

## 4 Residual Analysis

### 4.1 Introduction

a

Assumptions:  $E_i \sim \mathcal{N}(0, \sigma^2)$

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

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

(c) normal distribution,

(d)  $E_i$  independent

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

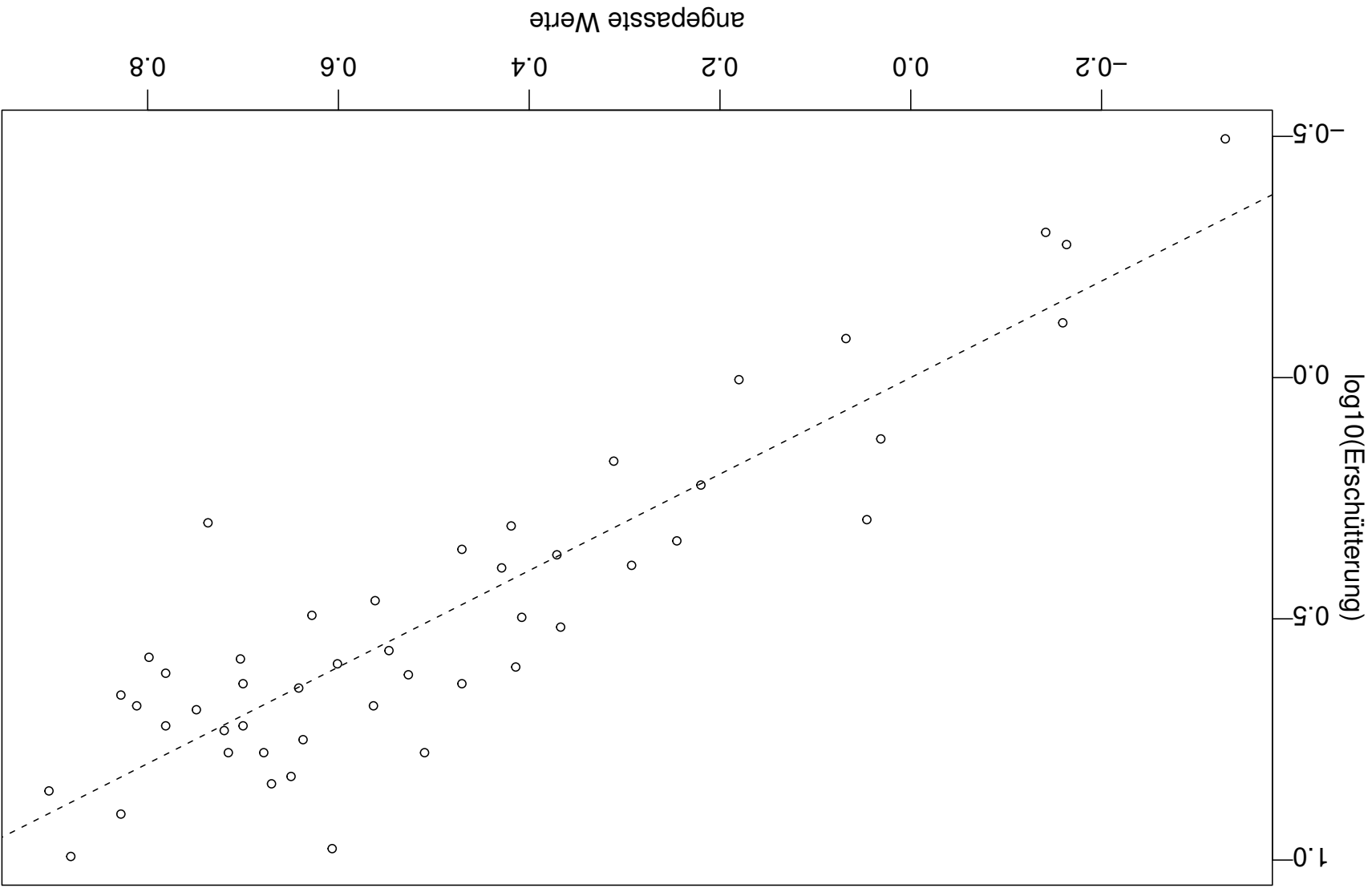
Find better model by

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

