

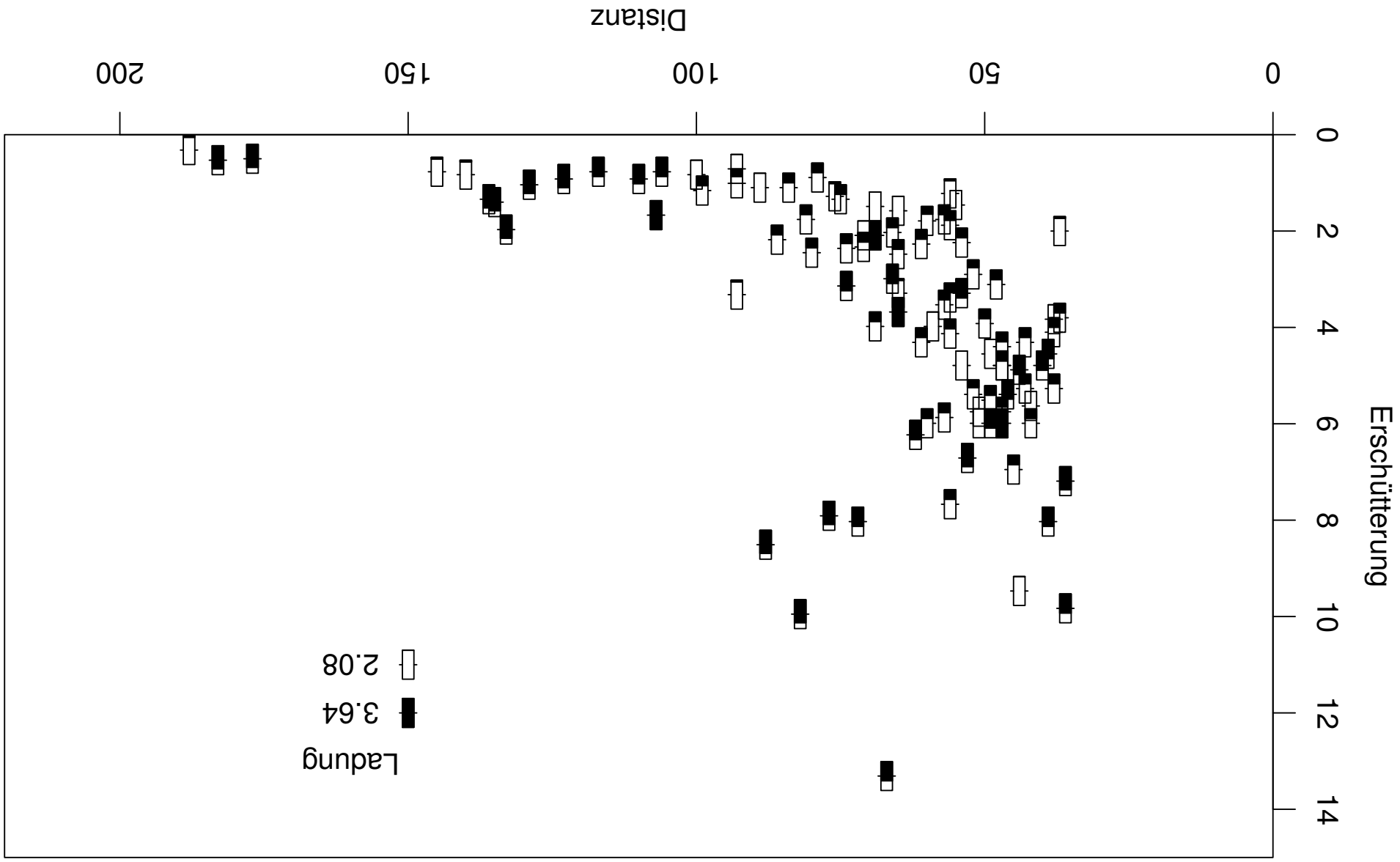
REGRESSION 1: Multiple lineare Regression

1 Einführung in die statistische Regressionsrechnung

1.1 Beispiele zur linearen Regression

b Beispiel Sprengungen.

Erschütterung \approx Funktion \langle Ladung, Distanz, Spreng-Sit., Untergrundart \rangle



c Y : Zielgröße

$x^{(1)}, x^{(2)}, \dots, x^{(m)}$: **Ausgangsgrößen** oder erklärende Variable

$$Y_i = h(x_i^{(1)}, x_i^{(2)}, \dots, x_i^{(m)}) + E_i$$

h : Regressionsfunktion, E_i : **Zufallsabweichung**

Einfachster Fall: 1 erklärende Variable

linearer Zusammenhang: $h(x) = \alpha + \beta x$

$$Y_i = \alpha + \beta x_i + E_i$$

d **Beispiel Schadstoffe im Tunnel.**

Y_i : Schadstoff-Emission pro Kilometer,
 $x_i^{(1)}$: Anzahl "Nicht-Lastwagen",
 $x_i^{(2)}$: Anzahl Lastwagen.

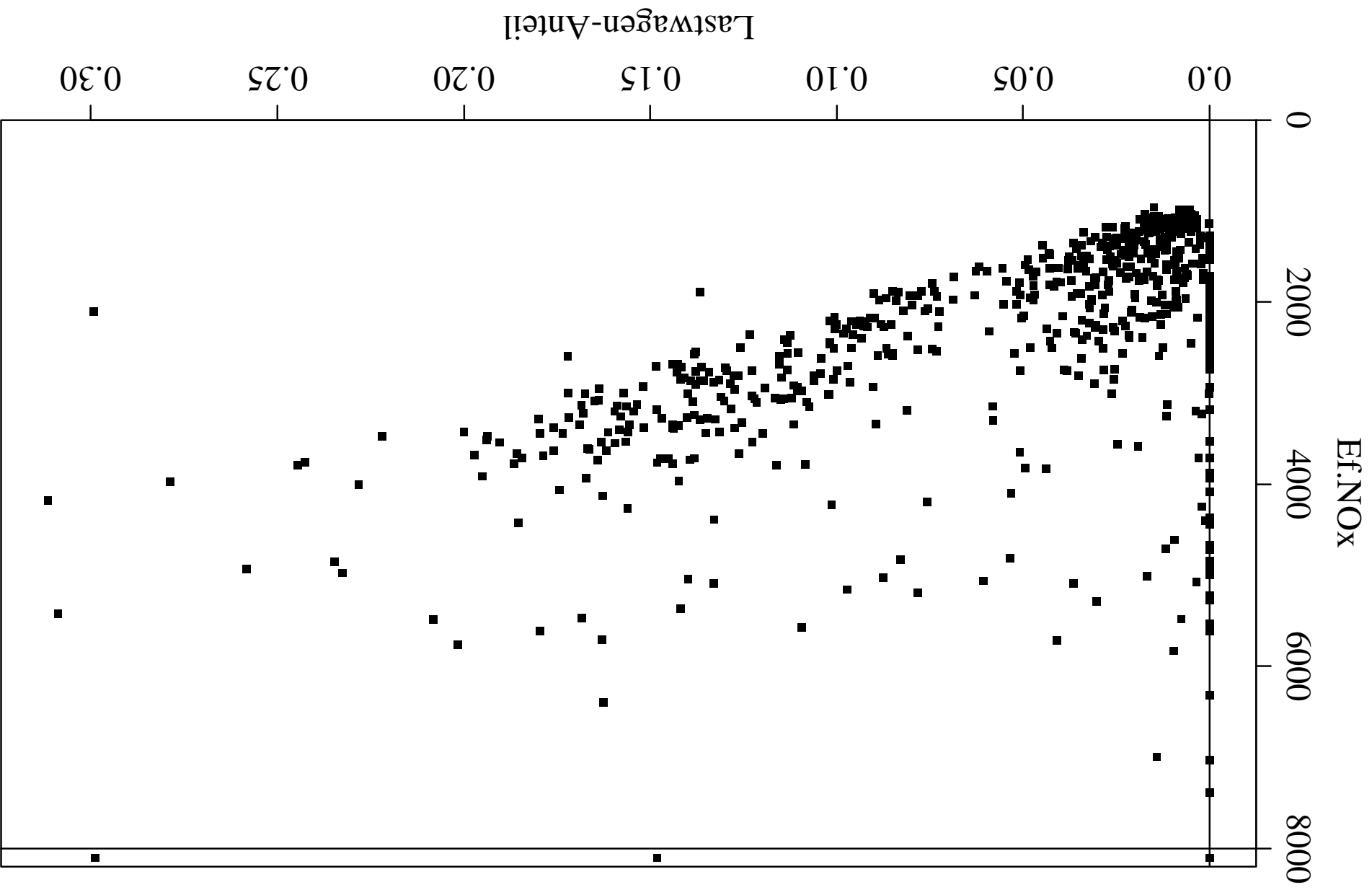
$$Y_i = \beta_1 x_i^{(1)} + \beta_2 x_i^{(2)} + E_i$$

β_1 (β_2) : durchschnittl. Emission pro Nicht-Lastwagen (Lastwagen)

Dividieren durch Fahrzeugzahl $x_i^{(1)} + x_i^{(2)}$

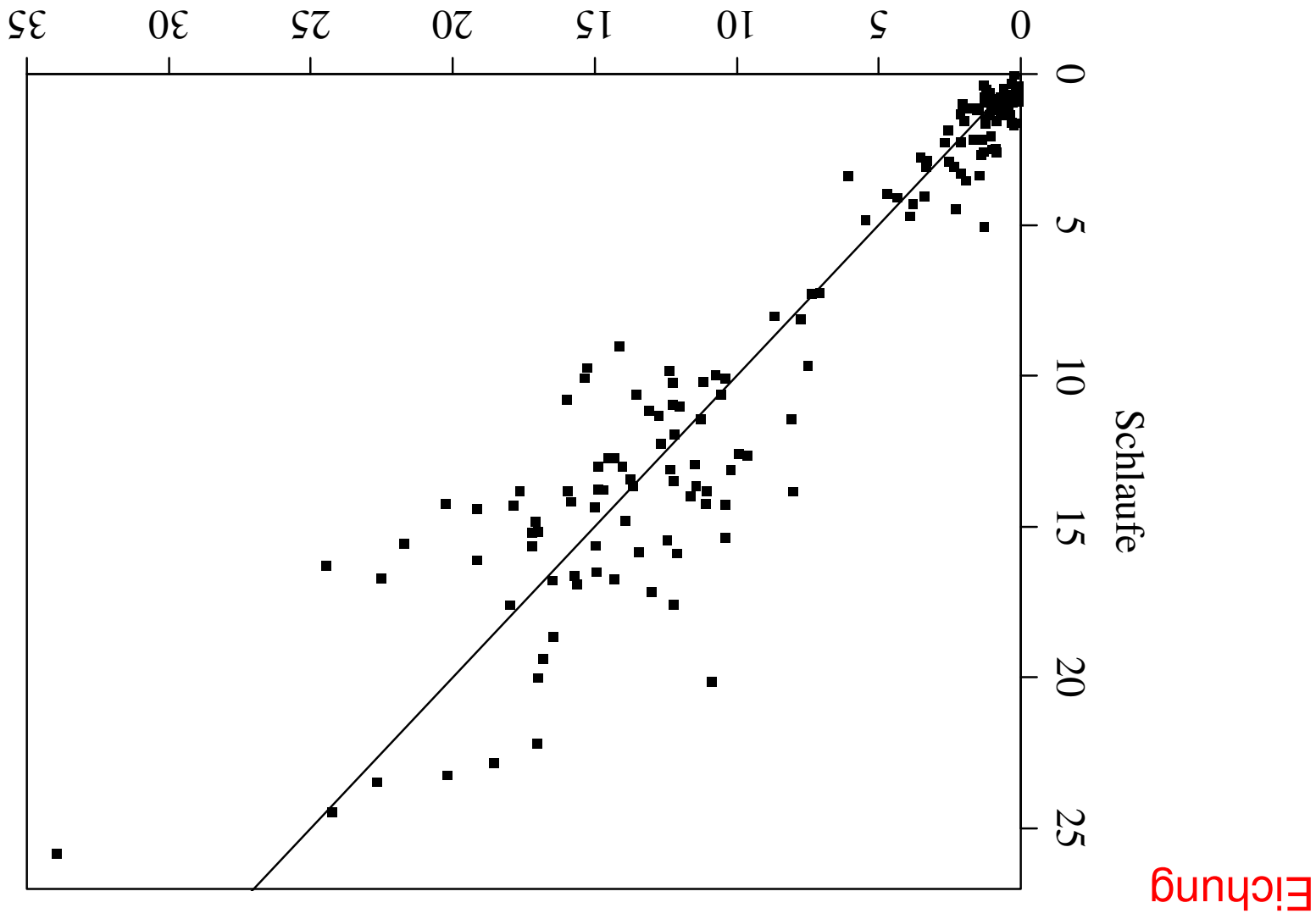
$$\begin{aligned} \tilde{Y}_i &= \beta_1 \tilde{x}_i^{(1)} + \beta_2 \tilde{x}_i^{(2)} + \tilde{E}_i = \beta_1 (1 - x_i^{(2)}) + \beta_2 \tilde{x}_i^{(2)} + \tilde{E}_i \\ &= \beta_1 + (\beta_2 - \beta_1) \tilde{x}_i^{(2)} + E_i = \alpha + \beta \tilde{x}_i^{(2)} + E_i \end{aligned}$$

← **einfache lineare Regr!**



e Beispiel Lastwagen-Anteil.

Y_i Anteil der Lastwagen gemäss Schlaufen-Detektor
 x_i Anteil der Lastwagen gemäss Video-Auszählung



f **Beispiel Antikörper-Produktion.**

Y : Produktion von Antikörpern in Zellen,
die Mäusen injiziert werden.

