4 Residual Analysis

4.1 Introduction

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

(a) $\mathcal{E}\langle E_i \rangle = 0$: Linearity, Additivity,

(b) equal variances: $\mathrm{var}\langle E_i \rangle = \sigma^2$,

(c) normal distribution,

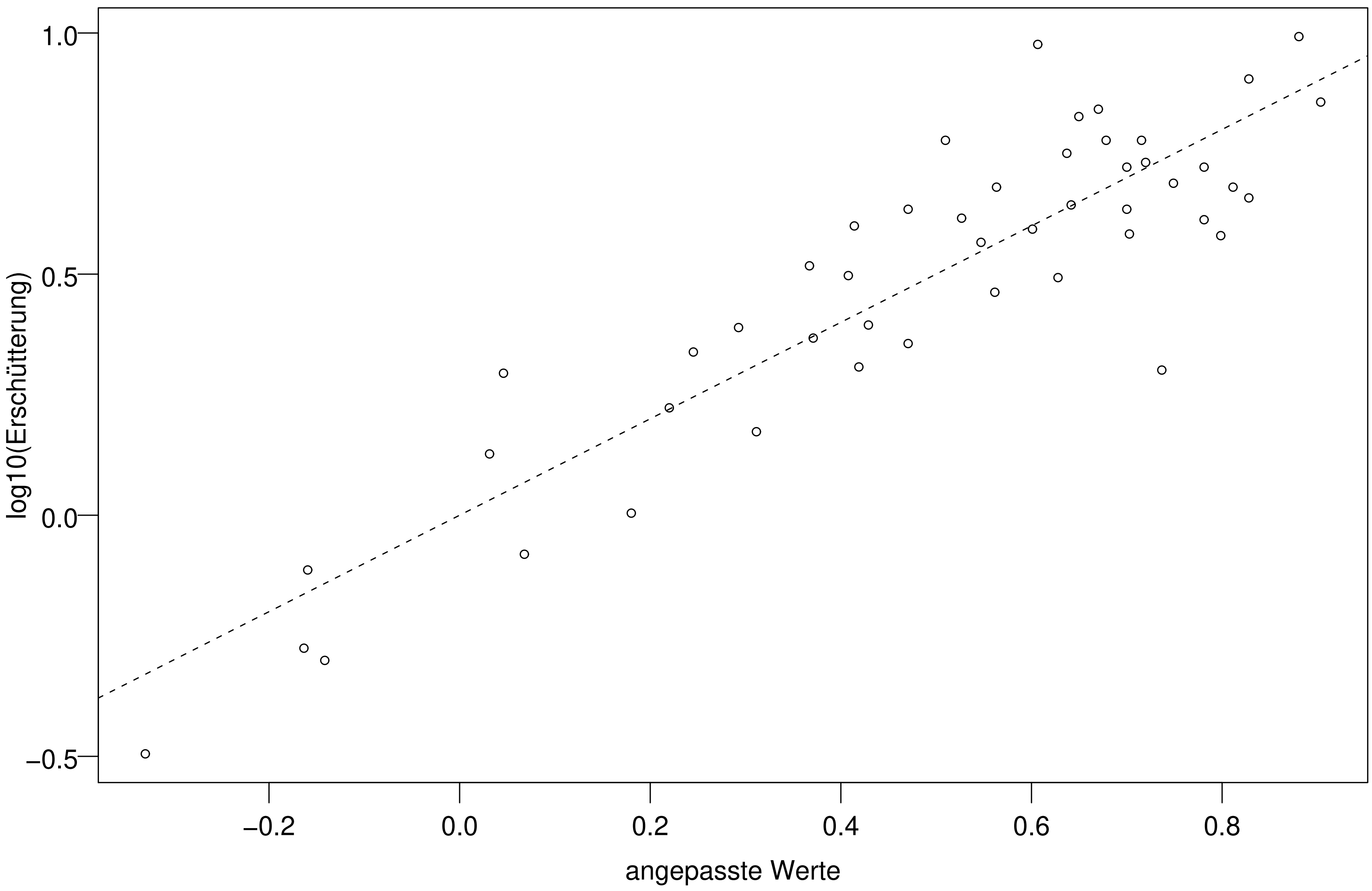
(d) E_i independent

Check model assumptions! $\longrightarrow$ find better model!

