

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.

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

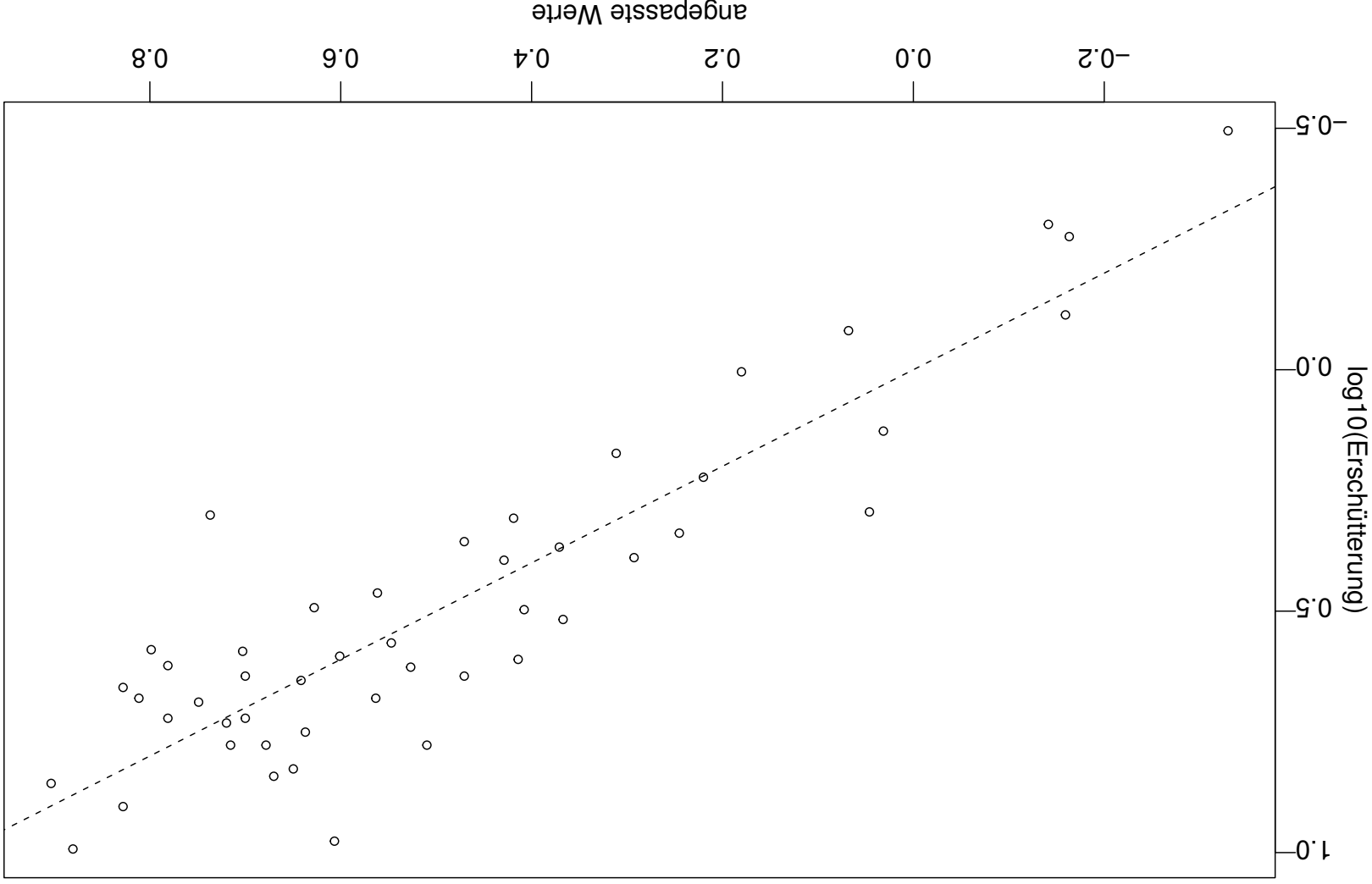
Abweichungen = Symptome

versch. Abweichungen \longrightarrow **Syndrom** \longrightarrow **Diagnose** \longrightarrow

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?

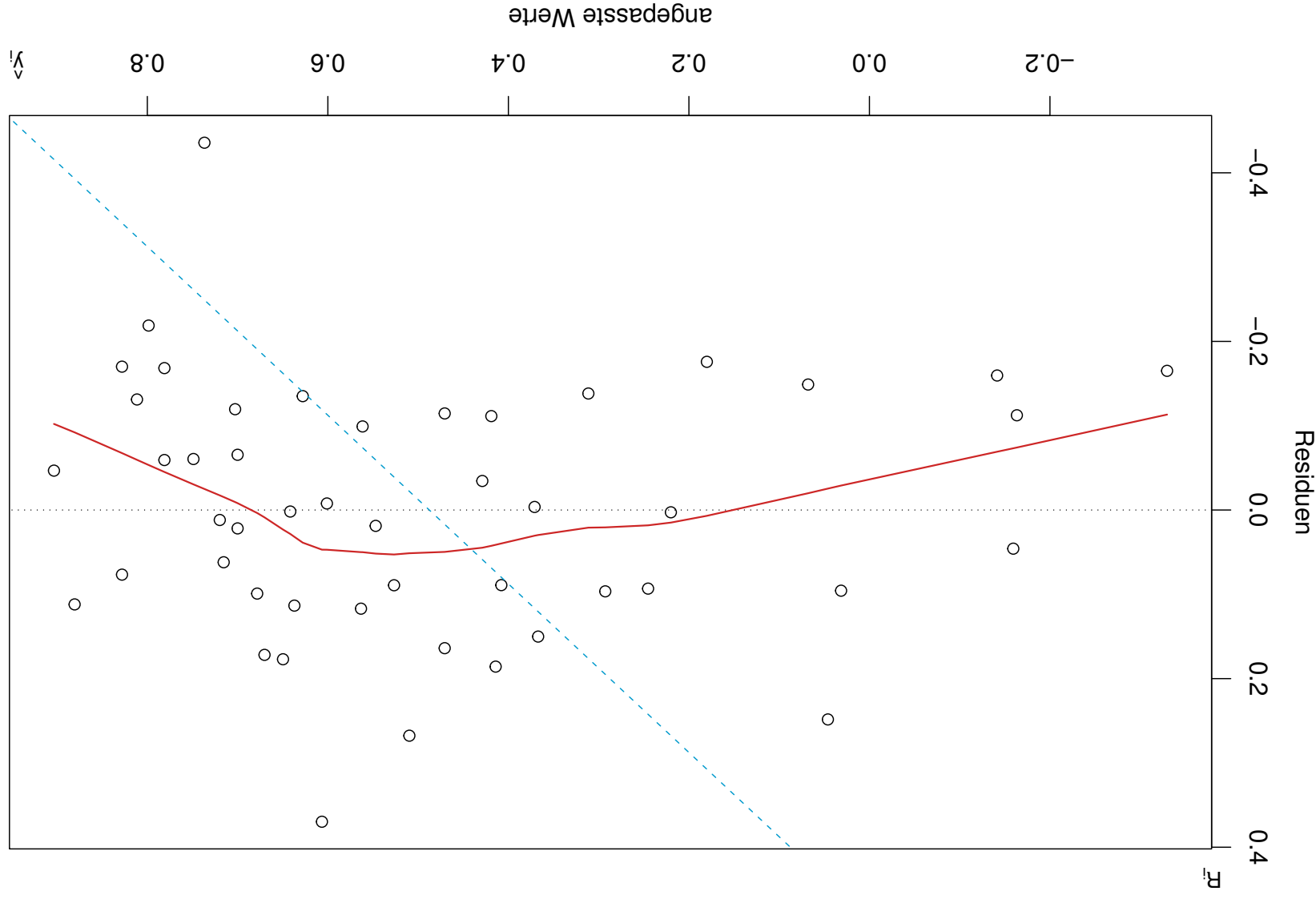
Ausreisser?

e **Wie beurteilen?**

• Abweichung im Bereich des Zufalls?

• Abweichung gefährlich? Antwort abhängig vom Zweck der Studie!

