6 KORRELATION UND REGRESSION

Inhalt:

Lernziele:

1. Die Abhängigkeit von zwei mehrstufig skalierten Variablen mit dem χ^2-Test prüfen können.
2. Den Zusammenhang zwischen zwei 2-stufig skalierten Variablen mit dem Chancenverhältnis (Odds Ratio) schätzen können.
3. Den Korrelationskoeffizienten ρ als Parameter der 2-dimensionalen Normalverteilung interpretieren können.
4. Einen Schätzwert und ein Konfidenzintervall für den Korrelationskoeffizienten ρ bestimmen können.
5. Die Abhängigkeit der zweidimensional-normalverteilten Variablen X und Y mit einem geeigneten Test prüfen können.
6. Die Parameter der Regression von Y auf X im Modell A mit zweidimensional-normalverteilten Variablen schätzen und die Abhängigkeitsprüfung durchführen können.
7. Die Parameter der Regression von Y auf X im Modell B (mit zufallsgestörter linearer Regressionsfunktion) schätzen und die Abhängigkeitsprüfung durchführen können.
8. Linearisierende Transformationen anwenden können, um nichtlineare Abhängigkeiten (allometrische, exponentielle bzw. gebrochen lineare) mit Hilfe von linearen Regressionsmodellen erfassen zu können.
9. Regressionsgeraden durch den Nullpunkt bestimmen können.
10. Probenmesswerte mit Hilfe von linearen Kalibrationsfunktionen schätzen können.

6.1Korrelation bei mehrstufig skalierten Variablen

Lernziel 6.1:
Die Abhängigkeit von zwei mehrstufig skalierten Variablen mit dem χ^2-Test prüfen können.

Ablaufschema:

- Beobachtungsdaten und Modell:
  X, Y: diskrete Merkmale mit k ≥ 2 Werten $a_1, a_2, \ldots, a_k$ bzw. m ≥ 2 Werten $b_1, b_2, \ldots, b_m$;
  Beobachtung der Variablen an n Untersuchungseinheiten
  → n Wertepaare (a_i, b_j);
  → Zusammenfassen der k x m Häufigkeiten n_{ij} der Wertepaare (a_i, b_j) in einer (zweidimensionalen) Kontingenztafel (=Rechteckschema aus k Zeilen und m Spalten) :

	Werte von Y						
Werte von X	b_1	b_2	...	b_j	...	b_m	Σ
a_1	n_{11}	n_{12}	...	n_{1j}	...	n_{1m}	$n_{1.}$
a_2	n_{21}	n_{22}	...	n_{2j}	...	n_{2m}	$n_{2.}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
a_i	n_{i1}	n_{i2}	...	n_{ij}	...	n_{im}	$n_{i.}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
a_k	n_{k1}	n_{k2}	...	n_{kj}	...	n_{km}	$n_{k.}$
Σ	$n_{.1}$	$n_{.2}$	...	$n_{.j}$	...	$n_{.m}$	n

  Es seien $p_{i.}$ und $p_{.j}$ die Wahrscheinlichkeiten, dass X den Wert a_i bzw. Y den Wert b_j annimmt. Bei Unabhängigkeit von X und Y ist die Wahrscheinlichkeit der Wertekombination (a_i,b_j) durch $p_{ij}=p_{i.}p_{.j}$ und die erwartete Häufigkeit von Untersuchungseinheiten mit dieser Wertekombination durch $np_{i.}p_{.j}$ gegeben. Die erwartete Häufigkeit wird durch $E_{ij} = n_{i.}n_{.j}/n$ geschätzt.

- *Hypothesen und Testgröße:*
  H_0 : „*X* und *Y* sind unabhängig“ gegen H_1 : „*X* und *Y* sind abhängig“.
  Die Abweichung zwischen den beobachteten Häufigkeiten n_{ij} und den bei Unabhängigkeit von X und Y (Nullhypothese H_0) zu erwartenden Häufigkeiten E_{ij} wird mit der Goodness of fit-Statistik

