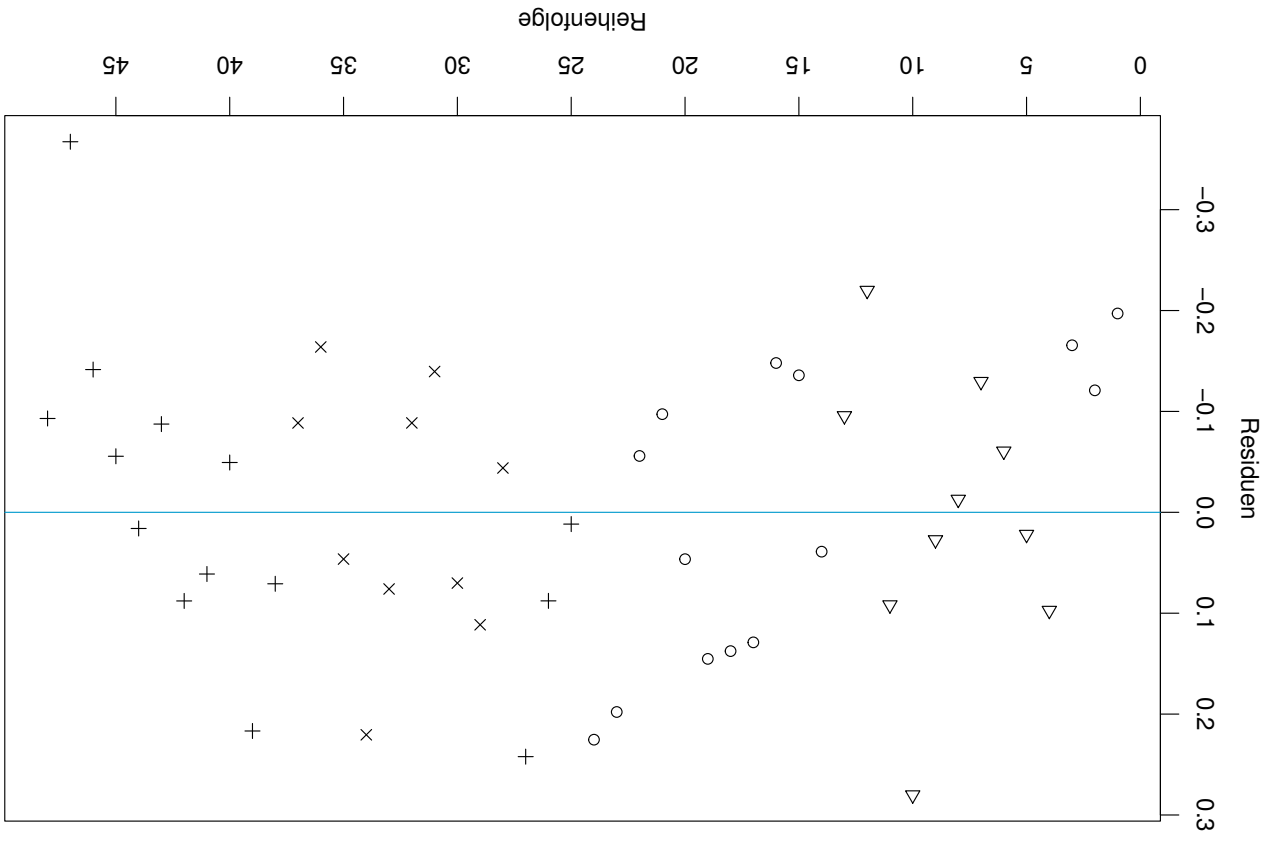
c 
$$\bar{\hat{\beta}} = (\widetilde{\mathbf{X}}^T \mathbf{W} \widetilde{\mathbf{X}})^{-1} \widetilde{\mathbf{X}}^T \mathbf{W} \bar{Y}, \quad \mathcal{E}\langle \hat{\beta}_j \rangle = \beta_j, \quad \text{var}\langle \hat{\beta}_j \rangle = \sigma^2 (\widetilde{\mathbf{X}}^T \mathbf{W} \widetilde{\mathbf{X}})^{-1}_{jj}.$$

d Überprüfung der Wahl der Gewichte:  $\sqrt{|R_i|}$  gegen Gewichte.