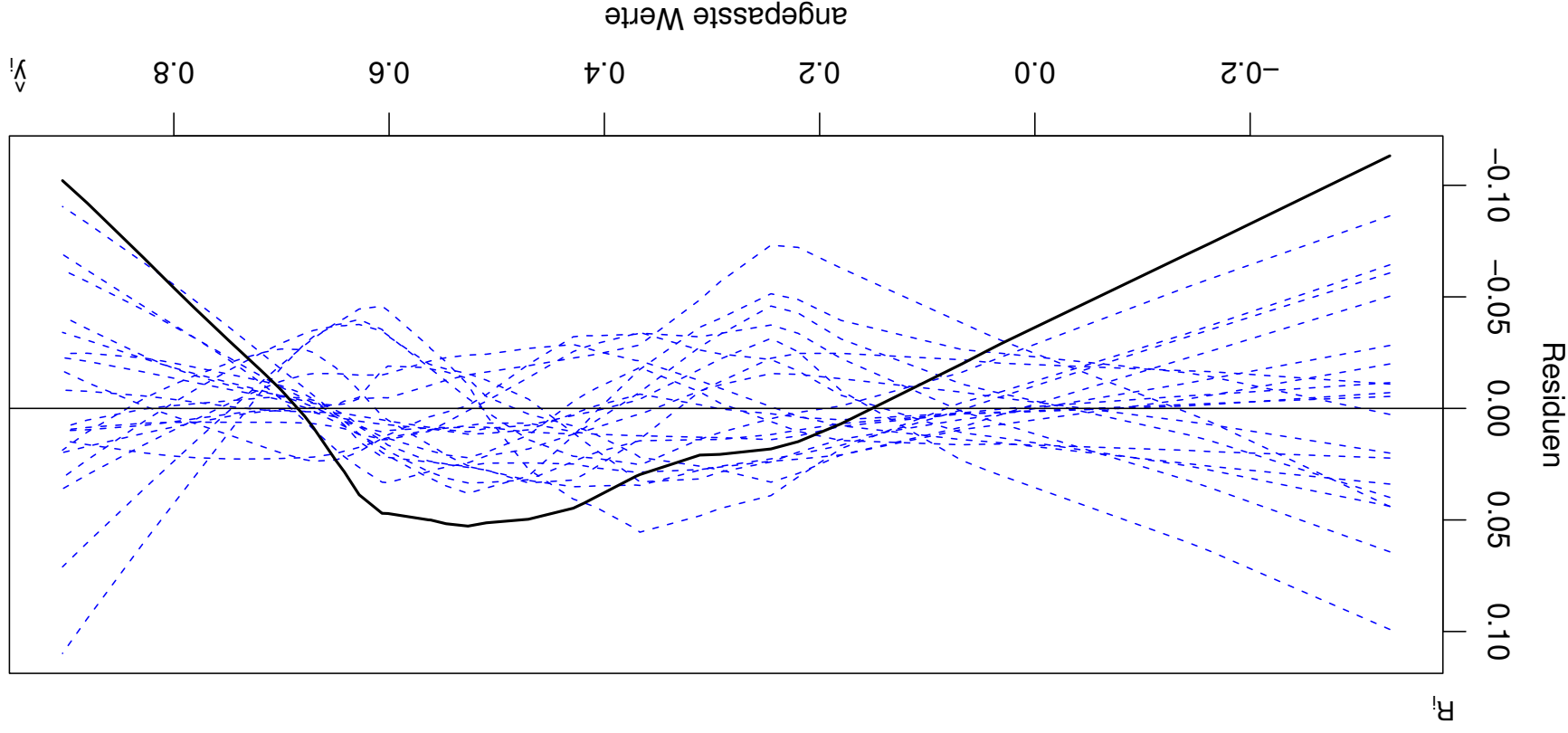
f Variante des Diagramms Y gegen \hat{y} zeigt Abweichungen genauer:
Tukey-Anscombe-Diagramm: Residuen gegen angepasste Werte



9 (a) **Regressionfunktion:** $\mathcal{E}\langle E_i \rangle = 0$. Mittelwert der R_i über „Fenster“

der \hat{y}_i

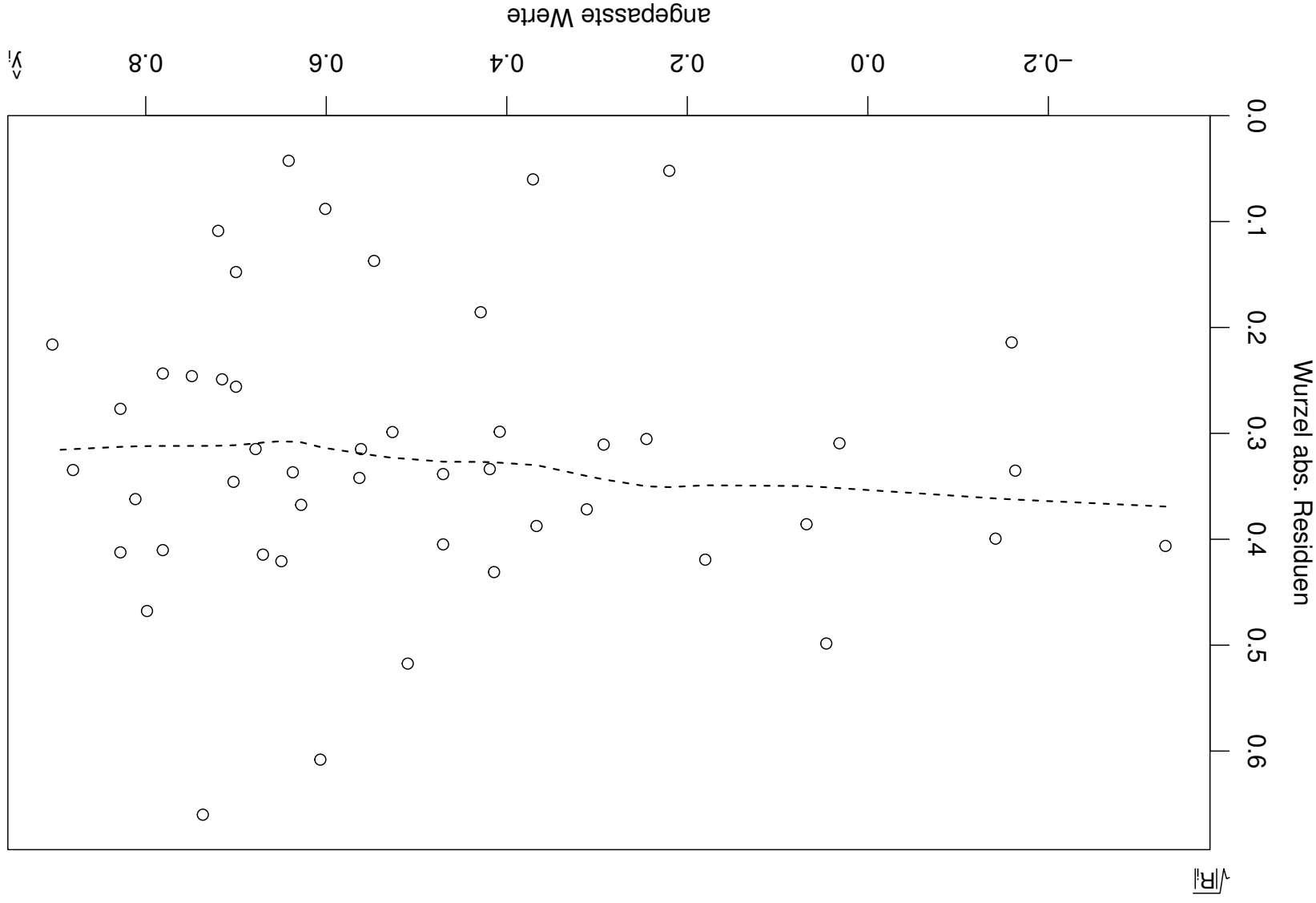
→ gleichendes Mittel → Glättung „lowess“, **robust**



h Abweichung zufällig? → Simulation

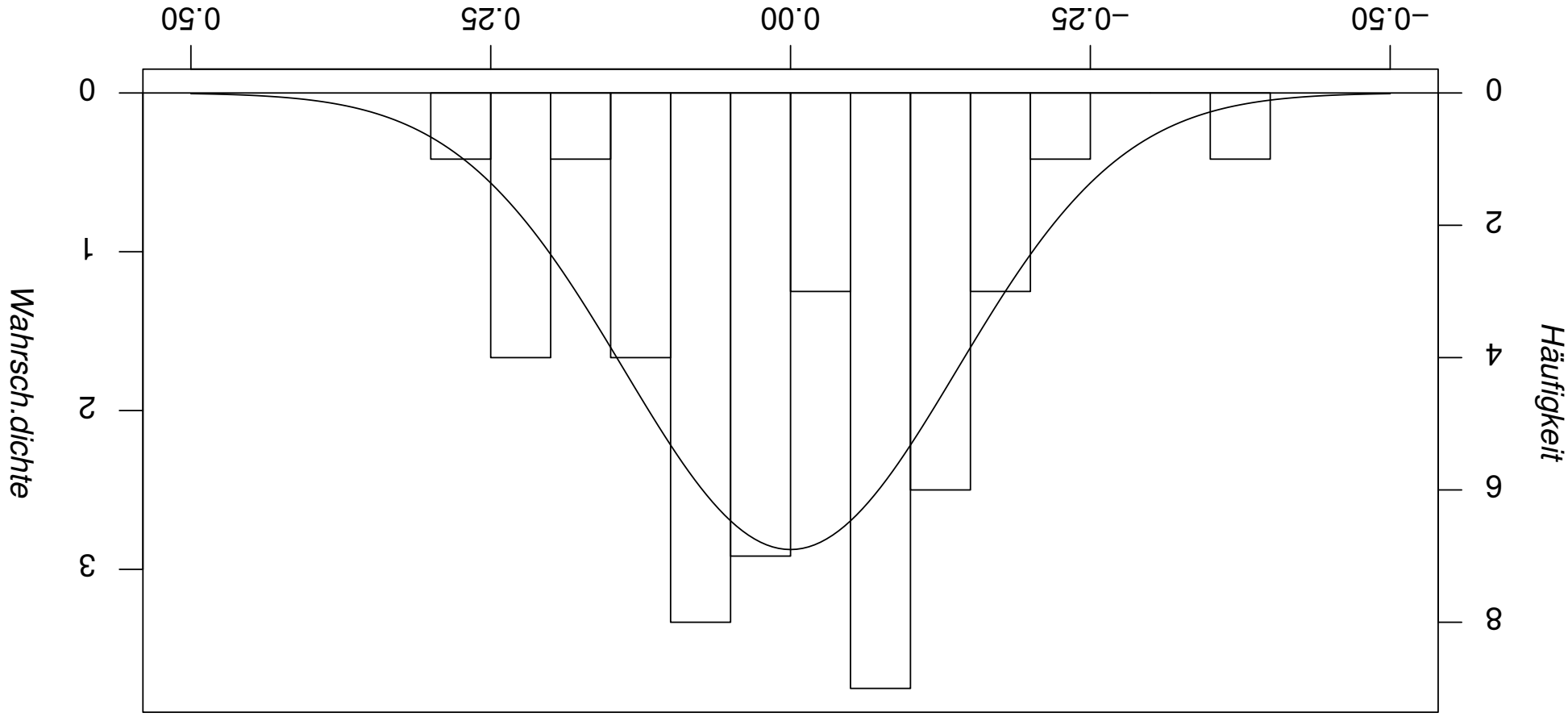
→ „grafischer Test“
19 zusätzliche Kurven. Ist die beobachtete „die extremste“?