Simple regression: Examine scatterplot  $Y$  vs.  $X$

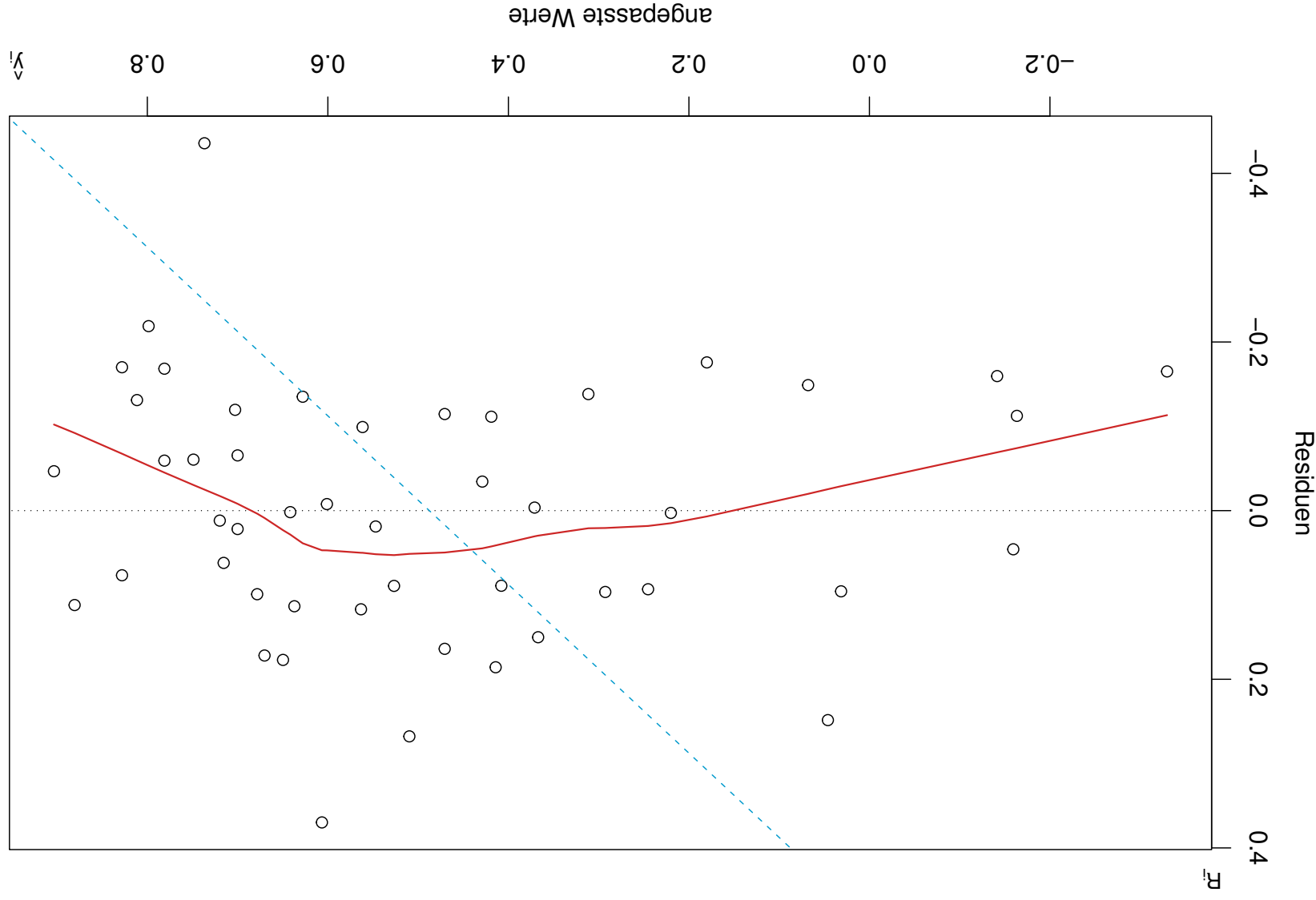
Multiple explanatory variables: use a linear combination of  $X$ 's for horizontal axis

→ use  $\hat{\beta}_0 + \hat{\beta}_1 X^{(1)} + \hat{\beta}_2 X^{(2)} + \dots = \hat{y}$ , "fitted values".



## 4.2 Residuals and fitted values

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



b What kinds of deviations from assumptions can show up?

