Biostatistik

Lineare Regression

Matthias Birkner

JOHANNES GUTENBERG UNIVERSITÄT MAINZ

Inhalt

1. Lineare Regression: wozu und wie?
2. t-Test für lineare Zusammenhänge
3. Weitere Beispiele und Anmerkungen
   - K. Pearsons Größen von Vätern und Söhnen
   - Körper- und Gehirngewicht: Transformation der Daten
4. Lineare Regression mit `R`

photo (c) by Jörg Hempel

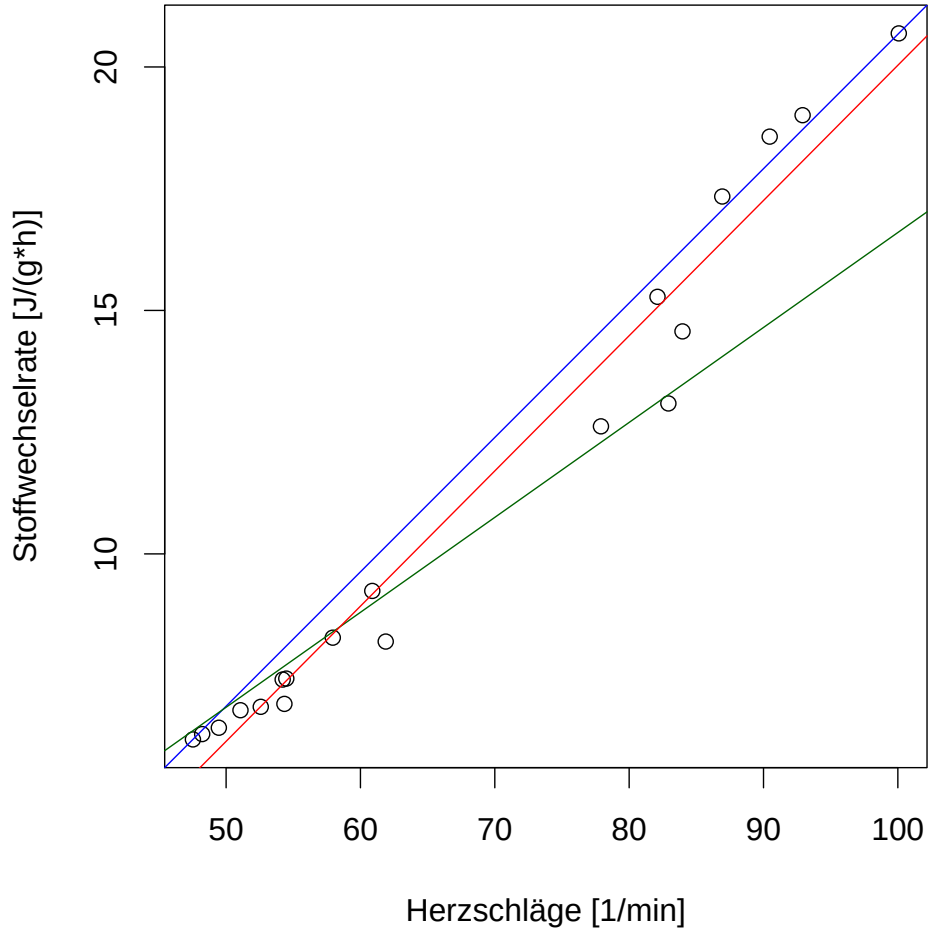
Gypus fulvus
Gänsegeier

Prinzinger, R., E. Karl, R. Bögel, Ch. Walzer (1999): Energy metabolism, body temperature, and cardiac work in the Griffon vulture Gyps vulvus - telemetric investigations in the laboratory and in the field.
Zoology **102**, Suppl. II: 15

- Daten aus der Arbeitsgruppe Stoffwechselphysiologie (Prof. Prinzinger) der Frankfurter Goethe-Universität.
- Telemetrisches System zur Messung der Herzfrequenz bei Vögeln auch während des Fluges.
- Wichtig für ökologische Fragen: die Stoffwechselrate
- Messung der Stoffwechselrate aufwändig und nur im Labor möglich.
- Können wir von der Herzfrequenz auf die Stoffwechselrate schließen?

Die Daten:

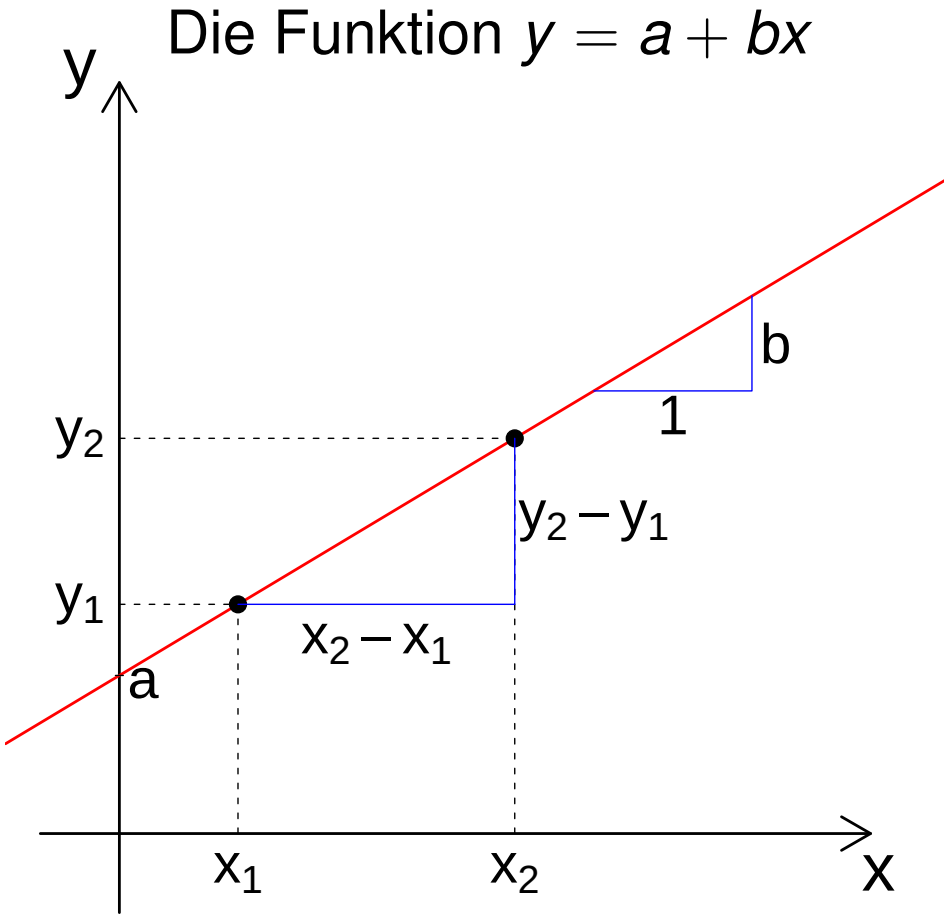
19 Beobachtungen eines Geiers
im Labor bei konstanter Temperatur (16° C)



Die Beobachtungen legen einen linearen Zusammenhang zwischen Herzfrequenz und Stoffwechselrate nahe.

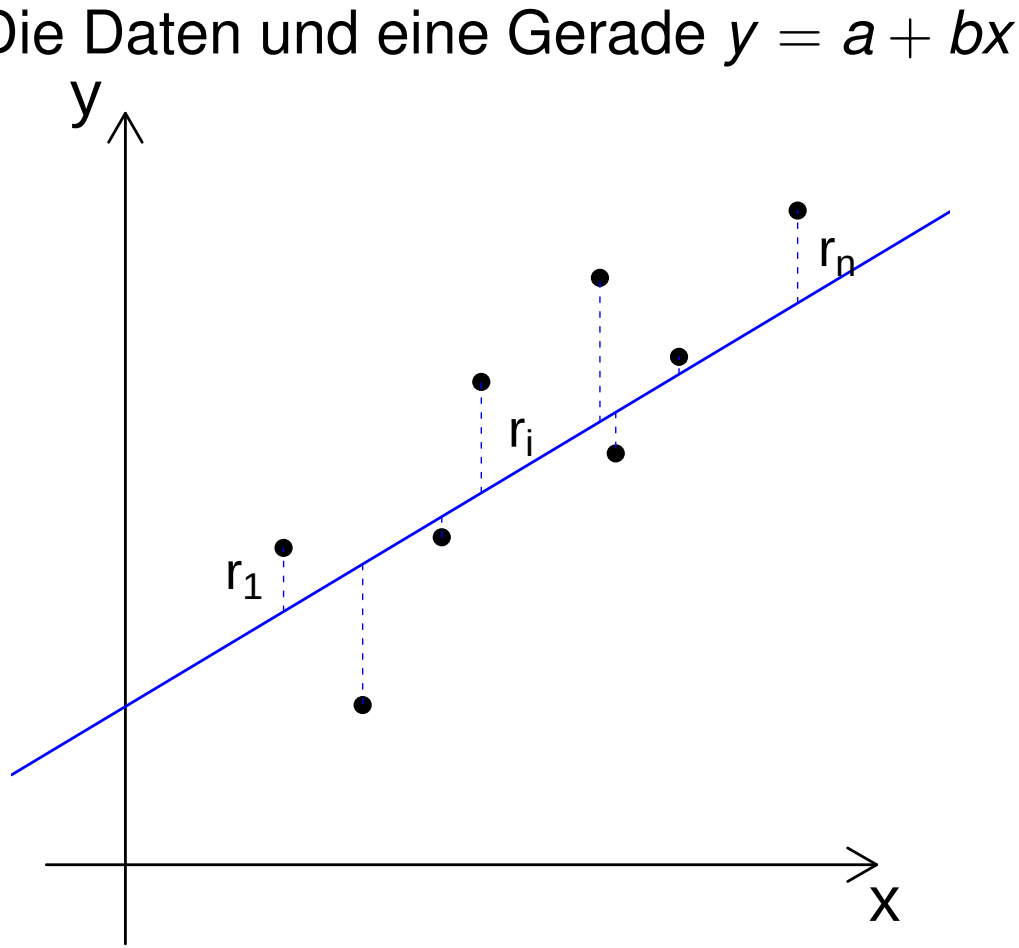
Welche Gerade passt „am Besten“?

Erinnerung: Lineare Funktionen



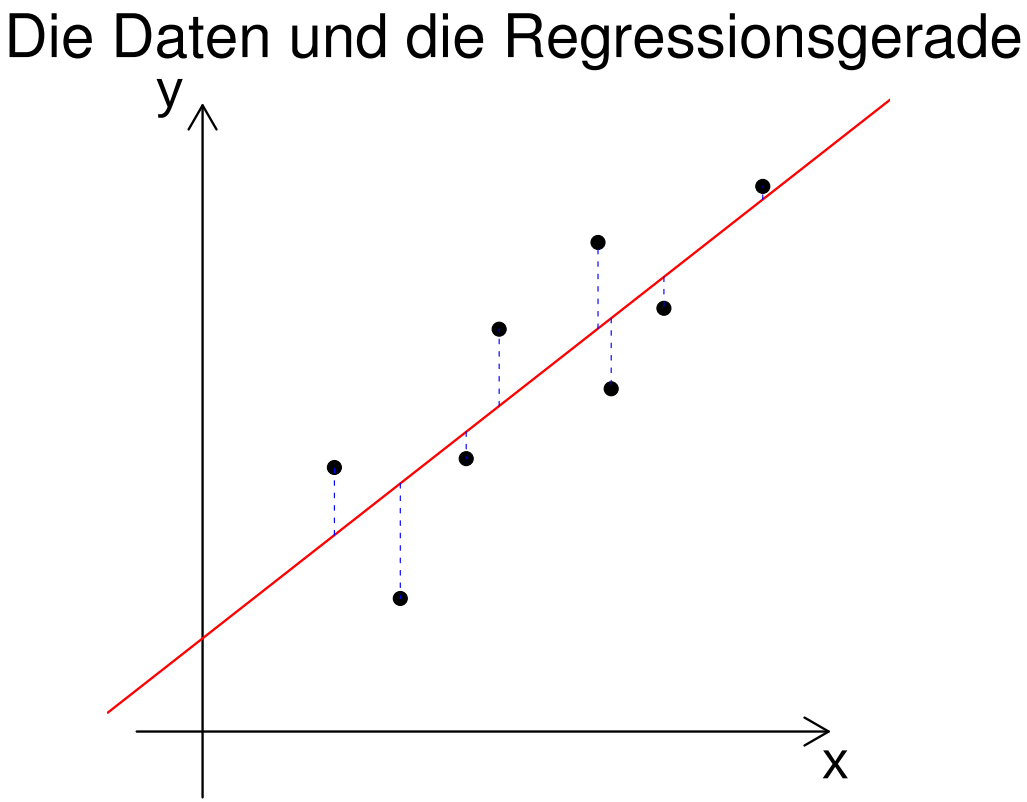
Hier ist a der Achsenabschnitt und b die Steigung.
Für jedes Paar von Punkten $(x_1, y_1) \neq (x_2, y_2)$ auf der Geraden gilt $\frac{y_2 - y_1}{x_2 - x_1} = b$.