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

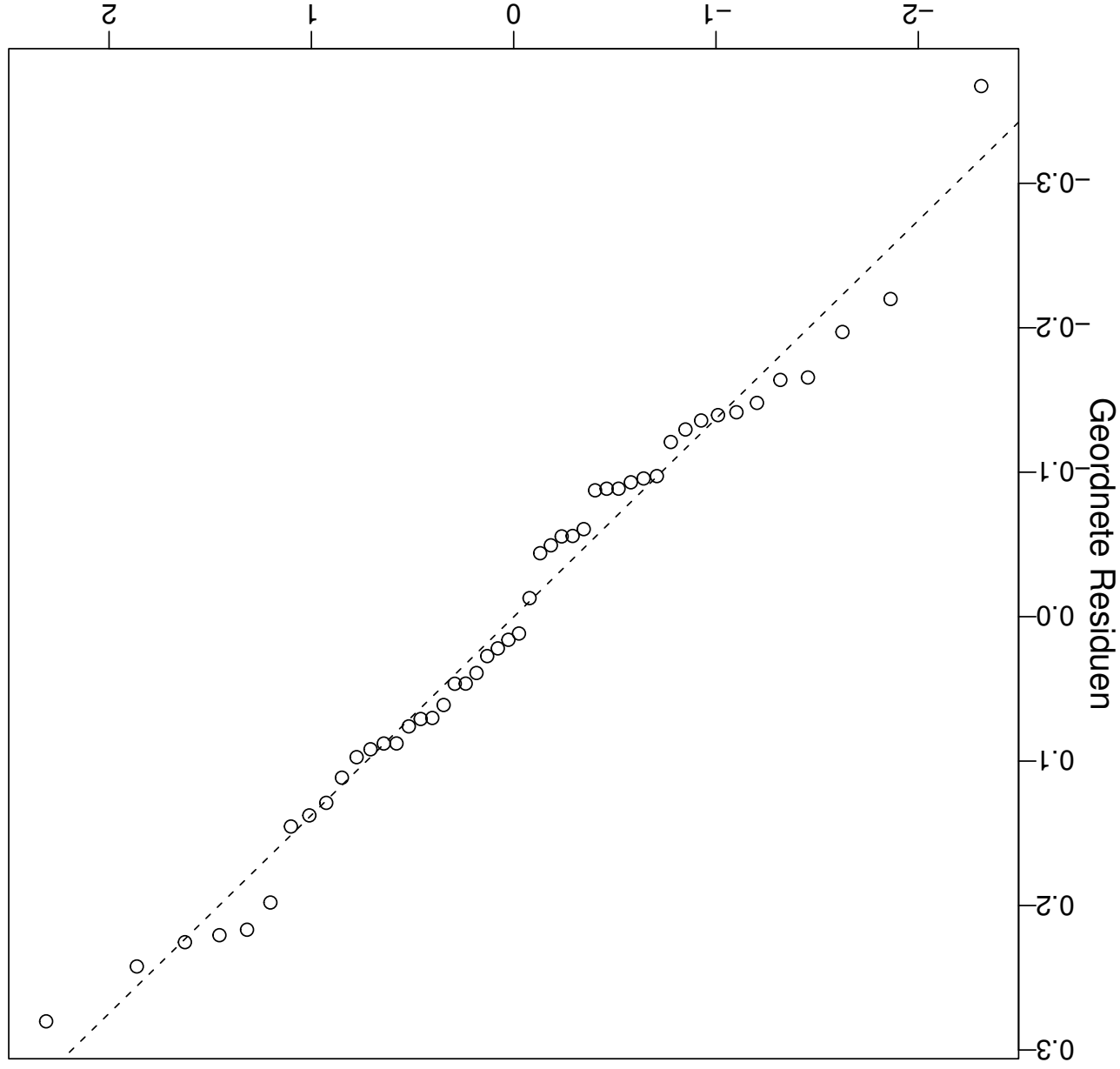


4.3 Verteilung der Fehler

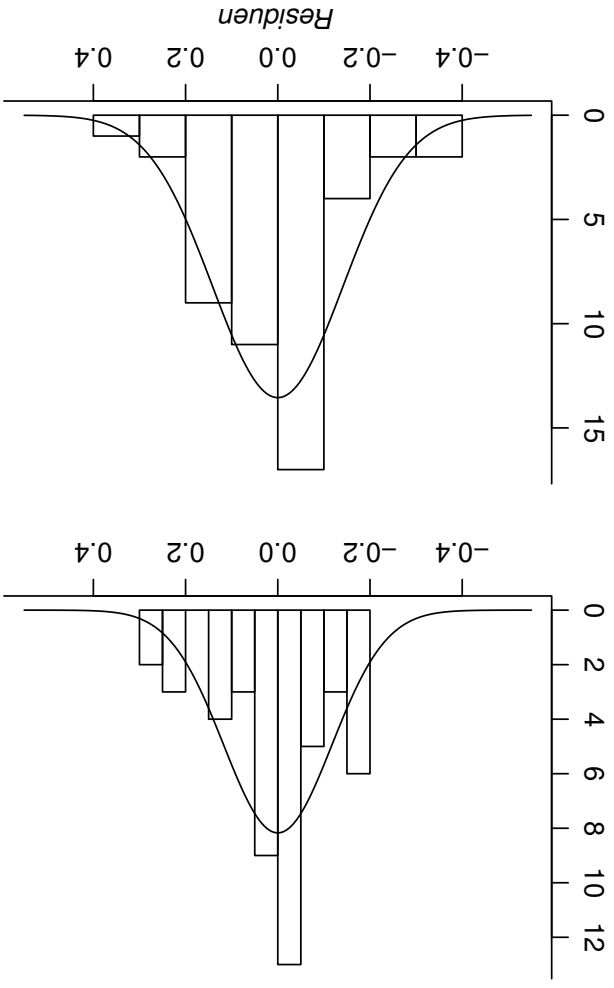
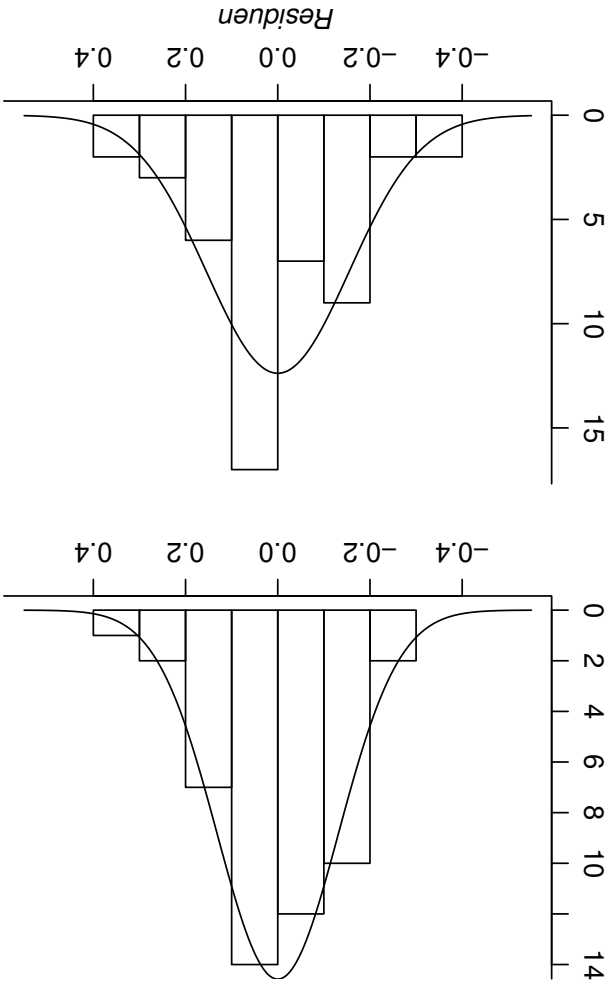
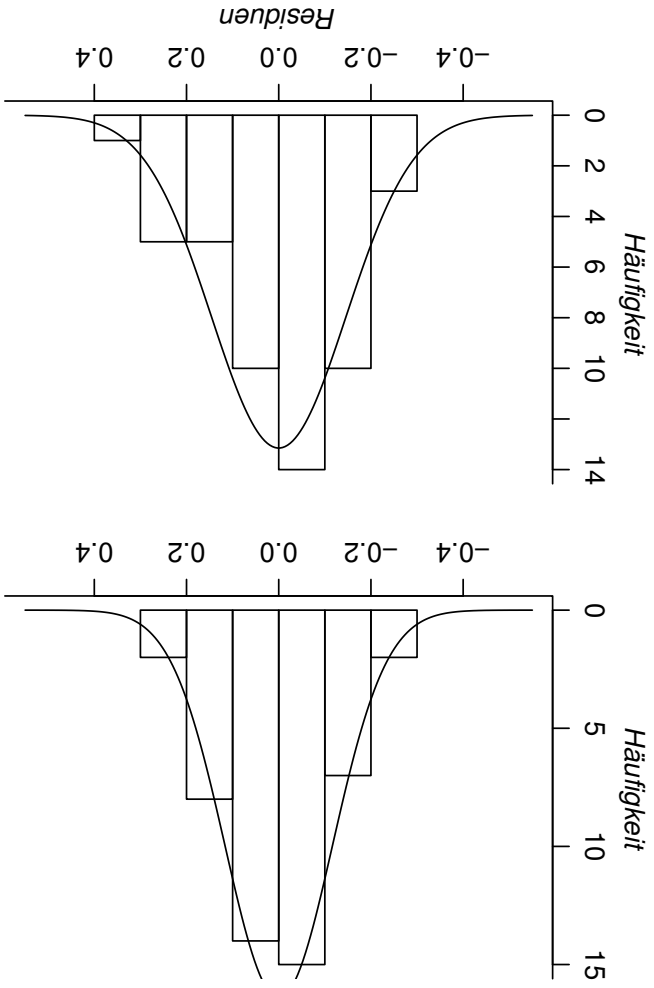
- a (c) Normalverteilung? Histogramm der E_i resp. Residuen R_i ! Die Zielgröße Y muss nicht normalverteilt sein !