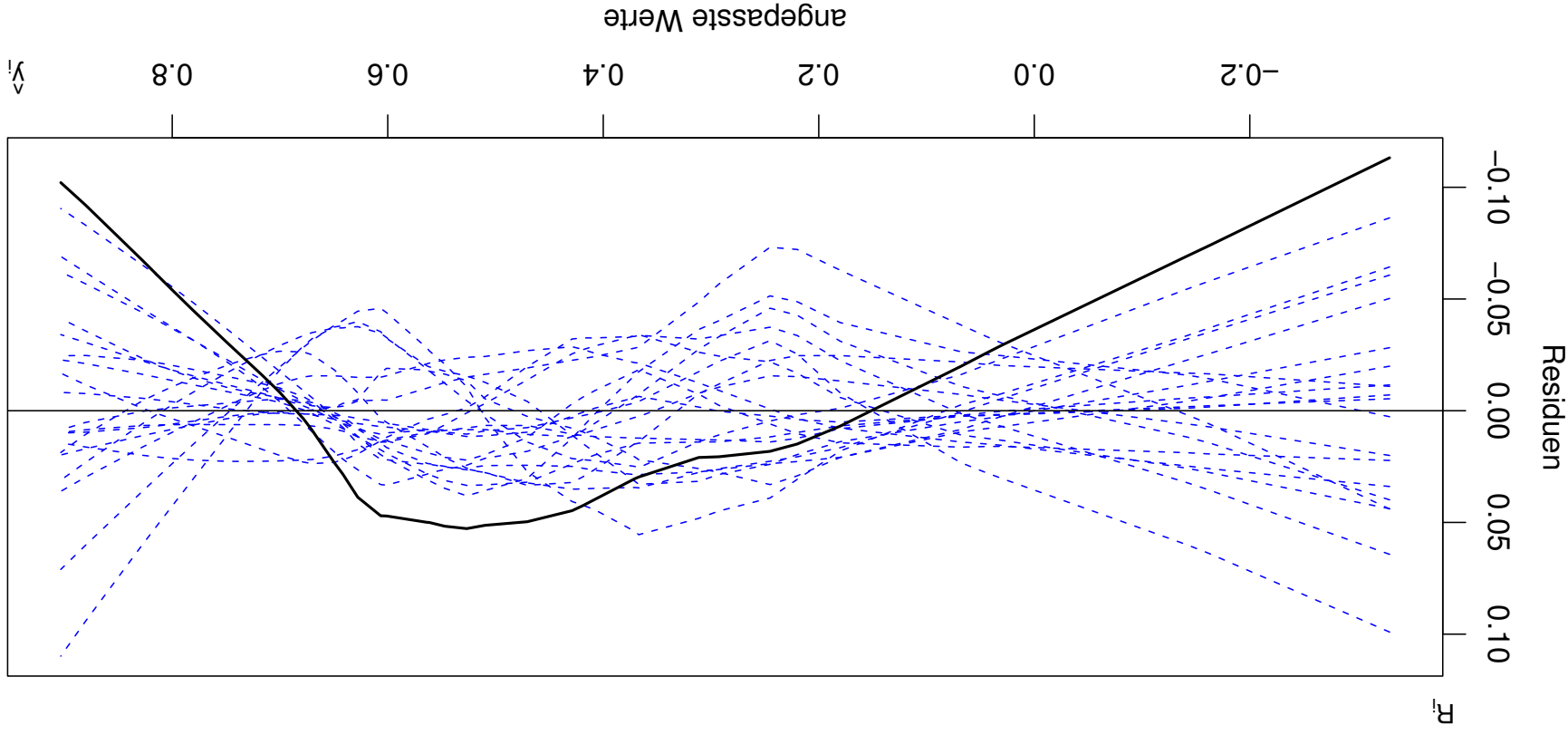
(a) **Regression function: Pattern** of the points:  
Sliding mean (smoother) may show curvature.

(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?**

c How to judge?