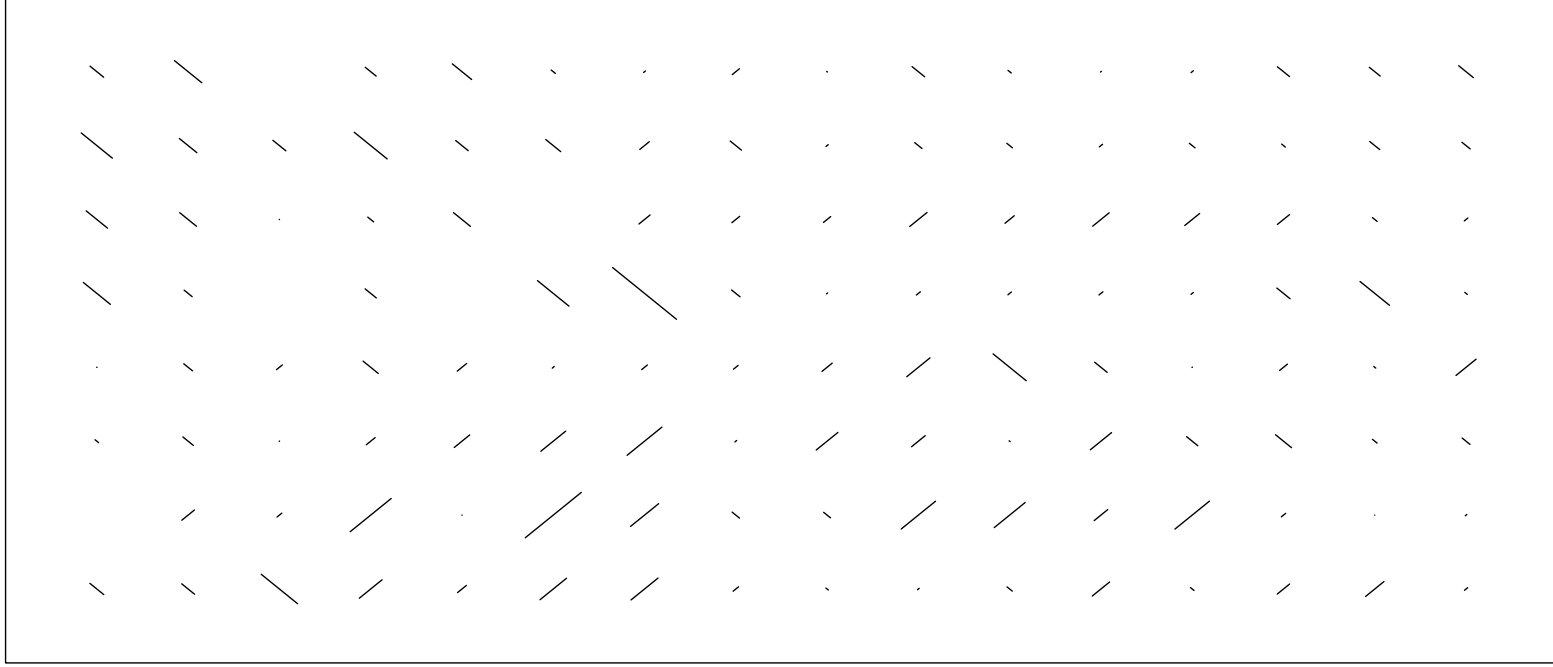
4.8 Gesamthafte Überprüfung Geschätzte Streuung der Fehler vergleichen mit Schätzung aus „benachbarten“ Beobachtungen (bezüglich erklärenden Var.)

## 4.8 Unabhängigkeit

a  $R_i$  auftragen gegen • Zeit, • Ort, • Gruppierungs-Variable.



d Räumliche Abhängigkeit, Bsp. basische Böden



Durbin-Watson-Test.

$$T = \frac{\sum_{i=2}^n (R_i - R_{i-1})^2}{\sum_{i=1}^n R_i^2}$$

Entscheidung: Unabhängigkeit

- Verwertung, falls  $T < c'$ ,
- Beibehaltung, falls  $T > c''$ ,
- gar nichts (unentscheidbar), falls  $T$  dazwischen liegt.

f Wenn Korrelationen vorliegen, dann sind

die P-Werte der üblichen Tests häufig grob falsch.

