Weiterbildungskurs “Stochastik”

Hansruedi Künsch

Seminar für Statistik
Departement Mathematik, ETH Zürich

24. Juni 2009

Inhalt

Fragestellungen

Typische Fragestellungen

- Mendel'sche Theorie postuliert in einem bestimmten Kreuzungsversuch von Erbsen 25% grüne und 75% gelbe Keimblätter. Experiment ergibt unter 8023 Pflanzen 2001 mit grünen Keimblättern.
  Spricht das für oder gegen Mendel's Theorie ?
- Aus einer Umfrage der New York Times:
  "In theory, in 19 cases out of 20 the results based on such a sample (of 1154 adults) will differ by no more than three percentage points in either direction from what would have been obtained by seeking out all American adults."
  Wie kommt man zu so einer Aussage ?

Die Binomialverteilung

Verteilung der Anzahl “Erfolge” bei n unabhängigen Wiederholungen eines Experimentes mit zwei möglichen Ausgängen “Erfolg/Misserfolg” unter gleichen Bedingungen.

- Mendel’s Experiment: Jede Pflanze stellt eine Wiederholung dar (n=8023), und ein “Erfolg” ist eine mit grünen Keimblättern. Man hat also 2001 Erfolge.
- Meinungsumfragen: Die Befragung einer Person stellt eine Wiederholung dar (n=1154), und ein “Erfolg” ist eine “Ja”-Antwort. Im Beispiel 56% Befürworter, also etwa 646 Erfolge.

Formeln für die Binomialverteilung

Sei p die Wahrscheinlichkeit für einen Erfolg bei einer bestimmten Wiederholung.

$$p_n(k) = P[\text{ genau } k \text{ Erfolge }] = \binom{n}{k} p^k (1-p)^{n-k}.$$

Rekursiver Zusammenhang:

$$p_n(k) = \frac{n-k+1}{k} \frac{p}{1-p} \quad p_n(k-1).$$

Folgerung:

- $p_n(k)$ wachsend im Bereich $k \leq (n+1)p$, fallend im Bereich $k \geq (n+1)p$.
- Auf beiden Seiten des Maximums fällt $p_n(k)$ schneller als exponentiell ab.

Konzentration der Binomialverteilung

Untersuchung der Binomialverteilung für verschiedene n und p (siehe Figur) zeigt:
Die Wahrscheinlichkeiten $p_n(k)$ sind wesentlich verschieden von null nur in einem relativ kleinen Bereich um $k = np$ herum (relativ zur Anzahl $n+1$ der prinzipiell möglichen Werte).

Bestimme einen **Bereich von plausiblen Werten für die Anzahl Erfolge**

$$A(n,p) = \{k_1(n,p), k_1(n,p)+1, \ldots k_2(n,p)\},$$

dessen Wahrscheinlichkeit (mindestens) 95% beträgt.

Berechnung von $A(n,p)$

Wir schneiden dazu auf beiden Seiten der Verteilung je 2.5% ab:

$$\sum_{j=0}^{k_1(n,p)-1} p_n(j) \leq 0.025 < \sum_{j=0}^{k_1(n,p)} p_n(j)$$

bzw.

$$\sum_{j=k_2(n,p)+1}^{n} p_n(j) \leq 0.025 < \sum_{j=k_2(n,p)}^{n} p_n(j).$$

Aus Symmetriegründen

$$k_2(n,p) = n - k_1(n,1-p).$$

$A(n,p)$ ist der Annahmebereich für einen zweiseitigen Test zum Niveau 5%.

Beispiel: Daten von Mendel

Für $n = 8023$ und $p = 0.25$ ist $A(n, p) = \{1930, \ldots, 2082\}$.
2001 beobachtete Erfolge liegen klar innerhalb des plausiblen Bereichs, also kein Widerspruch zur Mendel'schen Theorie. Im Gegenteil, Abweichungen eher zu klein.

Vom Test zum Vertrauensintervall

Bei Meinungsumfragen entspricht p dem Anteil Befürworter unter allen amerikanischen Erwachsenen, und das will man ja gerade mit der Umfrage herausfinden !
Wenn die erhaltenen 646 Erfolge im plausiblen Bereich $A(n,p)$ liegen, können wir das unbekannte p einschränken auf

$$\begin{aligned} I(n,646) &= \{p \mid A(n,p) \text{ enthält } 646\} \\ &= \{p \mid k_1(n,p) \leq 646 \leq k_2(n,p)\}. \end{aligned}$$

Gleiches Vorgehen für einen anderen beobachteten Wert x.