(Chiquadrat-Summe)

$$GF_s = \sum_{i=1}^{k}\sum_{j=1}^{m} \frac{\left(n_{ij} - E_{ij}\right)^2}{E_{ij}} \text{ mit } E_{ij} = \frac{n_{i.} n_{.j}}{n}$$

gemessen. Wenn H_0 (Unabhängigkeit) zutrifft, kann GF_s bei "großem" Stichprobenumfang n als Realisierung einer näherungsweise χ^2-verteilten Zufallsvariablen mit f= (k-1)(m-1) Freiheitsgraden aufgefasst werden.[1]

- Entscheidung mit dem P-Wert:
  $P < \alpha \Rightarrow H_0$ ablehnen; dabei ist $P=1-F_f(GF_s)$ mit F_f als Verteilungsfunktion der χ^2-verteilten Zufallsvariablen mit f= (k-1)(m-1) Freiheitsgraden.

- Entscheidung mit dem Ablehnungsbereich:
  H_0 wird abgelehnt, wenn $GF_s > \chi^2_{f,1-\alpha}$; dabei bezeichnet $\chi^2_{f,1-\alpha}$ das $1-\alpha$ - Quantil der χ^2-Verteilung mit dem Freiheitsgrad f=n-2.

Hinweis:
Der χ^2-Test zur Prüfung der Abhängigkeit zweier Variablen X und Y zu kann formal auch zur Prüfung der Homogenität von Populationen bezüglich eines Merkmals X mit den Werten $a_1, a_2, \ldots, a_k$ verwendet werden. In diesem Fall hat Y die Bedeutung eines Gliederungsmerkmals mit den Werten $b_1, b_2, \ldots, b_m$, durch die die zu vergleichenden m ≥2 Populationen unterschieden werden. Man bezeichnet die Populationen als homogen bezüglich X, wenn die Wahrscheinlichkeiten, mit denen die X-Werte $a_1, a_2, \ldots, a_k$ auftreten, in allen Populationen im selben Verhältnis stehen. Um Abweichungen von der Homogenität zu prüfen, wird in der Nullhypothese angenommen, dass die Populationen homogen sind. Die rein technische Durchführung des Tests ist dieselbe wie bei der Abhängigkeitsprüfung.

Lernziel 6.2:
Den Zusammenhang zwischen zwei 2-stufig skalierten Variablen mit dem Chancenverhältnis (Odds Ratio) schätzen können.

Bei zwei zweistufig skalierten Variablen X und Y (k = m = 2) reduziert sich die k x m-Kontingenztafel auf eine sogenannte Vierfeldertafel mit den im Folgenden zusammengefassten Häufigkeiten n_{ij} und Wahrscheinlichkeiten p_{ij} der Merkmalskombinationen (a_i, b_j):

[1] Um den Approximationsfehler klein zu halten, wird bei Anwendung der Chiquadrat-Approximation verlangt, dass alle erwarteten Häufigkeiten $E_{ij} > 5$ sind.

X	Y: b_1	Y: b_2	Σ
a_1	n_{11}	n_{12}	$n_{1.}$
a_2	n_{21}	n_{22}	$n_{2.}$
Σ	$n_{.1}$	$n_{.2}$	n

X	Y: b_1	Y: b_2	Σ
a_1	p_{11}	p_{12}	$p_{1.}$
a_2	p_{21}	p_{22}	$p_{2.}$
Σ	$p_{.1}$	$p_{.2}$	1

Definition:

- Das Chancen-Verhältnis OR (auch relative Chance genannt, engl. odds ratio) der zweistufig skalierten variablen X und Y ist gleich dem Verhältnis

$$OR = \frac{p_{11} / p_{21}}{p_{12} / p_{22}} = \frac{p_{11} p_{22}}{p_{12} p_{21}}$$

der Chance des Ereignisses „$X=a_1$“ (gegen „$X=a_2$“) unter der Bedingung „$Y=b_1$“) zur Chance des Ereignisses „$X=a_1$“ (gegen „$X=a_2$“) unter der Bedingung „$Y=b_2$“).

- Eigenschaften:
  - Wenn X und Y unabhängig sind, gilt $p_{11}:p_{21} = p_{12}:p_{22}=p_{1.}:p_{2.}$, d.h., das Chancen-Verhältnis den Wert OR = 1 an.
  - Indem man für die Einzelwahrscheinlichkeiten p_{ij} die entsprechenden relativen Häufigkeiten n_{ij}/n einsetzt, erhält man die Schätzfunktion

$$\widehat{OR} = \frac{n_{11} n_{22}}{n_{12} n_{21}}$$

bzw. einen Schätzwert für OR, wenn unter den n_{ij} die konkret beobachteten Werte der Zellenhäufigkeiten verstanden werden.