$x^{(j)}$: 4 "Prozess-Faktoren"

Experiment braucht viele Mäuse,

ist zeitaufwendig und kostet Geld.

Sparen durch **Versuchsplanung!**

Phase 1 : Wichtige Faktoren finden.

g Phase 2 : Optimale Einstellungen dieser Faktoren finden.

h Fragestellungen:

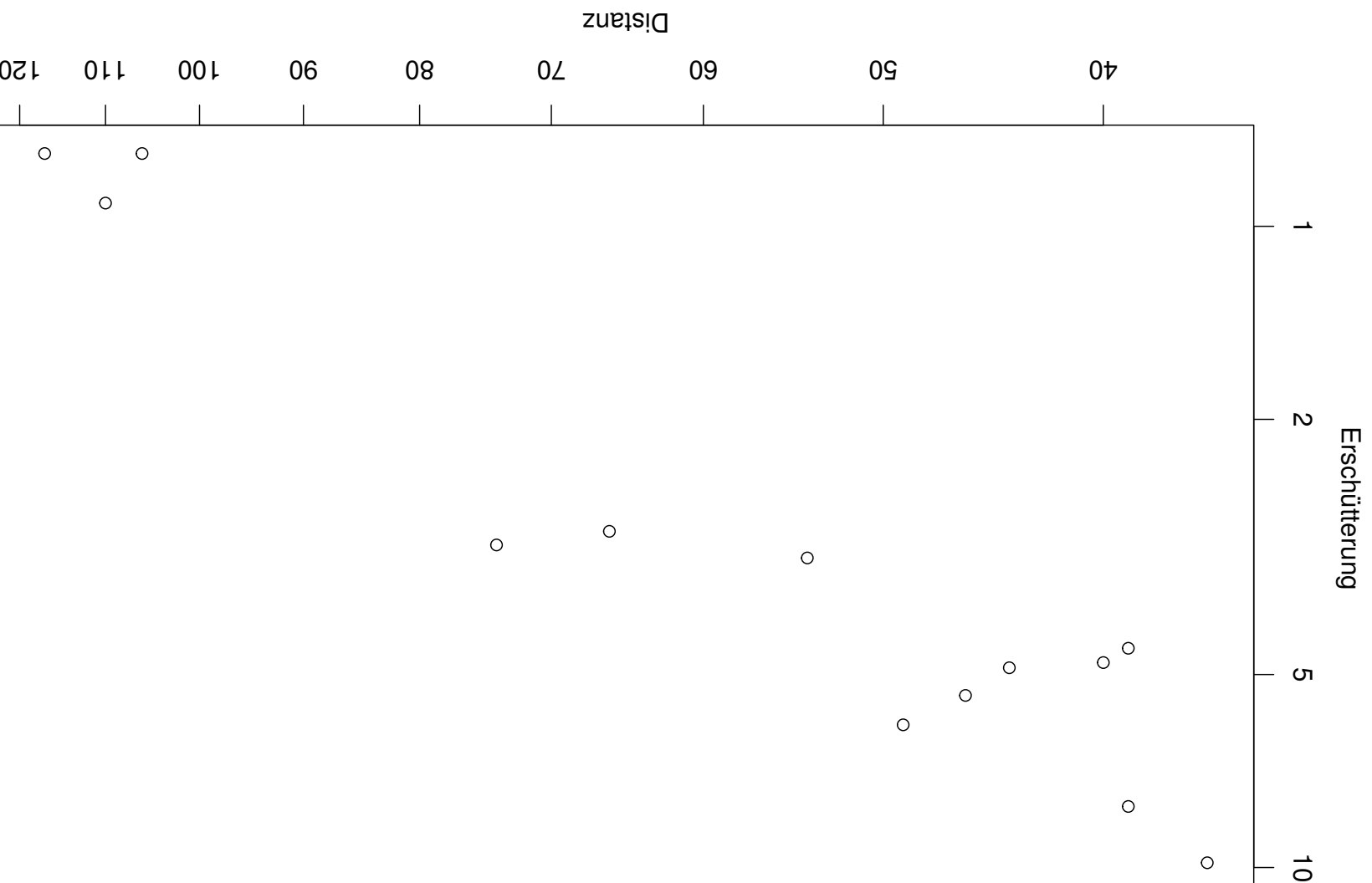
1. **Vorhersage, Prognose, Interpolation.**
Sprengungen: für geg. Distanz und Ladung die Erschütterung „vorher-sagen“.
Obere Grenze?
2. **Schätzung von Parametern.**
Gubrist-Tunnel: Emissionsfaktoren für Lastwagen und für übrige Fahrzeuge.
3. **Bestimmung von Einflussgrößen.**
Antikörper-Produktion, Phase 1: Wichtige Faktoren finden.
Forschungs-Projekte: Von welchen Grössen wird Y eigentlich beein-flusst?

4. **Optimierung**.
Antikörper-Produktion, Phase 2: Optimale Einstellungen.

5. **Eichung**.
Beispiel Lastwagen-Anteil:
Systematische Überschätzung durch Schlaufen-Detektor korrigieren.
Häufig: Messinstrumente.

2 Einfache lineare Regression

2



2 Einfache lineare Regression

2.1 Das Modell

a **Beispiel Sprengungen.**

$\log\langle\text{Erschütterung}\rangle \approx \alpha + \beta \log\langle\text{Distanz}\rangle$

c Erschütterung $\approx \gamma$ Distanz $^\beta$ mit $\gamma = 10^\alpha$.

d Gerade $y = \alpha + \beta x$.

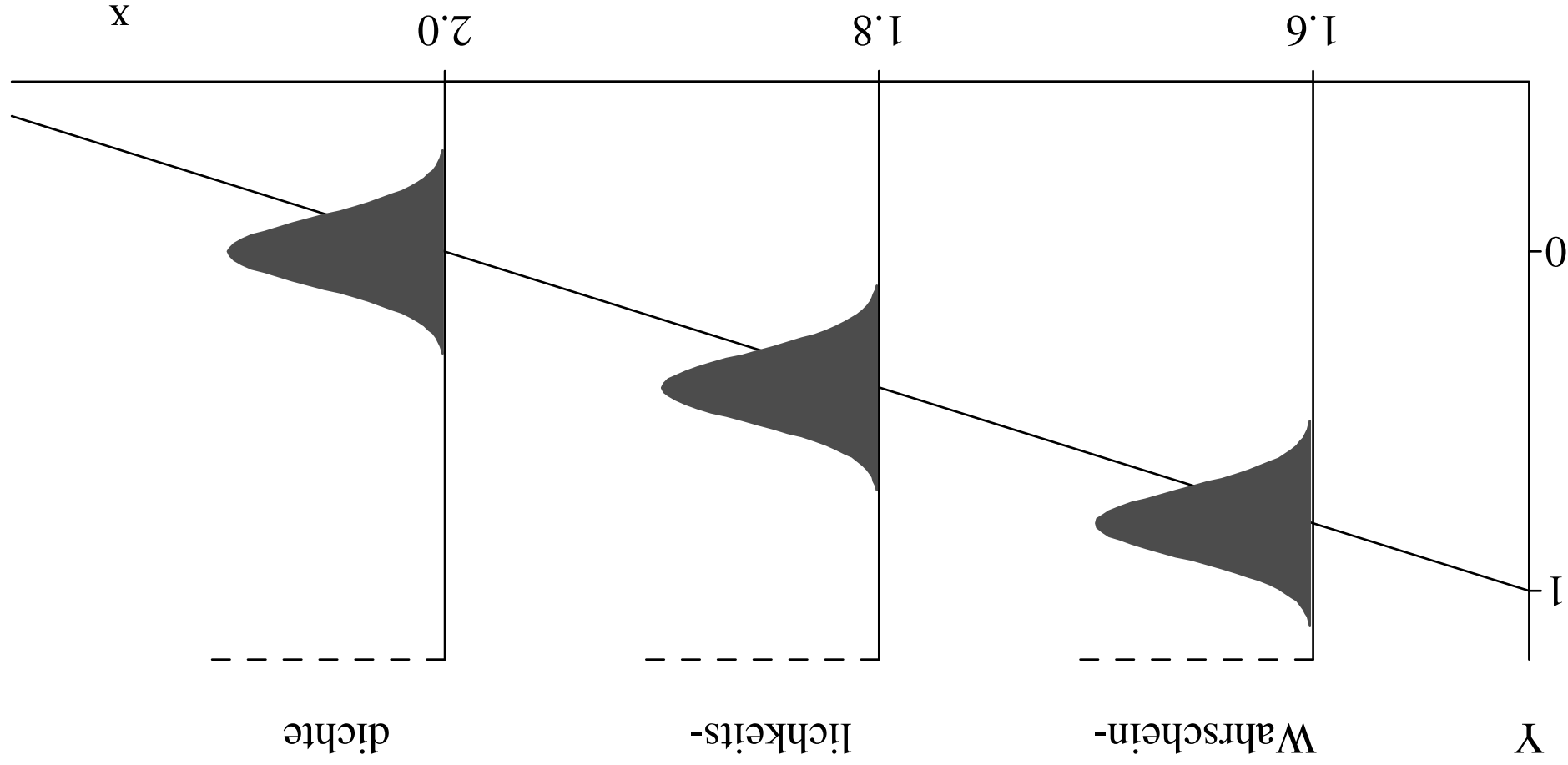
α : Achsenabschnitt, β : Steigung

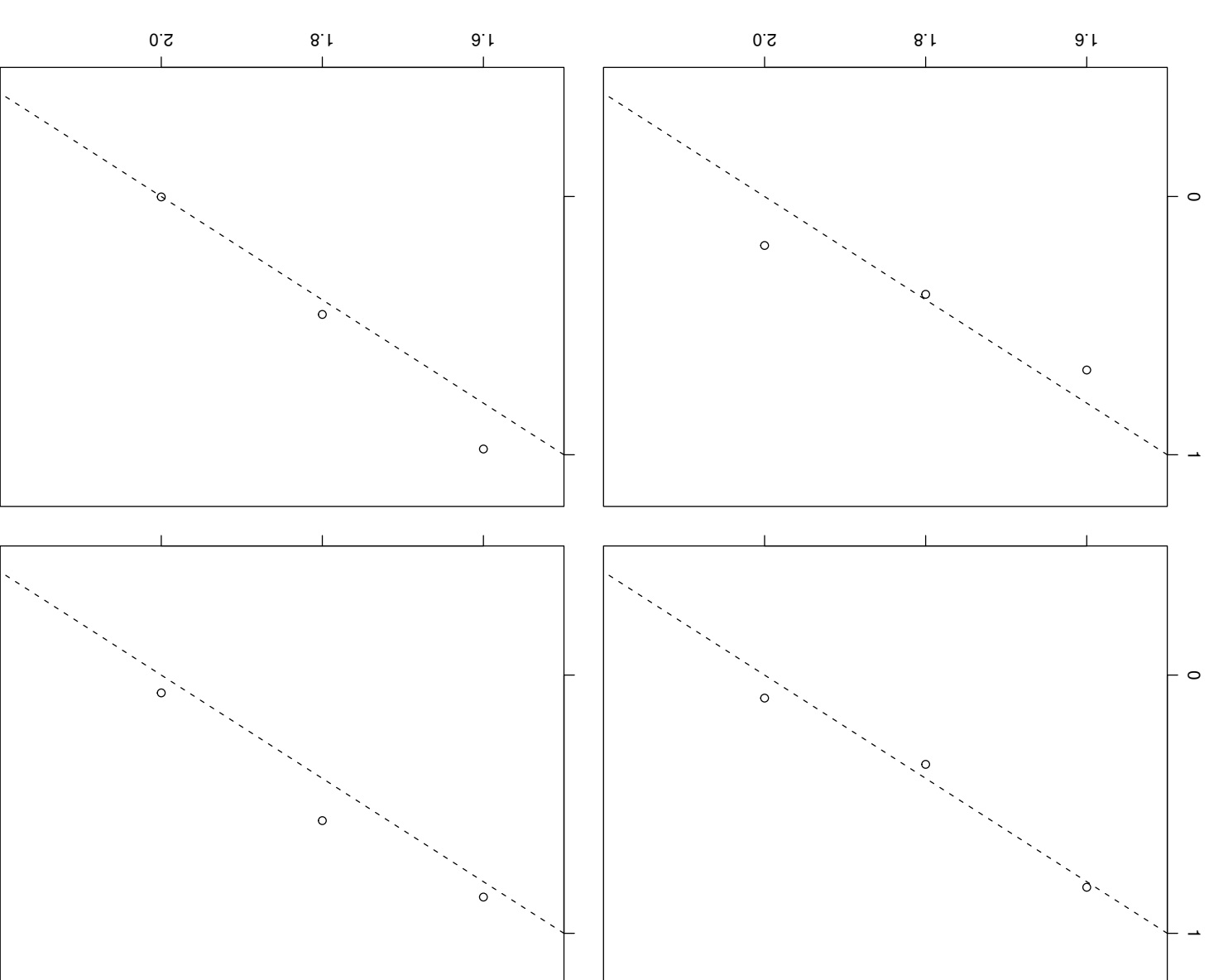
e Im Streudiagramm: Gerade legen!

f Modell: $Y_i = \alpha + \beta x_i + E_i$ $E_i \sim \mathcal{N}\langle 0, \sigma^2 \rangle$