Find better model by

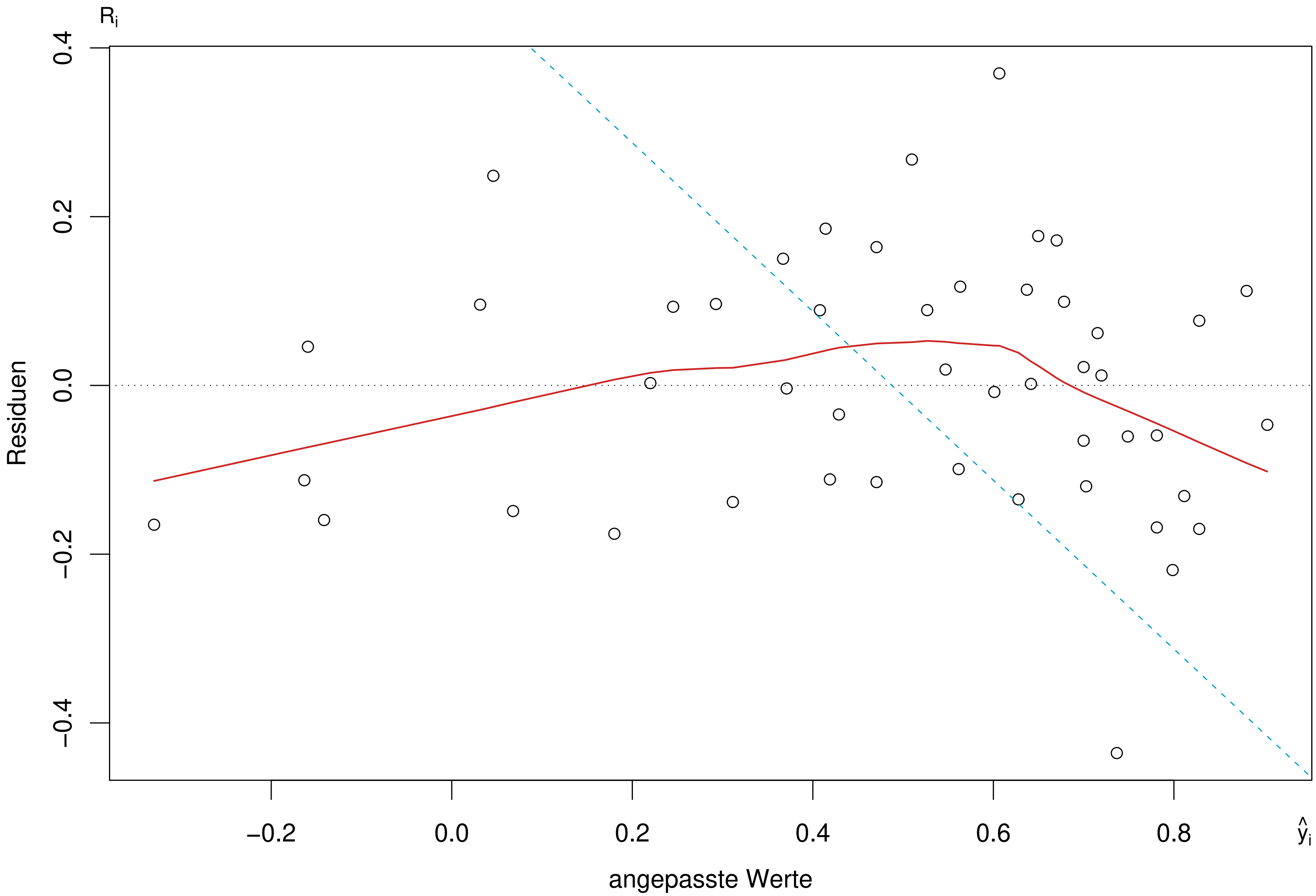
- transformaton of variables,
- additional terms, like interactions,
- weights for observations,
- use alternative methods for estimation and inference

Simple regression: Examine scatterplot Y vs. X
Multiple explanatory variables: use a linear combination of X's for horizontal axis
$\longrightarrow$ use $\widehat{\beta}_0 + \widehat{\beta}_1 X^{(1)} + \widehat{\beta}_2 X^{(2)} + \dots = \widehat{y}$, "fitted values".


1.0
0.5
0.0
–0.5
log10(Erschütterung)
–0.2
0.0
0.2
0.4
0.6
0.8
angepasste Werte

4.2 Residuals and fitted values

a Tukey-Anscombe Plot: Res. $R_i = y_i - \widehat{y}_i$ vs. fitted values $\widehat{y}_i$



4.2

b What kinds of deviations from assumptions can show up?