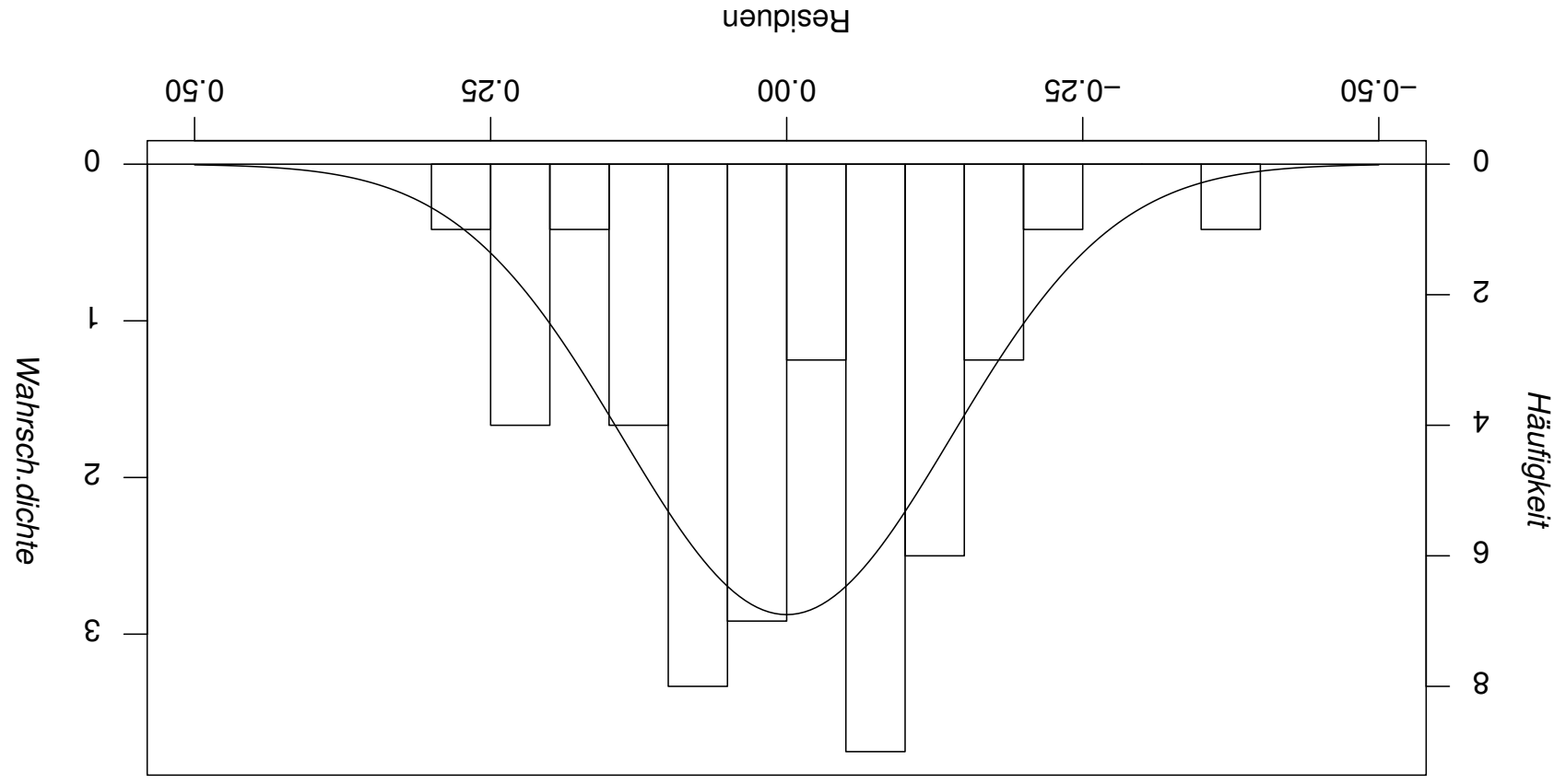
- Are deviations in range of chance? ← **Simulated curves**