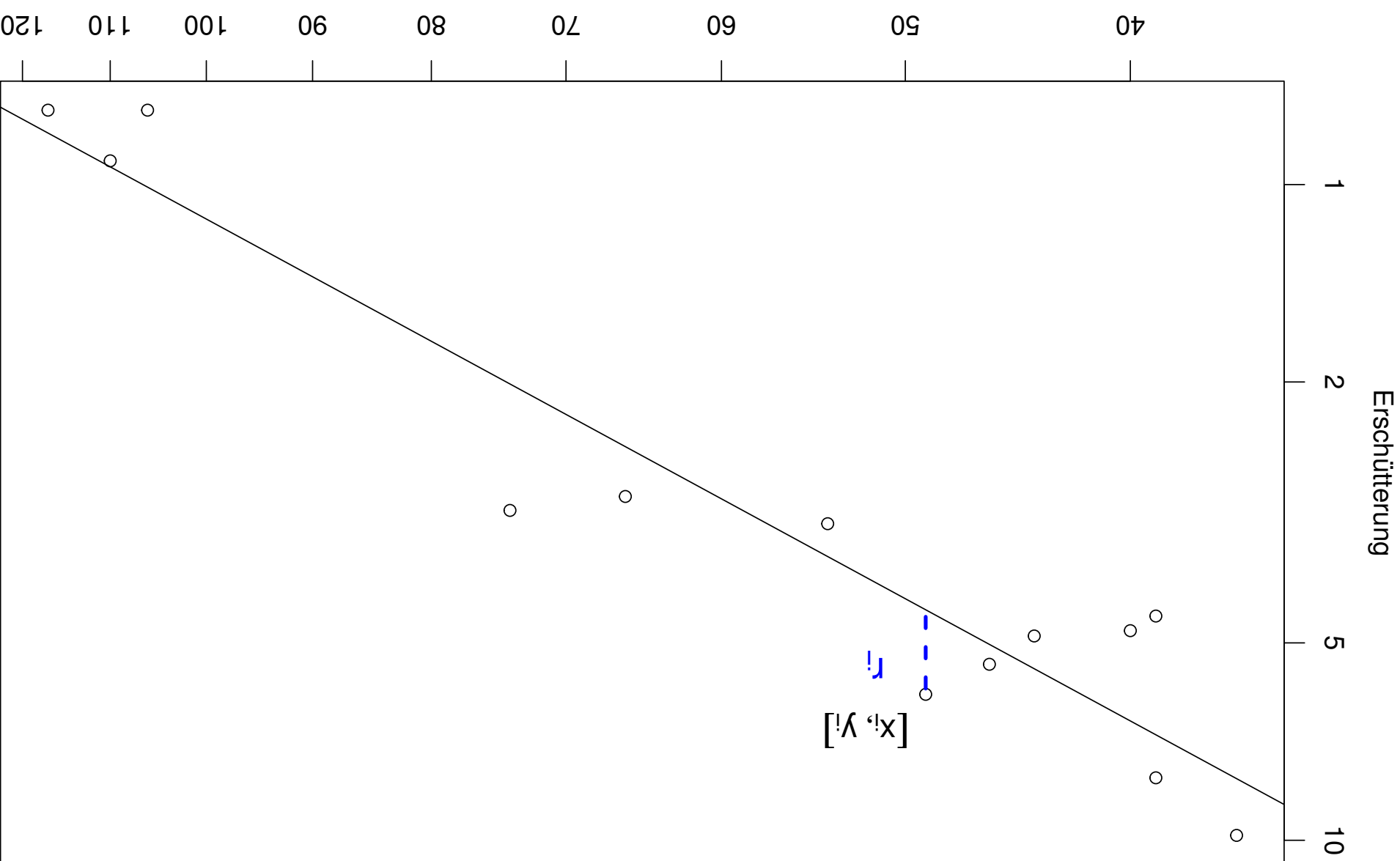
g Die drei Größen α , β und σ sind die **Parameter** des Modells.

h Veranschaulichung des Modells.





2.2 Schätzung der Parameter



c **Kleinste Quadrate:** Die Parameter werden so bestimmt,

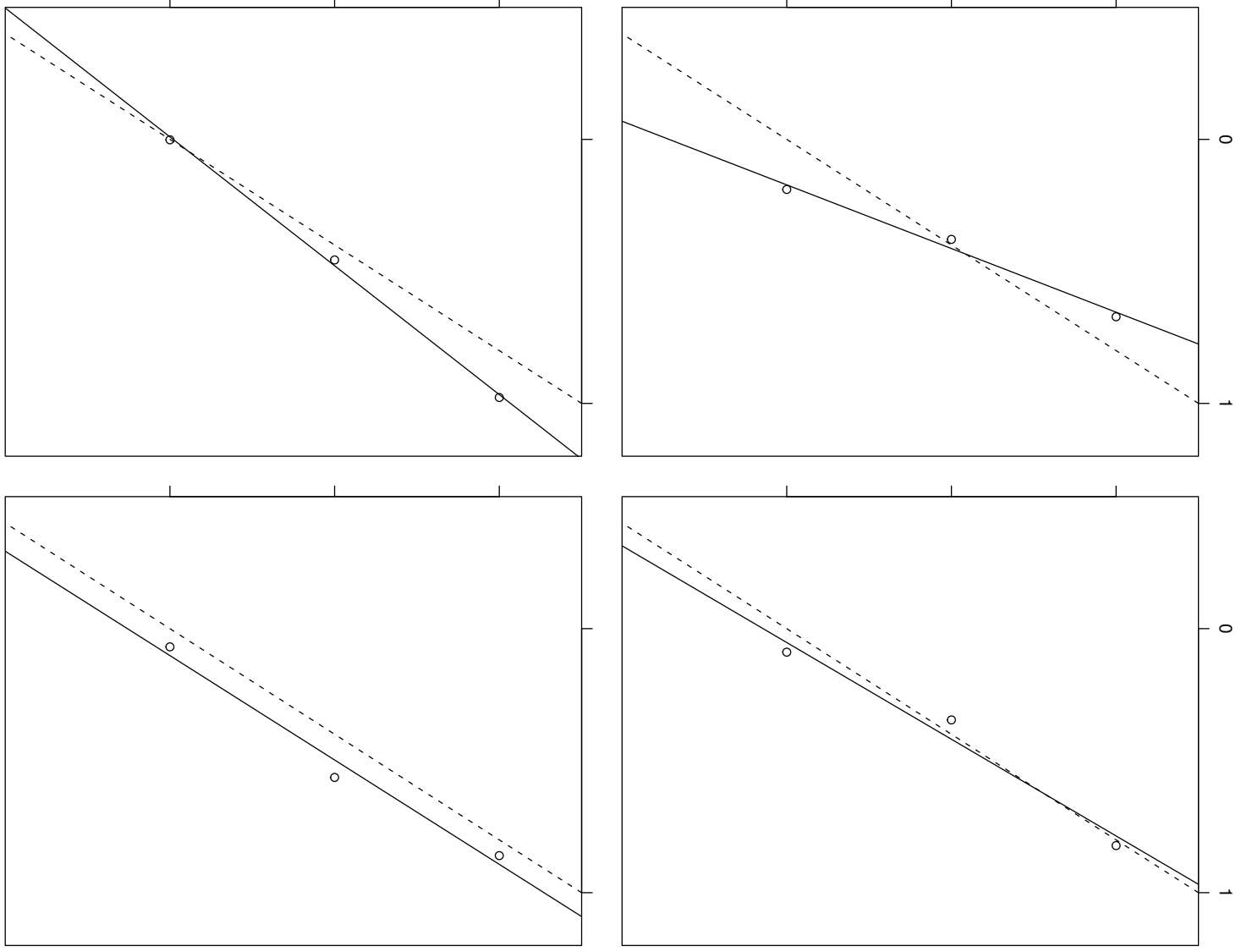
dass die Summe der quadrierten Abweichungen r_i ,

$$\sum_{i=1}^n r_i^2, \quad r_i = y_i - (\alpha + \beta x_i)$$

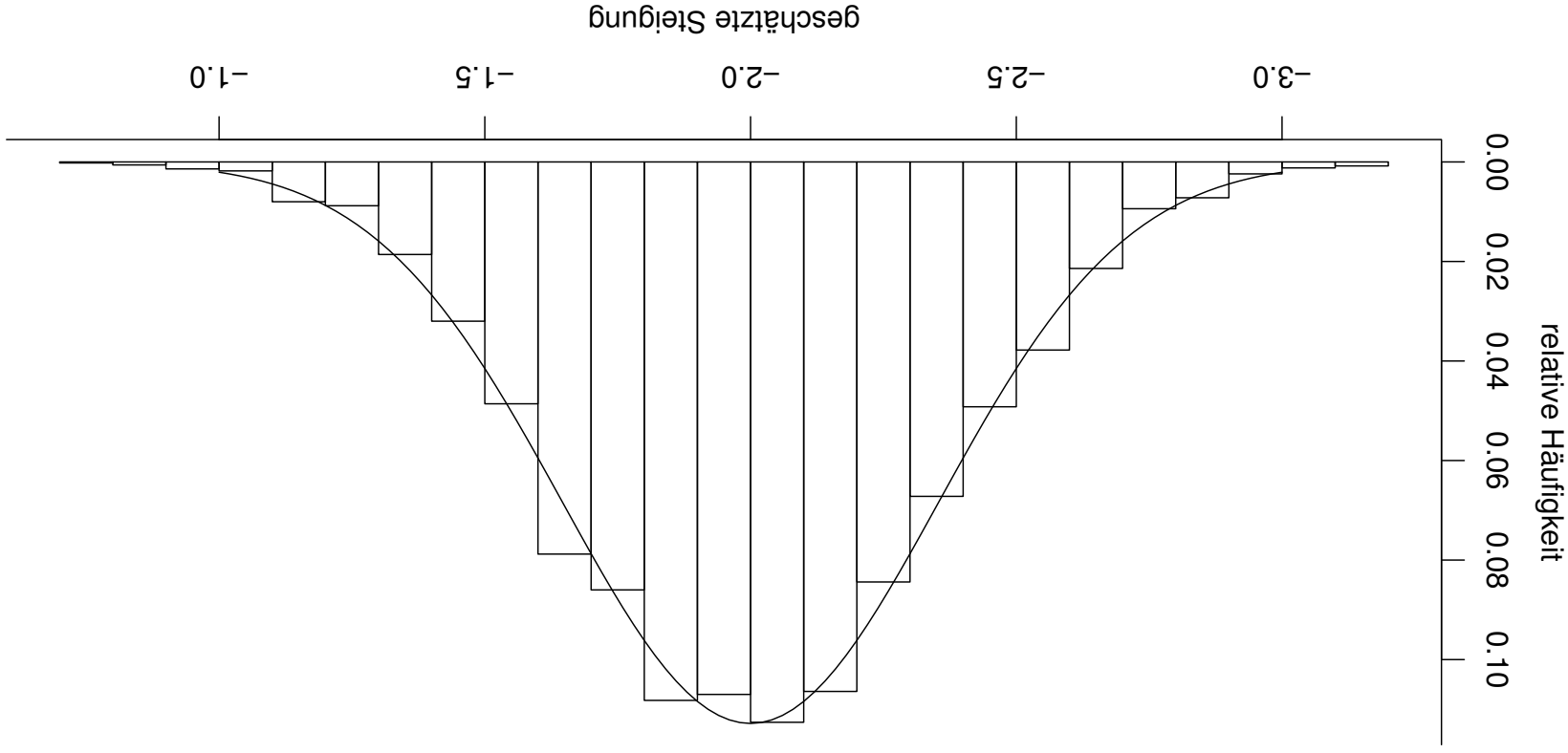
minimal wird. ←

$$\hat{\beta} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad \hat{\alpha} = \bar{y} - \hat{\beta} \bar{x} = \text{Maximum Likelihood}$$

f Schätzungen sind **Zufallsvariable**.
 Bezeichnung: $\hat{\alpha}$, $\hat{\beta}$.
 Zufallsvariable streuen.



h Simulierte Verteilung



! Theoretische Verteilungen:

$$\text{SSQ}(X) = \sum_{i=1}^n (x_i - \bar{x})^2$$

$$\hat{\alpha} \sim \mathcal{N}\left(\alpha, \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\text{SSQ}(X)}\right)\right)$$

$$\hat{\beta} \sim \mathcal{N}\left(\beta, \sigma^2 / \text{SSQ}(X)\right)$$

k Eigenschaften von Schätzungen:

Erwartungstreue?

l Mittlerer quadratischer Fehler, Varianz der Schätzung?

n Die Kleinste-Quadrate-Schätzungen $\hat{\alpha}$ und $\hat{\beta}$ sind

- erwartungstreu & normalvert. mit den angeg. Varianzen,

- die **besten Schätzungen**

sofern die Zufallsfehler unabhängig sind und

alle die gleiche Normalverteilung $\mathcal{N}(0, \sigma^2)$ haben.

Sonst sind andere Schätzungen besser!

- o Schätzung von $\sigma^2 = \text{var}\langle E_i \rangle$.

$$E_i \approx R_i = Y_i - (\hat{\alpha} + \hat{\beta}x_i) = Y_i - \hat{y}_i$$

Residuen R_i , angepasste Werte (fitted values, fit) \hat{y}_i

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n R_i^2$$

Drei Grundfragen der Schliessenden Statistik

1. Welcher Wert ist für den (jeden) Parameter am plausibelsten? ← Schätzung

2. Ist ein bestimmter Wert plausibel? ← Test.

3. Welche Werte sind insgesamt plausibel? ← Vertrauens- oder Konfidenzintervall

2.3 Tests und Vertrauensintervalle

b **Nullhypothese** $H_0: \beta = -2$... oder vollständig:

Die Beobachtungen folgen dem Modell der einfachen linearen Regression mit $\beta = -2$ und beliebigem α und σ .