Berechnung des Vertrauensintervalls

$I(n,x)$ enthält die **plausiblen Werten für die unbekannte Wahrscheinlichkeit** p. $I(n,x)$ heisst das **Vertrauensintervall** für p zum 95%-Niveau. (Illustration in Figur).

Grenzen des Vertrauensintervalls durch Auflösen der folgenden Gleichungen nach p:

$$k_1(n,p) = x, \quad k_2(n,p) = x.$$

also

$$\sum_{j=0}^{x-1} \binom{n}{j} p^j (1-p)^{n-j} = 0.025, \quad \sum_{j=x+1}^{n} \binom{n}{j} p^j (1-p)^{n-j} = 0.025.$$

Keine explizite Lösungen.

Interpretation des Vertrauensintervalls

Bei andern n Wiederholungen des Experiments (z.B. andere Personen befragt), ergäbe sich eine andere Anzahl x von Erfolgen, und damit auch ein anderes Vertrauensintervall. Das Vertrauensintervall ist also zufällig.
Wann liegt der unbekannte wahre Wert p nicht im Intervall ?
Genau, dann wenn die Anzahl Erfolge nicht in $A(n, p)$ liegt.
Nach Konstruktion geschieht das (höchstens) mit Wahrscheinlichkeit 0.05.
Das zufällige Vertrauensintervall enthält die unbekannte Wahrscheinlichkeit p in etwa 19 von 20 Fällen

Näherungen für grosses n

Für grosses n kann man $A(n,p)$ und $I(n,x)$ approximieren, indem man die Binomial- durch eine Normalverteilung ersetzt. Alternativ verfiziert man empirisch

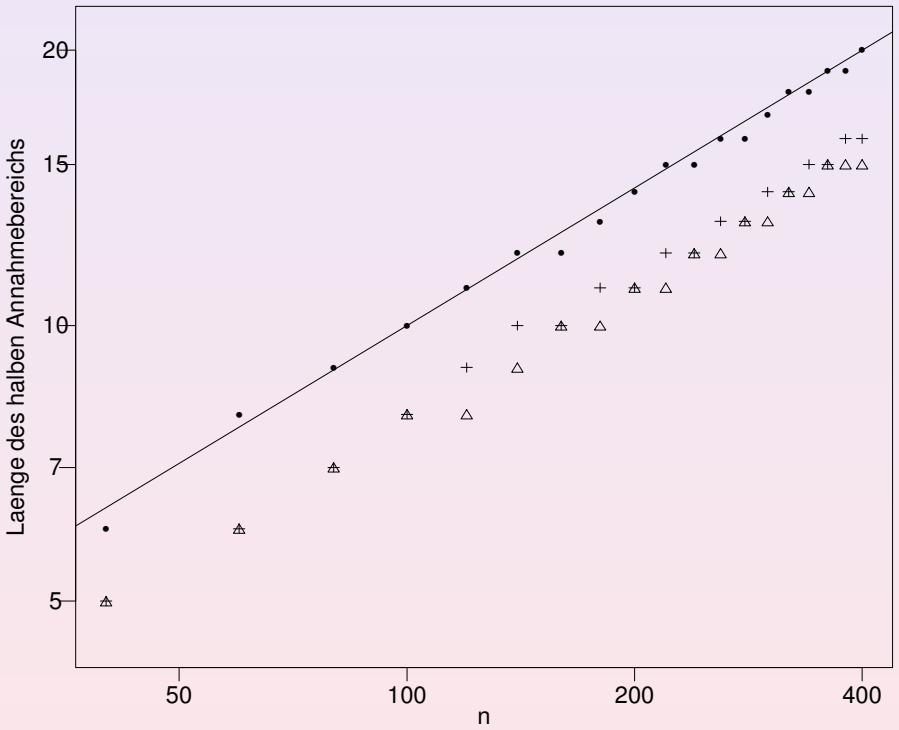
- $A(n,p)$ ist ungefähr symmetrisch um np.
- $np - k_1(n,p)$ ist am grössten für $p = 0.5$.
- $np - k_1(n,p) \approx const.\sqrt{n}$, und für $p = 0.5$ ist $const. \approx 1$.

Daraus folgt

$$I(n,x) \subseteq \left[\frac{x}{n} - \frac{1}{\sqrt{n}}, \frac{x}{n} + \frac{1}{\sqrt{n}}\right].$$

Illustration der Asymptotik

Länge des linken halben Annahmebereichs als Funktion von n.