  - Ein approximatives (1-α)-Konfidenzintervall für den (näherungsweise normalverteilten) Logarithmus von OR ist[2]:

$$\ln \frac{n_{11} n_{22}}{n_{12} n_{21}} \pm z_{1-\alpha/2} \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}$$

Durch Entlogarithmieren der Grenzen erhält man schließlich die entsprechenden Grenzen für OR.

Beispiel 6.1:

In einer Geburtenstation wurden von 50 Müttern unter 20 Jahren 28 Mädchen und 22 Knaben zur Welt gebracht. Von 70 Müttern über 20

[2] Schätzwert und Konfidenzintervall für das Chancenverhältnis können mit der R-Funktion oddsratio() - in Verbindung mit summary() und confint() - im Paket "vcd" (Visualizing Categorical Data) bestimmt werden.

Jahren gab es 37 Mädchen- und 33 Knabengeburten.
a) Man zeige, dass das Geschlecht der Kinder auf 5%igem Testniveau nicht vom Alter der Mütter abhängt.
b) Man beschreibe den Zusammenhang zwischen dem Geschlecht des Kindes und dem Alter der Mutter mit dem Chancen-Verhältnis und bestimme für diese Maßzahl ein 95%iges Konfidenzintervall.

Lösung mit R:

```
> # Beispiel 6.1 (Chiquadrat-Test, Schätzung des OR)
> # Dateneingabe
> nij <- matrix(c(28, 22, 37, 33), ncol=2,
+         dimnames=list(Geschlecht=c("maennl.", "weibl."),
+                       Alter=c("<20", ">=20"))); nij
          Alter
Geschlecht <20 >=20
   maennl.  28   37
   weibl.   22   33
> options(digits=4)
> # a) Abhängigkeitsprüfung
> # H0: "Geschlecht der Kinder hängt nicht vom Alter der Mütter ab"
> # gegen H1: ... hängt ab ...
> test <- chisq.test(nij, correct=F); test

        Pearson's Chi-squared test

data:  nij
X-squared = 0.116, df = 1, p-value = 0.7334

> test$expected # unter H0 zu erwartende Häufigkeiten
          Alter
Geschlecht   <20  >=20
   maennl. 27.08 37.92
   weibl.  22.92 32.08
> # b) Chancenverhältnis (Schätzwert, 95%-Konfidenzintwervall)
> OR <- (nij[1,1]/nij[2,1])/(nij[1,2]/nij[2,2])
> lnOR <- log(OR)
> alpha <- 0.05; zq <- qnorm(1-alpha/2)
> se_lnOR <- sqrt(1/nij[1,1]+1/nij[1,2]+1/nij[2,1]+1/nij[2,2])
> ug_lnOR <- lnOR-zq*se_lnOR; og_lnOR <- lnOR+zq*se_lnOR
> print(cbind(lnOR, se_lnOR, ug_lnOR, og_lnOR))
       lnOR se_lnOR ug_lnOR og_lnOR
[1,] 0.1268  0.3722 -0.6027  0.8562
> ug_OR <- exp(ug_lnOR); og_OR <- exp(og_lnOR)
> print(cbind(OR, ug_OR, og_OR))
        OR  ug_OR og_OR
[1,] 1.135 0.5474 2.354
> # OR mit R-Funktion oddsratio()
> library(vcd)
> or <- oddsratio(nij, log=F)
> summary(or); confint(or)
     Odds Ratio
[1,]       1.14
        lwr   upr
[1,] 0.5474 2.354
```

6.2 Korrelation bei metrischen Variablen

Lernziel 6.3:
Den Korrelationskoeffizienten ρ als Parameter der 2-dimensionalen Normalverteilung interpretieren können.