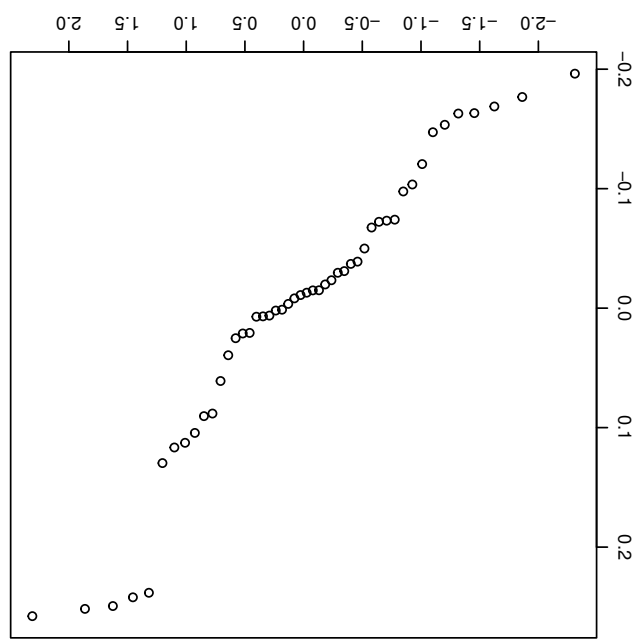
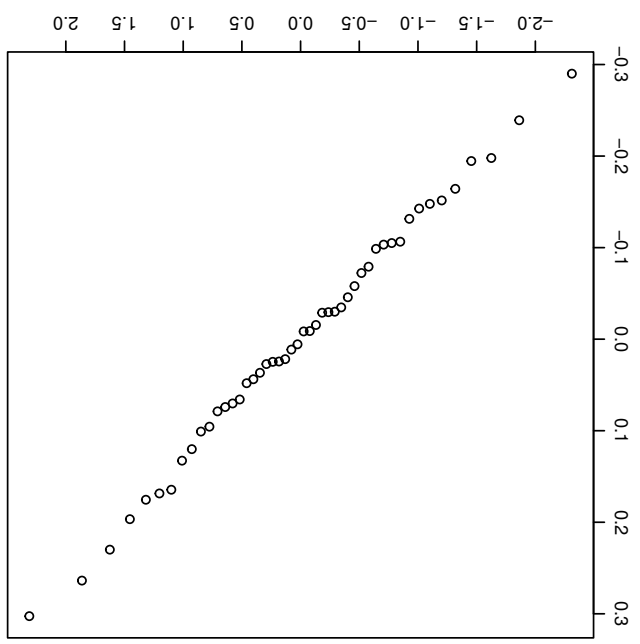
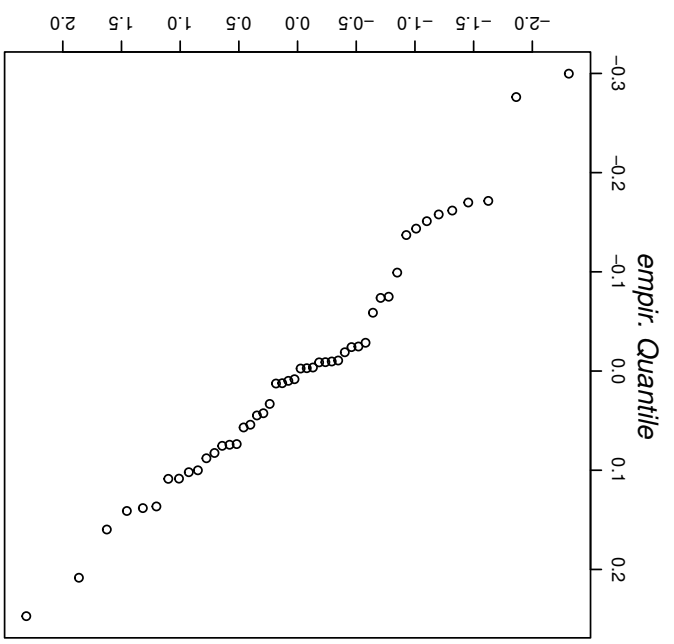
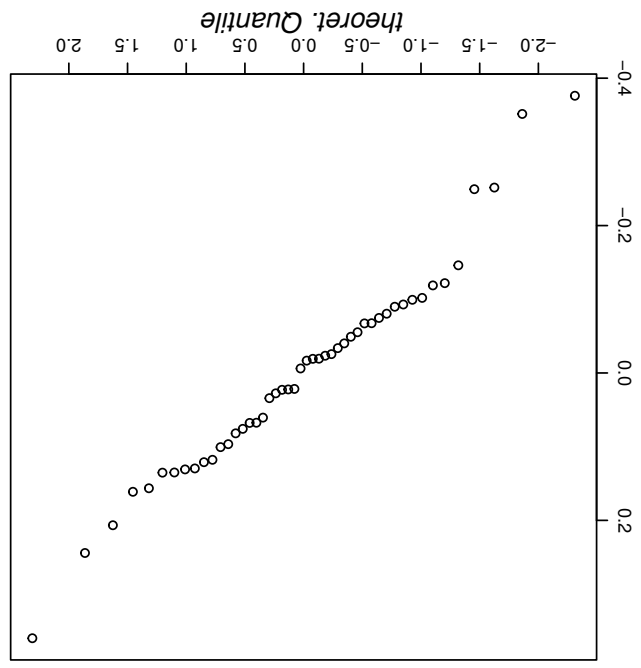
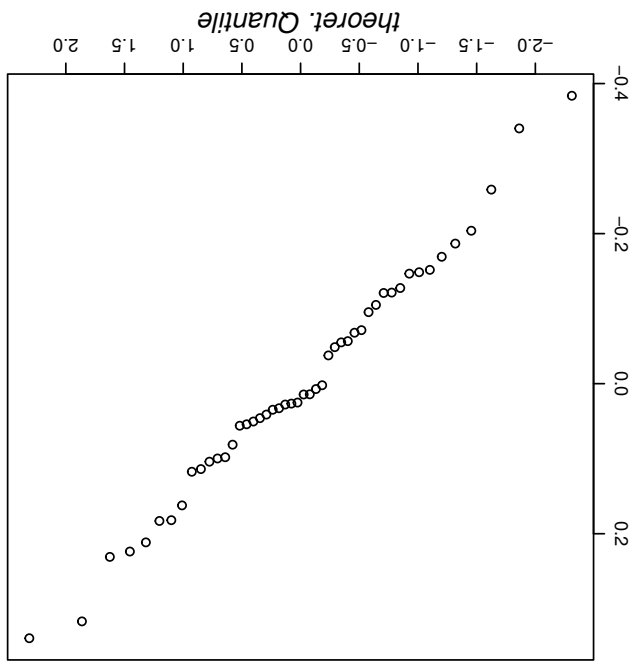
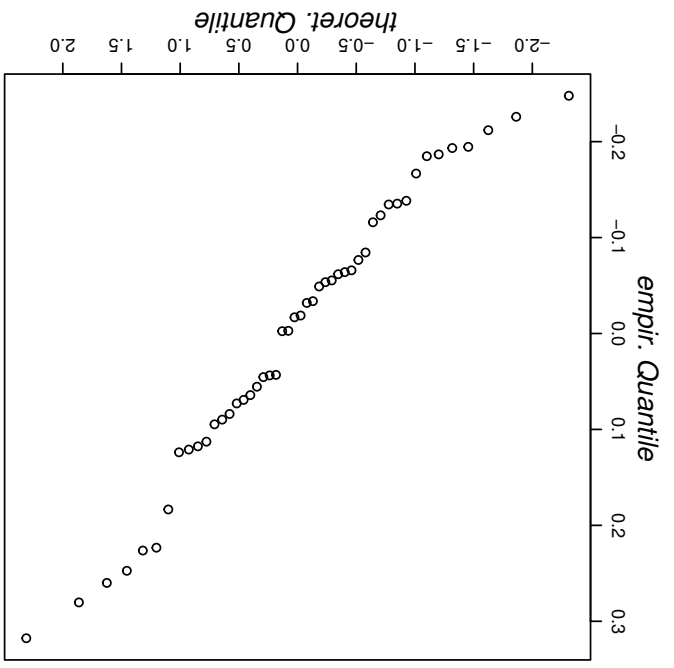


b Raffinierter: Quantil-Quantil-Diagramm (QQ-Plot, normal plot)



- 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

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

$\bullet \quad H_i$ misst den „Abstand“ zwischen \bar{x}_i und \bar{x} .

$$\text{einfache R.: } H_i = (1/n) + (x_i - \bar{x})^2 / \text{SSQ}(X).$$

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

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

g Residuen **standardisieren**, damit sie alle die gleiche Verteilung haben:

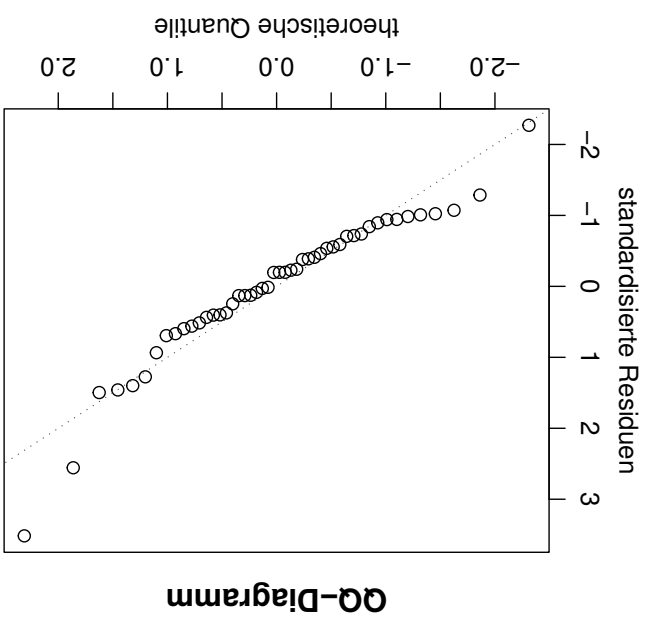
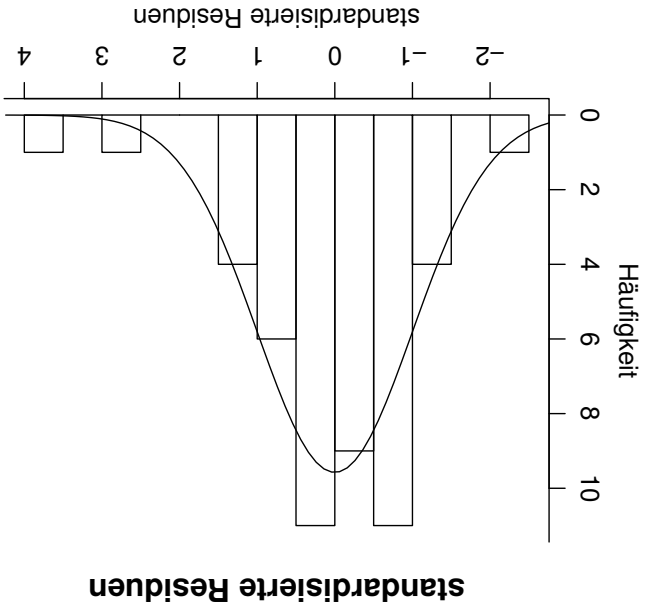
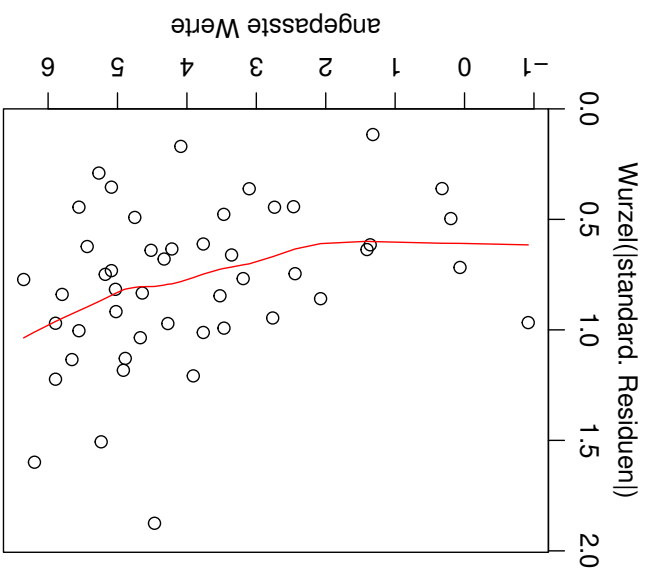
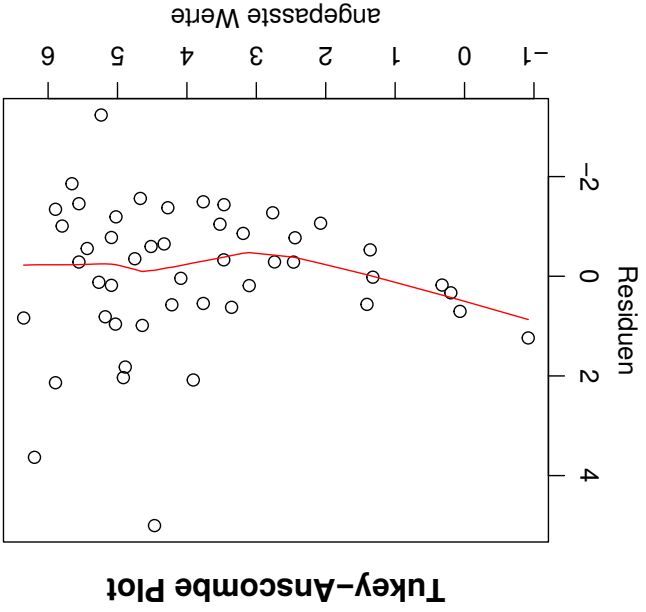
$$\tilde{R}_i = R_i / \left(\hat{\sigma} \sqrt{1 - H_i} \right)$$

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.

4.4 Zielgrösse transformieren?

- a Symptome → Syndrom → Diagnose → Therapie.
Umgekehrt: Krankheit → Syndrom
- Falsche / fehlende Transformation der Zielgrösse → ???
- Bsp. Sprengungen: Fehlende log-Transformation → ???



b Syndrom:

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

c = „Transformations-Syndrom“

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!

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

j 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 →

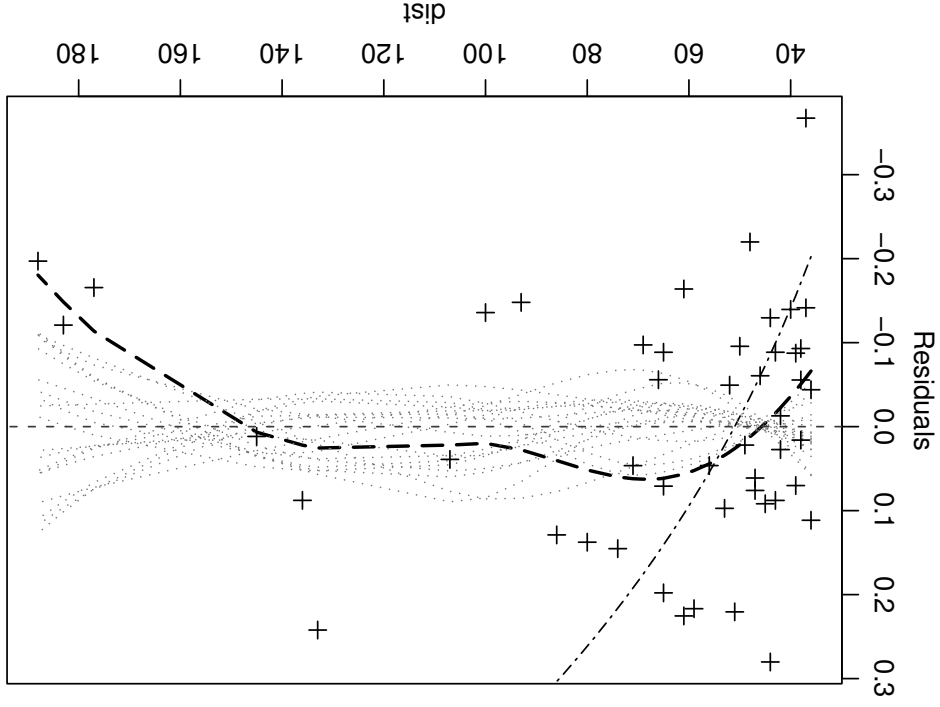
c **Langschwänzige Verteilung**. Kleinste Quadrate nicht optimal.

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

→ Block robuste Regression.