Problemstellung

Quantifizierung des Zusammenhangs zwischen zwei Grössen X und Y auf Grund von n Messungen (x_i, y_i). Falls Zusammenhang approximativ linear, Bestimmung von Steigung und Achsenabschnitt. Beispiele

- $X =$ Umgebungstemperatur, $Y =$ Frequenz im Balzruf einer Froschart.
- $X =$ Gehalt eines Lebensmittelfarbstoffs, $Y =$ Fläche unter dem Gipfel im Chromatogramm.
- $X =$ Luftdruck, $Y =$ Siedepunkt von Wasser.
- etc.

Die Methode der Kleinsten Quadrate

Minimiere die Summe der quadrierten Abweichungen in y-Richtung

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

bezüglich a und b. Lösung

$$a_0 = \bar{y} - b_0\bar{x}, \quad b_0 = \frac{\sum_i x_i(y_i - \bar{y})}{\sum_i x_i(x_i - \bar{x})} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}.$$

Didaktische Herleitung über die Spezialfälle $b = 0$, bzw. $a = 0$.

Übergang zur Statistik

Naheliegende Fragen

- Wie genau ist die Steigung der Geraden bestimmt ?
- Wie genau sind Vorhersagen des Y-Werts mit Hilfe der Regressionsgeraden bei gegebenem Wert x ?
- Können wir entscheiden, ob der Zusammenhang zwischen den Variablen tatsächlich linear ist ?

Beantwortung mit Methoden der schliessenden Statistik. Dazu brauchen wir ein Modell, wie die Fehler zustande kommen.

Das lineare Modell

Annahme: Es gibt eine unbekannte wahre Beziehung

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

Bei gegebenem Wert x_i streut der Wert y_i zufällig um den Funktionswert $\alpha + \beta x_i$:

$$y_i = \alpha + \beta x_i + E_i.$$

Wiederholung der Messung ergäbe einen andern Wert für E_i.

Annahme: Verteilung der Abweichungen E_i gleich für alle Werte x_i.