- Definition:
  X und *Y* heißen 2-dimensional normalverteilt mit den Mittelwerten μ_X, μ_Y, den Standardabweichungen $\sigma_X > 0$, $\sigma_Y > 0$ und dem Korrelationskoeffizienten ρ ($|\rho| < 1$), wenn sie mit Hilfe von 2 unabhängigen, N(0,1)-verteilten Zufallsvariablen Z_1, Z_2 wie folgt erzeugt werden können:

$$X = \sigma_X Z_1 + \mu_X, \quad Y = \sigma_Y \rho Z_1 + \sigma_Y \sqrt{1-\rho^2} Z_2 + \mu_Y$$

- Bezeichnungen:
  Im Falle ρ = 0 sind die Variablen X und Y nicht korreliert; sie variieren voneinander unabhängig. In den Fällen ρ = +1 oder ρ = -1 liegt eine perfekte(positive bzw. negative) Korrelation vor, d.h., die Variable X ist bis auf eine multiplikative (positive oder negative) Konstante gleich der Variablen Y.

- Standardform:
  Die Bedeutung des Parameters ρ = 0 kann man besser erkennen, wenn man in den Definitionsgleichungen die Variablen und X und Y durch die standardisierten Variablen $X' = (X-\mu_X)/\sigma_X$ bzw. $Y' = (Y-\mu_Y)/\sigma_Y$ ersetzt; es folgt:

$$X' = Z_1 \text{ und } Y' = \rho Z_1 + \sqrt{1-\rho^2} Z_2$$

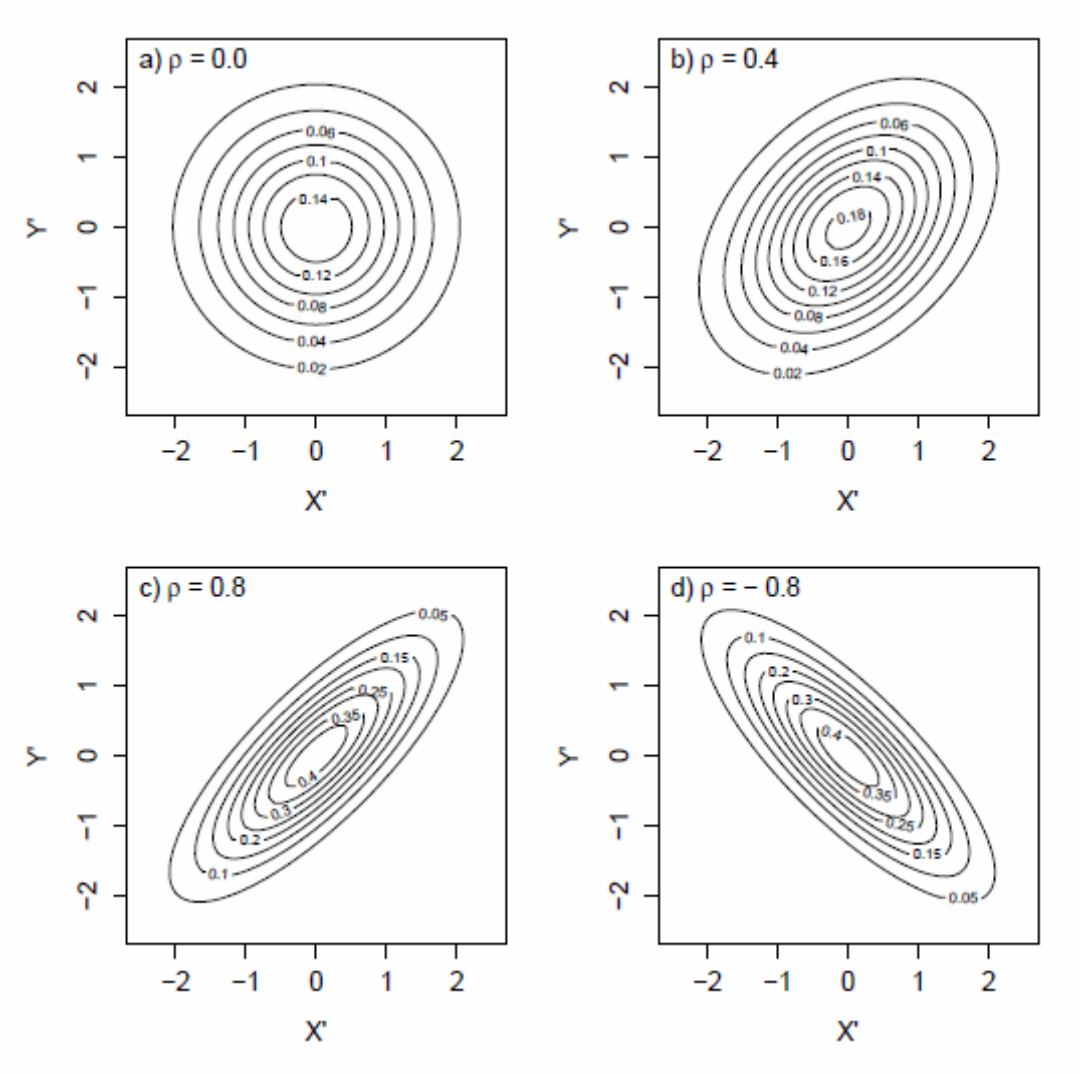
  Die gemeinsame Verteilung der standardisierten Variablen X´ und Y´ ist die Standardform der 2-dimensionalen Normalverteilung.