(a) Regression function: Pattern of the points:
Sliding mean (smoother) may show curvature.

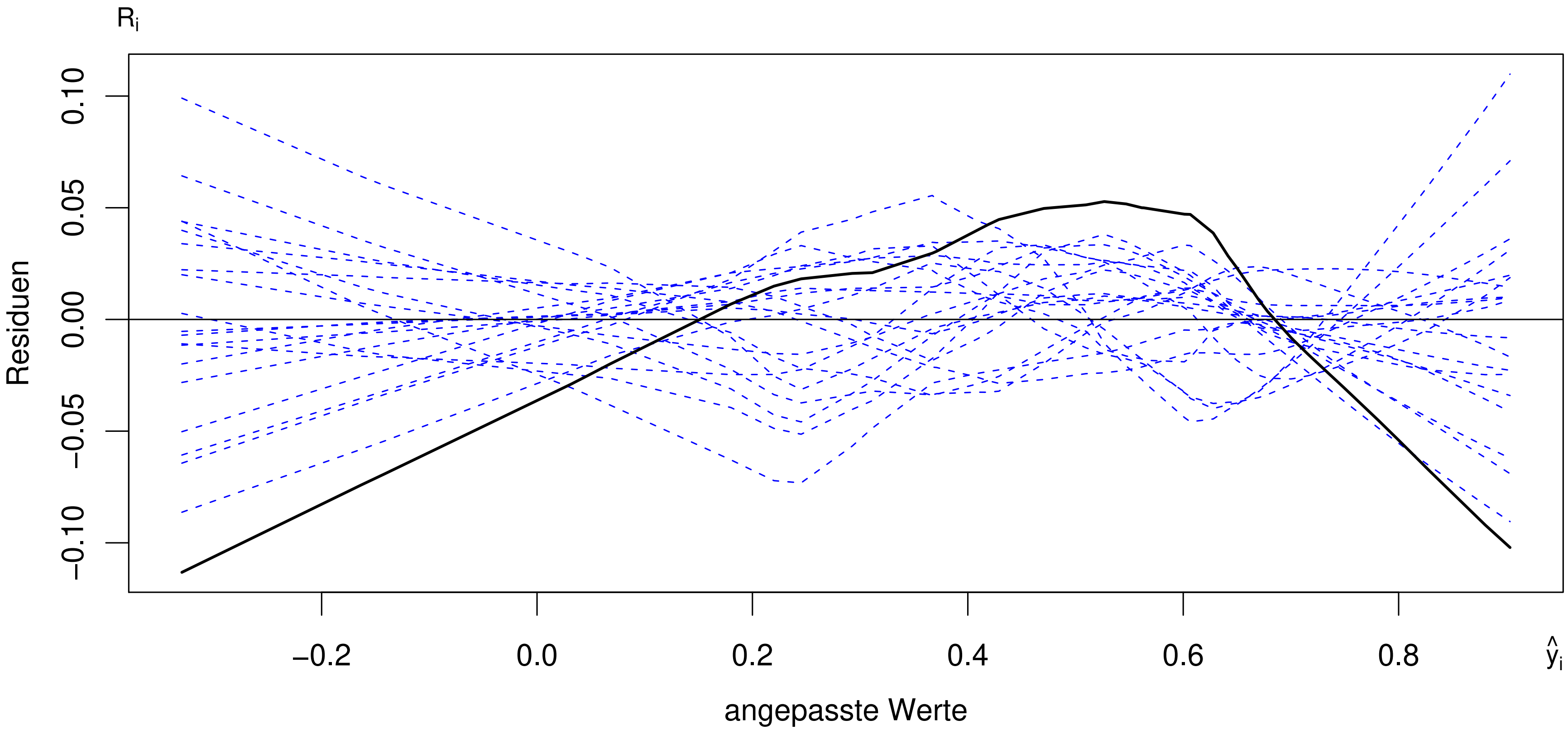
(b) Equality of variances: (vert.) variation of pts around smooth.
Points may "fan out" to the right.
More precisely seen in plot of absolute residuals against fit.

(c) Distribution of errors: Do points scatter symmetrically
around the 0 line (or the smooth)? Outliers?

4.2

c How to judge?