Unsere Situation: n beobachtete Paare (x_i, y_i), $i = 1, \ldots, n$



Für keine gegebene Steigung b und Achsenabschnitt a liegen die Beobachtungen genau auf der Geraden $y = a + bx$.
Es bleiben stets sogenannte *Residuen* $r_i := y_i - a - bx_i$ $(\neq 0)$.

kleinste-Quadrate-Schätzer $\widehat{a}$, $\widehat{b}$ und Regressionsgerade



Ansatz: Finde $\widehat{a}$ und $\widehat{b}$ derart, dass $\sum_{i=1}^{n} \left(y_i - \widehat{a} - \widehat{b}\, x_i\right)^2 \stackrel{!}{=}$ minimal

(wobei über alle Wahlen von $a, b \in \mathbb{R}$ minimiert wird).

Die Gerade $y = \widehat{a} + \widehat{b}\, x$ heißt die *Regressionsgerade*.

Man kann $\widehat{a}$, $\widehat{b}$ folgendermaßen berechnen:

Seien $\overline{x} := \frac{1}{n}\sum_{i=1}^{n} x_i$, $\overline{y} := \frac{1}{n}\sum_{i=1}^{n} y_i$ die (empirischen) Mittelwerte der x- bzw. der y-Werte,

$\sigma_x^2 := \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$, $\sigma_y^2 := \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2$, die zugehörigen („unkorrigierten Stichproben“-)Varianzen und

$\mathrm{cov}(x,y) := \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$ die (empirische) Kovarianz der x- und der y-Werte.

Satz (Koeffizienten der Regressionsgerade)

$$\widehat{b} = \frac{\mathrm{cov}(x,y)}{\sigma_x^2}, \quad \widehat{a} = \overline{y} - \widehat{b}\,\overline{x}.$$

Bemerkungen zu den Regressionskoeffizienten

$$\widehat{b} = \frac{\operatorname{cov}(x,y)}{\sigma_x^2}, \quad \widehat{a} = \overline{y} - \widehat{b}\,\overline{x}$$

1. Beobachtung: Die Regressionsgerade $y = \widehat{a} + \widehat{b}\,x$ geht durch den Schwerpunkt der Daten $(\overline{x}, \overline{y})$.
   (Dies kann eine Merkhilfe für die Formeln bilden.)

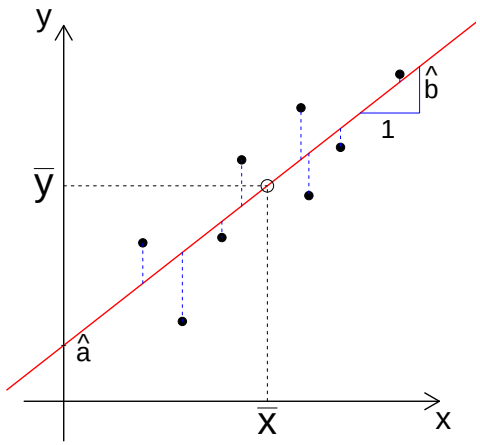
2. Man kann $\widehat{b}$ auch anders ausdrücken:

$$\frac{\operatorname{cov}(x,y)}{\sigma_x^2} = \frac{\sum_{i=1}^n (y_i - \overline{y})(x_i - \overline{x})}{\sum_{i=1}^n (x_i - \overline{x})^2} = \frac{\sum_{i=1}^n y_i (x_i - \overline{x})}{\sum_{i=1}^n (x_i - \overline{x})^2}$$

(Für das 1. Gleichheitszeichen erweitere mit n, für das 2. beachte $\overline{y} \sum_{i=1}^n (x_i - \overline{x}) = 0$. Die varianten Formeln können ggf. zum Rechnen angenehmer sein.)

Bemerkungen zu den Regressionskoeffizienten

$$\widehat{b} = \frac{\mathrm{cov}(x,y)}{\sigma_x^2}, \quad \widehat{a} = \overline{y} - \widehat{b}\,\overline{x}$$

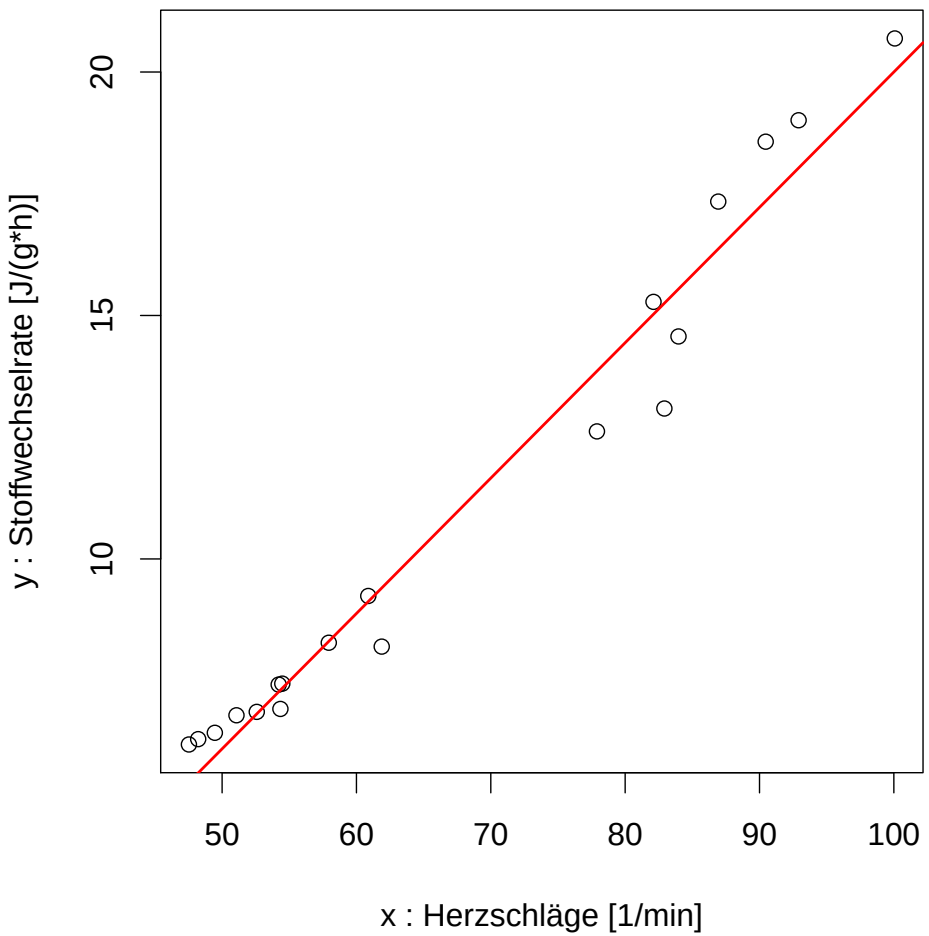


1. Beobachtung: Die Regressionsgerade $y = \widehat{a} + \widehat{b}\,x$ geht durch den Schwerpunkt der Daten $(\overline{x}, \overline{y})$.
   (Dies kann eine Merkhilfe für die Formeln bilden.)
2. Man kann $\widehat{b}$ auch anders ausdrücken:
$$\frac{\mathrm{cov}(x,y)}{\sigma_x^2} = \frac{\sum_{i=1}^n (y_i - \overline{y})(x_i - \overline{x})}{\sum_{i=1}^n (x_i - \overline{x})^2} = \frac{\sum_{i=1}^n y_i (x_i - \overline{x})}{\sum_{i=1}^n (x_i - \overline{x})^2}$$
   (Für das 1. Gleichheitszeichen erweitere mit n, für das 2. beachte $\overline{y} \sum_{i=1}^n (x_i - \overline{x}) = 0$. Die varianten Formeln können ggf. zum Rechnen angenehmer sein.)

Für das Geier-Beispiel ist mit $x \widehat{=}$Herzfrequenz, $y \widehat{=}$Stoffwechselrate:
$\overline{x} = 67.9$, $\overline{y} = 11.1$, $\text{cov}(x, y) = 83.5$, $\sigma_x^2 = 300.7$, also

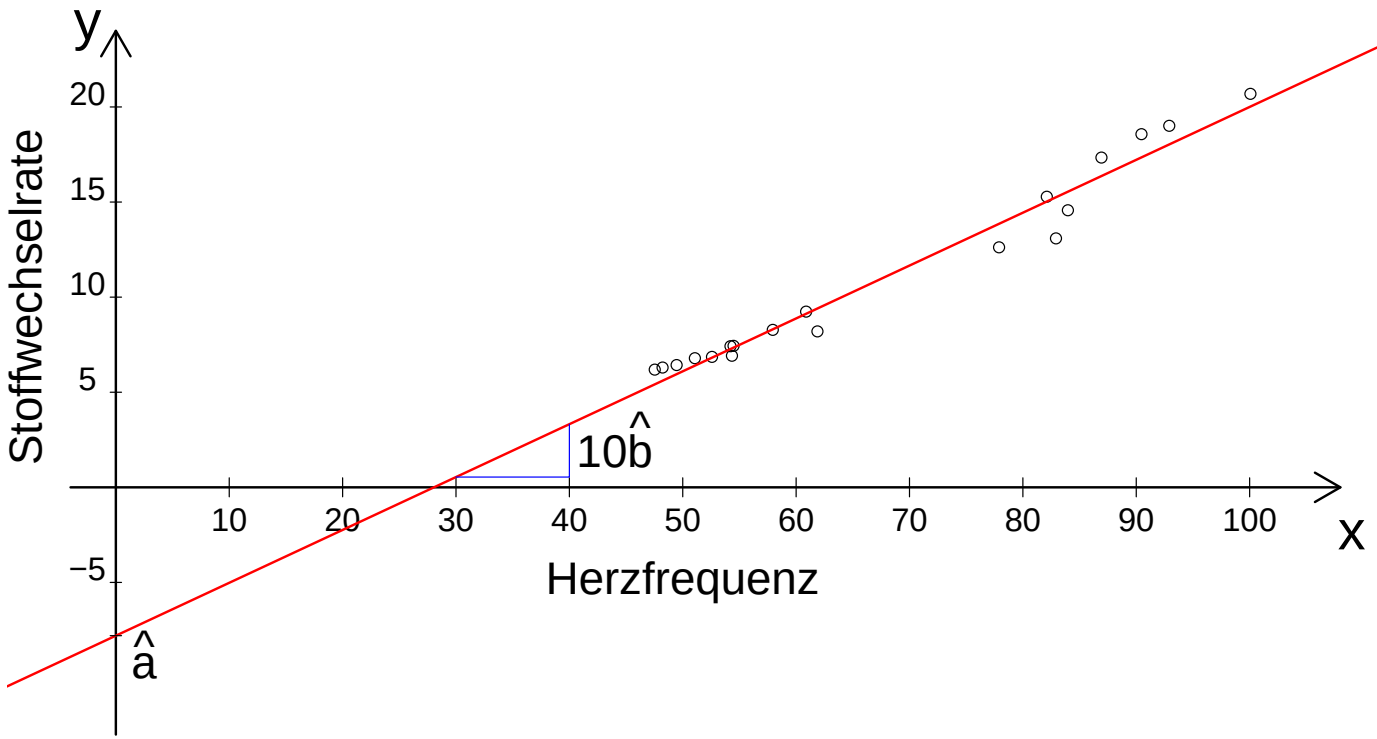
$$\widehat{b} = \frac{83.5}{300.7} = 0.278, \quad \widehat{a} = 11.1 - 67.9 \cdot 0.278 = -7.8$$

$n = 19$ Datenpunkte und
Regressionsgerade $y = -7.8 + 0.278x$



Interpretation der Regressionskoeffizienten

$n = 19$ Datenpunkte und Regressionsgerade $y = -7.8 + 0.278x$



$\widehat{b} = 0.278$: Erhöhung der Herzfrequenz um 10 erhöht die Stoffwechselrate im Mittel um $10\widehat{b} = 2.78$.