- Dichtefunktion der Standardform:
  rscheinlichkeitsdichte

$$z' = f_{X'Y'}(x', y') = \frac{1}{2\pi\sqrt{1-\rho}} \exp\left[-\frac{1}{2(1-\rho^2)} \left(x'^2 - 2\rho x' \acute{y} + y'^2 \right) \right]$$

  zu. Die grafische Darstellung der Dichtefunktion nehmen wir in einem aus den Merkmalsachsen (X', Y') und der Dichteachse (Z') aufgespannten dreidimensionalen, rechtwinkeligen Koordinatensystem vor. Der Graph von $f_{X'Y'}$ ist eine Fläche, die den

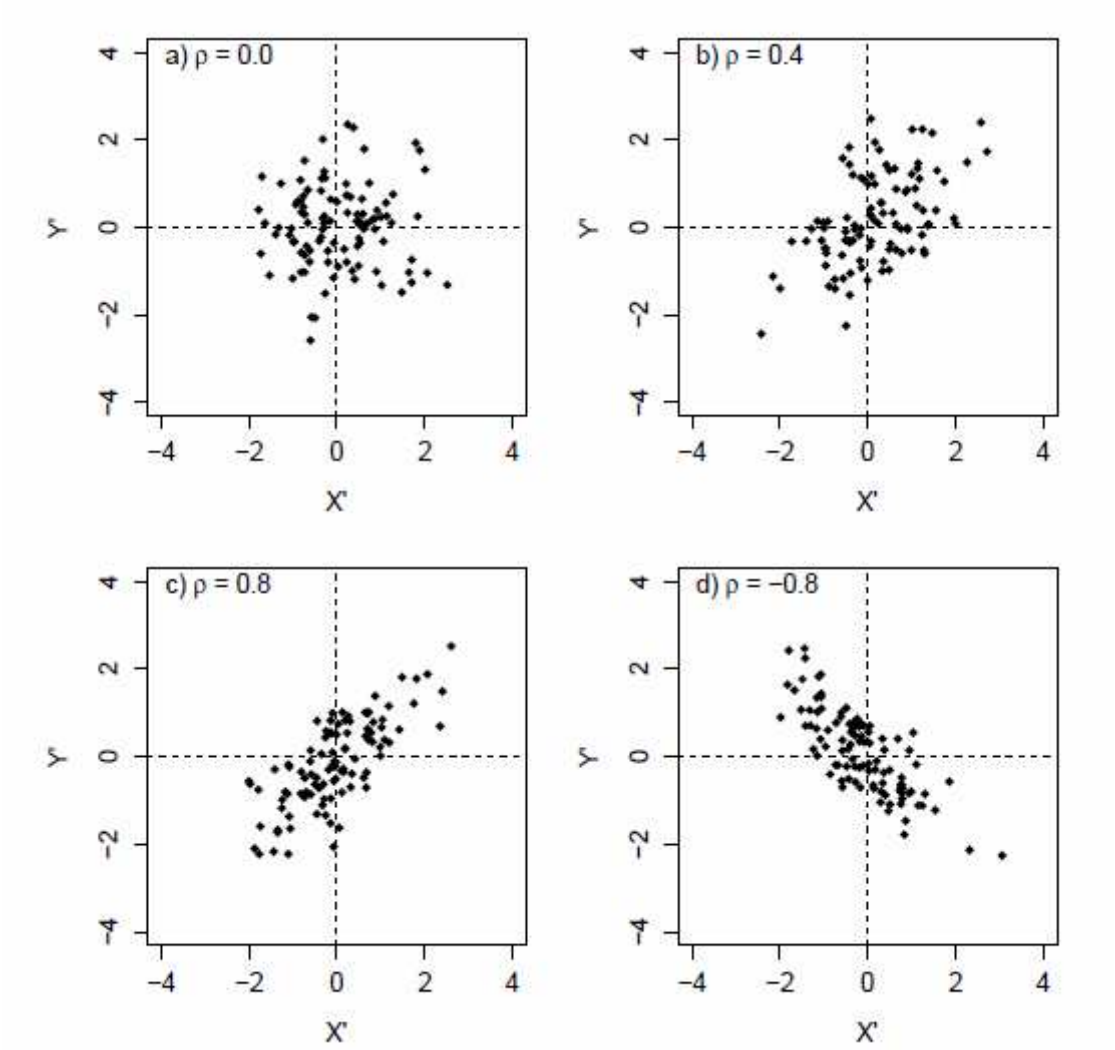
höchsten Wert an der Stelle x'=y'=0 annimmt und nach allen Seiten abfällt. Die Form der Dichtefläche hängt wesentlich vom Parameter ρ ab. Die folgende Grafik zeigt die Höhenlinien der Dichteflächen für verschiedene Korrelationskoeffzienten.