- Are deviations in range of chance? ⟶ Simulated curves

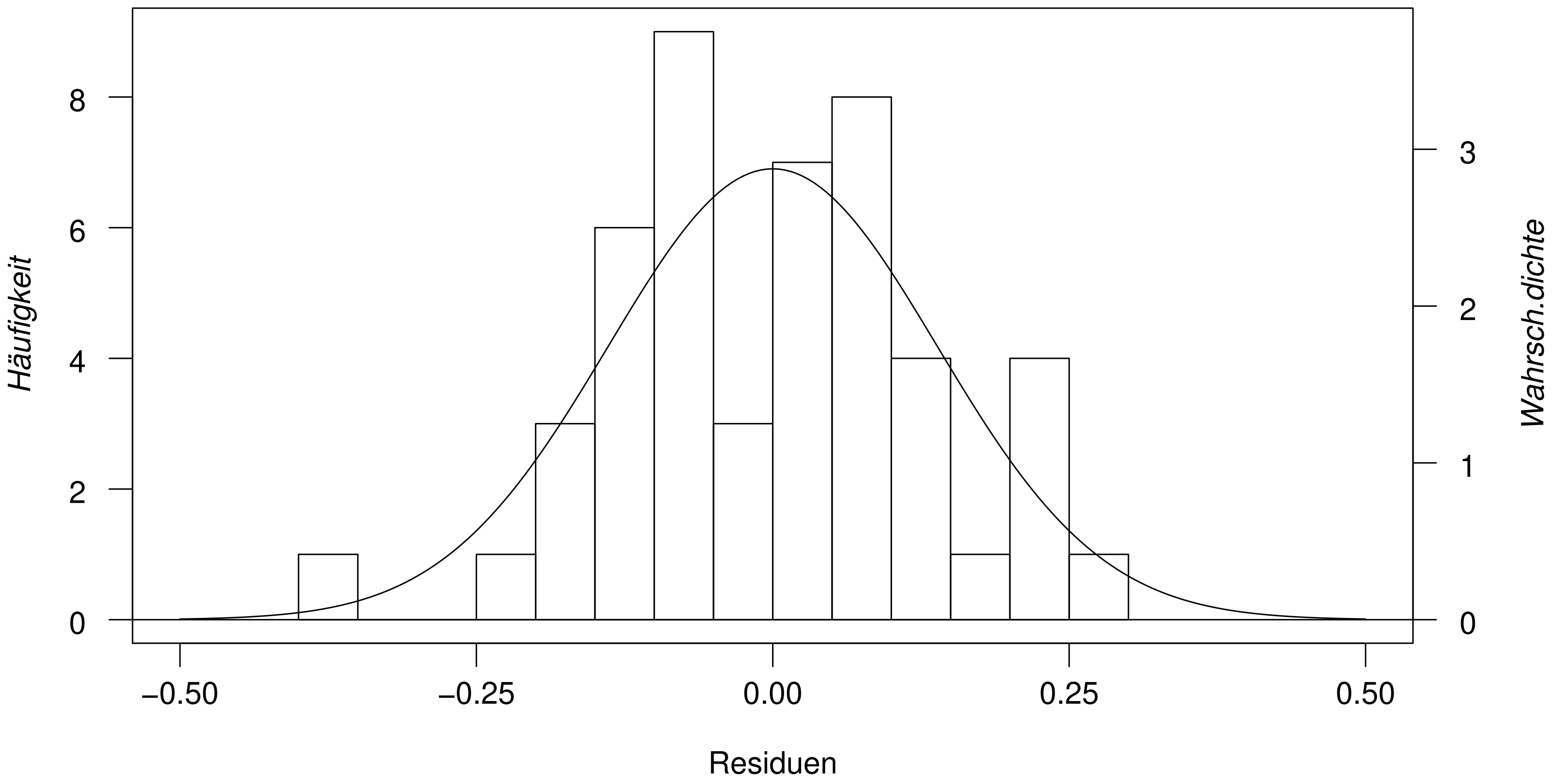


- Are deviations dangerous? Answer depends on purpose.