$\widehat{a} = -7.8$: Dies wäre die Stoffwechselrate eines hypothetischen Geiers mit Herzfrequenz 0.

(Offensichtlich kein sinnvoller Wert: Die Regressionsgerade ist nur in dem Bereich plausibel, in dem tatsächlich Beobachtungen vorliegen; Extrapolation auf eigene Gefahr!)

Regressionskoeffizienten: Woher kommen die Formeln?

$$\frac{1}{n}\sum_{i=1}^{n} r_i^2 = \frac{1}{n}\sum_{i=1}^{n} \left(y_i - a - bx_i\right)^2$$

$$= \left(\frac{1}{n}\sum_{i=1}^{n} \left(y_i - \overline{y} - b(x_i - \overline{x})\right)^2\right) + \left(\overline{y} - a - b\overline{x}\right)^2$$

denn

$$\left(y_i - \overline{y} - b(x_i - \overline{x})\right)^2 = \left(y_i - a - bx_i - (\overline{y} - a - b\overline{x})\right)^2$$

$$= \left(y_i - a - bx_i\right)^2 - 2\left(y_i - a - bx_i\right)\left(\overline{y} - a - b\overline{x}\right) + \left(\overline{y} - a - b\overline{x}\right)^2$$

und wenn man über i summiert und durch n teilt, ergeben die beiden letzten Terme zusammen gerade $-\left(\overline{y} - a - b\overline{x}\right)^2$.

Regressionskoeffizienten: Woher kommen die Formeln?

$$
\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} r_i^2 &= \frac{1}{n}\sum_{i=1}^{n} \left(y_i - a - bx_i\right)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n} \left(y_i - \overline{y} - b(x_i - \overline{x})\right)^2 + \left(\overline{y} - a - b\overline{x}\right)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n} (y_i - \overline{y})^2 - 2b\frac{1}{n}\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}) \\
&\qquad + b^2\frac{1}{n}\sum_{i=1}^{n} (x_i - \overline{x})^2 + \left(\overline{y} - a - b\overline{x}\right)^2 \\
&= \sigma_y^2 - 2b\,\mathrm{cov}(x,y) + b^2\sigma_x^2 + \left(\overline{y} - a - b\overline{x}\right)^2 \\
&= \sigma_y^2 - \frac{(\mathrm{cov}(x,y))^2}{\sigma_x^2} + \sigma_x^2\left(b - \frac{\mathrm{cov}(x,y)}{\sigma_x^2}\right)^2 + \left(\overline{y} - a - b\overline{x}\right)^2
\end{aligned}
$$

Regressionskoeffizienten: Woher kommen die Formeln?

Demnach:

$$\frac{1}{n}\sum_{i=1}^{n} r_i^2$$

$$= \underbrace{\sigma_y^2 - \frac{(\mathrm{Cov}(x,y))^2}{\sigma_x^2}}_{\text{hängt nicht von } a,\, b \text{ ab (und ist } \geq 0)} + \sigma_x^2 \underbrace{\left(b - \frac{\mathrm{Cov}(x,y)}{\sigma_x^2}\right)^2}_{\geq 0} + \underbrace{\left(\overline{y} - a - b\overline{x}\right)^2}_{\geq 0}$$

Die Summe der Residuenquadrate wird also minimiert durch die Wahlen

$$\widehat{b} = \frac{\mathrm{Cov}(x,y)}{\sigma_x^2}, \quad \widehat{a} = \overline{y} - \widehat{b}\,\overline{x}.$$

(Denn dann sind die beiden letzten Terme oben $= 0$.)

Lineare Regression: Modellvorstellung

Wir haben die Regressionsgerade

$$y = \hat{a} + \hat{b} \cdot x$$

durch die Minimierung der Summe der quadrierten Residuen definiert:

$$(\hat{a}, \hat{b}) = \arg\min_{(a,b)} \sum_i (y_i - (a + b \cdot x_i))^2$$

Dahinter steckt die Modellvorstellung, dass Werte a, b existieren, so dass für alle Datenpaare (x_i, y_i) gilt

$$y_i = a + b \cdot x_i + \varepsilon_i,$$

wobei alle ε_i unabhängig und normalverteilt sind mit Mittelwert 0 und derselben Varianz σ^2.

gegebene Daten:

Y	**X**
y_1	x_1
y_2	x_2
y_3	x_3
$\vdots$	$\vdots$
y_n	x_n

Modell: es gibt Zahlen a, b, σ^2, so dass

$$\begin{aligned} y_1 &= a + b \cdot x_1 + \varepsilon_1 \\ y_2 &= a + b \cdot x_2 + \varepsilon_2 \\ y_3 &= a + b \cdot x_3 + \varepsilon_3 \\ \vdots &\quad \vdots \\ y_n &= a + b \cdot x_n + \varepsilon_n \end{aligned}$$

Dabei sind $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ unabhängig $\sim \mathcal{N}(0, \sigma^2)$.

$\Rightarrow$ $y_1, y_2, \ldots, y_n$ sind unabhängig $y_i \sim \mathcal{N}(a + b \cdot x_i, \sigma^2)$.

a, b, σ^2 sind unbekannt, aber **nicht zufällig**.

Lineare Regression: Theorie

Modell: $y_i = a + b \cdot x_i + \varepsilon_i$ mit $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$,
fest (aber z.T. unbekannt): a, b, x_i, σ^2 zufällig: ε_i, y_i

Die kleinste-Quadrate-Schätzer

$$\hat{b} = \frac{\sum_i y_i (x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2}, \quad \hat{a} = \bar{y} - b\bar{x}$$

erfüllen

$$\mathbb{E}\left[\hat{a}\right] = a, \quad \mathbb{E}\left[\hat{b}\right] = b.$$

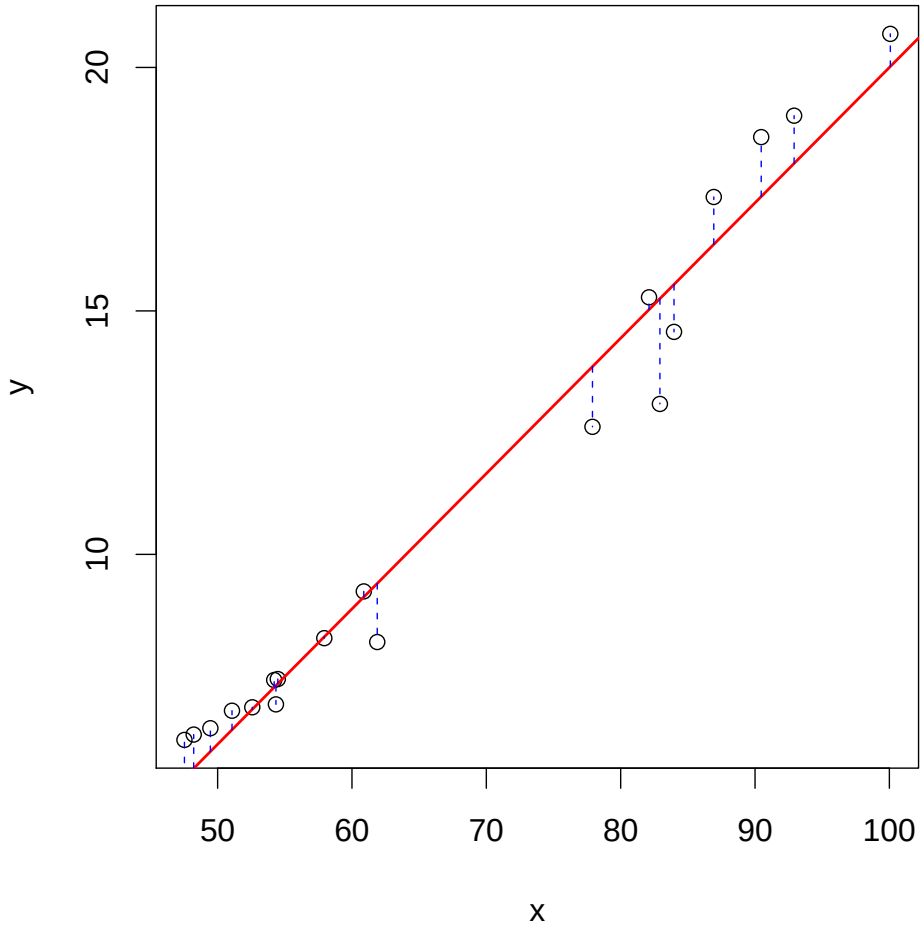
(D.h. sie sind sog. *erwartungstreue Schätzer*.)

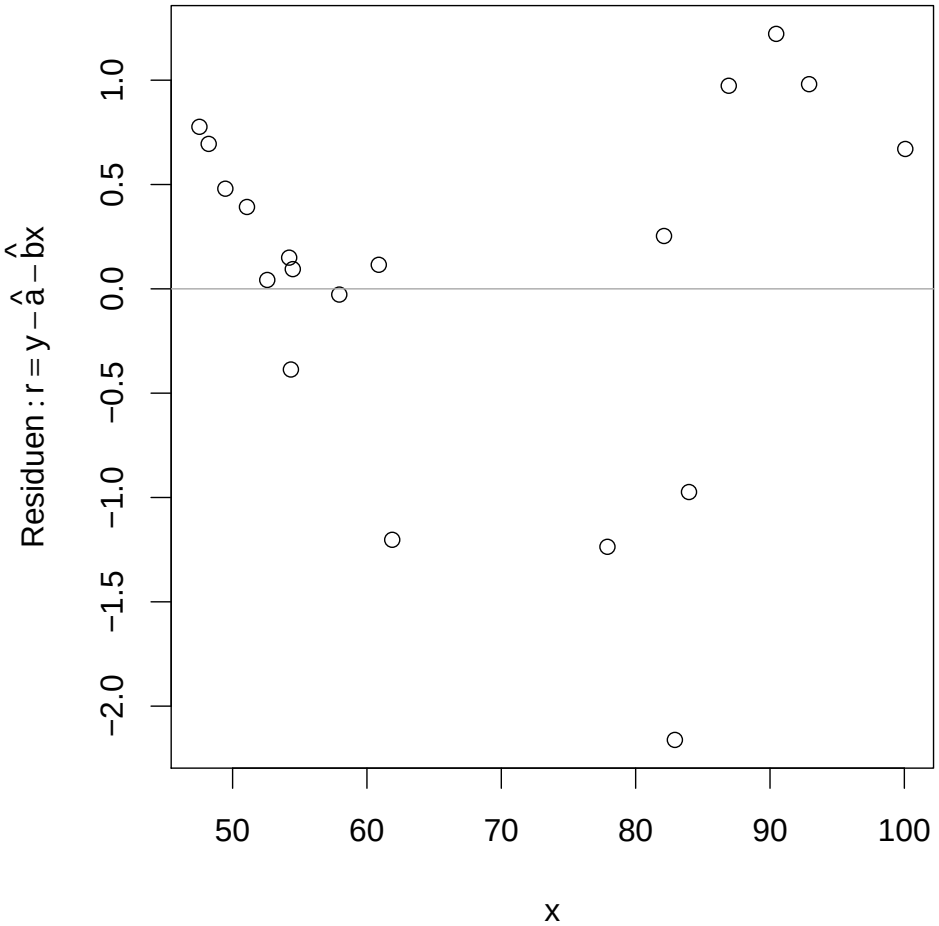
Wir schätzen σ^2 mit Hilfe der beobachten Residuenvarianz durch

$$s^2_{\text{res}} := \frac{1}{n-2} \sum_{i=1}^{n} \left(y_i - \hat{a} - \hat{b} \cdot x_i \right)^2. \qquad \text{Es gilt } \mathbb{E}\left[s^2_{\text{res}}\right] = \sigma^2.$$

(Beachte, dass durch $n - 2$ geteilt wird. Das hat damit zu tun, dass zwei Modellparameter a und b bereit geschätzt wurden, und somit 2 Freiheitsgrade verloren gegangen sind.)