a) X' und Y' sind nicht korreliert, man hat eine Drehfläche von der Form einer "Glockenfläche"; in den Fällen b) und c) sind X' und Y' positiv korreliert und in der Folge die Dichteflächen in Richtung gleicher X'- und Y'-Werte gedehnt und normal dazu gestaucht. Die Interpretation der zweidimensionalen Dichte ist analog zur eindimensionalen Dichtefunktion vorzunehmen. Bezeichnet $\Delta x' \Delta y'$ den Inhalt eines (kleinen) Rechtecks um den Punkt (x',y') der Merkmalsebene, dann wird die Wahrscheinlichkeit, dass die Variablen X' und Y' einen Wert in diesem Rechteck annehmen, durch das Volumen $f_{X'Y'}(x', y')\Delta x'\Delta y'$ der über dem Rechteck errichteten "Säule" bis zur Dichtefläche dargestellt. Realisierungen von X' und Y' fallen also mit größerer Wahrscheinlichkeit in Bereiche mit hohen Dichtewerten als in Bereiche mit niedrigen Dichtewerten. Der Inhalt

des gesamten unter der Dichtefläche liegenden Körpers ist auf den Wert 1 normiert.

- Visualisierung im Streudiagramm:
  Die folgende Grafik zeigt Streudiagramme von Zufallsstichproben (n=100) aus zweidimensional-normalverteilten Grundgesamtheiten.



a) $\rho_{XY} = 0$: X' und Y' sind nicht korreliert, die 100 Punkte streuen regellos in horizontaler und vertikaler Richtung. b) $\rho_{XY} = 0.4$: X' und Y' sind positiv korreliert, die Punktewolke zeigt eine erkennbare lineare Tendenz in dem Sinne, dass größere (kleinere) X'-Werte mit größeren (kleineren) Y'-Werten gepaart sind. c) $\rho_{XY} = 0.8$: Wegen der stärkeren positiven Korrelation ist die lineare Ausformung der Punkteverteilung deutlicher als im Falle $\rho_{XY} = 0.4$. d) $\rho_{XY} = -0.8$: X' und Y' sind negativ korreliert, die Punktewolke weist eine fallende linearеTendenz auf; größere (kleinere) X'-Werte sind nun mit kleineren (größeren) Y'-Werten gepaart.