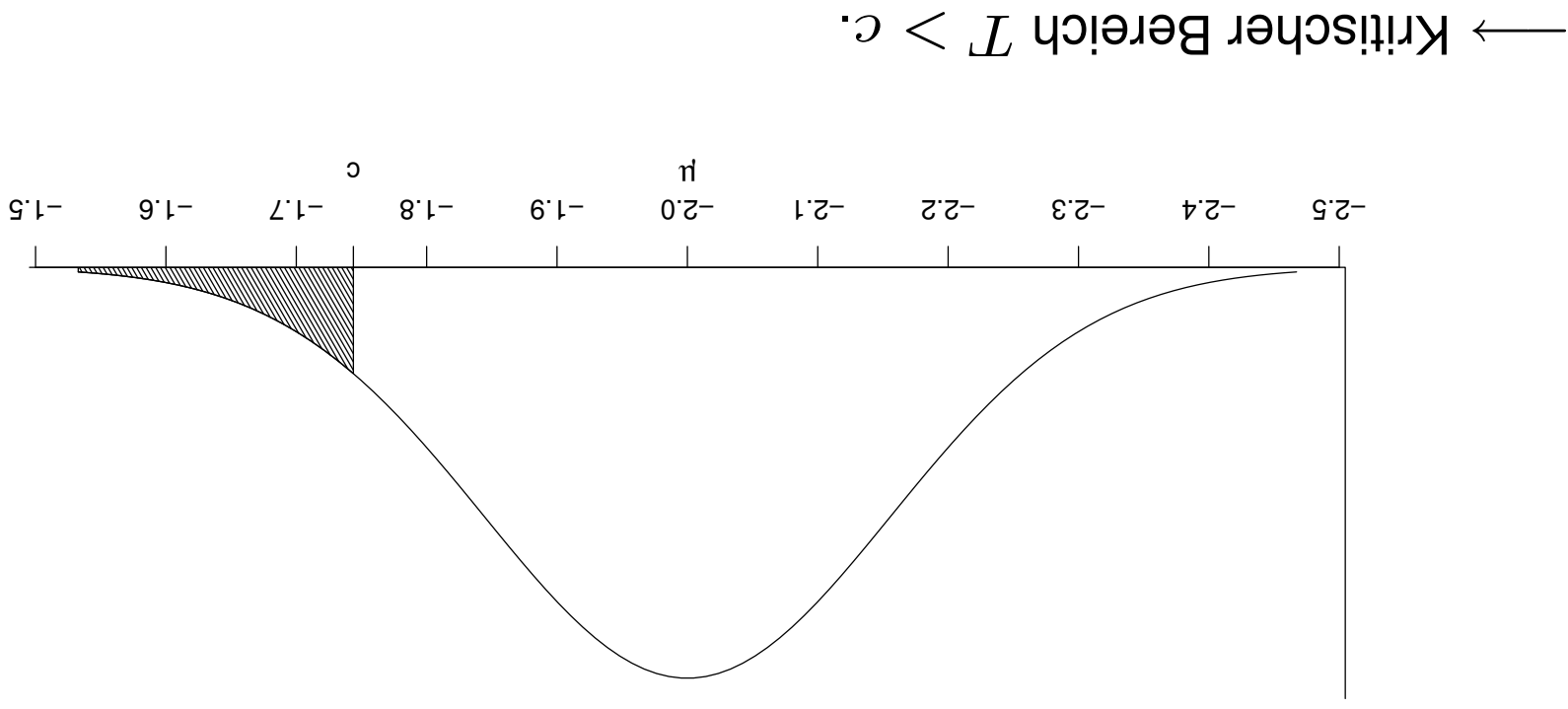
Alternativen H_A : einseitig $\beta > -2$ oder zweiseitig $\beta \neq -2$.

Test-Statistik: = Schätzung $\hat{\beta}$

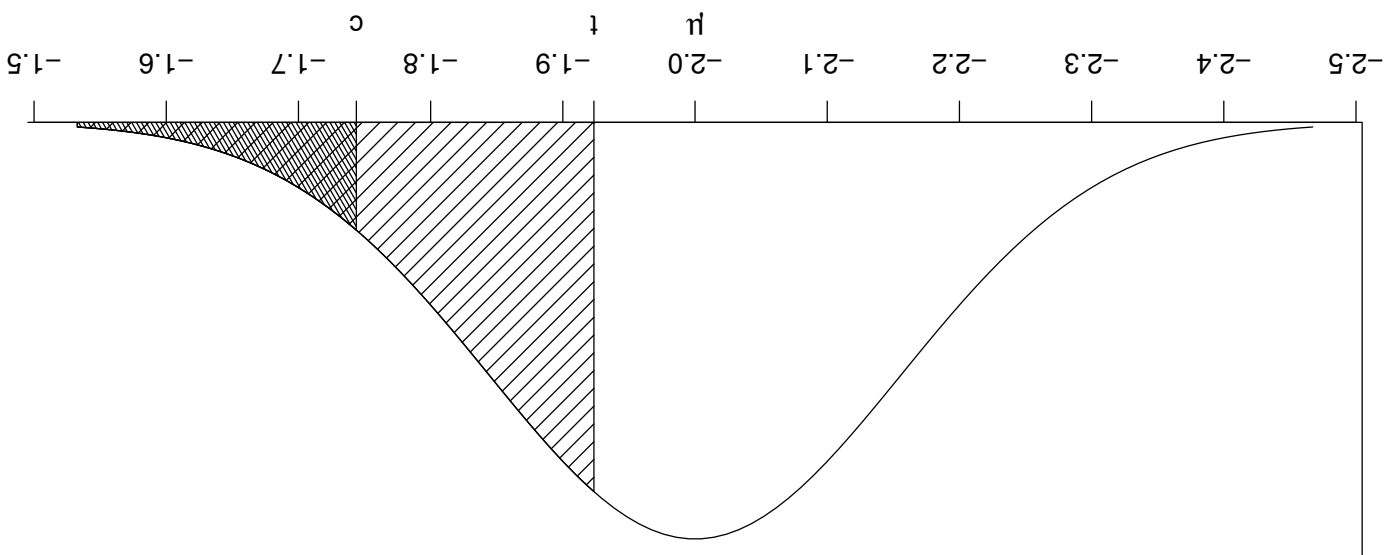
$$(\hat{\beta} - \beta_0) \sim \mathcal{N}(0, \text{se}(\hat{\beta})) \quad \text{se}(\hat{\beta})^2 = \hat{\sigma}^2 / \text{SSQ}(X)$$

$$T = (\hat{\beta} - \beta_0) / \text{se}(\hat{\beta}) \sim \text{t-Vert.}_{n-2}$$

Verteilung der Test-Statistik unter der Nullhypothese
 $T = (\hat{\beta} - \beta_0) / \text{se}(\hat{\beta}) \sim t\text{-Vert. } n-2$



$$c \quad T = (\hat{\beta} - \beta_0) / \text{se}(\hat{\beta}) \sim t\text{-Vert.}_{n-2} \longrightarrow \text{P-Wert.}$$



H_0 meistens: $\beta = 0 \longrightarrow$ P-Wert aus Programm.

d Computer-Output.

Regression Analysis - Linear model: $Y = a + bX$

Dependent variable: log10(ersch)		Independent variable: log10(dist)	
Parameter	Estimate	Standard Error	T Value
Intercept	$\hat{\alpha} = 3.8996$	$SE(\alpha) = 0.3156$	$T(\alpha) = 12.36$
Slope	$\hat{\beta} = -1.9235$	$SE(\beta) = 0.1783$	$T(\beta) = -10.79$

R-squared = $0.9136 = r^2_{XY}$

Std.dev. of Error = $\hat{\sigma} = 0.1145$ on $n - 2 = 11$ degrees of freedom

F-statistic: 116.4 on 1 and 11 degrees of freedom, the p-value is 3.448e-07

f **Vertrauensintervall**: Der Annahmebereich war

$$-q_{0.975}^{t_{n-2}} \leq T \leq q_{0.975}^{t_{n-2}}$$

$$-q_{0.975}^{t_{n-2}} \widehat{se}(\hat{\beta}) \leq \hat{\beta} - \beta \leq q_{0.975}^{t_{n-2}} \widehat{se}(\hat{\beta})$$

Die linke Ungleichung ergibt

$$\hat{\beta} \leq \beta + q_{0.975}^{t_{n-2}} \widehat{se}(\hat{\beta})$$

analog: untere Grenze für β . – Zusammen:

$$\hat{\beta} - q_{0.975}^{t_{n-2}} \widehat{se}(\hat{\beta}) \leq \beta \leq \hat{\beta} + q_{0.975}^{t_{n-2}} \widehat{se}(\hat{\beta})$$

$$\hat{\beta} \pm q_{0.975}^{t_{n-2}} \widehat{se}(\hat{\beta})$$

g Vertrauensintervall für β im Beispiel:

Regression Analysis - Linear model: $Y = a + bX$

Dependent variable: log10(ersch)		Independent variable: log10(dist)	
Parameter	Estimate	Standard Error	T Value Prob. Level
Intercept	$\hat{\alpha} = 3.8996$	$SE(\beta) = 0.3156$	$T(\alpha) = 12.36$ 0
Slope	$\hat{\beta} = -1.9235$	$SE(\beta) = 0.1783$	$T(\beta) = -10.79$ 0

R-squared = $0.9136 = r^2_{XY}$

Std.dev. of Error = $\hat{\sigma} = 0.1145$ on $n - 2 = 11$ degrees of freedom

F-statistic: 116.4 on 1 and 11 degrees of freedom, the p-value is 3.448e-07

$-1.9235 \pm 2.20 \cdot 0.178 = [-2.32, -1.53]$

2.4 Vertrauens- und Prognose-Bereiche

a Wie gross ist die Erschütterung bei Distanz 50m?

Erwartungswert der Erschütterung bei 50m Distanz?

Vertrauensintervall dafür?

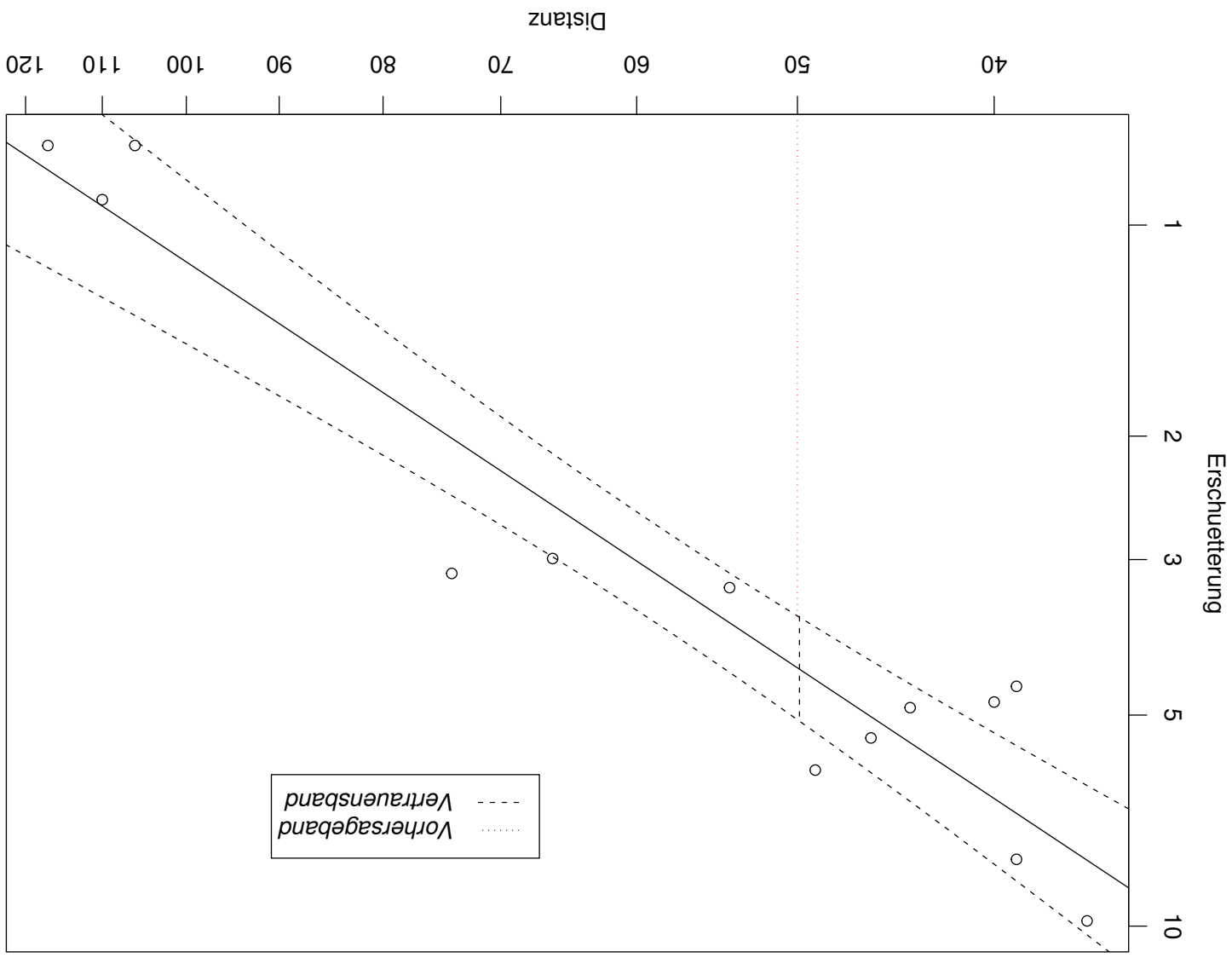
b Testgrösse für $h(x_0) = y_0$:

$$T = \frac{\widehat{y}_0 - y_0}{se(y_0)}, \quad se(y_0) = \widehat{\sigma} \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS(X)}}$$