Geier-Beispiel: Regressionsgerade, Residuen gegen *x*-Werte





Wir hatten $\hat{a} = -7.8$, $\hat{b} = 0.278$, man findet

$$s_{\text{res}} = \sqrt{\frac{1}{n-2} \sum_{i=1}^{n} \left(y_i - \hat{a} - \hat{b} \cdot x_i \right)^2} = 0.91.$$

Wir haben $\hat{a} = -7.8$, $\hat{b} = 0.278$, $s_{\text{res}} = 0.91$ gefunden.

Mit diesen Informationen können wir die Genauigkeit unserer Schätzung beurteilen

und auch einschätzen, wie genau wir einen *neuen* Beobachtungswert vorhersagen könnten.

Beispiel: Angenommen, bei einer weiteren Messung wurde bei Herzfrequenz $x = 76$ [1/min]
eine Stoffwechselrate von $y = 14.3$ [J/(g·h)] gemessen.

Die Vorhersage der Regressionsgerade wäre
$-7.8 + 0.278 \cdot 76 = 13.33$, d.h. sie weicht um
$14.3 - 13.33 = 0.97$ von der Messung ab.
Ist das ein Grund, an unserem Modell zu zweifeln?

Nein: Wenn $\sigma \approx s_{\text{res}} = 0.91$ gilt, so beobachten wir ein ε_{n+1}, das von derselben Größenordnung wie seine Streuung ist — was im Modell mit W'keit ca. $1/3$ passieren kann.
(Wenn wir dagegen $y = 16.3$ gemessen hätten, wären wir schon beunruhigt ...)

Beispiel: Rothirsch (*Cervus elaphus*)

photo (c) BS Thurner Hof

Hängt der Anteil männlicher Nachkommen einer Hirschkuh mit ihrem sozialen Rang zusammen?

Frage: Hängt der Anteil männlicher Nachkommen einer Hirschkuh mit ihrem sozialen Rang zusammen?

Betrachten wir folgende Theorie:
Hirschkühe können das Geschlecht ihrer Nachkommen beeinflussen.

Unter dem Gesichtspunkt evolutionär stabiler Strategien ist zu erwarten, dass schwache Tiere eher zu weiblichem und starke Tiere eher zu männlichem Nachwuchs tendieren.

Folgender Artikel berichtet über eine Langzeitstudie, bei der eine Gruppe Rothirsche auf der schottischen Insel Rùm über 15 Jahre beobachtet wurde, und die zu obiger Frage Daten gesammelt hat:

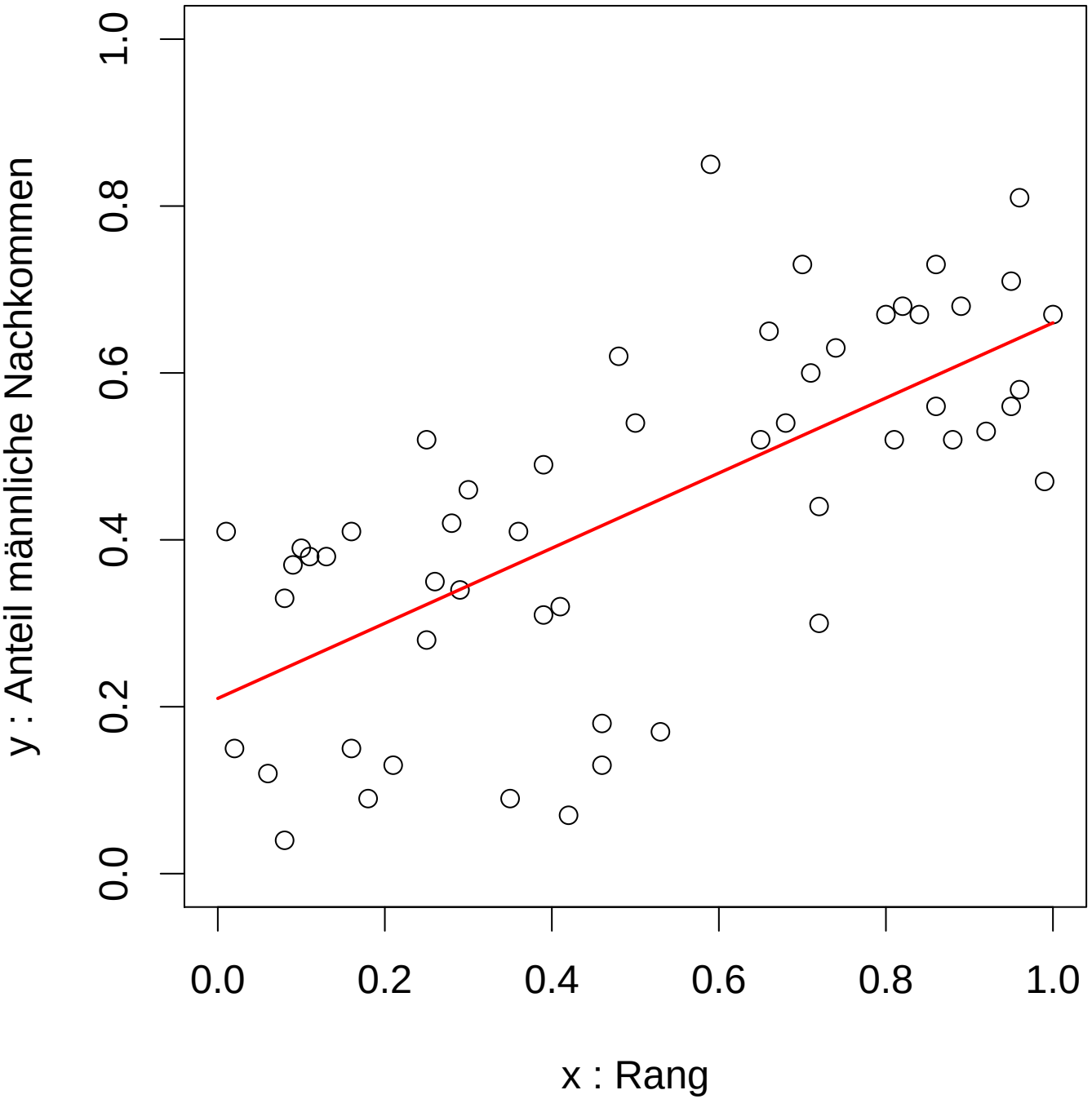
Clutton-Brock, T. H. , Albon, S. D., Guinness, F. E. (1986) Great expectations: dominance, breeding success and offspring sex ratios in red deer.
Anim. Behav. **34**, 460—471.

	Rang	Anteil männl. Nachkommen
1	0.01	0.41
2	0.02	0.15
3	0.06	0.12
4	0.08	0.04
5	0.08	0.33
6	0.09	0.37
.	.	.
.	.	.
.	.	.
52	0.96	0.81
53	0.99	0.47
54	1.00	0.67

Die Beobachtungen:

Für 54 Hirschkühe wurde der Rang (normiert auf einen Wert in $[0, 1]$, grob gesprochen ein Schätzwert für die Wahrscheinlichkeit, dass die betreffende Kuh ein „Duell“ mit einer zufällig ausgewählten anderen Kuh „gewinnt“) und der Anteil männlicher Nachkommen beobachtet.

(Simulierte Daten, die sich an den Werten aus der Originalpublikation orientieren.)



Es ist

$\overline{x} = 0.51$
(mittl. Rang)

$\overline{y} = 0.44$
(mittl. Ant. männl. Nachk.)

$\sigma_x^2 = 0.097$

$cov(x, y) = 0.044$

$\hat{b} = \frac{0.044}{0.097} \doteq 0.45$

$\hat{a}$
$= 0.44 - 0.51 \cdot 0.45$
$\doteq 0.21$

Wir beobachten einen wachsenden (und ungefähr linearen) Zusammenhang zwischen Rang und Anteil männlicher Nachkommen einer Hirschkuh. Ist das ein systematischer Effekt oder könnte reiner Zufall diese Beobachtung genausogut erklären?

Dazu betrachten wir unser Modell:

$$Y = a + b \cdot X + \varepsilon \qquad \text{mit } \varepsilon \sim \mathcal{N}(0, \sigma^2)$$

Wie berechnet man die Signifikanz eines Zusammenhangs zwischen dem *erklärenden Merkmal X* und der *Zielgröße Y*?

Wir haben b durch $\hat{b} = 0.45 \neq 0$ geschätzt. Könnte das wahre b auch 0 sein?

Anders formuliert: Mit welchem Test können wir der Nullhypothese $b = 0$ zu Leibe rücken?

Wie groß ist der Standardfehler unserer Schätzung $\hat{b}$?

Modell:

$$y_i = a + b \cdot x_i + \varepsilon_i \qquad \text{mit } \varepsilon_i \sim \mathcal{N}(0, \sigma^2)$$

nicht zufällig: a, b, x_i, σ^2 zufällig: ε_i, y_i

$$\text{var}(y_i) = \text{var}(a + b \cdot x_i + \varepsilon_i) = \text{var}(\varepsilon_i) = \sigma^2$$

und $y_1, y_2, \ldots, y_n$ sind stochastisch unabhängig.

$$\hat{b} = \frac{\sum_i y_i (x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2}$$

Demnach

$$\begin{aligned}
\text{Var}(\hat{b}) &= \text{Var}\left(\frac{\sum_i y_i (x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2}\right) = \frac{\text{Var}\left(\sum_i y_i (x_i - \bar{x})\right)}{\left(\sum_i (x_i - \bar{x})^2\right)^2} \\
&= \frac{\sum_i \text{Var}\left(y_i\right) (x_i - \bar{x})^2}{\left(\sum_i (x_i - \bar{x})^2\right)^2} = \sigma^2 \cdot \frac{\sum_i (x_i - \bar{x})^2}{\left(\sum_i (x_i - \bar{x})^2\right)^2} \\
&= \sigma^2 \Big/ \sum_i (x_i - \bar{x})^2
\end{aligned}$$

Tatsächlich ist $\hat{b}$ Normalverteilt mit Mittelwert b und

$$\mathrm{Var}(\hat{b}) = \sigma^2 \Big/ \sum_i (x_i - \bar{x})^2$$

Problem: Wir kennen σ^2 nicht.
Wir schätzen σ^2 mit Hilfe der beobachten Residuenvarianz durch

$$s^2_{\text{res}} := \frac{\sum_i \left(y_i - \hat{a} - \hat{b} \cdot x_i\right)^2}{n-2}$$

Erinnerung: Hier wird durch $n-2$ geteilt. Das hat damit zu tun, dass zwei Modellparameter a und b bereit geschätzt wurden, und somit 2 Freiheitsgrade verloren gegangen sind.

$$\mathrm{Var}(\hat{b}) = \sigma^2 \Big/ \sum_i (x_i - \bar{x})^2$$

Schätze σ^2 durch

$$s^2_{\mathrm{res}} = \frac{\sum_i \left(y_i - \hat{a} - \hat{b} \cdot x_i\right)^2}{n-2}.$$

Dann ist der Standardfehler von $\hat{b}$ gegeben durch $\frac{s_{\mathrm{res}}}{\sqrt{\sum_i (x_i - \bar{x})^2}}$
(unser Schätzwert für die Streuung von $\hat{b}$)
und (unter den Modellannahmen)

$$\frac{\hat{b} - b}{s_{\mathrm{res}} \Big/ \sqrt{\sum_i (x_i - \bar{x})^2}} = \frac{\hat{b} - b}{s_{\mathrm{res}} \Big/ \sigma_x \sqrt{n}}$$

Student-t-verteilt mit $n-2$ Freiheitsgraden. Wir können also den t-Test anwenden, um die Nullhypothese $b = 0$ zu testen.