Beispiel 6.2:

R-Skript zur Erzeugung der Streudiagramme:

```
par(mfrow=c(2, 2))
par(pin=c(6, 4), mai=c(0.8, 0.9, 0.2, 0.1))
par(cex.axis=1.2, cex.lab=1.2)
# 2-dimensionale Normalverteilung, rho=0
x <- rnorm(100, 0, 1); y <- rnorm(100, 0, 1)
plot(x, y, type="p", col="black", pch=18, xlab="X'", ylab="Y'",
  xlim=c(-4, 4), ylim=c(-4, 4), frame.plot=T)
abline(h=0, lty=2); abline(v=0, lty=2)
text(-4.2, 3.8, col="black", expression("a) "*rho*"=0.0"), pos=4, cex=1.2)
# 2-dimensionale Normalverteilung, rho=0.4
rho <- 0.4
x <- rnorm(100, 0, 1); y <- rho*x+sqrt(1-rho^2)*rnorm(100, 0, 1)
plot(x, y, type="p", col="black", pch=18, xlab="X'", ylab="Y'",
  xlim=c(-4, 4), ylim=c(-4, 4), frame.plot=T)
abline(h=0, lty=2); abline(v=0, lty=2)
text(-4.2, 3.8, col="black", expression("b) "*rho*"=0.4"), pos=4, cex=1.2)
# 2-dimensionale Normalverteilung, rho=0.8
rho <- 0.8
x <- rnorm(100, 0, 1); y <- rho*x+sqrt(1-rho^2)*rnorm(100, 0, 1)
plot(x, y, type="p", col="black", pch=18, xlab="X'", ylab="Y'",
  xlim=c(-4, 4), ylim=c(-4, 4), frame.plot=T)
abline(h=0, lty=2); abline(v=0, lty=2)
text(-4.2, 3.8, col="black", expression("c) "*rho*"=0.8"), pos=4, cex=1.2)
# 2-dimensionale Normalverteilung, rho=-0.8
rho <- -0.8
x <- rnorm(100, 0, 1); y <- rho*x+sqrt(1-rho^2)*rnorm(100, 0, 1)
plot(x, y, type="p", col="black", pch=18, xlab="X'", ylab="Y'",
  xlim=c(-4, 4), ylim=c(-4, 4), frame.plot=T)
abline(h=0, lty=2); abline(v=0, lty=2)
text(-4.2, 3.8, col="black", expression("d) "*rho*"=-0.8"), pos=4, cex=1.2)
```

Lernziel 6.4:
Einen Schätzwert und ein Konfidenzintervall für den Korrelationskoeffizienten ρ bestimmen können.

- Definitionen:
  Es sei (x_i,y_i) (i=1,2,...,n) eine 2-dimensionale Zufallsstichprobe der 2-dimensional normalverteilten Zufallsvariablen X und Y. Dann bezeichnet man

  $$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) \quad \text{und} \quad r_{xy} = \frac{s_{xy}}{s_x s_y}$$

  als Kovarianz bzw. Produktmomentkorrelation (oder Pearson-Korrelation) der X- und Y-Stichprobe.

- Eigenschaften der Produktmomentkorrelation:
  - Es gilt $-1 \le r_{XY} \le +1$.
  - r_{XY} (kurz r) ist die klassische Schätzfunktion für ρ. Die Verteilung von r ist kompliziert. Wendet man auf r die Fisher-Transformation

$$Z : r \to Z(r) = \frac{1}{2}\ln\frac{1+r}{1-r}$$

- an, so ist - wenn ρ nicht zu nahe bei -1 oder +1 liegt - die neue Variable Z bereits für kleine n näherungsweise normalverteilt mit den Parametern

$$\mu_Z \approx \frac{1}{2}\ln\frac{1+\rho}{1-\rho} + \frac{\rho}{2(n-1)} \quad \text{und} \quad \sigma_Z^2 \approx \frac{1}{n-3}.$$

- approximatives (1-α)-Konfidenzintervall $[z_u, z_o]$ für $Z(\rho)=\mu_Z-\rho/[2(n-1)]$ mit

$$z_u = Z(r) - \frac{r}{2(n-1)} - z_{1-\alpha/2}\sigma_Z \quad \text{und} \quad z_o = Z(r) - \frac{r}{2(n-1)} + z_{1-\alpha/2}\sigma_Z$$

- Rücktransformation von der Z- auf die r-Skala → (1-α)-Konfidenzintervall für ρ

$$\left[\frac{\exp(2z_u)-1}{\exp(2z_u)+1}, \frac{\exp(2z_o)-1}{\exp(2z_o)+1}\right]$$

Lernziel 6.5:
Die Abhängigkeit der zweidimensional-normalverteilten Variablen X und Y mit einem geeigneten Test prüfen können.