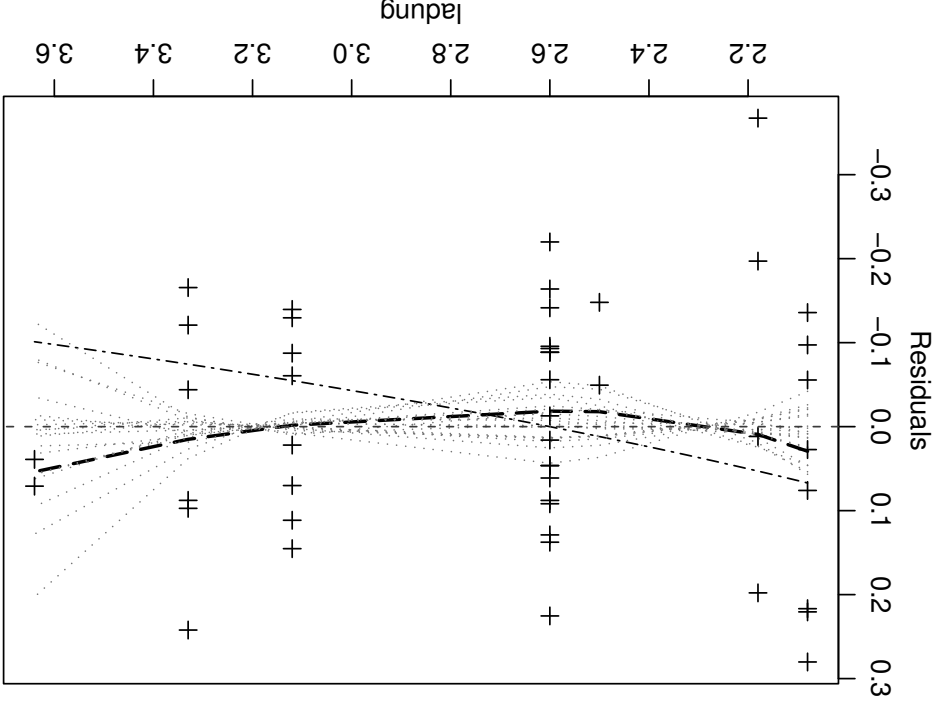
4.6 Residuen und erklärende Variable

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

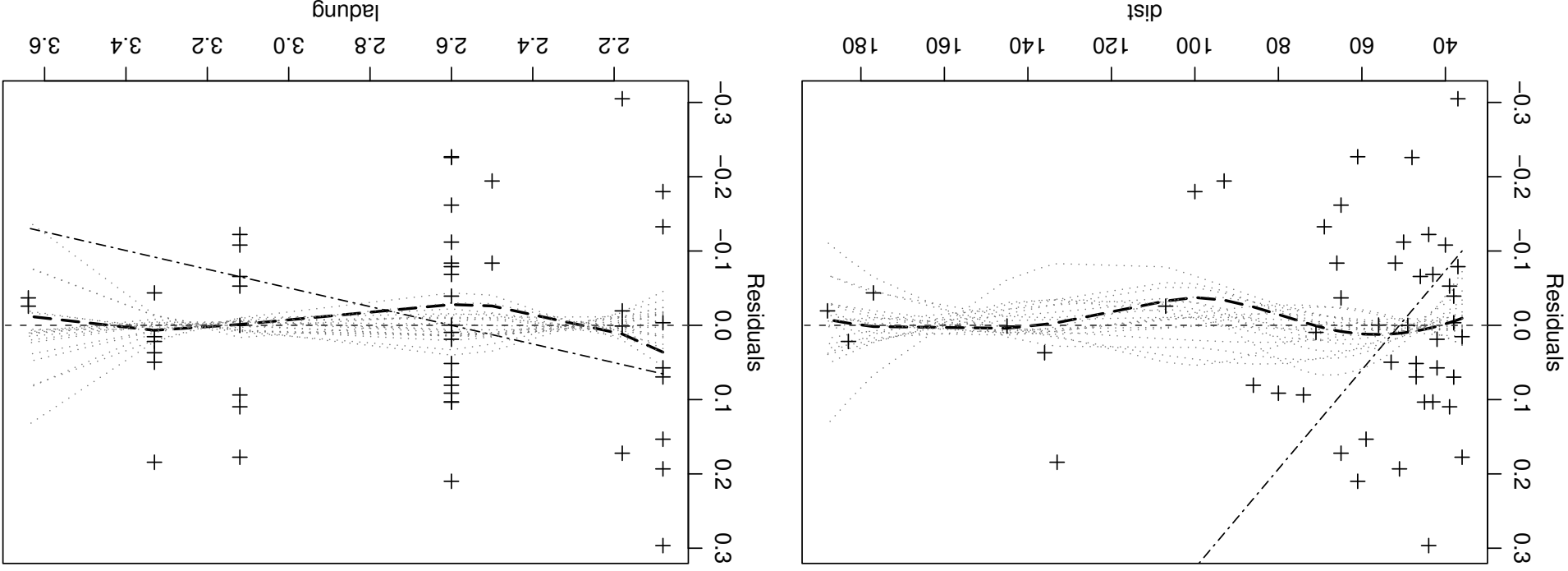


$\log_{10}(\text{ersch}) \sim \text{Stelle} + \log_{10}(\text{dist}) + \log_{10}(\text{ladung})$

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



$\log_{10}(\text{ersch}) \sim \text{Stelle} + \text{dist} + \text{ladung}$



c Wenn Transformation nicht hilft: **quadratischer Term**,

oder glatte Funkt. statt linearer \rightarrow Glättung, Nichtparametrische Regr.

4.7 Gewichete lineare Regression

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

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

Formal: Maximum likelihood

\rightarrow Gewichete Kl. Quadrata = minimiere $\sum_i w_i R_i^2$, $w_i = 1/\sigma_i^2$

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

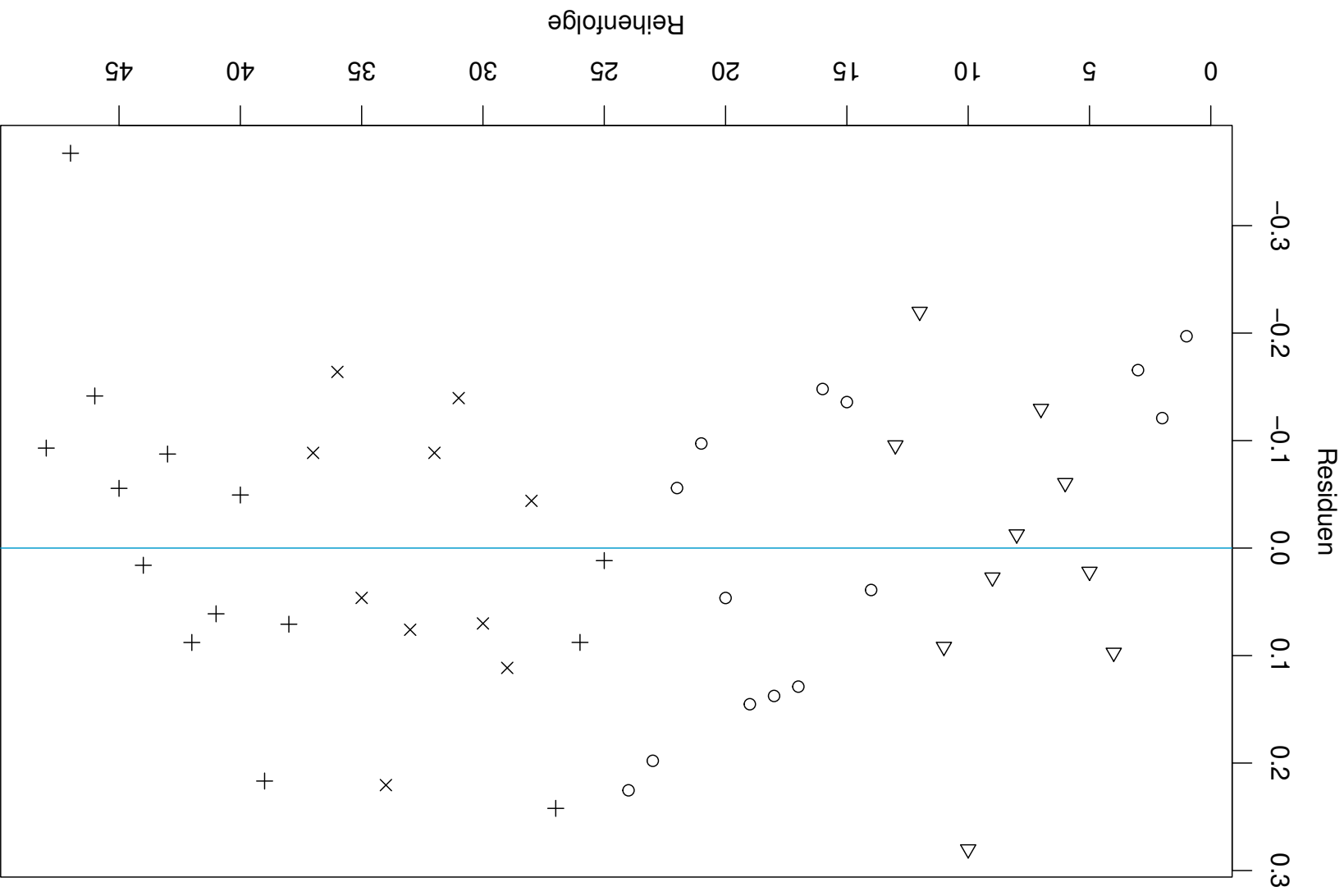
σ_i Funktion von $x_i^{(j)}$, $\approx \sigma^2 v\langle x_i^{(j)} \rangle \rightarrow w_i = 1/v\langle x_i^{(j)} \rangle$.

Achtung: σ_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!)

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

4.8 Unabhängigkeit

a R_i auftragen gegen ● Zeit, ● Ort, ● Gruppierungs-Variablen.



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"

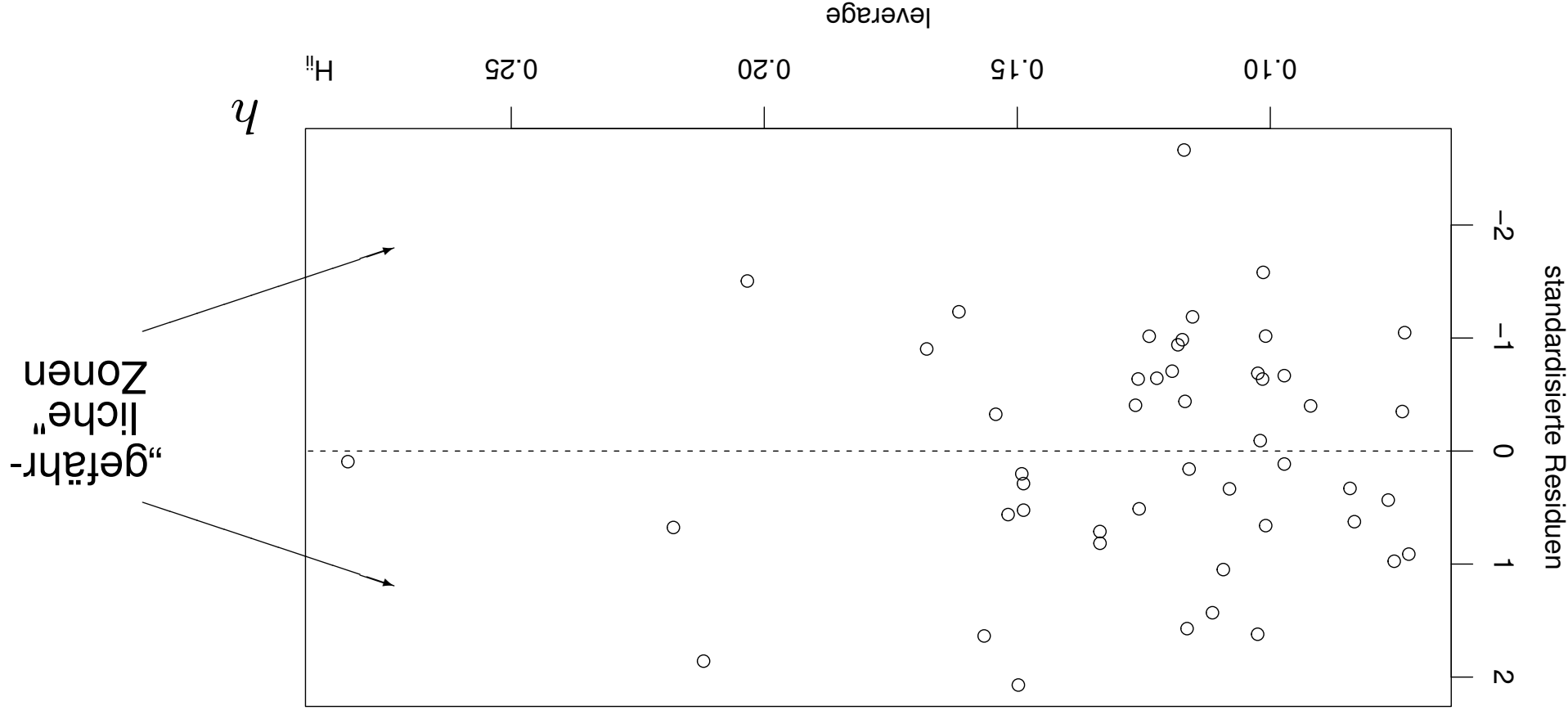
b Analyse ohne t te Beobachtung
"(influence) **diagnostics**":