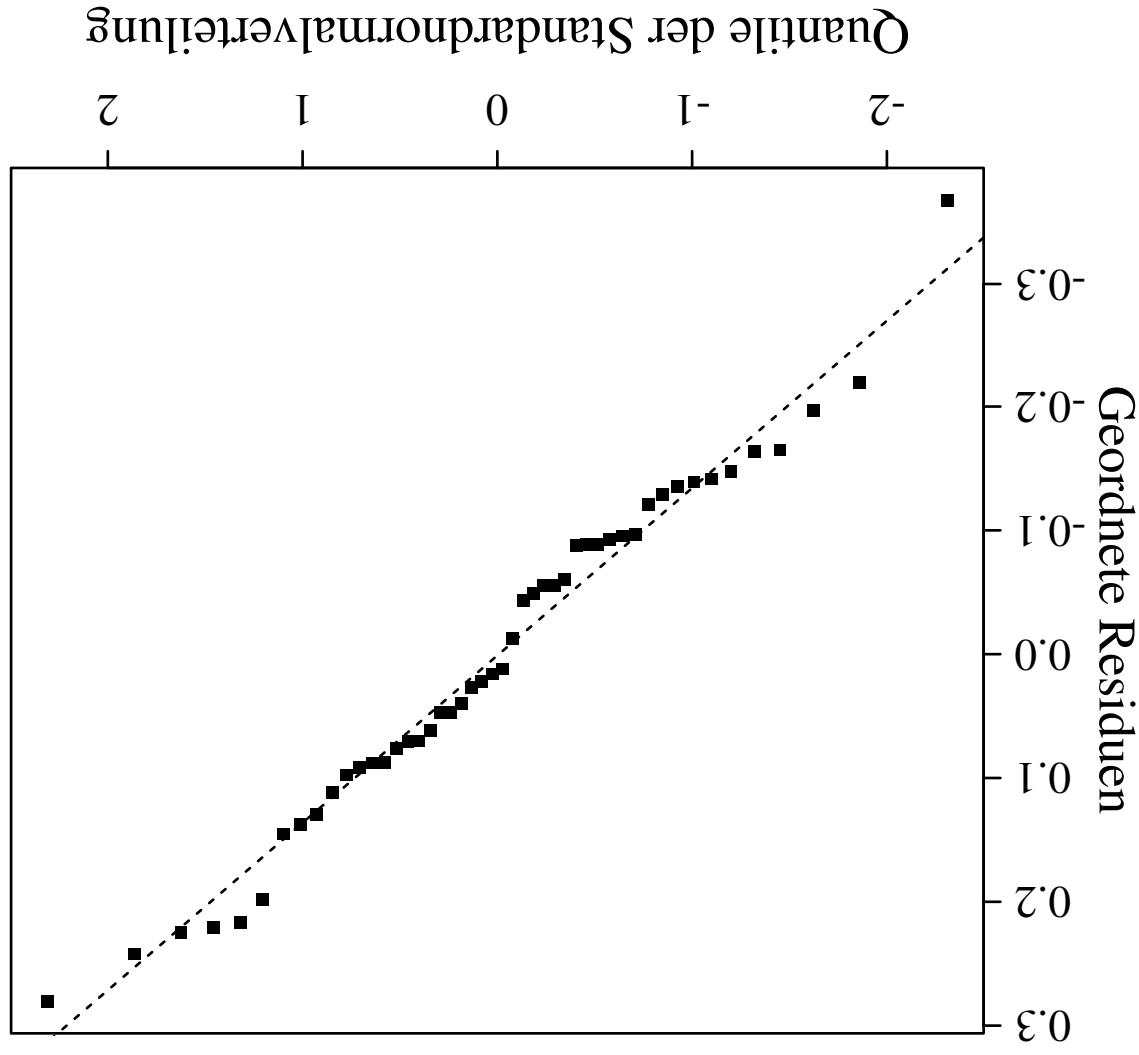
- Are deviations dangerous? Answer depends on **purpose**.