4.3 Distribution of errors

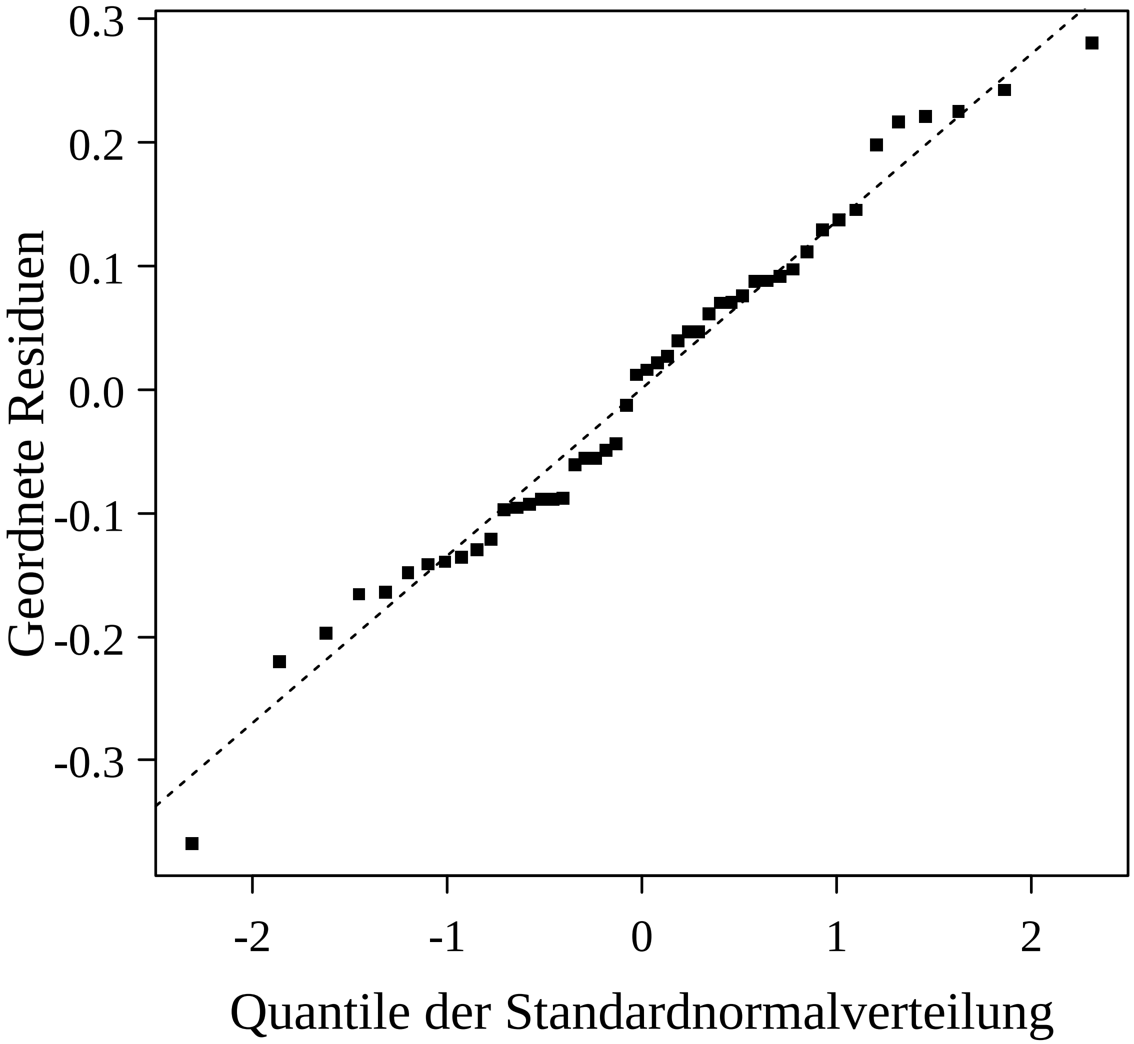
a (c) normal distribution ? Histogram of E_i ... $\longrightarrow$ residuals R_i

! Note that Y does NOT need to be normally distributed !