Illustration I

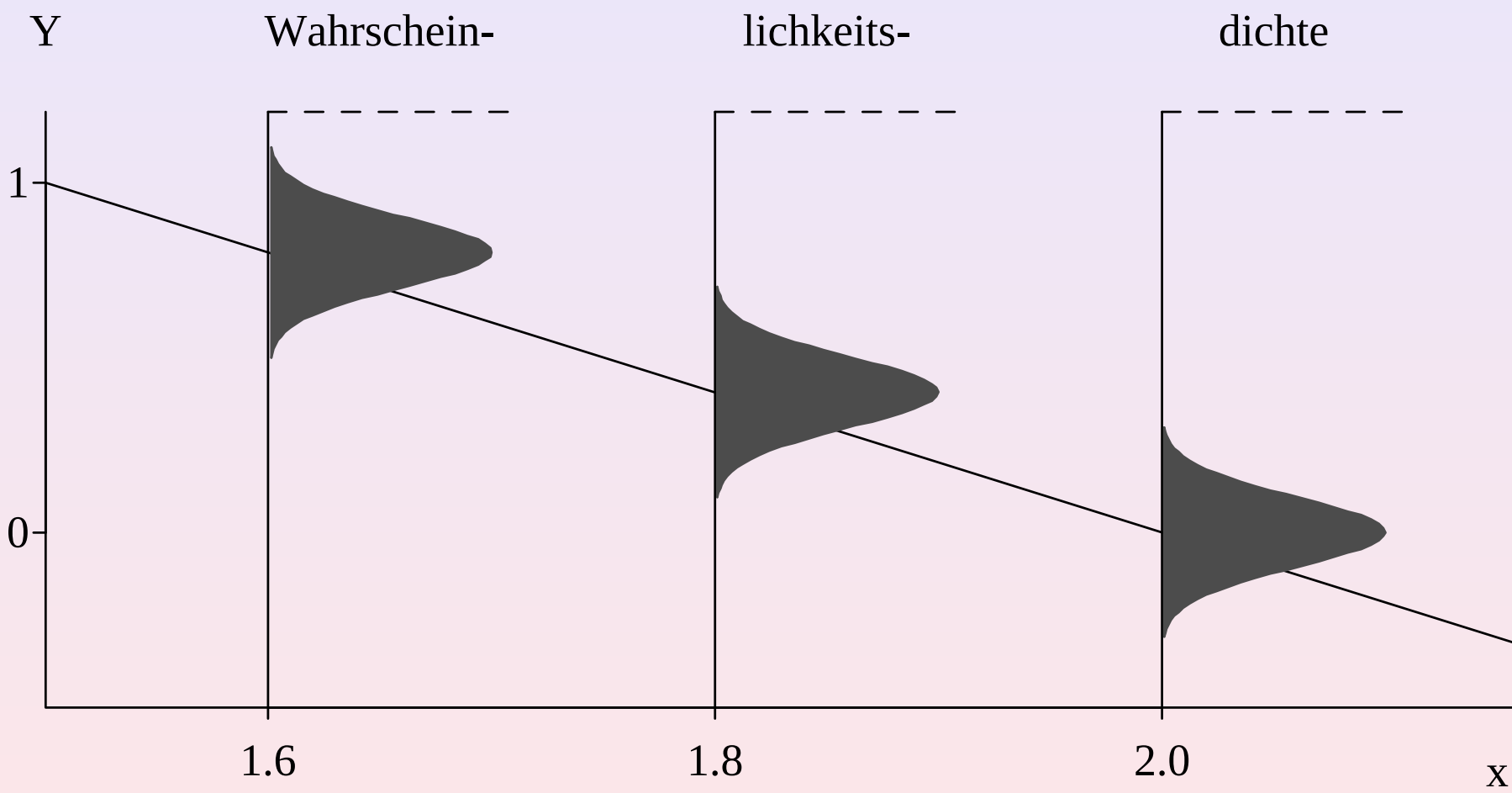


Illustration II

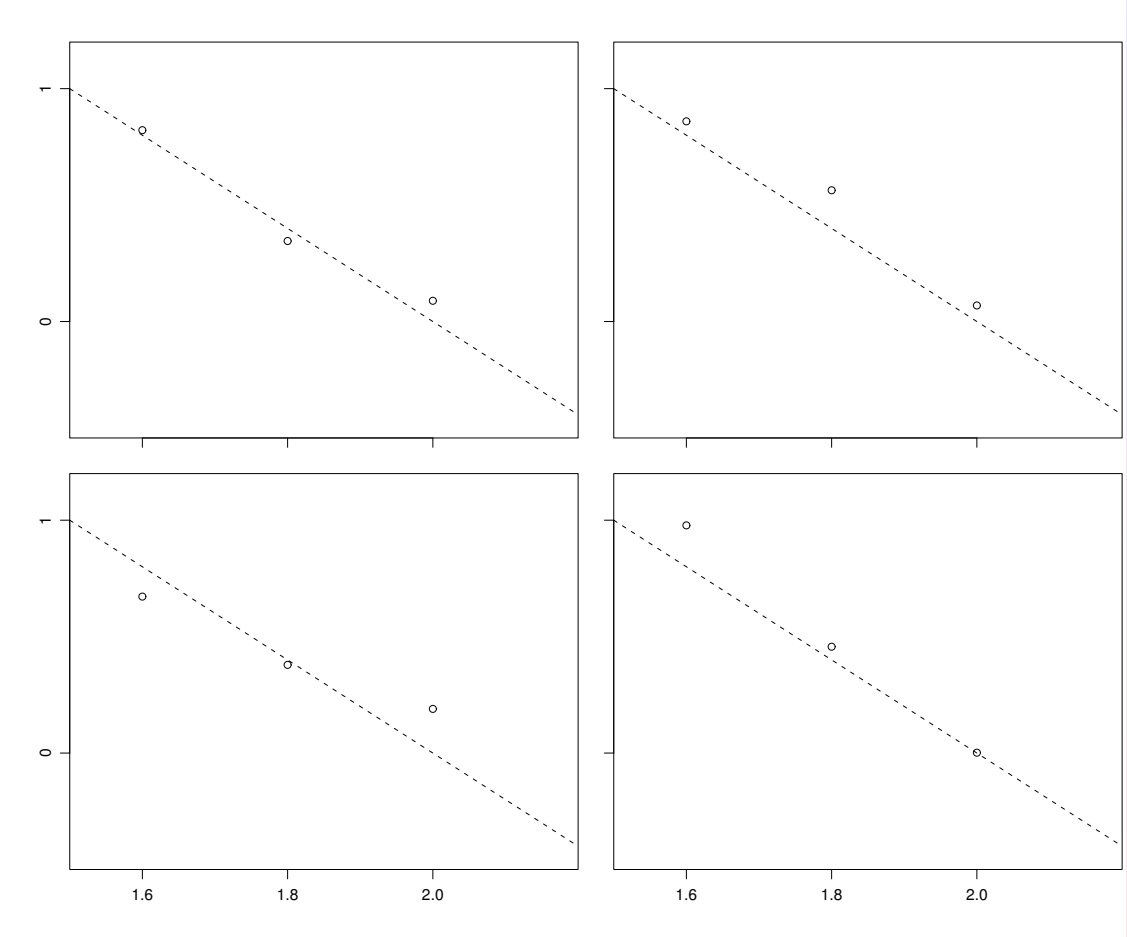
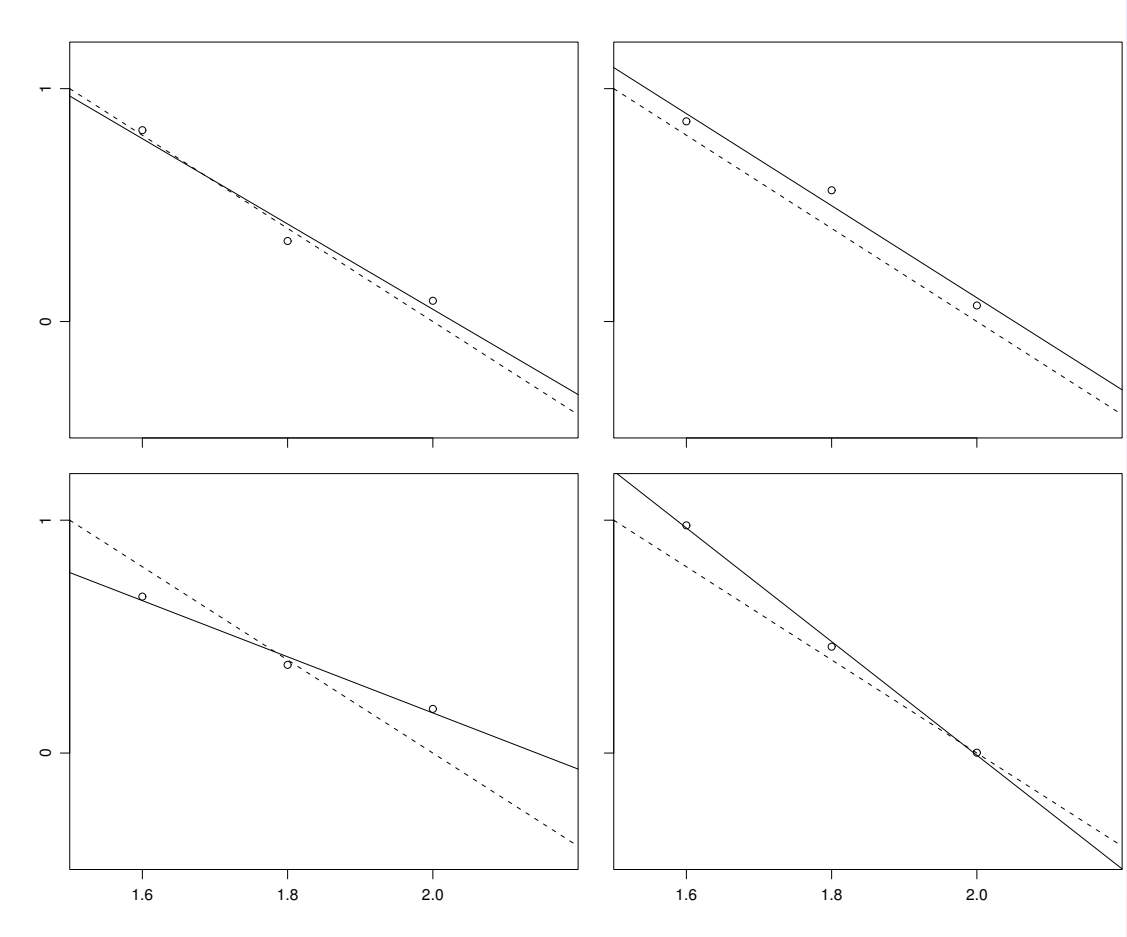


Illustration III



Verteilung von a_0, bzw. b_0

Wenn die y_i zufällig schwanken, dann tun dies auch die geschätzten Parameter a_0 und b_0. Herleitung der Verteilung von a_0 und b_0 zu kompliziert für die Mittelschule.

Ausweg: Simulation.