## 4.3 Distribution of errors

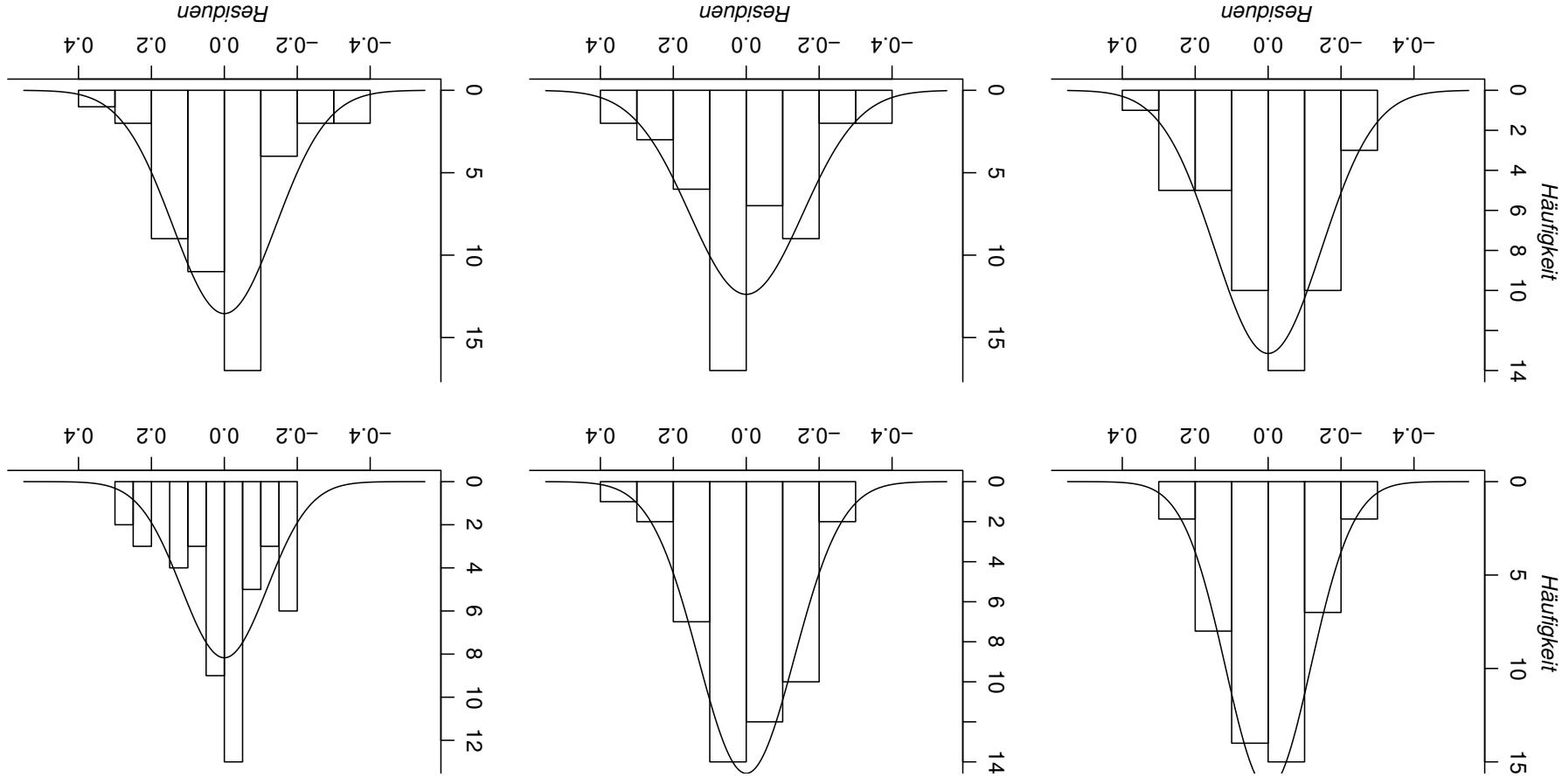
a (c) normal distribution ? Histogram of  $E_i$  ... → residuals  $R_i$  ! Note that  $Y$  does NOT need to be normally distributed !