Im Rothirschkühe-Beispiel:

$\hat{b} = 0.45$, $\frac{s_{\text{res}}}{\sqrt{\sum_i (x_i - \bar{x})^2}} = 0.0673$,

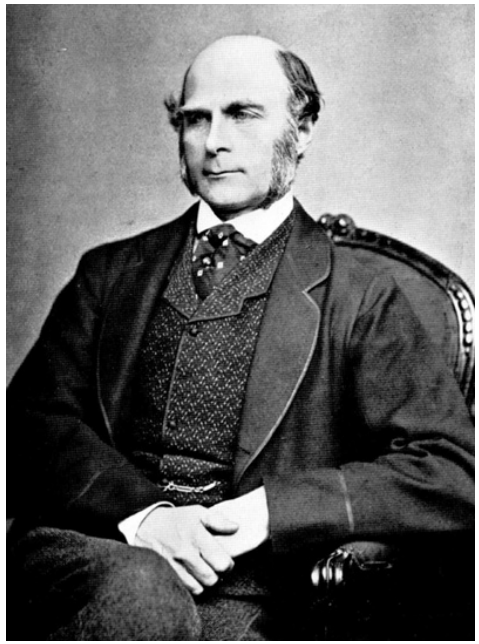
also beobachten wir $t = \frac{\hat{b} - 0}{s_{\text{res}} \Big/ \sqrt{\sum_i (x_i - \bar{x})^2}} = 6.7$

Einer Tabelle entnehmen wir: Das 99.95%-Quantil der Student-Verteilung mit 50 Freiheitsgraden ist 3.496 (und das der Student-Vert. mit 60 Freiheitsgraden ist 3.460).

Wir können also die Nullhypothese „das wahre $b = 0$“ zum Signifikanzniveau 0.1% ablehnen.

Bemerkung: Das beweist natürlich nicht, dass Hirschkühe das Geschlecht ihrer Nachkommen willentlich bestimmen können.
Es scheint eher plausibel anzunehmen, dass es Faktoren gibt, die den Rang und die Geschlechterverteilung der Nachkommen zugleich beeinflussen, siehe die Diskussion in dem zitierten Artikel von T. H. Clutton-Brock et. al.

Woher kommt der Name „Regression“ (nach lat. regressio, Zurückkommen)?



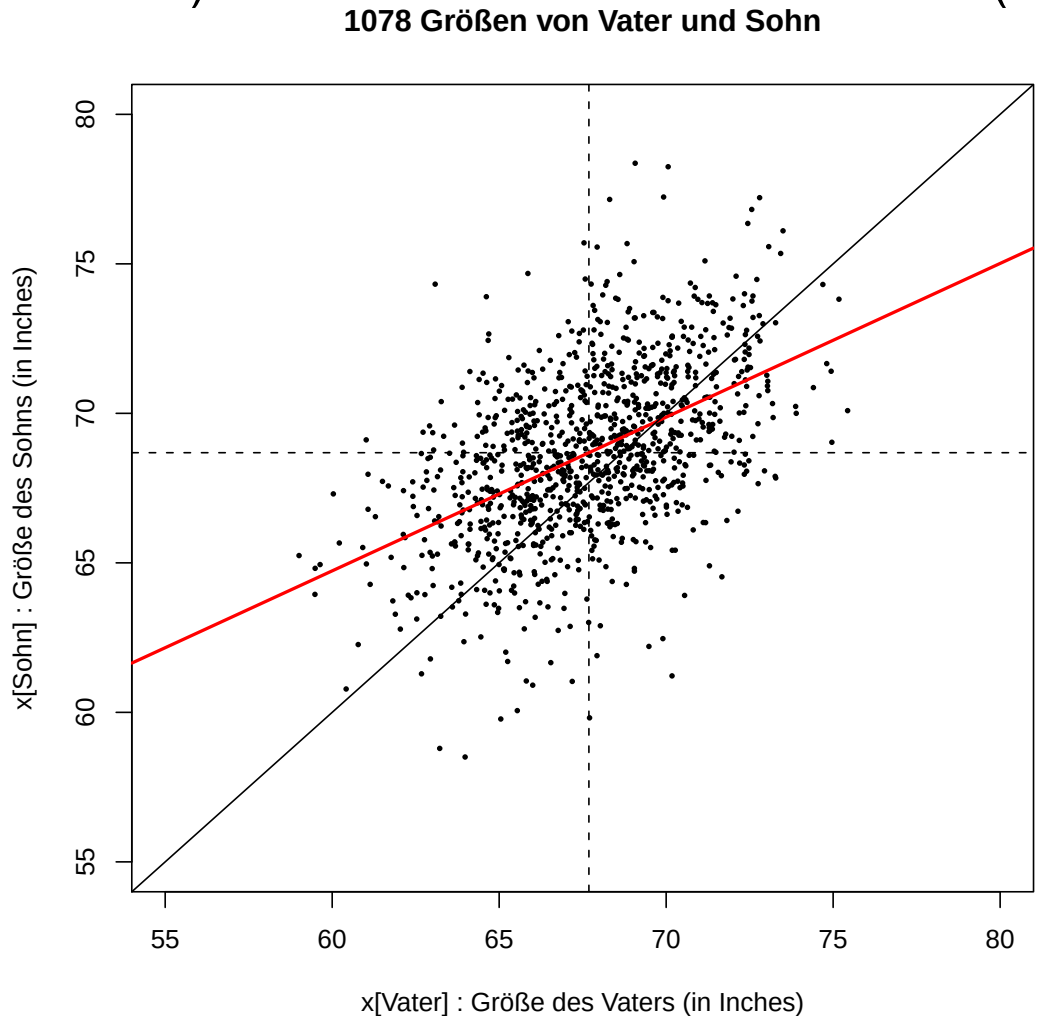
Francis Galton (1822–1911, engl. Wissenschaftler*)

hat angesichts biometrischer Beobachtungen den Ausdruck “regression towards the mean” geprägt.

---

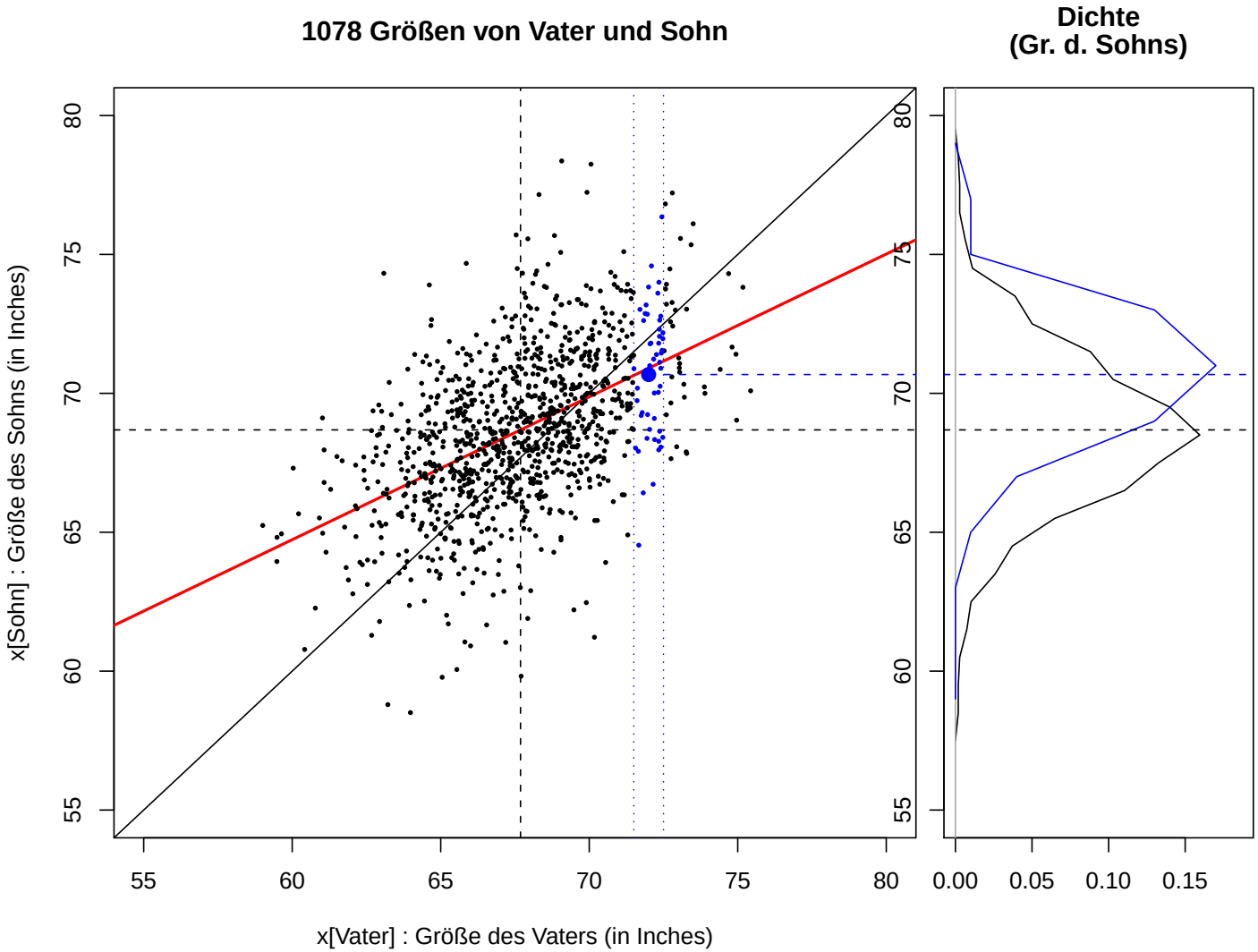
*Siehe auch Robert Langkjaer-Bain, The troubling legacy of Francis Galton, *Significance* 16(3), 2019.
https://doi.org/10.1111/j.1740-9713.2019.01275.x

Ein (relativ berühmter) Datensatz von Karl Pearson (1858-1936)