Hauptresultat

Wenn die Abweichungen E_i unabhängig und $\mathcal{N}(0, \sigma(E)^2)$-verteilt sind, dann kann man zeigen, dass die Kleinste-Quadrate-Schätzung b_0 ebenfalls normalverteilt ist, nämlich

$$b_0 \sim \mathcal{N}(\beta, \sigma(b_0)^2), \quad \sigma(b_0)^2 = \frac{\sigma(E)^2}{(n-1)s(x)^2}$$

Analoges Resultat auch für a_0.

Vertrauensintervall für die Steigung

Zunächst wird $\sigma(E)$ als bekannt vorausgesetzt.
Normalverteilung für b_0 impliziert, dass mit Wahrscheinlichkeit 95%

$$\beta - 1.96\frac{\sigma(E)}{\sqrt{n-1}\ s(x)} \leq b_0 \leq \beta + 1.96\frac{\sigma(E)}{\sqrt{n-1}\ s(x)}.$$

Auflösen nach β zeigt, dass das Intervall

$$b_0 \pm 1.96\frac{\sigma(E)}{\sqrt{n-1}\ s(x)}$$

die unbekannte wahre Steigung β mit Wahrscheinlichkeit 95% "einfängt".

Modifikation, wenn $\sigma(E)$ unbekannt

Schätze $\sigma(E)^2$ aus den Daten durch

$$s(E)^2 = \frac{1}{n-2} \sum_{i=1}^{n} (y_i - a_0 - b_0 x_i)^2$$

und berücksichtige die zusätzliche Unsicherheit durch eine eine andere, leicht grössere Konstante als 1.96.

Die richtige Konstante ist das 97.5%-Quantil der sogenannten t-Verteilung mit $n - 2$ Freiheitsgraden.

Die Rolle von Transformationen

Mit Hilfe von Transformationen der x_i und/oder der y_i lassen sich viele Beziehungen linearisieren:

$$\begin{aligned} y = \alpha x^{\beta} &\Leftrightarrow \log y = \log \alpha + \beta \log x \\ y = \alpha e^{\beta x} &\Leftrightarrow \log y = \log \alpha + \beta x. \end{aligned}$$

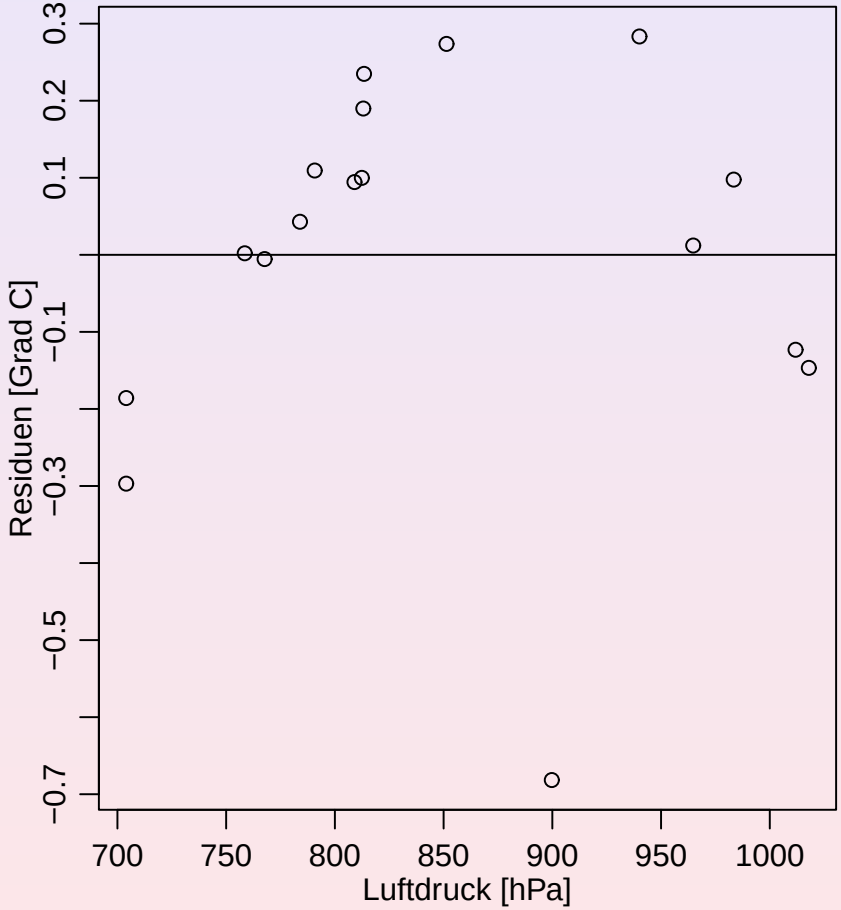
Additive Abweichungen vor, bzw. nach der Transformation sind nicht gleichwertig