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

$$d_i = \frac{R_i^2 H_{ii}}{p \hat{\sigma}^2 (1 - H_{ii})^2} = (1/p) \tilde{R}_i^2 H_{ii} / (1 - H_{ii})$$

Merkpunkte

Residuen-Analyse

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!

5 Modell-Entwicklung

5.1 Problemstellung

- a Welche Ausgangs-Variablen sollen in welcher Form in der Modell-Gleichung der linearen Regression erscheinen?
- b Beispiel Baukosten

Bez.	Bedeutung	Typ	Transf.
K	Baukosten	Betrag	log
G	Grösse	Betrag	log
D	Datum der Baubewilligung	kontin.	–
WZ	Wartezeit zwischen Antrag und Baubewilligung	Betrag	–
BZ	Bauzeit: Zeit bis Inbetriebnahme	Betrag	–
Z	Zweitwerk: früheres Werk auf gleichem Gelände	binär	–
NE	Werk steht im Nordosten der USA	binär	–
KI	Werk arbeitet mit Kühlturm	binär	–
BW	Reaktor hergestellt durch Babcock-Wilcox	binär	–
N	Anzahl Werke, die das gleiche Ingenieur-Team bereits erbaut hat, +1	Anzahl	Wurzel
KG	Partielle Kostengarantie des Generalunternehmers	binär	–

c First aid transformations:

$$d \log_{10}\langle K \rangle = \beta_0 + \beta_1 \log_{10}\langle G \rangle + \beta_2 D + \beta_3 WZ + \beta_4 BZ + \beta_5 Z + \beta_6 NE + \beta_7 K + \beta_8 BW + \beta_9 \sqrt{N} + \beta_{10} KG + \text{Fehler}$$

e **Ein einzelner Term.**

- t-Test für einzelnes β_j

- F-Test für mehrere β_j aufs Mal,

z. B. alle β_j für eine nominale Variable.

F-Test für den Vergleich von Modellen:

Coefficients:		Value	Std. Error	t value	Pr(> t)	Signif
(Intercept)		-6.02586	2.34729	-2.57	0.018	*
lg10(G)		0.69254	0.13713	5.05	0.000	***
D		0.09525	0.03580	2.66	0.015	*
WZ		0.00263	0.00955	0.28	0.785	.
BZ		0.00229	0.00198	1.16	0.261	.
Z		-0.04573	0.03561	-1.28	0.213	.
NE		0.11045	0.03391	3.26	0.004	**
KT		0.05340	0.02970	1.80	0.087	.
BW		0.01278	0.04537	0.28	0.781	.
sqrt(N)		-0.02997	0.01780	-1.68	0.107	.
KG		-0.09951	0.05562	-1.79	0.088	.

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

5.2 Automatisierte Verfahren zur Modellwahl

a Schrittweise rückwärts

b WZ weglassen!

BW, BZ, Z, \sqrt{N} und KT weglassen!

Coefficients:

	Value	Std. Error	t value	Pr(> t)	Signif
(Intercept)	-3.4612	1.1458	-3.02	0.005	**
log10(G)	0.6629	0.1295	5.12	0.000	***
D	0.0610	0.0160	3.82	0.001	***
NE	0.0831	0.0330	2.52	0.018	*
KG	-0.1844	0.0424	-4.35	0.000	***

c Schrittweise vorwärts ...

e „Alle Gleichungen“ – all subsets.

f Kriterien

1. „Bestimmtheitsmass“ R^2 oder multiple Korrelation R ,
Wert der Test-Statistik für das gesamte Modell (F-Test),
zur F-Test-Statistik gehöriger P-Wert,
 4. geschätzte Varianz $\hat{\sigma}^2$ der Fehler
 5. korrigiertes Bestimmtheitsmass R^2 (adjusted R^2):
$$R^2_{\text{adj}} = 1 - \frac{n-1}{n-p'}(1 - R^2)$$
 6. $C_p := \text{SSQ}_{(E)} / \hat{\sigma}^2_m + 2p' - n = n(\text{MS}_E / \hat{\sigma}^2_m - 1 + 2p'/n)$,
 7. Informations-Kriterium von Akaike $\text{AIC} \approx C_p$.
- Grössere Modelle sind nicht immer besser.

h C_p im Beispiel:

KT und \sqrt{N} dazunehmen!

P-Wert für KG: 0.049.

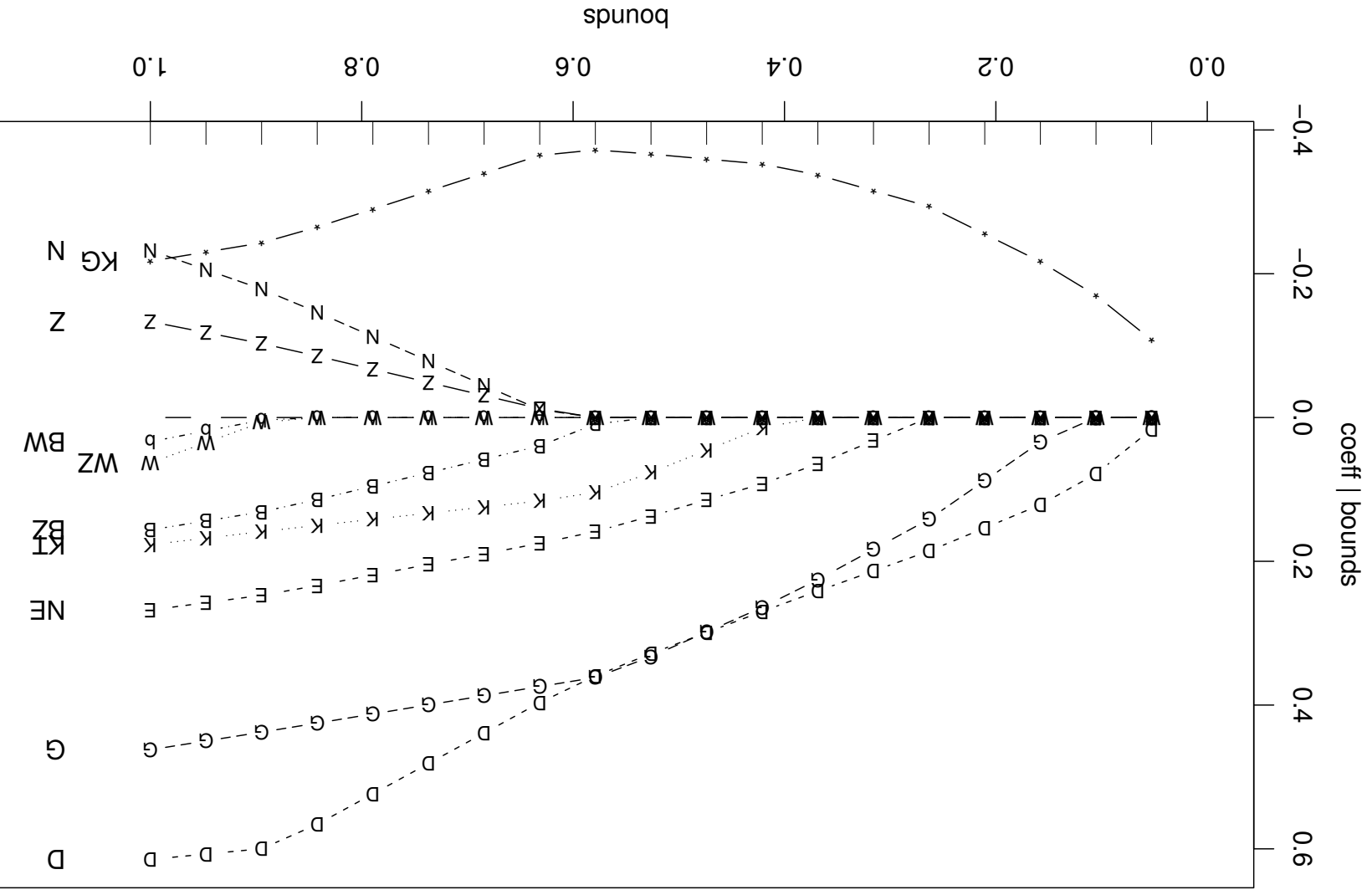
! **Lasso** Penalized Regression

$$\hat{Q}(\bar{\beta}; \lambda) = \sum_i R_i^2 + \lambda \sum_j |\beta_j^*|.$$

λ : wie stark bestrafen?