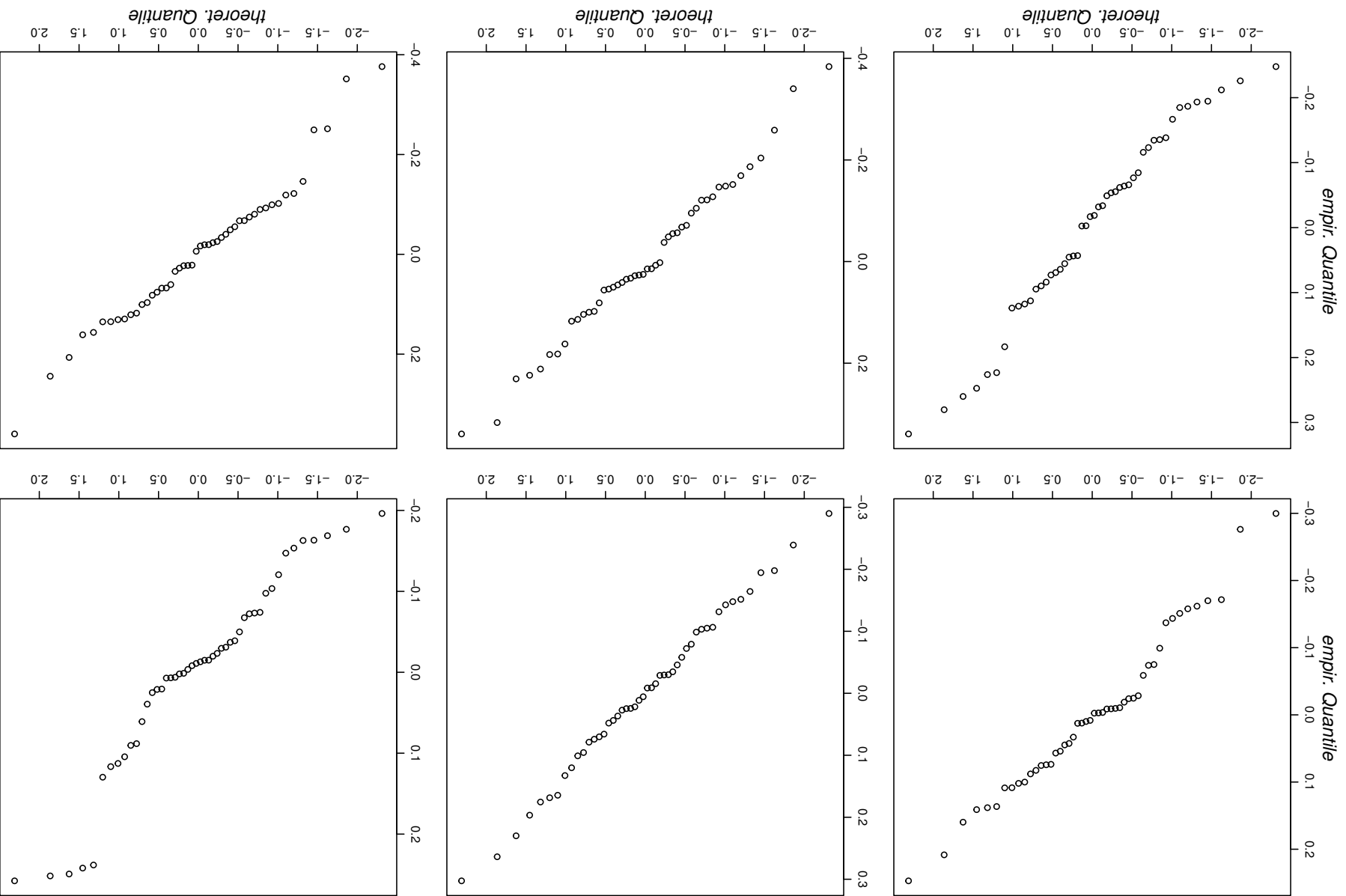


b Refinement: quantile-quantile-plot (qq-plot, normal plot)



- c Deviations significant? → goodness of fit test
- d ... or simulate!





e Distribution of errors? Errors  $E_i \neq$  residuals  $R_i$

$R_i = Y_i - \hat{y}_i$  both random.  $\hat{y}_i$  dependant of  $Y_i$ , hence of  $E_i$ .

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

$H_i$  leverage

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

$H_i$  measures "distance" between  $\bar{x}_i$  and  $\bar{x}$ .

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

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

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

**Standardize** residuals → identical distribution.

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

Use stand. residuals 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 diseases for calibration.

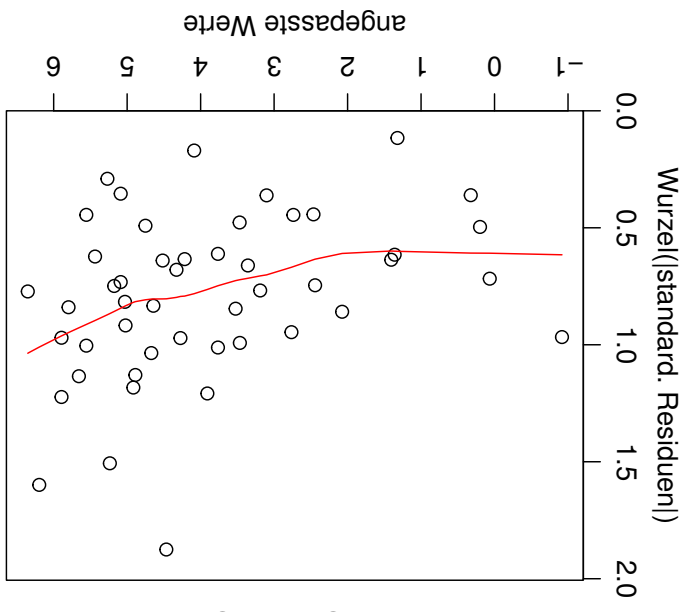
Disease: Missing log transformation ←

• curved smooth in TA plot (plate shape)