→ Verallgemeinerte Kleinste Quadrate, Regression von Zeitreihen

## 4.9 Einflussreiche Beobachtungen

a **Ausreisser**: Haben sie wesentlichen Einfluss auf die Analyse? „Sensitivitäts-Analyse“

b Analyse ohne *z*te Beobachtung

„(influence) **diagnostics**“:

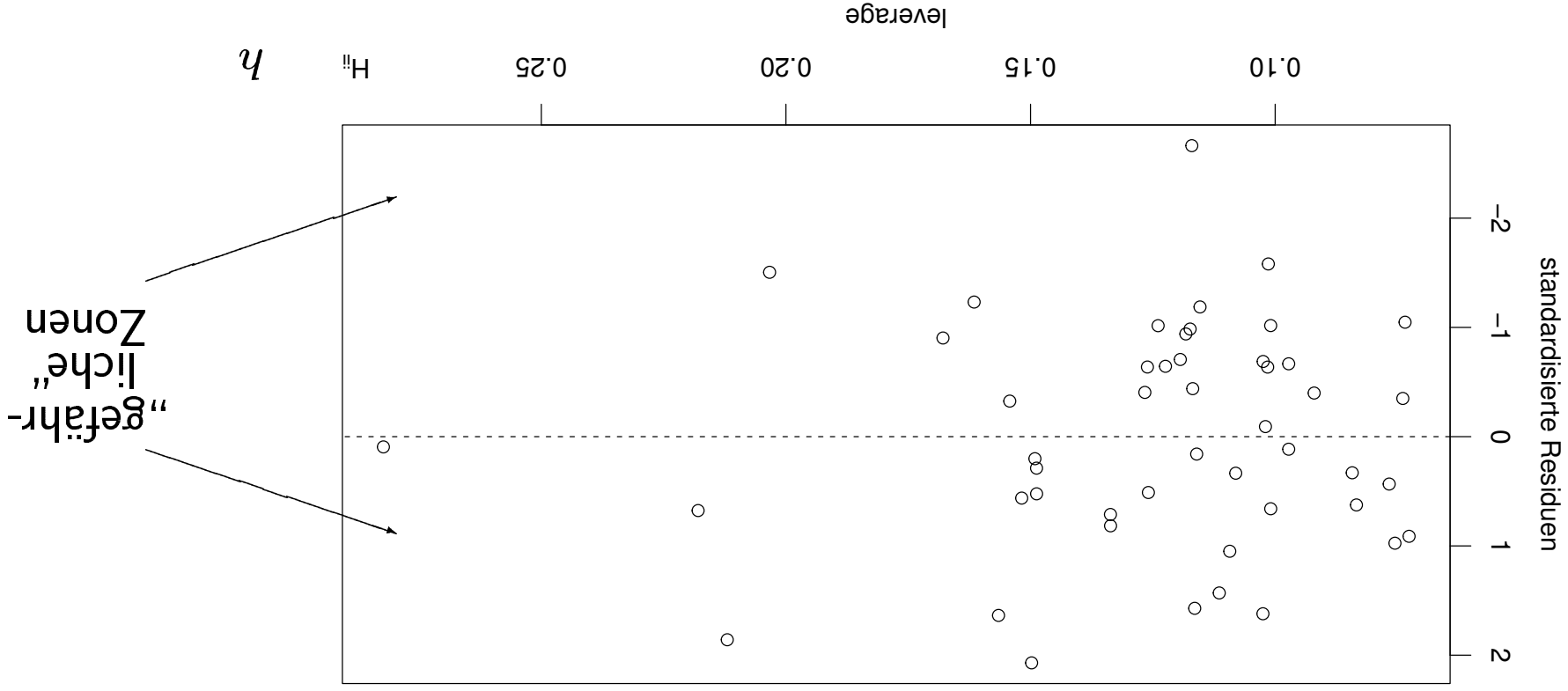
Veränderung von Schätzwerten, Test-Statistiken, ...

2 weglassen : nicht unbedingt additive Effekte.

„masking“, „swamping“:

c  $R_i$  gegen  $H_{ii}$  auftragen.