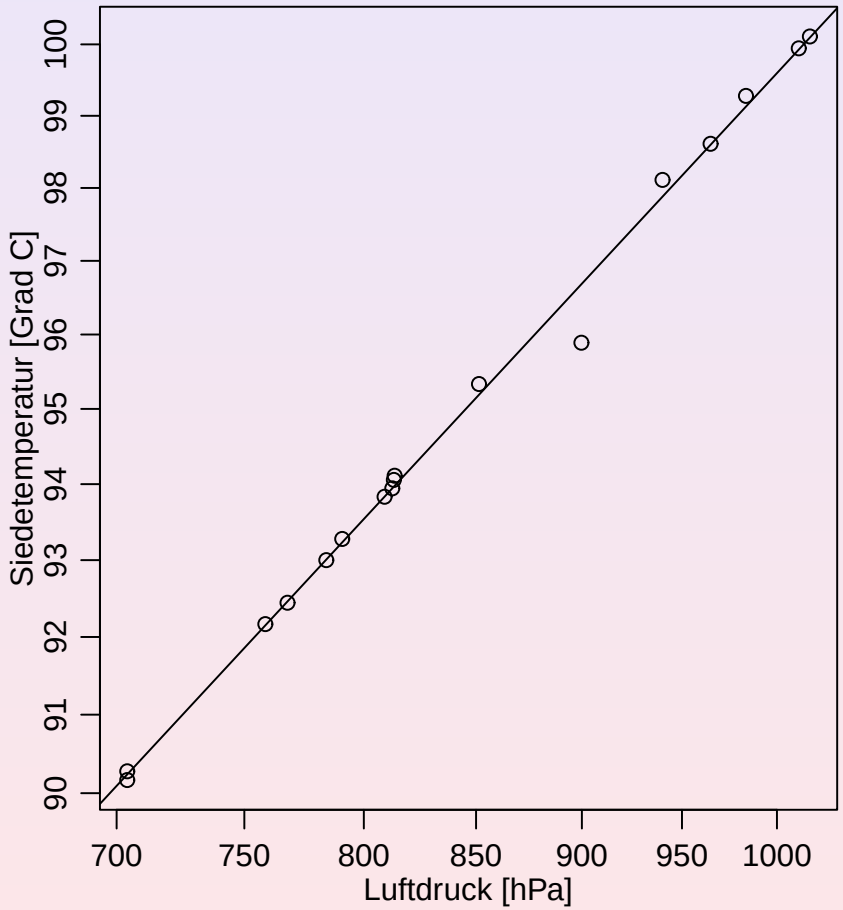
$$\log y_i = \log \alpha + \beta \log x_i + E_i \Leftrightarrow y_i = \alpha {x_i}^{\beta} \cdot \exp(E_i).$$

Linearisierung der Beziehung macht oft auch die Annahme von additiven Abweichungen gleicher Streuung plausibler.