4.3

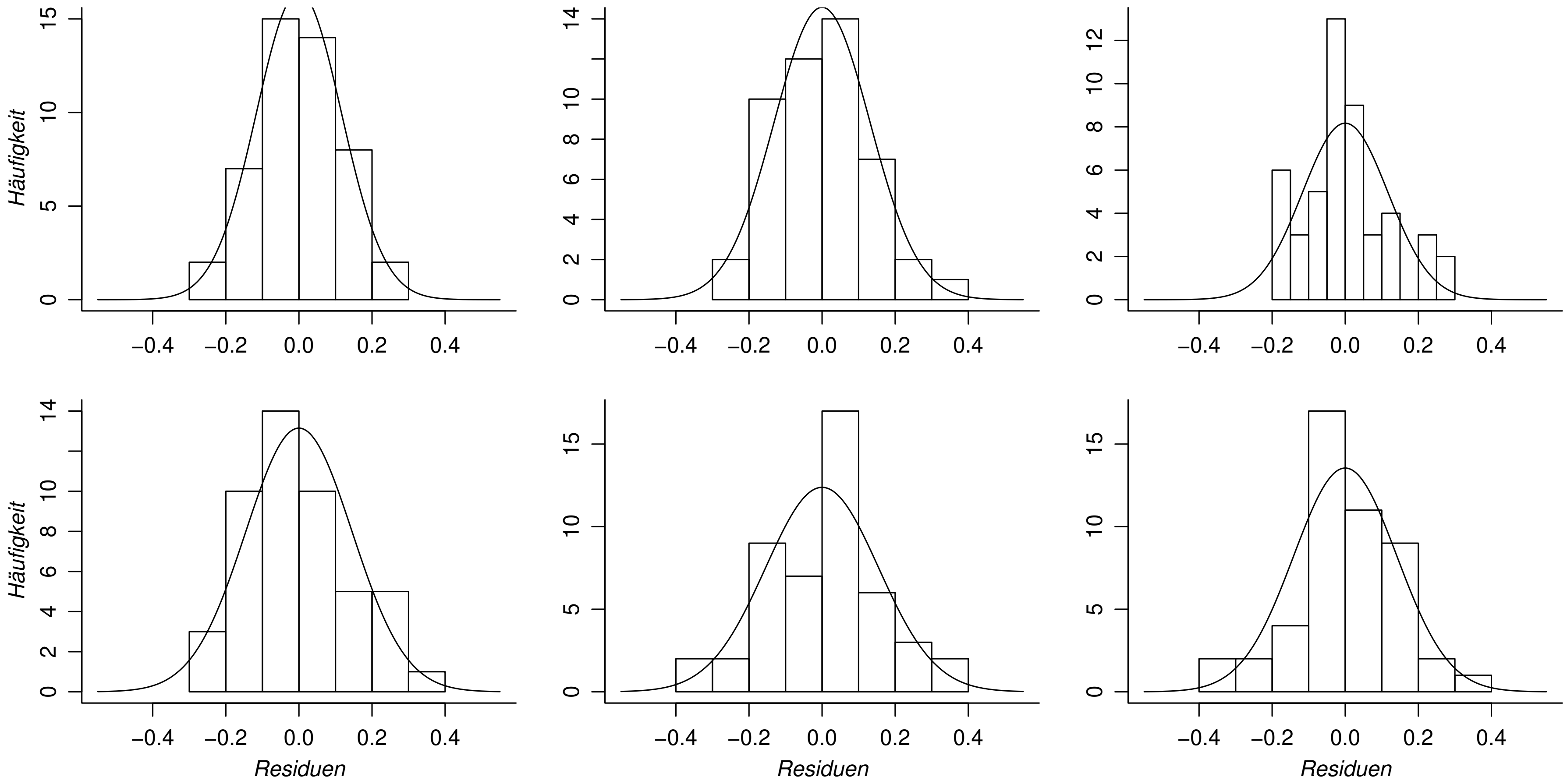
b Refinement: quantile-quantile-plot (qq-plot, normal plot)