Variation von λ \rightarrow Koeffizienten werden exakt 0.

\rightarrow Modellwahl



! bestes = wahres Modell ?

Mehrere als Ergebnis des Verfahrens deklarieren!

Unter den „guten“ Modellen soll mit Hilfe von

Plausibilitäts-Überlegungen und Fachwissen

ein geeignetes (oder wenige geeignete)

ausgewählt werden.

Explorative Datenanalyse findet NICHT

das richtige Modell, sondern

einige Modelle, die den Daten gut entsprechen.

k **Prioritäten von Termen** Prinzip: Wenn quadratischer Term im Modell ist,

lässt man den linearen nicht weg.

Wieso? ...

... ausser man habe gute Gründe, vom Prinzip abzuweichen.

Ebenso:

• Wechselwirkung 1. Ordnung $X_1 : X_2$ → beide Haupteffekte drin

lassen.

• Intercept immer drin behalten.

5.3 Kollinearität

a Modell $\bar{Y} = \mathbf{X}\bar{\beta} + \bar{E}$

\mathbf{X} ist **singular**, $X^{(j)}$'s kollinear, wenn

$$\mathbf{X} \text{ singular} \iff \det\langle \mathbf{X} \rangle = 0$$

$$\iff \text{es gibt } \bar{c} \text{ mit } \mathbf{X}\bar{c} = \bar{0} \quad (\bar{c} \neq \bar{0})$$

$$\iff \text{es gibt ein } j \text{ mit } x^{(j)} = \sum_{k \neq j} c_k x^{(k)} + c_0$$

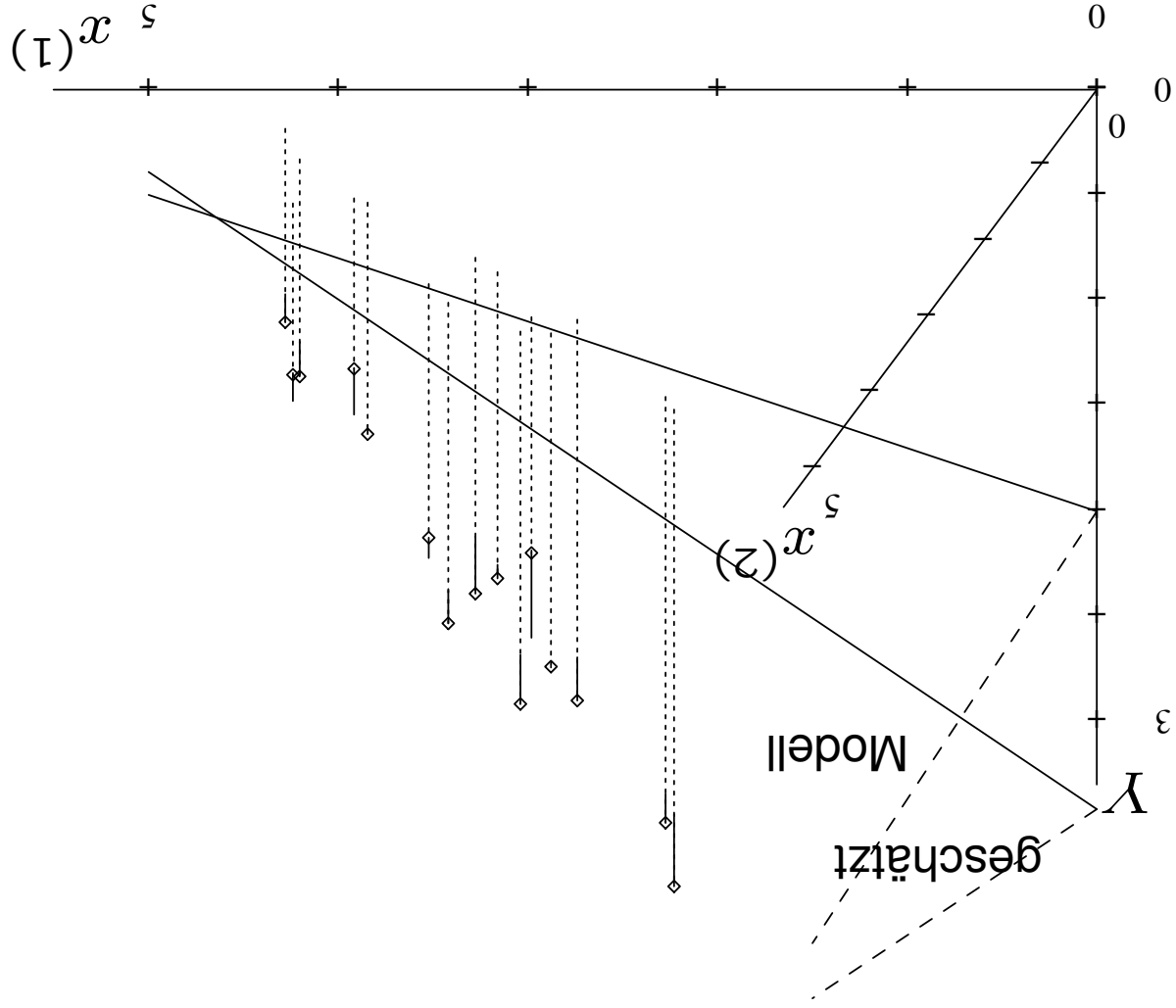
Parameter nicht eindeutig, da

$$\mathbf{X}\bar{\beta} = \mathbf{X}(\bar{\beta} + \gamma\bar{c}), \quad \gamma \text{ beliebig}$$

b Lösung: Kolonne streichen!

Achtung: Interpretation der Parameter kann sich ändern!

c Genäherte Kollinearität \rightarrow Parameter schlecht bestimmt.



d Grosse Standardfehler für Schätzungen \rightarrow Koeffizient nicht signif. $\neq 0$.

e Vorhersage im Bereich der Daten i.O.

f **Wie entdeckt man Kollinearität?**

– Standardfehler der $\hat{\beta}_j$

– Gibt es eine Beziehung $x_i^{(j)} \approx \tilde{c}_0 + \sum_{k \neq j} \tilde{c}_k x_i^{(k)}$?

= Regressionsproblem! Bestimmtheitsmass R_j^2

oder variance inflation factor $VIF_j = 1/(1 - R_j^2)$

g Was tun gegen Kollinearität?

– Wahl der Versuchsbedingungen.

h – x -Variable linear transformieren, z.B. Summe und Differenz

oder „wichtigere“ Variable und Residuen der anderen.

i – Variable mit dem höchsten R_j^2 wegi! (Meist sowieso nicht signifikant)

j* „Ridge Regression“

5.4 Strategien der Modell-Entwicklung

a Modellwahl ist ein Zusammenspiel von

- Vorwissen aus Anwendung und Statistik,
- Residuen-Analyse, „Detektivarbeit“,
- automatischen Modellwahl-Methoden,
- Residuen-Analyse, „Detektivarbeit“,
- Prinzip der Einfachheit,
- Beurteilung der Plausibilität vom Fachwissen her.

Modellwahl ist von der Fragestellung abhängig!

(a) Welche Variablen beeinflussen die Zielgröße?
→ Achtung! Es gibt nur Indizien!

(b) Vorhersage.

(c) Modell im Wesentlichen vorgegeben.
Allenfalls **Störvariable** einbeziehen.

Beispiel Medikamentenprüfung.

Im Folgenden soll (a) oder (b) oder (c) gefragt sein.

0. Daten einlesen, Variablennamen festlegen
 1. "First aid" Transformationen.
 2. Ein grosses Modell
- alle Variablen (Haupteffekte),
 - Ergebnis eines "Schrittweise-Vorwärts-Verfahrens"

3. Überprüfung des zufälligen Teils:

- Ausreisser in den Residuen,

- Verteilung der Residuen,

- Gleichheit der Varianzen,

- Unabhängigkeit der Fehler.

Es kann aufgrund der Ergebnisse angezeigt sein,

- die Zielgrösse zu transformieren,

- Gewichte einzuführen,
 - robuste(re) Methoden zu verwenden, soweit dies nicht schon sowieso geschieht.
 - 4. **Nicht-Linearitäten.** Residuen gegen Ausgangsgrößen.
 - 5. **Automatisierte Variablen-Wahl**
 - 6. **Variablen hinzufügen.**
 - 7. **Wechselwirkungen.** Erst nach Bereinigung der Nicht-Lin. durch entspr. Plots oder numerisch
- step (. . . , scope = (x1 + x2 + . . .) ^ 2)

8. Einflussreiche Beobachtungen.
9. Kritik mit Fachwissen.
10. Anpassung prüfen.
11. Revision.
12. Entfernte Terme überprüfen.
- Feiern!

b Beispiel der Baukosten

Frage nach dem Nutzen der Kostengarantie

Hier führt Detektivarbeit zur überzeugendsten Antwort!

Merkpunkte

1. **Automatisierte** Verfahren zur Variablenwahl sind ein nützliches Hilfsmittel – finden aber nicht „die Wahrheit“
2. Modellwahl ist ein **Zusammenspiel** von
 - Vorwissen aus Anwendung und Statistik,
 - **Residuen-Analyse, „Detektivarbeit“**,
 - automatischen Modellwahl-Methoden,
 - Residuen-Analyse, „Detektivarbeit“,
 - Prinzip der Einfachheit,
 - Beurteilung der Plausibilität vom Fachwissen her.

Modell-Entwicklung