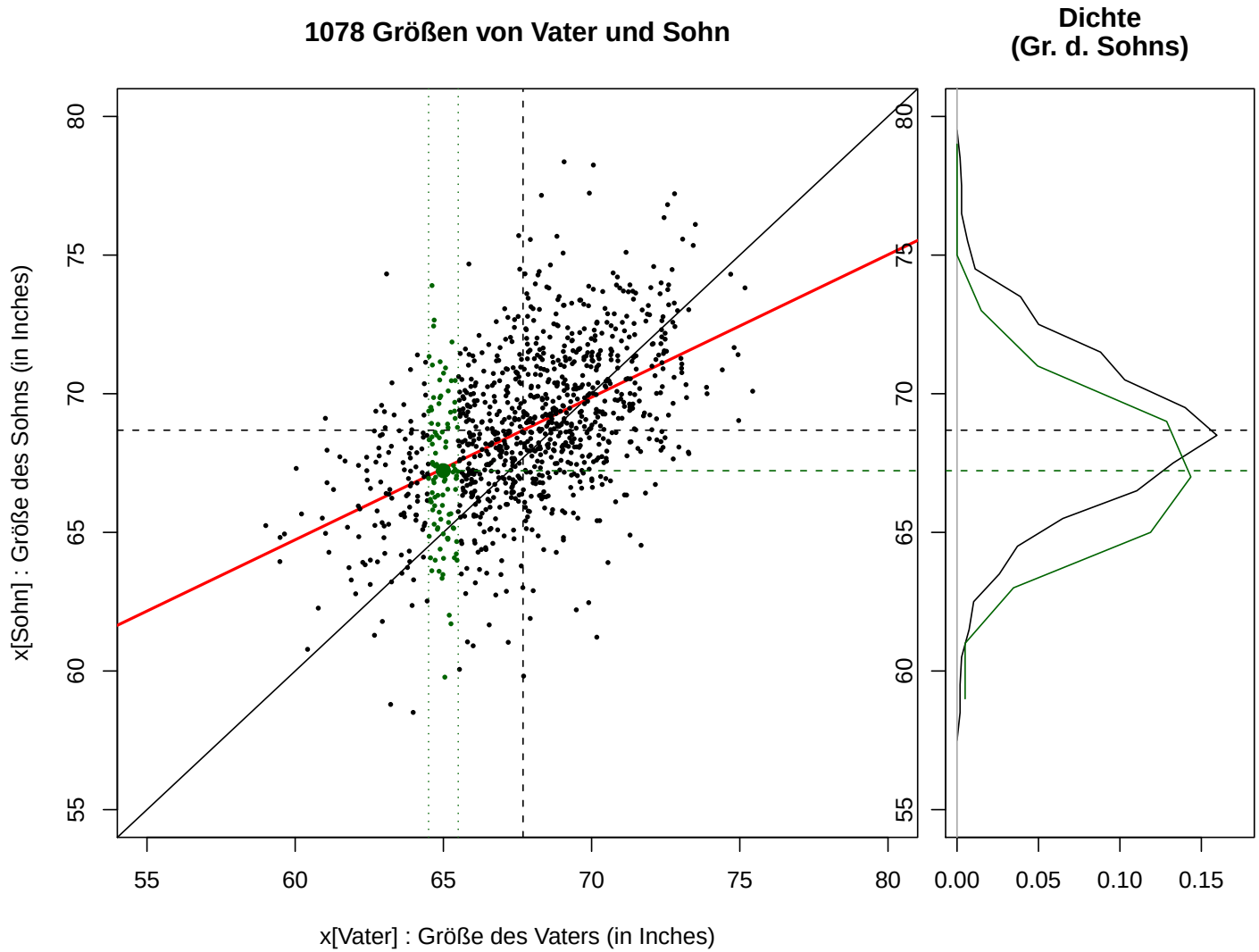


$\overline{x}_{\text{Vater}} = 67.7$, $\overline{x}_{\text{Sohn}} = 68.7$, $\sigma^2_{\text{Vater}} = 7.52$ ($\sigma_{\text{Vater}} = 2.74$, $\sigma_{\text{Sohn}} = 2.81$),
$\text{cov}(x_{\text{Vater}}, x_{\text{Sohn}}) = 3.87$
(Korrelationskoeffizient $\rho = \text{cov}(x_{\text{Vater}}, x_{\text{Sohn}})/(\sigma_{\text{Vater}}\,\sigma_{\text{Sohn}})) = 0.50$)

Regressionsgerade: $x_{\text{Sohn}} = 33.89 + 0.514 x_{\text{Vater}}$.

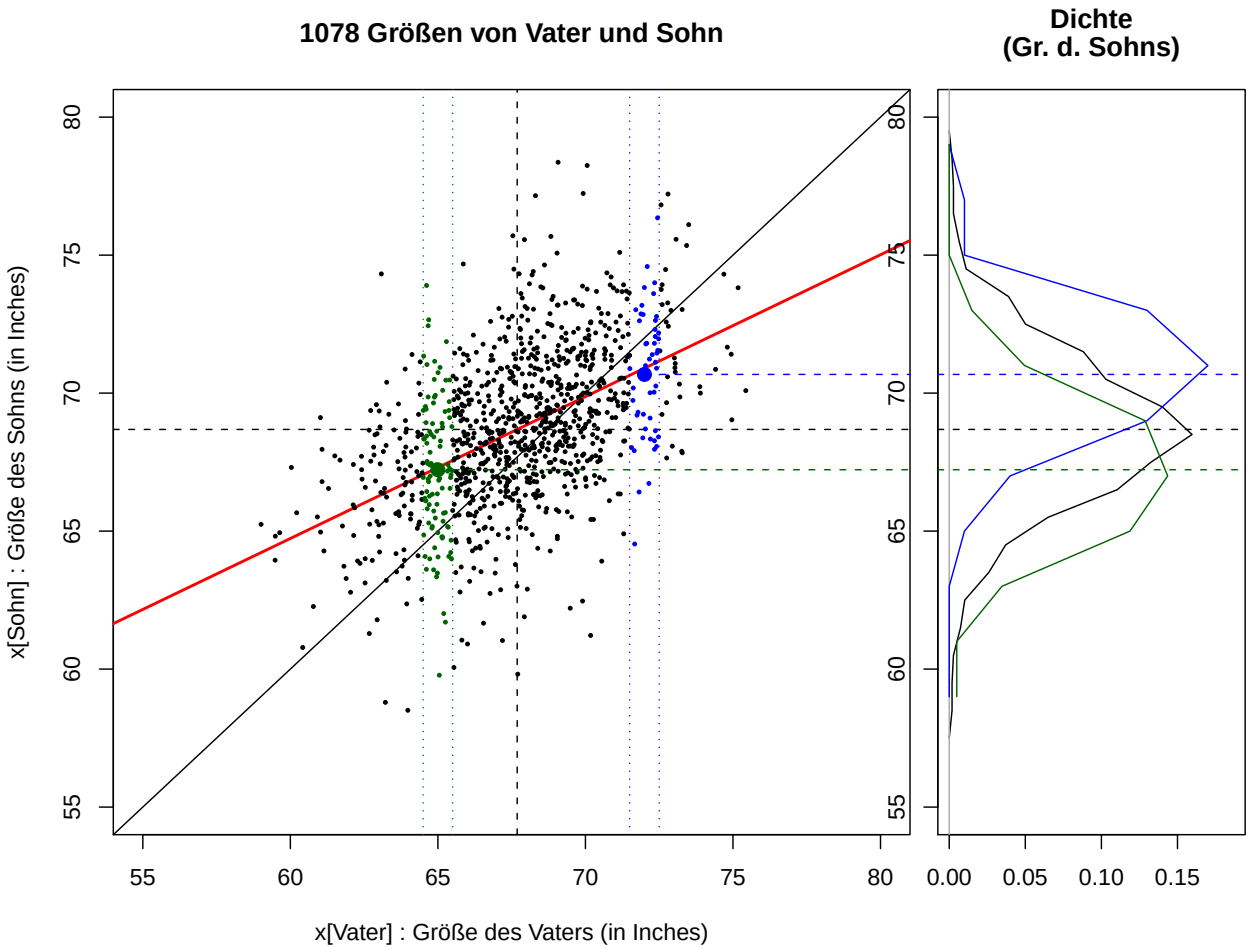


Betrachten wir die Söhne von überdurchschnittlich großen Vätern (z.B. Väter, die ca. 72 Inches groß sind):
Diese Söhne sind überdurchschnittlich groß (im Vergleich zu allen Söhnen), aber im Mittel kleiner als ihr Vater.



Betrachten wir andererseits die Söhne von unterdurchschnittlich großen Vätern (z.B. Väter, die ca. 65 Inches groß sind):
Diese Söhne sind unterdurchschnittlich groß (im Vergleich zu allen Söhnen), aber im Mittel größer als ihr Vater.

“Regression towards the mean”

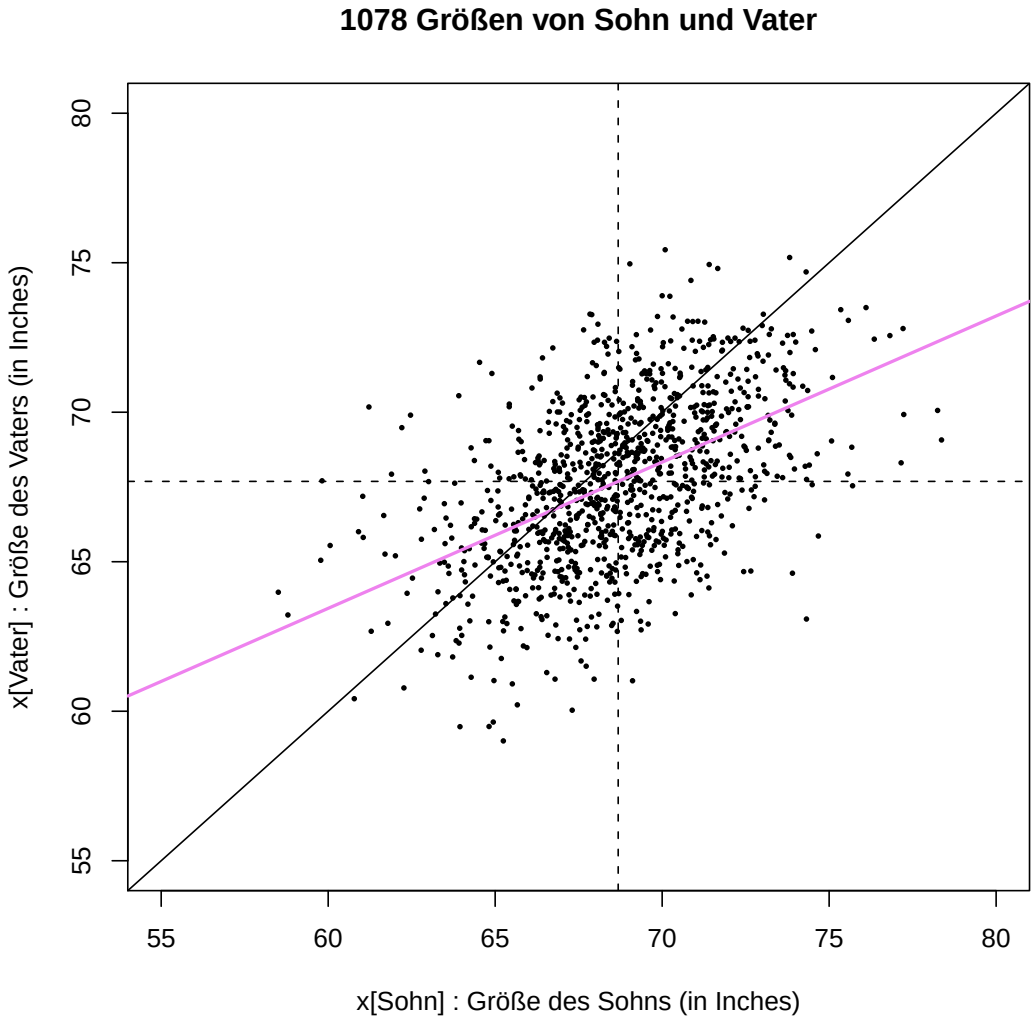


Wir sehen: Söhne überdurchschnittlich großer Väter sind im Mittel kleiner als ihr Vater (aber immer noch größer als der Populationsdurchschnitt), für Söhne unterdurchschnittlich großer Väter ist es umgekehrt: „Rückkehr zum Mittelwert“.

Bemerkung: Das beobachtete Phänomen der „Rückkehr zum Mittelwert“ bedeutet nicht notwendigerweise einen tieferen kausalen Zusammenhang, es tritt stets im Zusammenhang mit natürlicher Variabilität auf (technisch gesehen stets, wenn für den Korrelationkoeffizient ρ gilt $|\rho| < 1$).