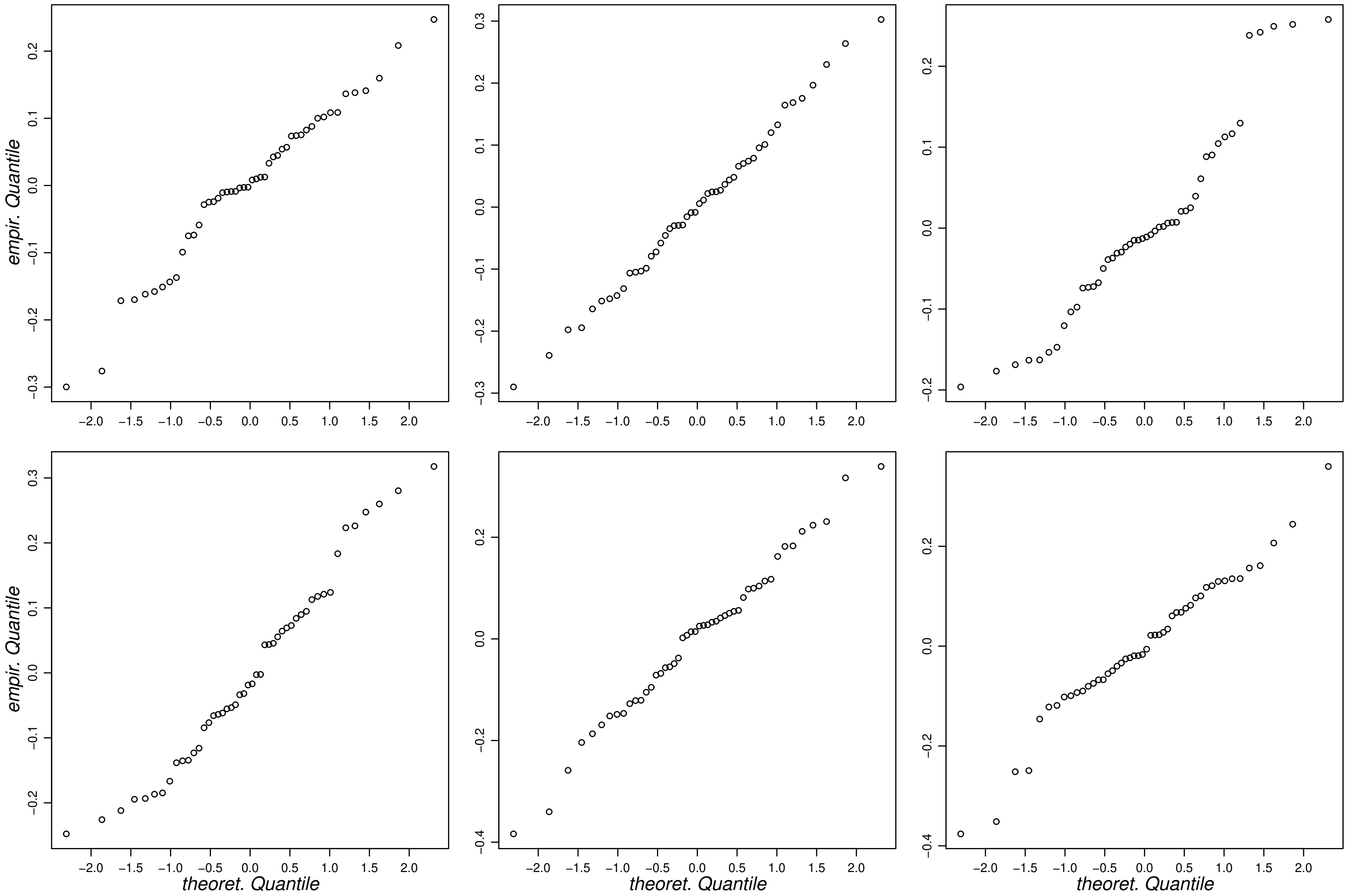


4.3

c Deviations significant? $\longrightarrow$ goodness of fit test

d ... or simulate!




empir. Quantile
theoret. Quantile

4.3

e Distribution of errors? Errors $E_i \neq$ residuals R_i
$R_i = Y_i - \widehat{y}_i$ both random. $\widehat{y}_i$ dependant of Y_i, hence of E_i.

f $R_i \sim \mathcal{N}\langle 0, \sigma^2(1 - H_{ii})\rangle$.
H_{ii} leverage

- $Y_i \to Y_i + \Delta y_i \quad \longrightarrow \quad \widehat{y}_i \to \widehat{y}_i + H_{ii}\Delta y_i$
- H_{ii} measures „distance" between $\underline{x}_i$ and $\overline{\underline{x}}_i$.
  simple regr.: $H_{ii} = (1/n) + (x_i - \bar{x})^2/\text{SSQ}^{(X)}$.
  multiple r.: $H_{ii} = (1/n) + d\langle x_i, \bar{x}\rangle^2$. d: Mahalanobis dist.
- $0 \le H_{ii} \le 1, \quad \text{ave}_i\langle H_{ii}\rangle = p/n$.

4.3

g Standardize residuals $\longrightarrow$ identical distribution.

$$\widetilde{R}_i = R_i \Big/ \left(\widehat{\sigma}\sqrt{1 - H_{ii}}\right)$$

Use stand. residals for checking the distribution!

Usually unimportant!

But: Notion of leverage will be used again!

4.4 Should we transform the target variable?

a What if deviations do show up?