Verteilung: t -Verteilung mit $n - 2$ Freiheitsgraden

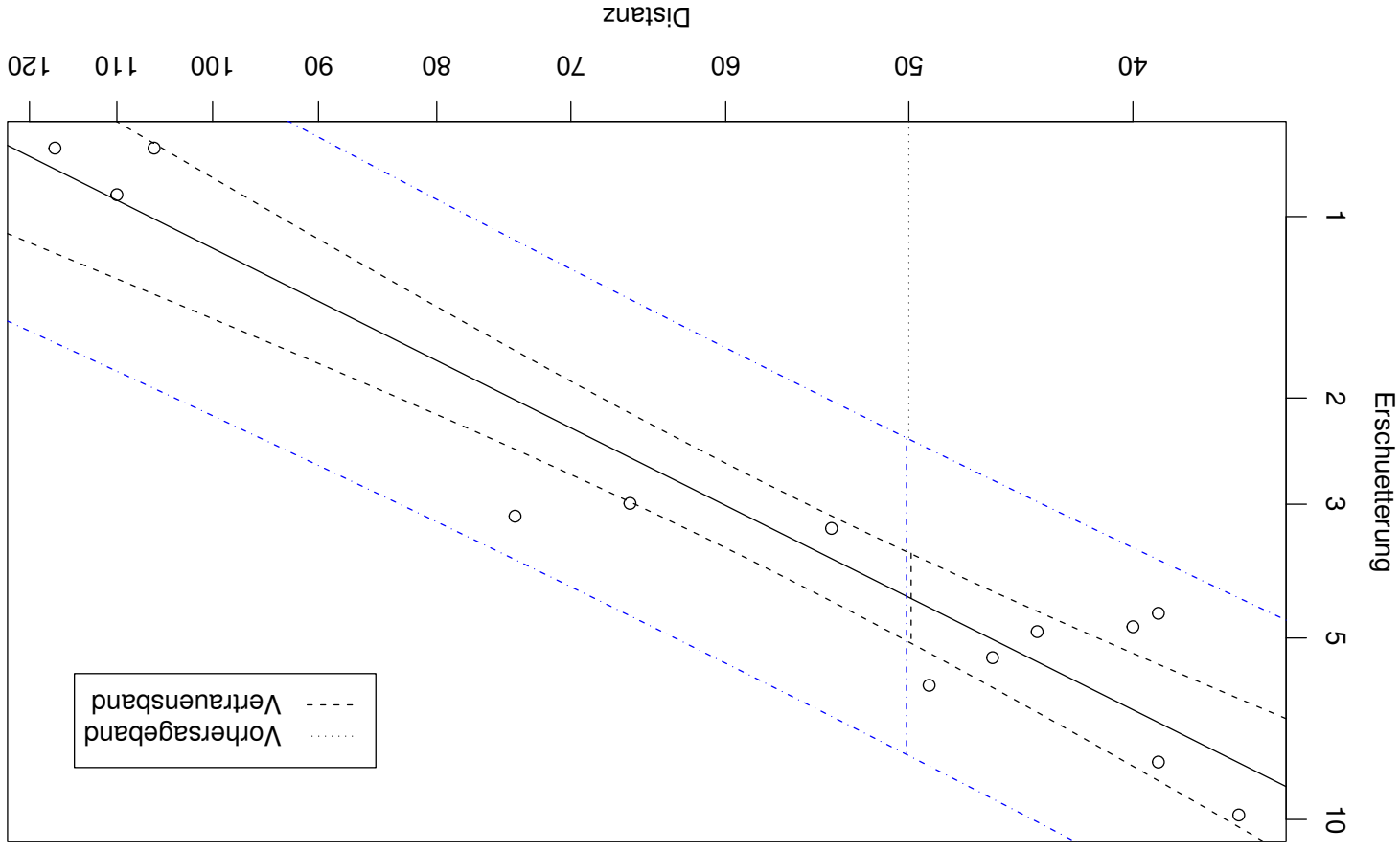
Vertrauensintervall:

$$(\widehat{\alpha} + \widehat{\beta}x_0) \pm t_{n-2}^{0.975} se(y_0)$$



d Das „Vertrauensband“ gibt an, wo die **idealen Funktionswerte** $h(x) = \mathcal{E}\langle Y \rangle$ bei gegebenen x liegen.

In welchem Bereich liegen **künftige Beobachtungen** (zu geg. x_0)? ... kein Vertrauensintervall, sondern ein „**Vorhersage-Intervall**“.



$$R_0 = Y_0 - (\hat{\alpha} + \hat{\beta}x_0) = \left(Y_0 - (\alpha + \beta x_0) \right) - \left((\hat{\alpha} + \hat{\beta}x_0) - (\alpha + \beta x_0) \right) = \left(\dots \right) - \left(\dots \right) : \text{beide } \sim \mathcal{N}, \text{ unabhängig.} \longrightarrow R_0 \sim \mathcal{N}(0, \dots)$$

f Interpretation nicht einfach.

→ Toleranz-Intervall.

Merkpunkte

Einfache Regression

1. Regression ist die am meisten verbreitete Methodik der Statistik

2. Die **einfache** lineare Regression ist eine einfache Anwendung

des **Grundschemas**:

• **Modell**: $Y_i = \alpha + \beta x_i + E_i$, $E_i \sim \mathcal{N}\langle 0, \sigma^2 \rangle$, unabhängig

• **Schätzung**: Maximum likelihood führt auf kleinste Quadrate

• **Test**: Schätzung als Teststatistik, standardisiert mit

geschätzter Streuung \rightarrow t-Test

• **Vertrauensintervall**: aus dem t-Test

3 Multiple lineare Regression

3.1 Modell und Statistik

- a Zusammenhang zwischen einer Zielgröße Y und mehreren Ausgangsgrößen $X^{(1)}, X^{(2)}, \dots, X^{(m)}$
- $$Y_i = \beta_0 + \beta_1 x_i^{(1)} + \beta_2 x_i^{(2)} + \dots + \beta_m x_i^{(m)} + E_i$$
- Parameter: $\beta_0, \beta_1, \beta_2, \dots, \beta_m, \sigma^2$.
- „abhängige“ Variable = Zielvariable
- „unabhängige“ Variable = Ausgangs-, erklärende Variable
- b Beispiel Sprengungen: $Y = \log_{10}$ 〈Erschütterung〉, $X^{(1)} = \log_{10}$ 〈Distanz〉 und $X^{(2)} = \log_{10}$ 〈Ladung〉.

c **Schätzung**, Tests, Vertrauensintervalle:

Kleinste Quadrate. Theorie etwas später.

d Computer-Ergebnis

Coefficients:				
(Intercept)	2.8323	0.2229	12.71	0.000
log10(dist)	-1.5107	0.1111	-13.59	0.000
log10(ladung)	0.8083	0.3042	2.66	0.011
Value	Std. Error	t value	Pr(> t)	

Residual standard error: 0.1529 on 45 degrees of freedom
 Multiple R-Squared: 0.8048
 F-statistic: 92.79 on 2 and 45 degrees of freedom
 p-value 1.11e-16

e **Tests**: Welche Fragen sind zu stellen?

Frage **A**.

Beeinflusst die **Gesamtheit der Ausgangsgrößen** die Zielgröße?

← „F-Test“

Coefficients:

...

Residual standard error: 0.1529 on 45 degrees of freedom

Multiple R-Squared: 0.8048

F-statistic: 92.79 on 2 and 45 degrees of freedom

p-value 1.11e-16

Coefficients:

...

Residual standard error: 0.1529 on 45 degrees of freedom

Multiple R-Squared: 0.8048

F-statistic: 92.79 on 2 and 45 degrees of freedom

p-value 1.11e-16

„Multiple R-Squared“ ist das Quadrat der

multiplen Korrelation = $\text{corr}(Y_i, \text{angepasste Werte } \hat{y}_i)$

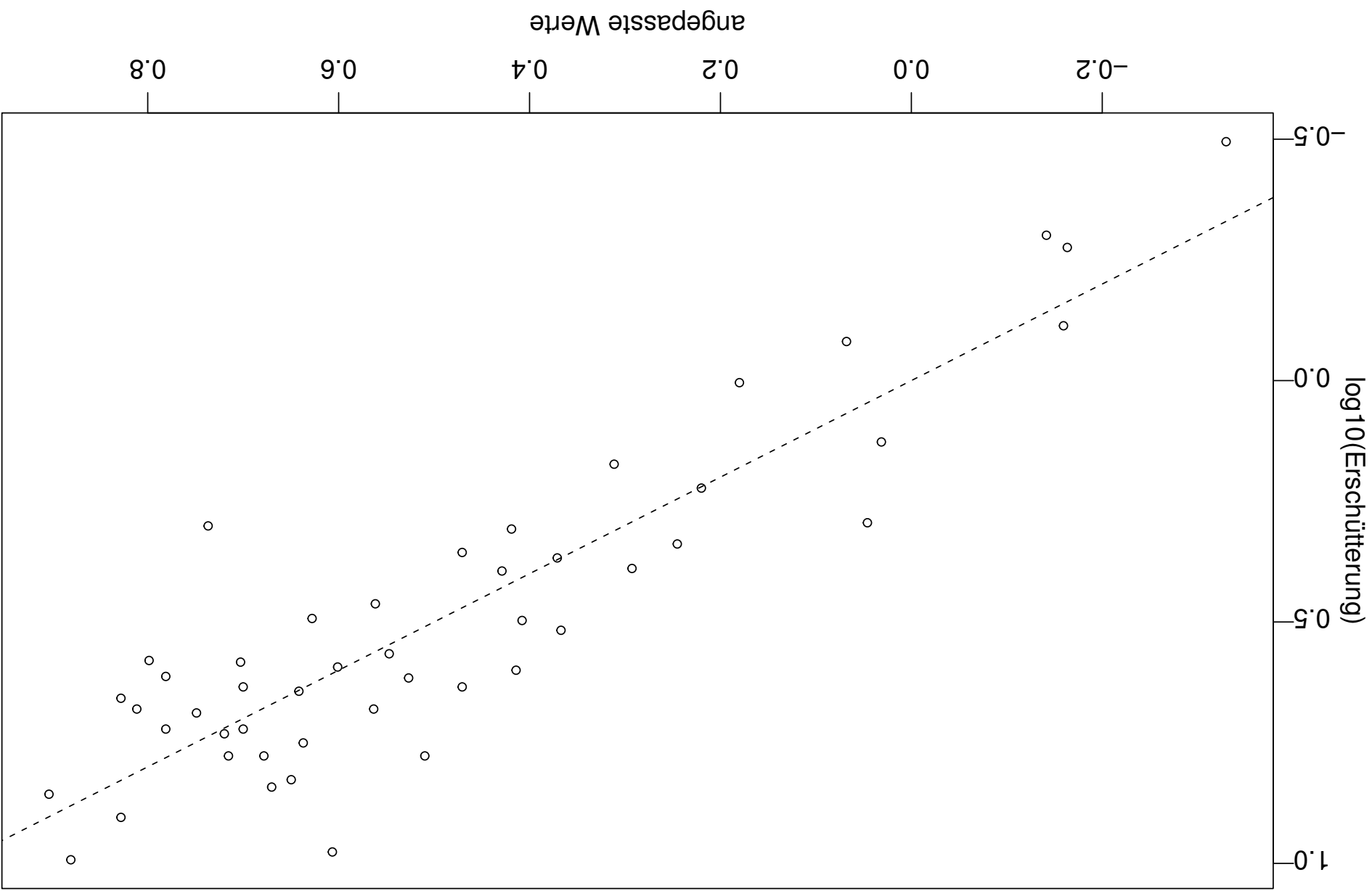
$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i^{(1)} + \hat{\beta}_2 x_i^{(2)} + \dots + \hat{\beta}_m x_i^{(m)}$$

 $R^2 =$ **Bestimmtheitsmass**,

misst den Anteil der erklärten Streuung

an der Streuung der Y -Werte,

$$R^2 = 1 - \text{SSQ}(E) / \text{SSQ}(Y).$$



Frage **B.**

Einfluss der einzelnen Variablen $X^{(j)}$?

Coefficients:				
	Value	Std. Error	t value	$\Pr(> t)$
(Intercept)	2.8323	0.2229	12.71	0.000
$\log_{10}(\text{dist})$	-1.5107	0.1111	-13.59	0.000
$\log_{10}(\text{ladung})$	0.8083	0.3042	2.66	0.011

Residual standard error: 0.1529 on 45 degrees of freedom

Multiple R-Squared: 0.8048

F-statistic: 92.79 on 2 and 45 degrees of freedom

p-value 1.11e-16

Der t-Wert und der P-Wert in derjenigen Zeile, die $X^{(j)}$ entspricht, prüft, ob die Variable $X^{(j)}$ **aus dem Modell weggelassen** werden kann: Nullhypothese $\beta_j = 0$.

h **Vertrauensintervall** für β_j : $\hat{\beta}_j \pm t_{(0.975)}^{n-2} \text{se}_{(\beta_j)}$

Coefficients:				
	Value	Std. Error	t value	Pr(> t)
(Intercept)	2.8323	0.2229	12.71	0.000
log10(dist)	-1.5107	0.1111	-13.59	0.000
log10(ladung)	0.8083	0.3042	2.66	0.011

Residual standard error: 0.1529 on 45 degrees of freedom

Multiple R-Squared: 0.8048

F-statistic: 92.79 on 2 and 45 degrees of freedom

p-value 1.11e-16

! Beispiel: $-1.5107 \pm 2.014 \cdot 0.1111$

$= -1.5107 \pm 0.2237 = [1.2869, 1.7345]$.

! **t-Quotient.** Kolonne t in üblichen Tabellen: Überflüssig?