Bestimmen wir (spaßeshalber) im Größen-Beispiel die Regressionsgerade für die Größe des Vaters als Funktion der Größe des Sohns:

Wir hatten $\overline{x}_{\text{Vater}} = 67.7$, $\overline{x}_{\text{Sohn}} = 68.7$, $\text{cov}(x_{\text{Vater}}, x_{\text{Sohn}}) = 3.87$, $\sigma^2_{\text{Sohn}} = 7.92$
und finden die Regressionsgerade $x_{\text{Vater}} = 34.1 + 0.489 x_{\text{Sohn}}$



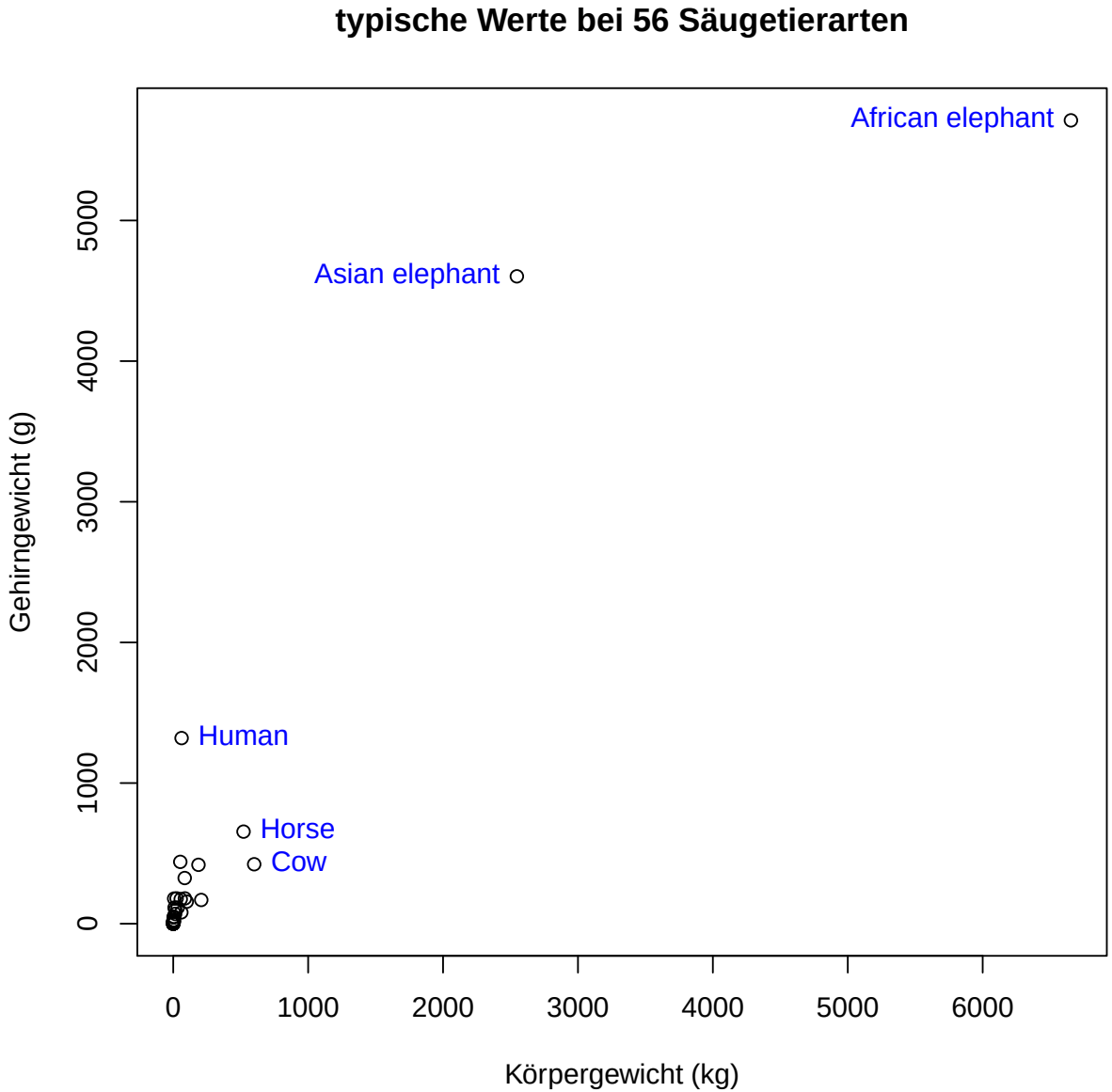
Regressionsgerade: $x_{\text{Vater}} = 34.1 + 0.489 x_{\text{Sohn}}$

Daten: Typisches Körpergewicht [kg] und Gehirngewicht [g] von 56 Säugetierarten

species	brain.weight.g	weight.kg
Short-nosed echidna	25.00	4.500
North American Opossum	6.30	1.700
Phalanger	11.40	1.620
European hedgehog	3.50	0.770
Desert hedgehog	2.40	0.550
Tenrec	2.60	0.900
Greater short-tailed shrew	0.29	0.019
Lesser short-tailed shrew	0.14	0.005
Musk shrew	0.33	0.048
Star-nosed mole	1.00	0.060
Eastern american mole	1.20	0.075
Tree shrew	2.50	0.104
Little brown bat	0.25	0.010
Big brown bat	0.30	0.023
Slow loris	12.50	1.400
Galago	5.00	0.200
Squirrel monkey	20.00	0.743
Owl monkey	15.50	0.480
Patas monkey	115.00	10.000
Macaque	179.00	6.800
Baboon	180.00	25.235
Chimpanzee	440.00	52.200
Human	1320.00	62.000
Giant armadillo	81.00	60.000
Long-nosed armadillo	10.80	3.500
Rabbit	12.10	2.500
Arctic ground squirrel	5.70	0.920
Thirteen-lined ground squirrel	4.00	0.101

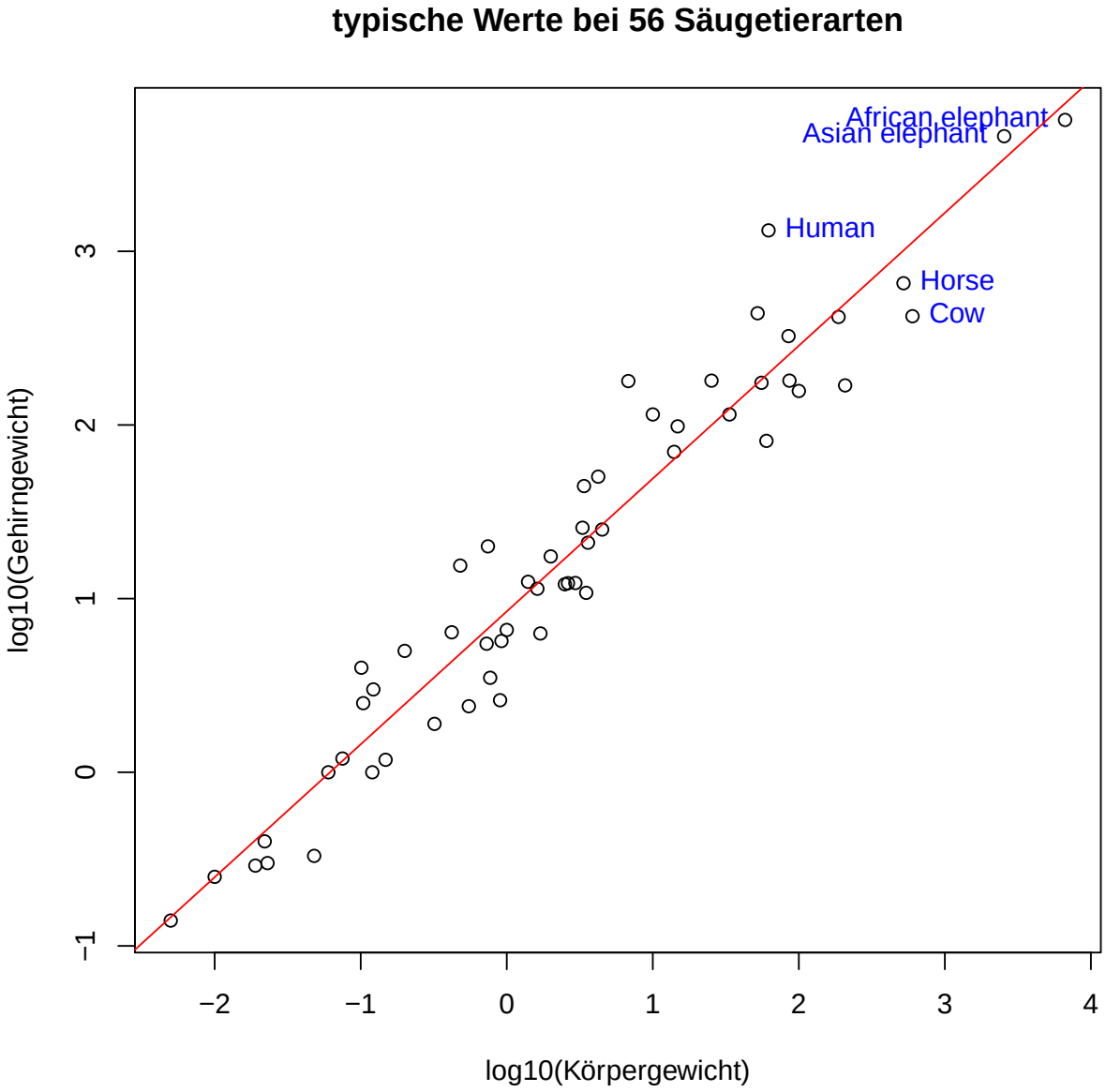
species	brain.weight.g	weight.kg
Cotton rat	1.18	0.148
Golden hamster	1.00	0.120
Mole rat	3.00	0.122
African giant pouched rat	6.60	1.000
Laboratory rat	1.90	0.320
House mouse	0.40	0.022
Guinea pig	5.50	0.728
Chinchilla	6.40	0.420
Red fox	50.40	4.230
Arctic fox	44.50	3.380
Dog	70.00	14.000
Genet	17.50	2.000
Domestic cat	25.60	3.300
Jaguar	157.00	100.000
Gray seal	325.00	85.000
Asian elephant	4603.00	2547.000
African elephant	5712.00	6654.000
Rock hyrax	21.00	3.600
Gray hyrax	12.27	2.625
Tree hyrax	12.30	2.950
Horse	655.00	521.000
Donkey	419.00	187.000
Brazilian tapir	169.00	207.501
Pig	180.00	86.250
Roe deer	98.20	14.800
Cow	423.00	600.000
Goat	115.00	33.500
Sheep	175.00	55.500

Daten entnommen aus V. M. Savage and G. B. West. A quantitative, theoretical framework for understanding mammalian sleep. Proceedings of the National Academy of Sciences, 104 (3):1051-1056, 2007
(nur Spezies, für die Information zum mittleren Gehirngewicht vorlag)



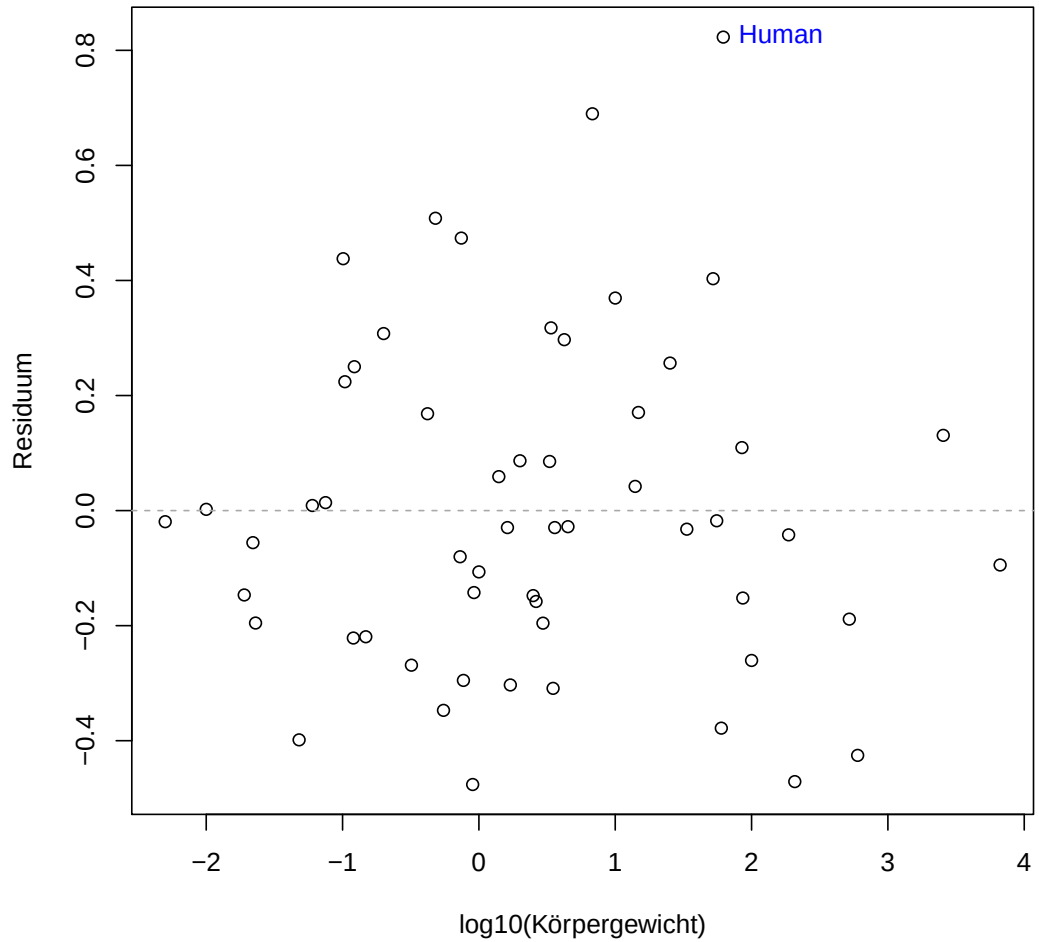
Wir sehen — wenig: Wenige „schwere“ Arten dominieren die Skala.