Effekt von Transformationen I

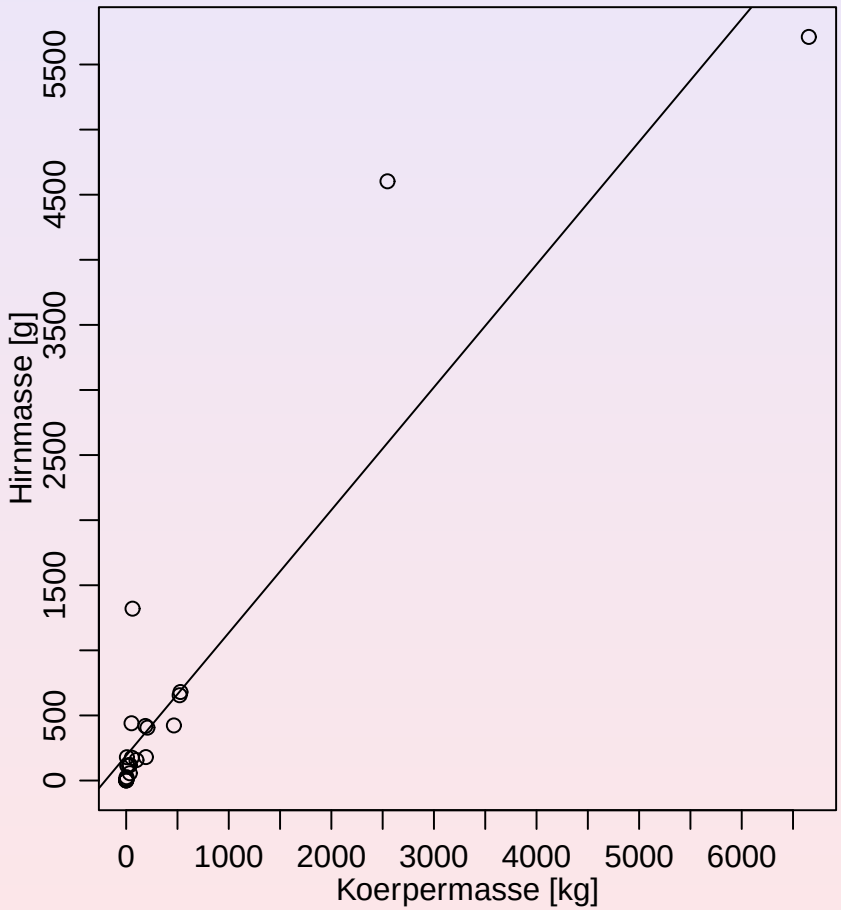
Luftdruck und Siedetemperatur von H_2O

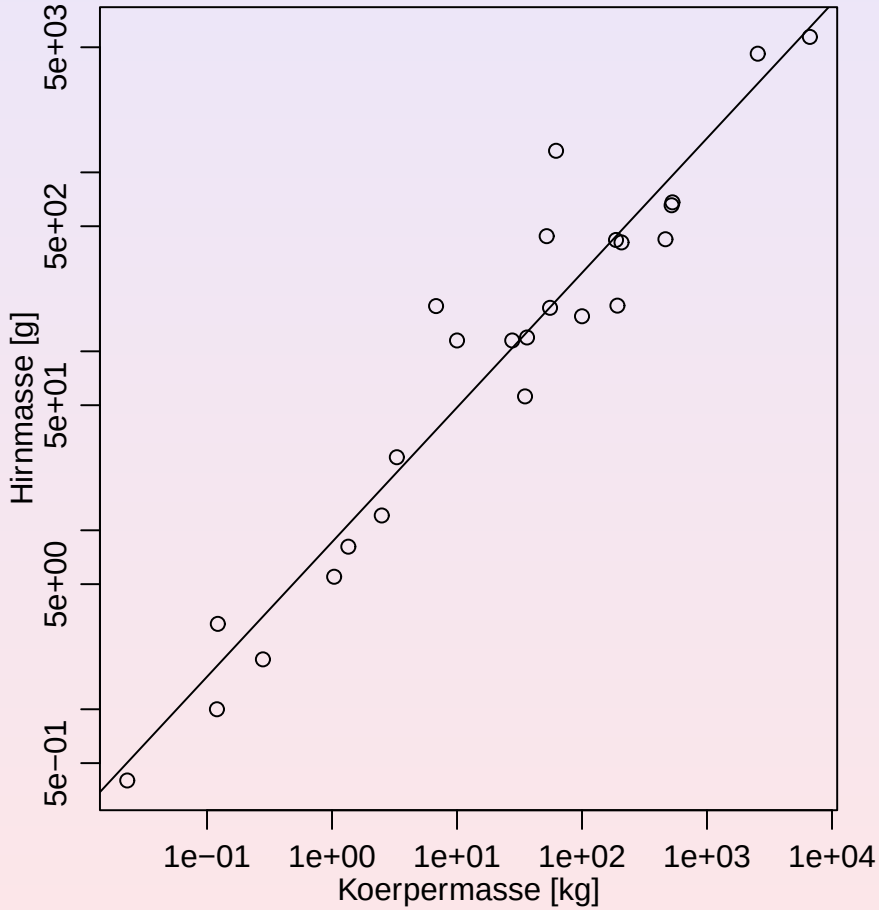




Effekt von Transformationen II

Körper- und Hirnmasse von Säugetieren





Regression als "Rückschritt zum Mittel"

Rechnung zeigt

$$\sum_{i=1}^{n}(y_i - a_0 - b_0 x_i)^2 = (n-1)s(y)^2\left(1 - r_{xy}^2\right).$$

Daraus folgt $|r_{xy}| < 1$, wenn die Punkte nicht auf einer Geraden liegen.

Damit sind bei der Regressionsgerade

$$\frac{y - \bar{y}}{s(y)} = r_{xy} \cdot \frac{x - \bar{x}}{s(x)}$$

die y-Werte näher beim Mittelwert als die x-Werte, wenn wir die Standardabweichungen als Einheit benutzen.