Mass für Signifikanz, anders als P-Wert

$$\tilde{T}_j = \frac{\hat{\beta}_j}{\text{se}(\hat{\beta}_j) \cdot q_{0.975}^{(t_k)}} = T / q_{0.975}^{(t_k)} \cdot$$

$\tilde{T}_j > 1$ bedeutet signifikanter Koeffizient

Coefficients:		coef	stcoef	signif	R2.x	df	p.value
(Intercept)	2.832	0.000	6.31	NA	1	1	0.000
log10(dist)	-1.511	-0.903	-6.75	0.01659	1	1	0.000
log10(ladung)	0.808	0.176	1.32	0.01659	1	1	0.011

St.dev. of Error = 0.1529 on 45 degrees of freedom
 Multiple R-Squared: 0.8048
 F-statistic: 92.79 on 2 and 45 degrees of freedom
 p-value 1.11e-16

Vertrauensintervall: Bis auf Faktor $\tilde{T}_j \neq 1$.

$$\hat{\beta}_j \cdot \frac{\tilde{T}_j}{T_j} \cdot (T_j \neq 1) = \hat{\beta}_j \cdot (1 \neq 1/\tilde{T}_j) \cdot$$

k **Standardisierte Koeffizienten.**

$$\hat{\beta}_j^* = \hat{\beta}_j \cdot \text{sd} \langle X^{(j)} \rangle / \text{sd} \langle Y \rangle \cdot$$

Einfache Regression: $\hat{\beta}_j^* =$ Korrelation.

Allg: Um wie viel verändert sich Y , gemessen in $\text{sd} \langle Y \rangle$ -Einh.,

wenn sich $X^{(j)}$ um eine $\text{sd} \langle X^{(j)} \rangle$ verändert?

→ Vergleiche der Einflussstärke von versch. Ausgangsgrößen.

- | **Kollinearitätsmass.** $R^2 \cdot x$ Bestimmtheitsmass für Regression von $X^{(j)}$ als Zielgrösse auf alle anderen Regressoren.
Soll niedrig sein, sonst sind Koeffizienten schlecht bestimmt.
Siehe später.

3.2 Vielfalt der Fragestellungen

a Im Modell der multiplen Regression werden

keine Annahmen über die X -Variablen gemacht. Beliebig:

- Datentyp: stetig, diskret, zweiwertig, später nominal.
- Verteilung der einzelnen Variablen: keine, Nicht zufällig.
- Gemeinsame Verteilung der Variablen: keine, Nicht zufällig.
Keine Unabhängigkeit vorausgesetzt!
Ein $X^{(j)}$ darf eine deterministische (nicht-lineare) Funktion einer anderen oder mehrerer anderer sein.

c **Binäre** Ausgangs-Variablen, $Y_i = \beta_0 + \beta_1 x_i + E_i$

→ $Y_i = \beta_0 + E_i$ für $x_i = 0$,

$Y_i = \beta_0 + \beta_1 + E_i$ für $x_i = 1$.

$\beta_0 = \mu_0 =$ Erwartungswert für Gr. $x_i = 0$,

$\beta_0 + \beta_1 = \mu_1 =$ Erwartungswert für Gr. $x_i = 1$.

→ Zwei-Gruppen-Problem = Spezialfall der (einfachen) Regression.

d **Beispiel Sprengungen**: Betrachte nur 2 Messstellen.

$$Y_i = \beta_0 + \beta_1 x_i^{(1)} + \beta_2 x_i^{(2)} + E_i,$$

$X^{(1)}$: log Distanz, $X^{(2)} = 0$ für eine Messst., = 1 für andere

→ Zwei Geraden $y = \beta_0 + \beta_1 x^{(1)}$, $y = (\beta_0 + \beta_2) + \beta_1 x^{(1)}$
 Gleiche Steigung β_1 , Geraden sind **parallel**.

e 4 Messstellen \rightarrow Indikatorvariable für Gruppen j :

$$x_i^{(j)} = \begin{cases} 1 & \text{falls } i\text{-te Beobachtung aus der } j\text{-ten Gruppe} \\ 0 & \text{sonst} \end{cases}$$

Modell:

$$Y_i = \mu_1 x_i^{(1)} + \mu_2 x_i^{(2)} + \dots + E_i$$

Setzt man $\mu_j = \beta_j$, so steht das multiple Regressionsmodell da,

allerdings ohne Achsenabschnitt β_0 .

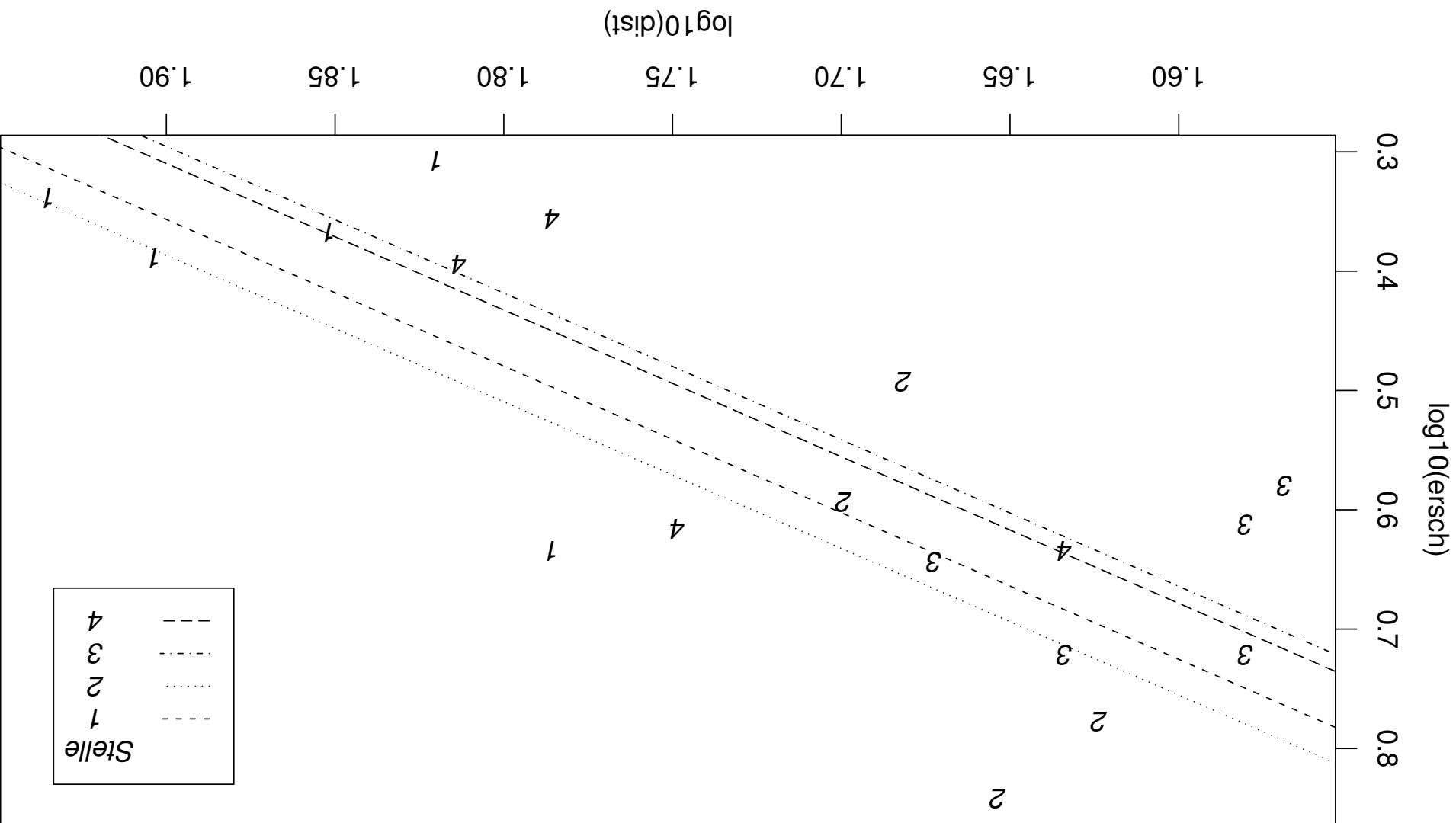
Nominale Ausgangs-Var., l Werte \rightarrow verwandeln in l **dummy variables**.

- f Modell mit Achsenabschnitt: Parameter nicht eindeutig. Lösung:
 – eine „Nebenbedingung“ einführen oder
 – eine Variable weglassen.

g Coefficients:

	Value	Std. Error	t value	Pr(> t)	Signif
(Intercept)	2.51044	0.28215	8.90	0.000	***
log10(dist)	-1.33779	0.14073	-9.51	0.000	***
log10(ladung)	0.69179	0.29666	2.33	0.025	*
St2	0.16430	0.07494	2.19	0.034	*
St3	0.02170	0.06366	0.34	0.735	
St4	0.11080	0.07477	1.48	0.146	

Residual standard error: 0.1468 on 42 degrees of freedom
 Multiple R-Squared: 0.8322
 F-statistic: 41.66 on 5 and 42 degrees of freedom, p-value 3.22e-15



! Frage C:

Unterscheiden sich die Stellen überhaupt
in bezug auf die Zielgröße?

Nullhypothese: Die Koeffizienten der Variablen St 2 bis St 4 sind alle = 0.

! F-Test zum Vergleich von Modellen

k		Df	Sum of Sq	RSS	F Value	Pr(F)
	log10(dist)	1	1.947	2.851	90.4	4.9e-12
	log10(ladung)	1	0.117	1.022	5.44	0.025
	Stelle	3	0.148	1.052	2.283	0.093

Funktion regr

```
Call:
regr(formula = log10(ersch) ~ log10(dist) + log10(ladung) + Stelle,
      data = t.d)
```

Terms:

	coef	stcoef	signif	R2.x	df	p.value
(Intercept)	2.5104436	0.0000000	4.408963	NA	1	0.0000
log10(dist)	-1.3377937	-0.7993097	-4.710628	0.24824540	1	0.0000
log10(ladung)	0.6917912	0.1510358	1.155520	0.02408888	1	0.0246
Stelle	NA	NA	1.322707	0.08883789	3	0.0930

Coefficients for factors:

\$Stelle	1	2	3	4
	0.0000000	0.1643009	0.0216981	0.1107950

St.dev.error: 0.1468 on 42 degrees of freedom

Multiple R^2: 0.8322 Adjusted R-squared: 0.8122

F-statistic: 41.66 on 5 and 42 d.f., p-value: 3.22e-15

- n Einfluss der Stelle: Je eine additive Konstante für jede Stelle.
 Verschiedene Steigungen für verschiedene Stellen?
 —→ Wechselwirkungen.

o Einfacher Fall: **Sind zwei Geraden gleich?**

$$Y_i = \alpha + \beta x_i + \Delta \alpha g_i + \Delta \beta x_i g_i + E_i$$

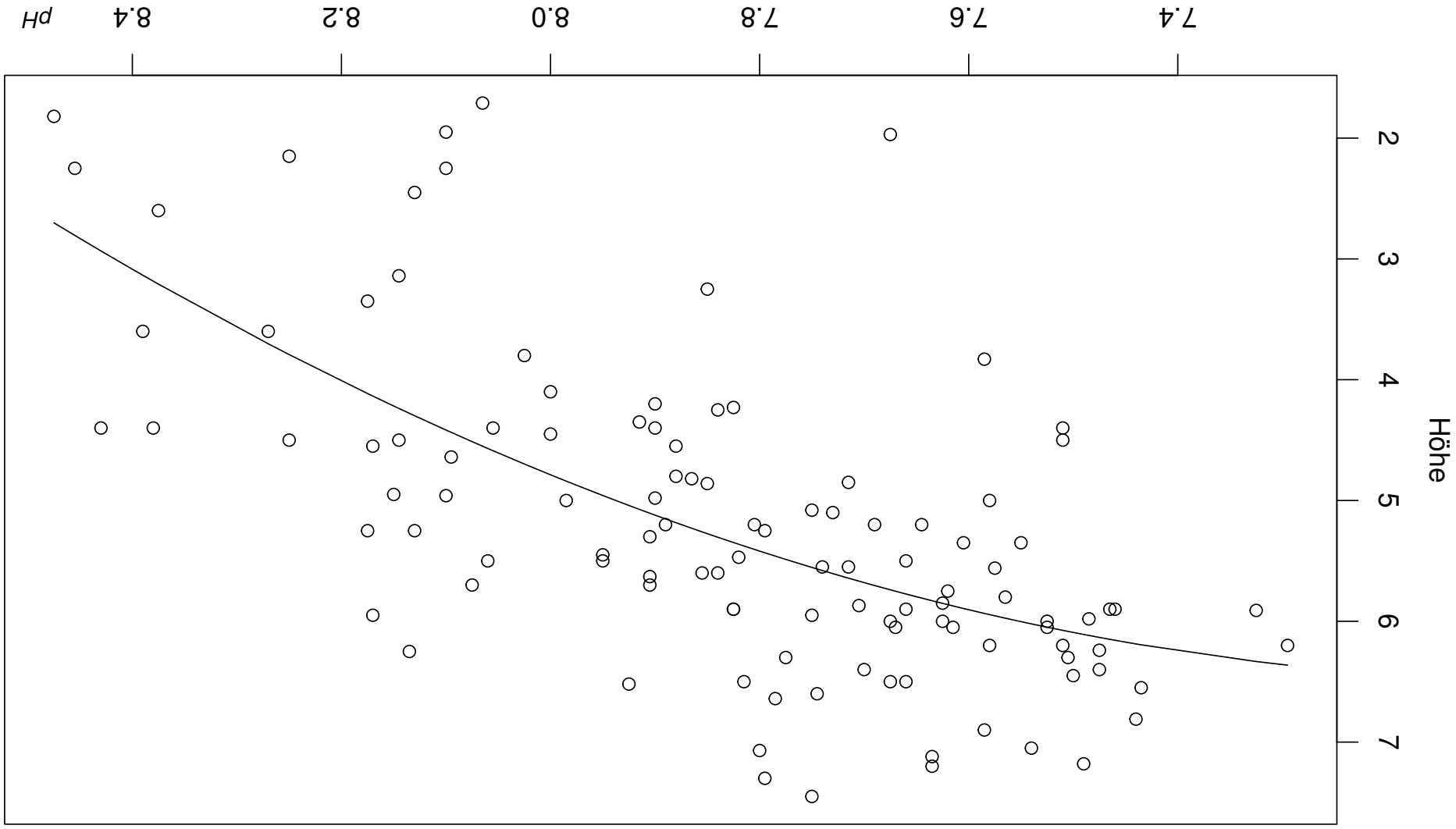
g_i : Gruppenzugehörigkeit

Multiple Regression?