Abhilfe: Logarithmieren!



Auf der log-Skala sieht ein linearer Zusammenhang plausibel aus.

Die Regressionsgerade ist $y = 0.926 + 0.765x$.

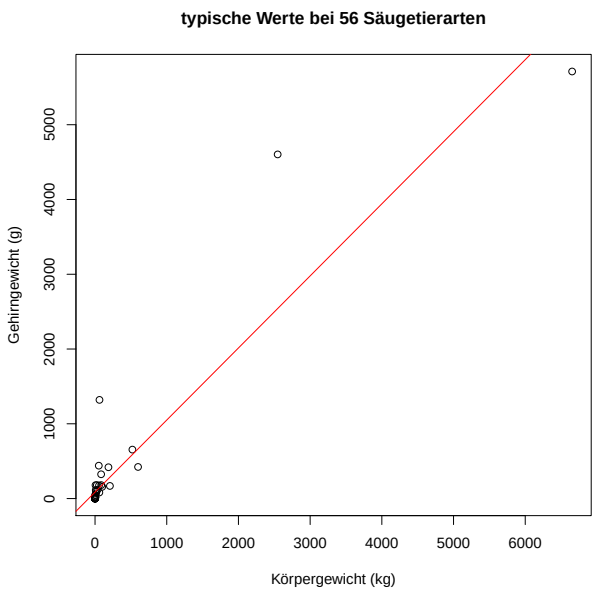


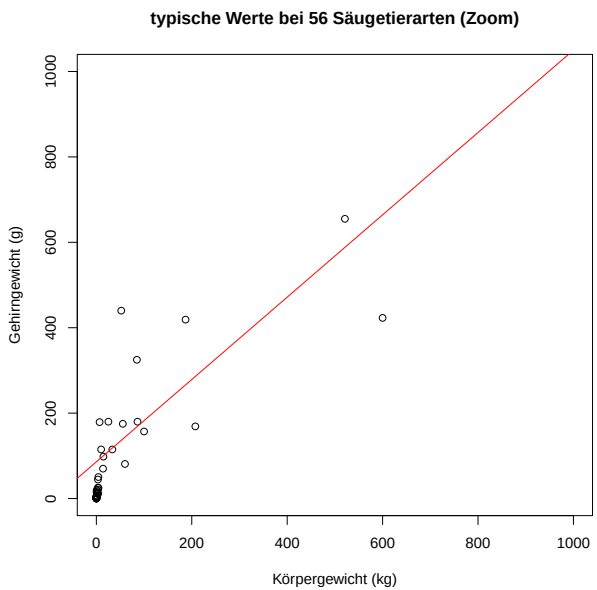
Die Residuen passen in etwa zur Modellvorstellung
„log(Gehirngewicht)=0.926+0.765 · log(Körpergewicht) + ε“,
wobei die Streuung des „Fehlerterms“ ε nicht von der Art abhängt.

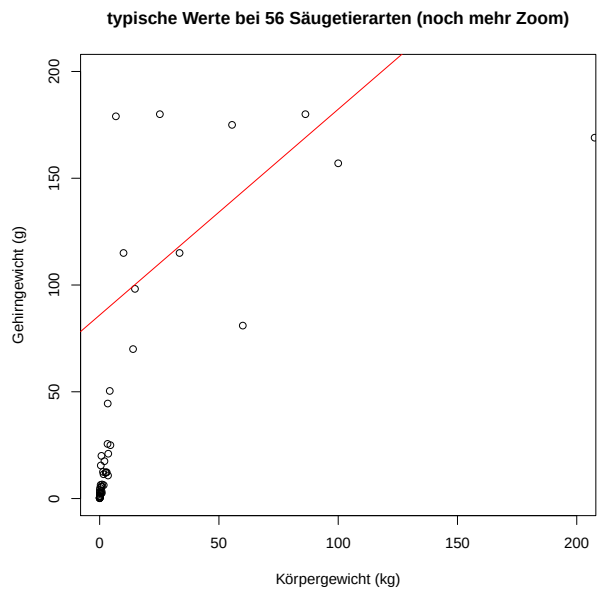
Wir haben zunächst die x-Werte (Körpergewichte) und die y-Werte (Gehirngewichte) logarithmiert, und dann die Regressionsgerade angepasst.

Frage: War das nur ein Trick, damit das Diagramm besser erkennbar erkennbar wird, oder gibt es eine weitere inhaltliche Bedeutung?

Passen wir den unlogarithmierten Wertepaaren ebenfalls eine Regressiongerade an
(es kommt $y_{\text{Gehirngew.}} = 85.92 + 0.964 x_{\text{Körpergew.}}$ heraus).

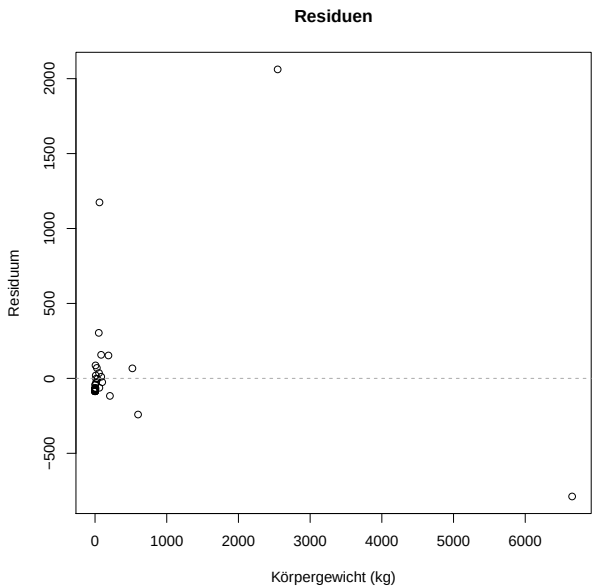


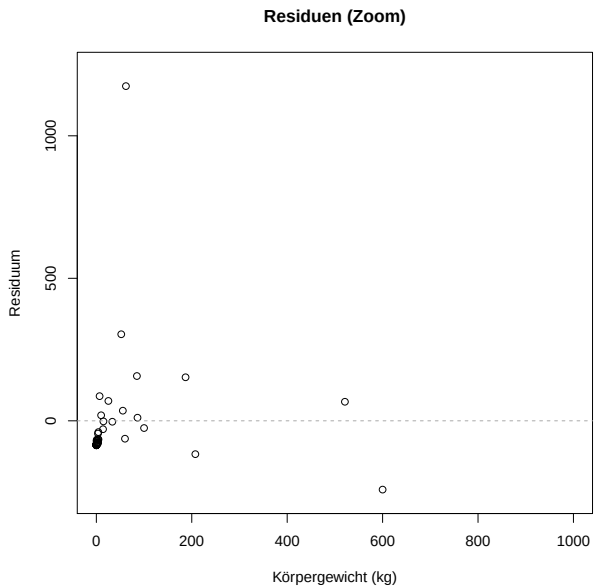


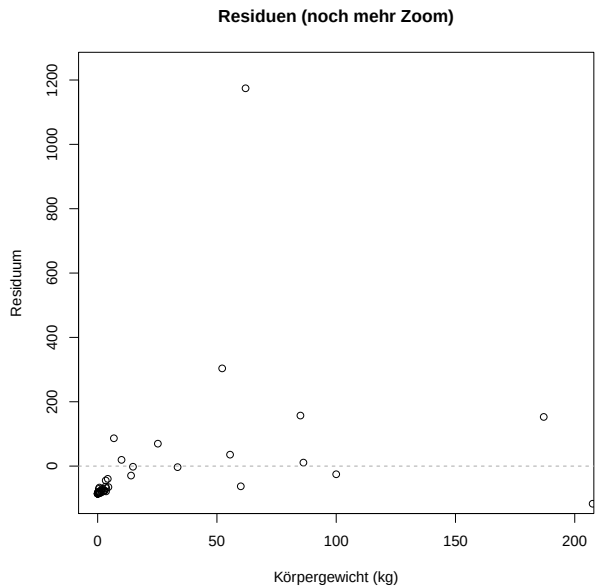


Die Anpassung wirkt schlechter. Wir betrachten die „leichteren“ Arten näher.

Und noch näher.







Die Residuen wirken nun sehr inhomogen, was nicht zur Modellvorstellung passt. (Der Eindruck bleibt auch bei näherer Betrachtung der leichten Arten bestehen und auch bei noch näherer.)

Wir sehen, dass die Varianz der Residuen von den angepassten Werten bzw. dem Körpergewicht abhängt. Man sagt, es liegt *Heteroskedastizität* vor.
Das Modell geht aber von *Homoskedastizität* aus, d.h. die Residuenvarianz soll von den erklärenden Merkmalen (dem Körpergewicht) und den angepassten Werten (annähernd) unabhängig sein.

Varianzstabilisierende Transformation:
Wie können wir die Körper- und Hirnmasse umskalieren, um Homoskedastizität zu erreichen?

Idee: Bei Elefanten kann das typischerweise 5 kg schwere Hirn je nach Individuum auch mal 500 g schwerer oder leichter sein. Wenn bei einer Tierart das Hirn typischerweise 5 g schwer ist, wird es nicht um 500 g variieren können, sondern vielleicht ebenfalls um 10%, also ± 0.5 g. Die Variabilität ist hier also nicht additiv, sondern multiplikativ:

$$\text{Hirnmasse} = (\text{erwartete Hirnmasse}) \cdot \text{„Rauschen“}$$

Das können wir aber in etwas mit additivem Zufallsterm umwandeln, indem wir auf beiden Seiten den Logarithmus anwenden:

$$\log(\text{Hirnmasse}) = \log(\text{erwartete Hirnmasse}) + \log(\text{„Rauschen“})$$

Lineare Regression mit R

Eingabe der Datensätze `x` und `y` (z.B. „von Hand“):

```
x <- c(1.0, 2.2, 2.7, 2.7, 3.5, 5.0)
y <- c(4.9, 7.2, 8.8, 8.4, 10.4, 11.2)
```

Ausgabe von Steigung und y-Achsenabschnitt der Ausgleichsgeraden, sowie Kovarianz und Korrelationskoeffizient

```
lm(y ~ x)
cov(x, y)
cor(x, y)
```

Grafische Darstellung

```
plot(x, y) # die Datenpunkte malen
abline(lm(y ~ x), col='red', lwd=3)
                    # und Ausgleichsgerade hinzufügen
```