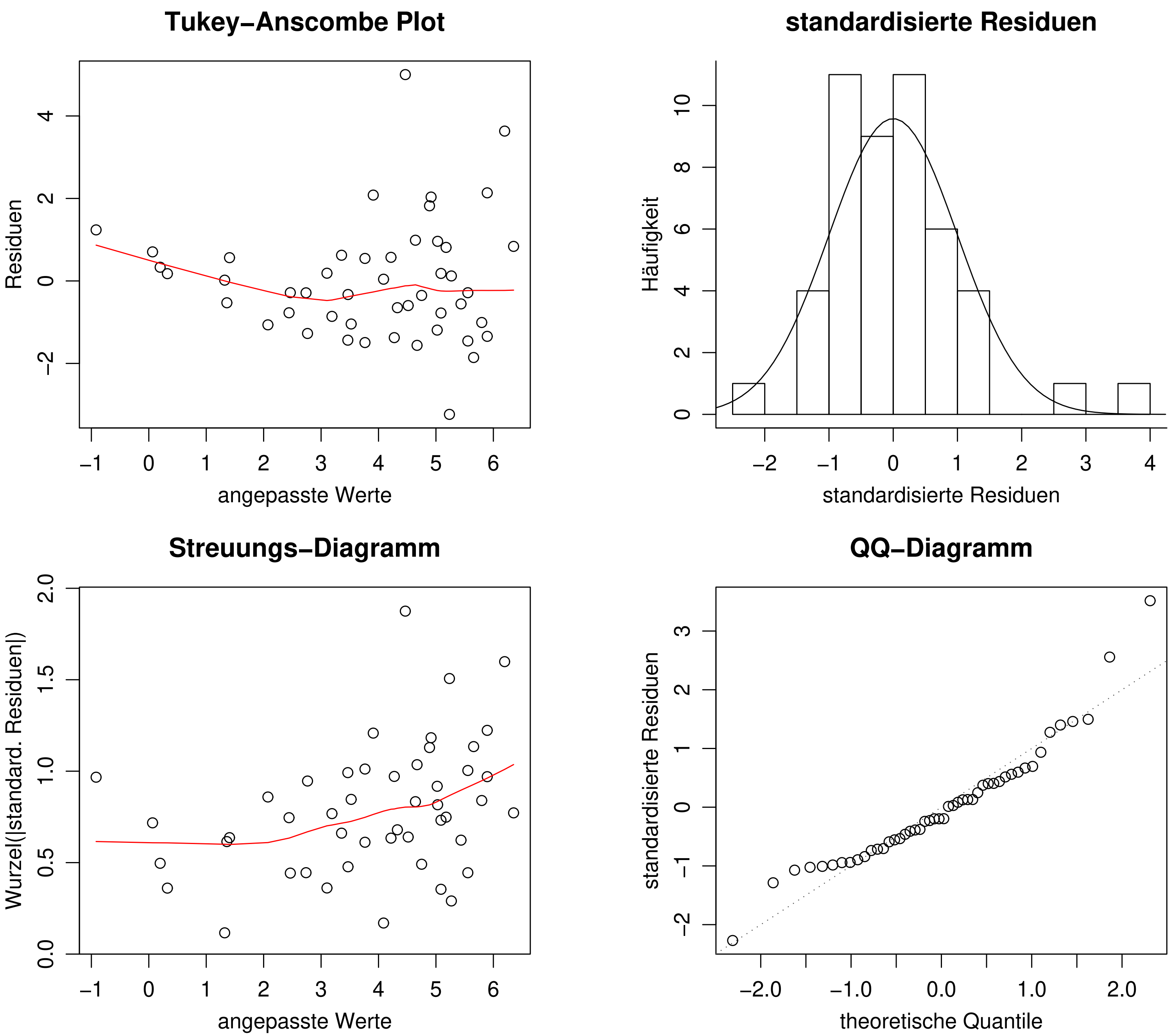
Cf. medical diagnosis on the basis of symptoms.

Study symptoms of known deseases for calibration.

Disease: Missing log transformation $\longrightarrow$

- curved smooth in TA plot (plate shape)
- Variation fanning out to the right
- Skewed distribution

$\longrightarrow$ Transformation Syndrom


Tukey−Anscombe Plot
Residuen
angepasste Werte
standardisierte Residuen
Häufigkeit
standardisierte Residuen
Streuungs−Diagramm
Wurzel(|standard. Residuen|)
angepasste Werte
QQ−Diagramm
standardisierte Residuen
theoretische Quantile

4.4

b **First Aid Transformations**

amounts $\longrightarrow$ `log`

counts $\longrightarrow$ `sqrt`

Percentages $\longrightarrow$ "arc sin"-Trsf. `asin(sqrt(p/100))`

c **Outliers**

d **Long-tailed distributions** $\longrightarrow$ robust methods.

4.5 Residuals and Explanatory Variables

a Plot Residuals against explanatory variables
$\longrightarrow$ Transfomation of xs, additional terms

b Non-constant variances $\longrightarrow$ weighted regression

c Influential Points

d Independence of Errors: Plot Residuals against Sequence