Test für $\Delta \beta = 0$ oder für $\Delta \alpha = 0$, $\Delta \beta = 0$.

p $X^{(2)} = (X^{(1)})^2$ → quadratische Regression

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + E_i.$$



q quadratische \rightarrow **polynomiale Regression**.

Spezialfall der multiplen **linearen** Regression!

Linear in den Koeffizienten!

$$\beta_0 + \beta_1 x_i^{(1)} + \beta_2 x_i^{(2)} + \dots + \beta_m x_i^{(m)}$$

r Optimum der Zielgröße? \rightarrow nicht monotone Regressionsfunktion

Einfachste Fn: Quadratisch.

2 Ausgangs-Variablen: Quadratische Fläche:

$$Y = \beta_0 + \beta_1 x^{(1)} + \beta_2 x^{(2)} + \beta_{11} x^{(1)2} + \beta_{22} x^{(2)2} + \beta_{12} x^{(1)} x^{(2)}$$

β s schätzen, Optimum bestimmen!

s Das Modell der multiplen linearen Regression

ist **sehr flexibel**:

- **Transformation** der X - (und Y -) Variablen:
Linearisieren des Zusammenhangs.
- Vergleich von zwei Gruppen.
- Zwei Geraden. „Wechselwirkungen“.
- Mehrere Gruppen, **nominale Ausgangs-Variablen**.
Vgl. Varianzanalyse.
- **Polynomiale** Regression.

3.3

a Einfluss mehrerer Ausgangsgrößen auf die Zielgröße

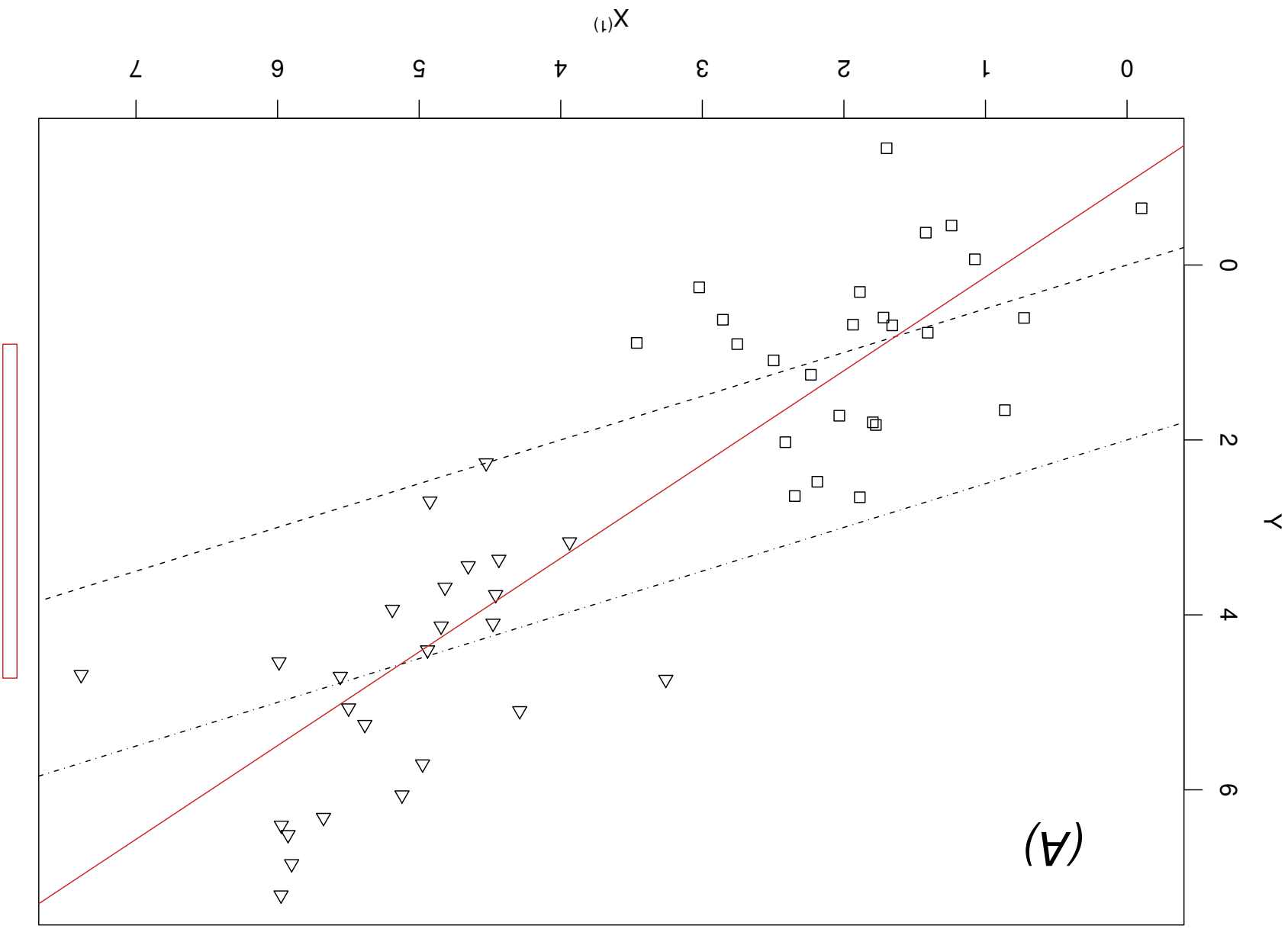
- Multiple Regression

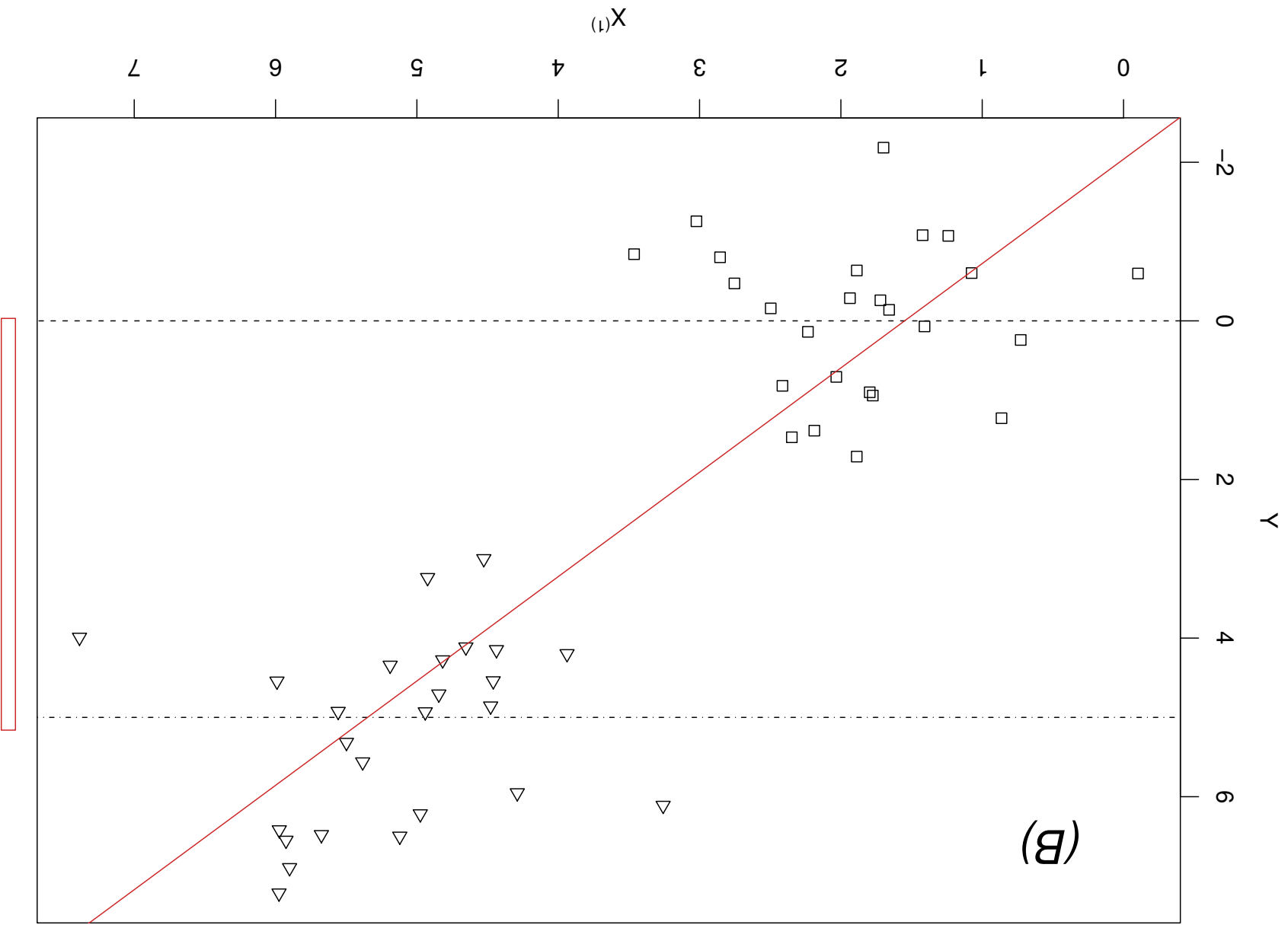
- Mehrere einfache Regressionen: einfacher zu verstehen!