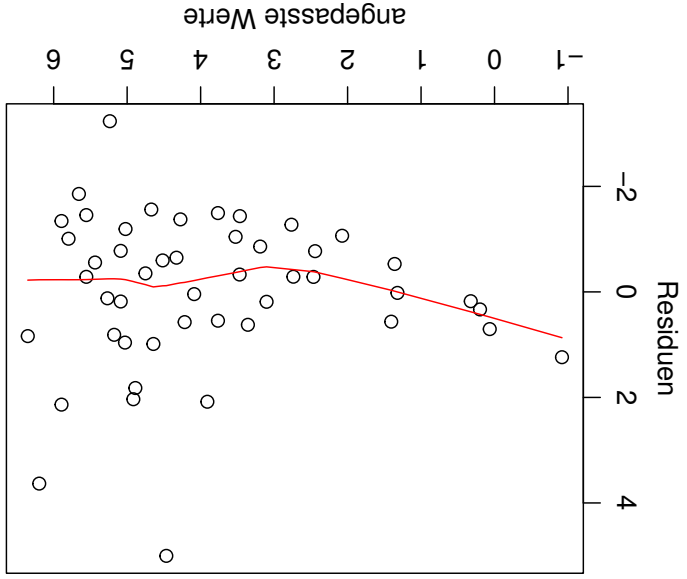
• Variation fanning out to the right

• Skewed distribution

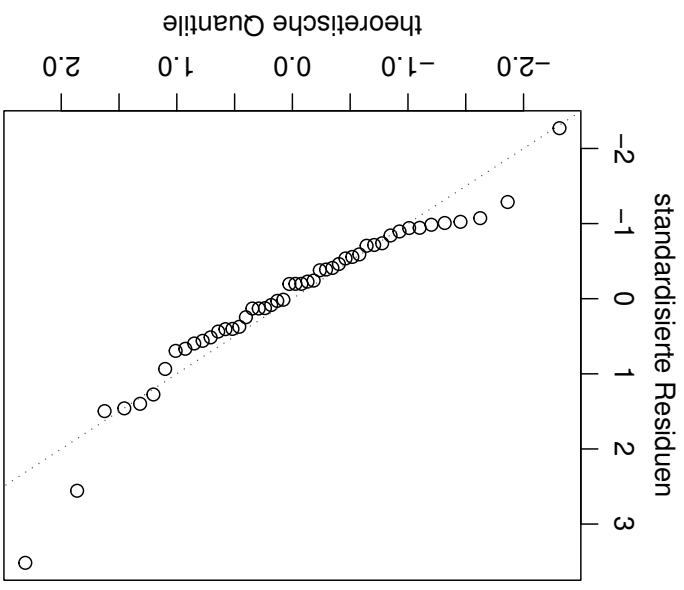
← Transformation Syndrom



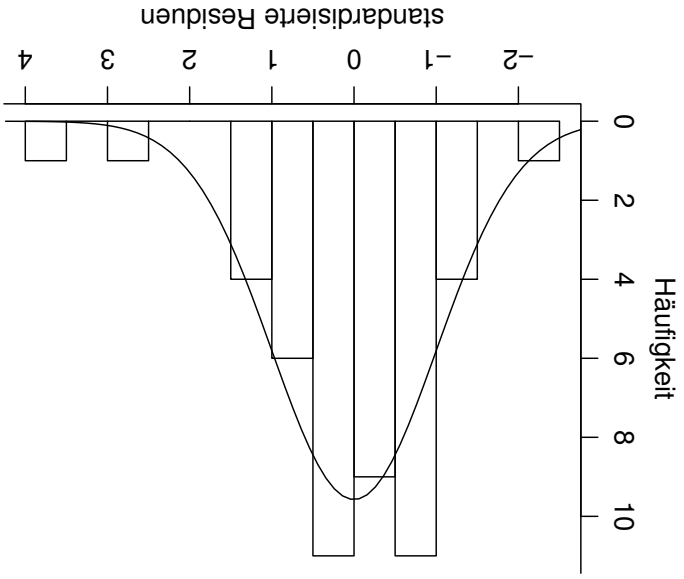
**Streuungs-Diagramm**



**Tukey-Anscombe Plot**



**QQ-Diagramm**



**standardisierte Residuen**

b **First Aid Transformations**

amounts  $\leftarrow$  log

counts  $\leftarrow$  sqrt

Percentages  $\leftarrow$  "arc sin"-Trsf. as in (sqrt(p/100))

c **Outliers**

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

## 4.5 Residuals and Explanatory Variables

a Plot Residuals against explanatory variables  
← Transformation of  $x$ s, additional terms

b Non-constant variances ← weighted regression

c Influential Points

d Independence of Errors: Plot Residuals against Sequence