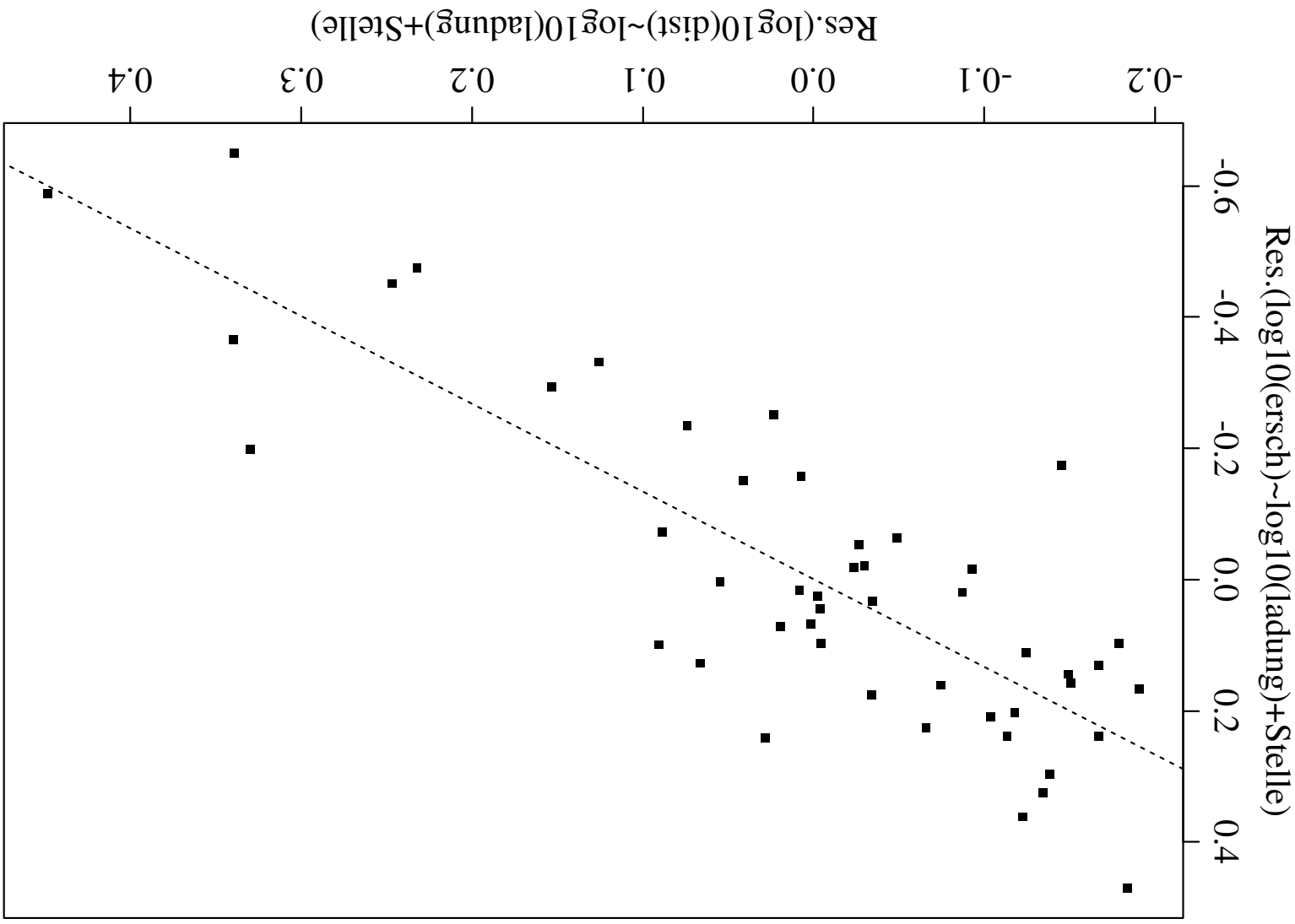
Influence diagnostics nehmen zu mit  $|R_i|$  und  $H_{ii}$ .



e Distanz von Cook

$$d_i = \frac{R_i^2 H_{ii}}{\widehat{\sigma}^2 (1 - H_{ii})^2} = (1/p) \frac{R_i^2 H_{ii}}{(1 - H_{ii})}$$

f  $R_i^{(Y|-j)}$  gegen  $R_i^{(X^{(j)}|-j)}$  anfragen  
 added variable plot oder partial regression leverage plot



1. Im Tukey-Anscombe-Diagramm sieht man Abweichungen von
  - der angenommenen Regressionsfunktion,
  - der Gleichheit der Varianzen (Scale Plot)
  - der Form der Verteilung der Fehler (genauer: QQ-Plot)

**Transformation** der Zielgröße hilft oft.

2. Residuen gegen erkl. Variable  $\rightarrow$  Transformation der erkl. Wechselwirkungen

3. **Einflussreiche** Beobachtungen

4. **Residuenanalyse** dient der **Verbesserung** eines Regressionsmodells.  
**Regression ohne Residuenanalyse ist unzulässig!**