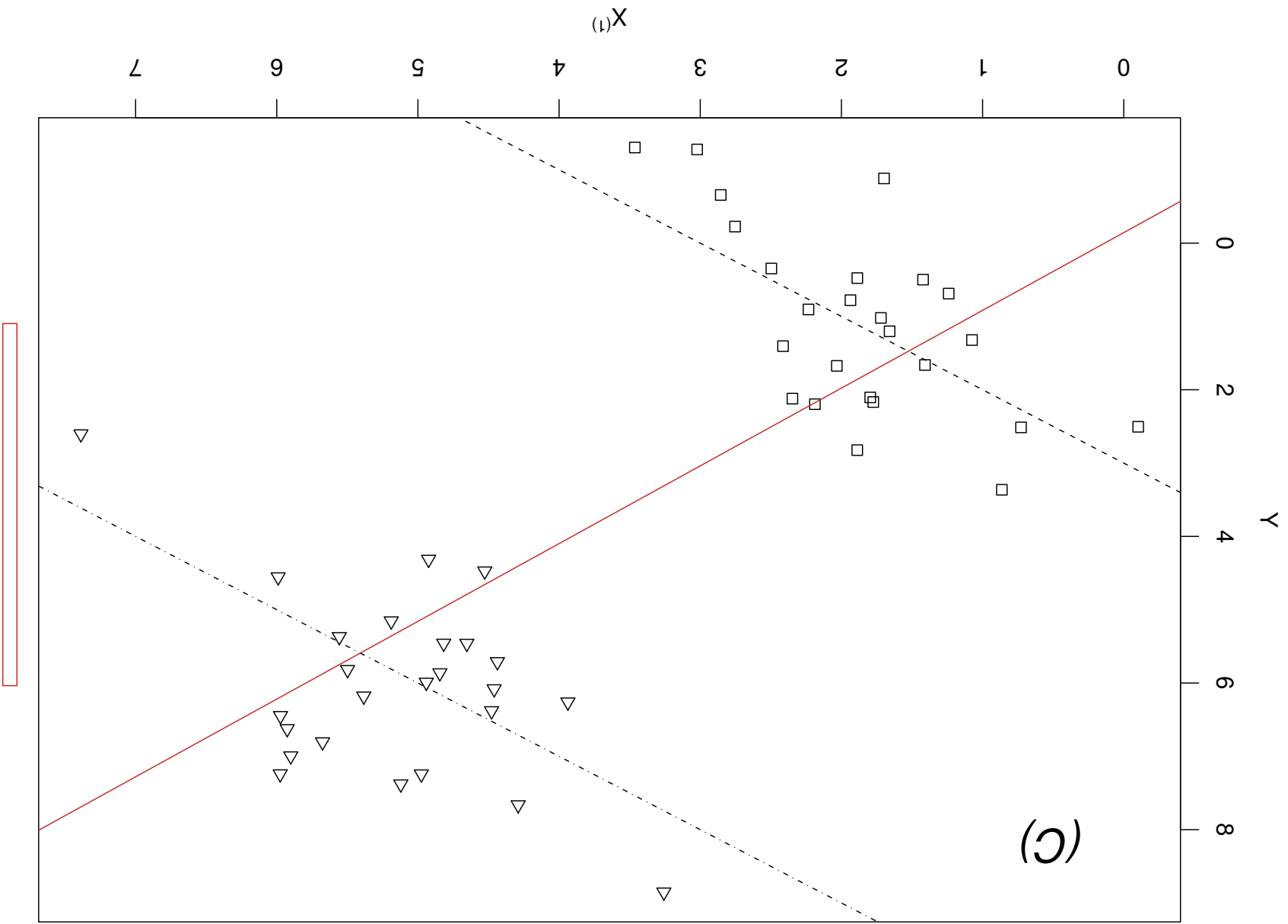
3.3 Ist multiple Regression mehr als

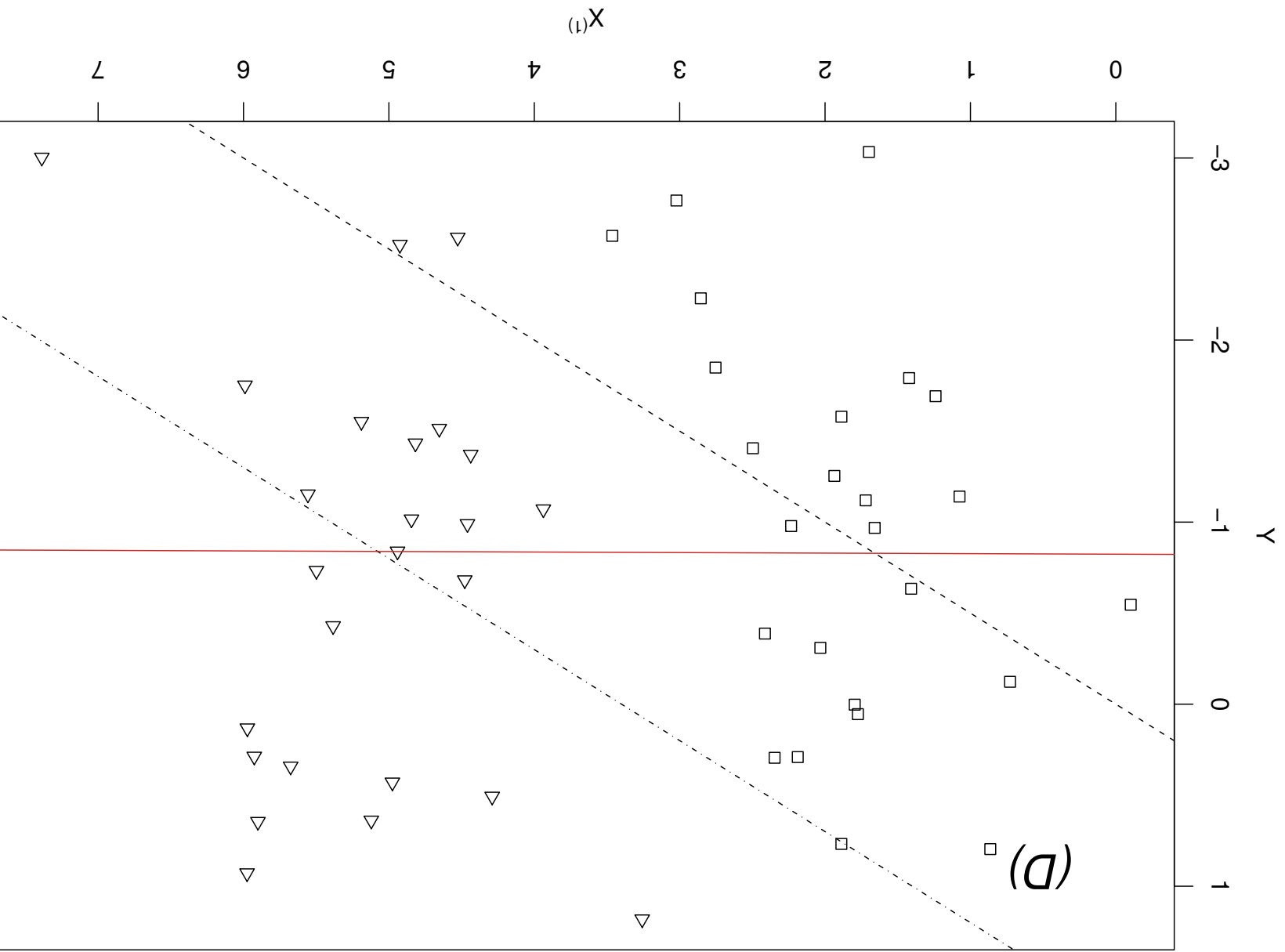
die Zusammenfassung von einfachen R.?

d Künstliches Bsp: Eine kontinuierliche $X^{(1)}$ und eine diskrete.



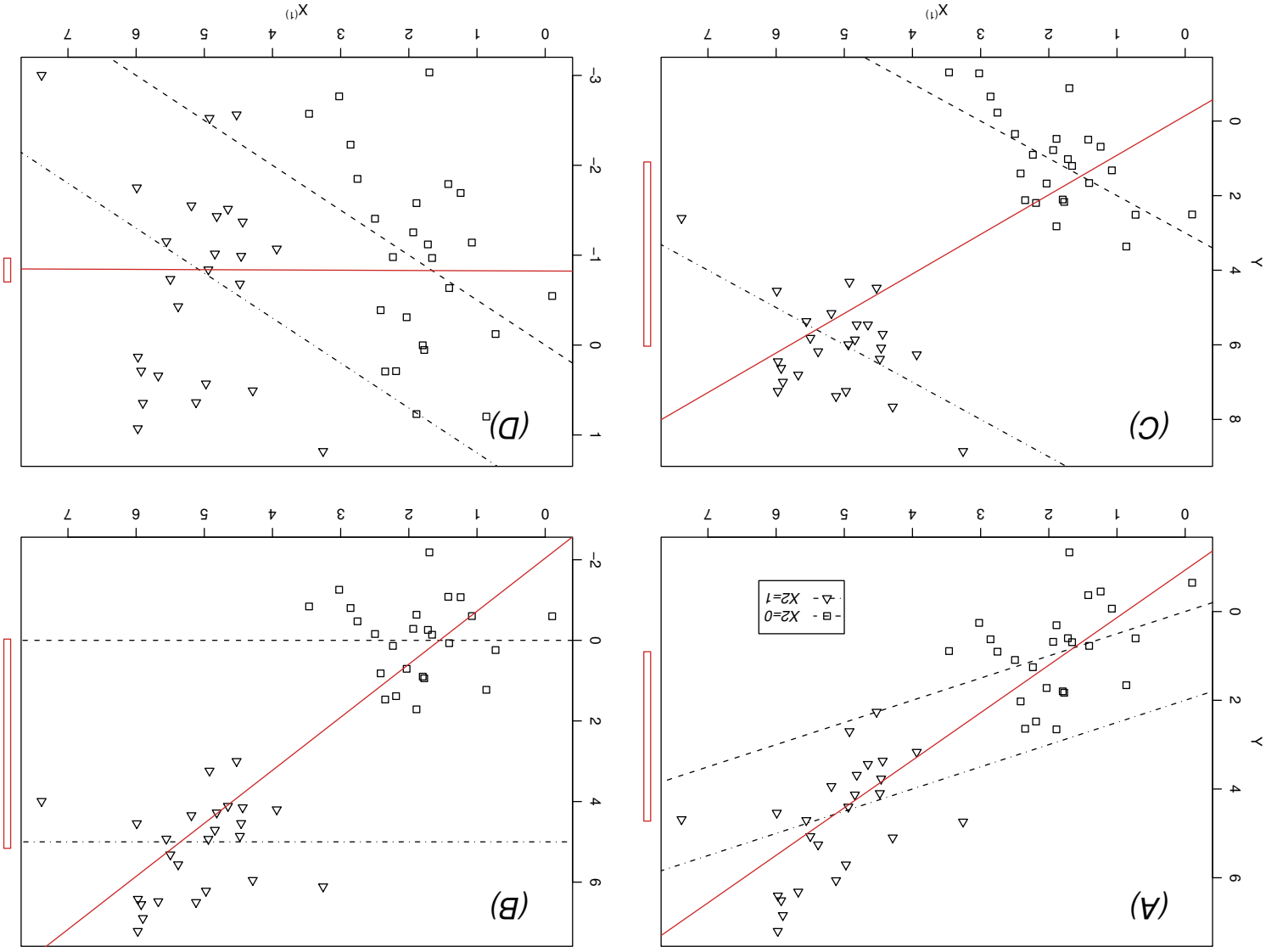






□

Die Bedeutung der Regressionskoeffizienten hängt prinzipiell davon ab, welche Ausgangsgrößen im Modell auftreten!



Ursache-Wirkungs-Beziehungen?!

Indizien für solche Beziehungen sammeln!

β_j signifikant, Ursache-Wirkungs-Beziehung plausibel

← "Nachweis" der Wirkung (?)

Achtung: indirekte Wirkungen sind möglich!

- $X^{(1)} \rightarrow Z \rightarrow Y$.

- $Z \rightarrow X^{(1)}; Z \rightarrow Y$.

Z im Modell → keine indirekten Wirkungen.

← "alle denkbaren" ursächlichen Var. ins Modell aufnehmen!

Besser:

- geplante Versuche,

- Nachweis eines Wirkungs-Mechanismus.

! β_j nicht signifikant \rightarrow kein Einfluss! ???

- Nullhypothese kann man nicht beweisen
- Ursächlicher Effekt kompensiert durch gegensätzlichen Effekt einer korrelierten Einflussgröße.
- Einfluss nicht-linear.

! Deshalb:

- möglichst alle möglichen ursächlichen Größen ins Modell aufnehmen,
- die Linearität der Zusammenhänge überprüfen (s. Residuenanalyse),
- ein **Vertrauensintervall** für den Koeffizienten liefern – statt eines P-Wertes.

k Indirekte Einflüsse können nicht vorkommen, wenn $X^{(j)}$ und Z nicht zusammenhängen (unkorreliert oder orthogonal sind).

Schätzung von β_j im multiplen und im einfachen Modell sind dann gleich.

l Multiples Modell ist trotzdem sehr nützlich: Kleinere Residuenstreuung $\hat{\sigma}$

→ kürzere Vertrauensintervalle.

m Zusammenfassend: Ein multiples Regressionsmodell sagt mehr aus als viele einfache Regressionen – im Falle von korrelierten Ausgangsgrößen viel mehr.

Merkpunkte

Multiple Regression

1. Die multiple lineare Regression bildet ein **reichhaltiges Modell** mit vielen Anwendungen.
2. **Multiple** Regression führt zu einer viel **aussagekräftigeren Analyse** als viele einfache Regressionen.
3. Regression allein kann man **keine Ursache – Wirkungsbeziehungen** beweisen.

3.4 S-Funktionen

a $> r.lm < -lm(\log10(ersch) \sim \log10(d\ddot{a}st),$
 $data = d.spreng)$

Modell-Formeln

$\log10(ersch) \sim \log10(d\ddot{a}st) + \log10(Ladung)$
 $+ St + St:\log10(d\ddot{a}st)$

Wechselwirkung!

Modell-Formeln allgemein

Klasse von S-Objekten, charakterisiert durch \sim
 Regression: ZielgröÙe \sim Regressor-Terme

$Y = X_1 + X_2$ sieht wie Mathematik aus!

Bedeutet in der lin.Regr.: $Y_i = \beta_0 + \beta_1 X_i^{(1)} + \beta_2 X_i^{(2)} + E_i$

Syntax hat eigene Regeln, die nicht immer den math. Zeichen entsprechen.
 Aufpassen (wie bei Matrizen)

Zielgröße ~ Regressor-Terme

Terme (rechte Seite):

– quantitative Variable

– Faktor

– Funktion von Ausgangsvariablen

– Wechselwirkung zwischen solchen Termen

Linke Seite:

– Funktionen von (einzelnen) Variablen

– mehrere Variable (logistische, multivariate, ... Regr.)

– fehlt für multivariate Verfahren (Hauptkomponenten, ...)

Viele Funktionen brauchen Formeln.

`plot(formula, ...)` benützt linke Seite vertikal, rechte horiz.

Erweiterung: $Y \sim X \mid Z$

– `cplot`

– gemischte Modelle der Varianzanalyse. Ausserdem:

$Y \sim X \mid Z$, $Y \sim X/Z$, $Y \sim X \% \text{in} \% Z$: Varianzanalyse

Wo werden Variable gesucht?

Die Funktionen, die `formula` als Argument haben, haben auch `data`.

Variable in der Formel sollen Spalten-Namen von `data` sein.

... sonst wird im `search`-Pfad gesucht, also zuerst im akt. workspace.

(`reg` erlaubt das nicht!)

Abkürzungen

- $Y \sim \cdot, \text{ data=t.d.} : \cdot$ steht für

„alle anderen Variablen“ (untransformiert)

- Wechselwirkungen:

$$X_1 * X_2 \iff X_1 + X_2 + X_1 : X_2$$

- $(X_1 + X_2 + X_3)^2 :$

alle Haupteffekte und alle Wechselwirkungen 1.O.

Komplikation: Die Zeichen +, *, ^ haben neue Bedeutung.

Manchmal möchte man die ursprüngliche Bedeutung haben.
→ Funktion $I(\dots) : \dots$ nicht als Formel interpretieren!

$$I(X_1^2), I(X_1 * (X_2 - 4))$$

(Innerhalb von Funktionen unnötig, z.B. $+ \text{sqrt}(X_1^2 + X_2^2)$)

b **Fehlende Werte**

Einfachste Behandlung: Zeilen mit ≥ 1 fehlenden Wert weglassen.
`lm(..., na.action=na.omit, ...)`

c `summary(r.lm, cor=FALSE)`

wird gebraucht, um Resultate anzuschauen.

d `drop1(r.lm, test="F")`. Faktoren prüfen. F-Test.

`anova, summary` für aov-Objekte macht andere Tests ...

e Funktion regr. Argumente wie Γ_m

- braucht kein summary,
- prüft Faktoren ohne Aufruf von `drop1`,
- zeigt neue Grösse „signif“, mit der man Vertrauensintervalle einfach berechnen kann,
- liefert weitere nützliche Grössen `signif` und `R2.x`,
- wird für viele weitere Modelle brauchbar sein.