Ablaufschema:

- Beobachtungsdaten und Modell:
  Die Variation der Variablen X und Y wird durch eine zweidimensionale Normalverteilung mit dem Korrelationsparameter ρ beschrieben. Von X und Y liegt eine zweidimensionale Zufallsstichprobe vor, die aus den an n Untersuchungseinheiten beobachteten Wertepaaren (x_i, y_i) (i=1,2, … ,n) besteht. Der Verteilungsparameter ρ wird mit der aus den Beobachtungswerten bestimmten Produktmomentkorrelation r geschätzt.

- Hypothesen und Testgröße:
  Der Vergleich des Parameters ρ mit dem Wert null (dieser Wert entspricht dem Fall zweier unabhängiger Variablen X und Y) erfolgt nach einer der folgenden Testvarianten:

  $H_0 : \rho = 0$ gegen $H_1 : \rho \neq 0$ (Variante II)
  $H_0 : \rho \leq 0$ gegen $H_1 : \rho > 0$ (Variante Ia),

$H_0 : \rho \geq 0$ gegen $H_1 : \rho < 0$ (Variante Ib)

Als Testgröße wird die Stichprobenfunktion

$$TG = \frac{r_{XY}\sqrt{n-2}}{\sqrt{1-r_{XY}^2}} \sim t_{n-2}$$

verwendet, die unter H_0: $\rho = 0$ einer t-Verteilung mit n-2 Freiheitsgraden folgt. Berechnet man r mit den konkreten Stichprobenwerten ein, erhält man die Realisierung TG_s der Testgröße.

- Entscheidung mit dem P-Wert:
  $P < \alpha \Rightarrow H_0$ ablehnen; dabei ist
  $P=2F_{n-2}(-|TG_s|)$ für die zweiseitige Testvariante II,
  $P=1-F_{n-2}(TG_s)$ für die Testvariante Ia bzw.
  $P=F_{n-2}(TG_s)$ für die Variante Ib;
  F_{n-2} bezeichnet die Verteilungsfunktion der t_{n-2}-Verteilung.[3]

- Entscheidung mit dem Ablehnungsbereich:
  H_0 wird abgelehnt, wenn
  $|TG_s| > t_{n-2,1-\alpha/2}$ (Variante II) bzw.
  $TG_s > t_{n-2-1,\ 1-\alpha}$ (Variante Ia) bzw.
  $TG_s < -\ t_{n-n,\ 1-\alpha}$ (Variante Ib) gilt;
  Dabei bezeichnet $t_{n-2,\gamma}$ das γ-Quantil der t-Verteilung mit dem Freiheitsgrad f=n-2.

Beispiel 6.3:

An 27 Leukämiepatienten wurden die in der folgenden Tabelle angeführten Expressionswerte der Gene A (Variable X) und B (Variable Y) ermittelt.[4] Man bestimme unter der Annahme, dass X und Y zweidimensional-normalverteilt sind,
a) einen Schätzwert und ein 95%iges Konfidenzintervall für die Produktmomentkorrelation ρ und zeige
b) auf 5%igem Signifikanzniveau, dass $\rho \neq 0$ ist.

```
x:   0.194, -0.011,  0.270, -0.248, -0.391,  0.005, -0.027,  0.363,
     -0.195, -0.123, -0.056, -0.138, -0.436,  0.002,
```

[3] Für die Abhängigkeitsprüfung mit der Produktmomentkorrelation ρ steht in R die Funktion cor.test() mit der Parametersetzung method="pearson" zur Verfügung. Neben dem P-Wert wird mit dieser Funktion auch der Schätzwert r und ein approximatives Konfidenzintervall für ρ auf der Grundlage der Fisher-Transformation berechnet.

[4] Die Stichproben sind dem Datensatz „golub“ im Paket „multtest“ aus der Software-Sammlung „bioconductor“ entnommen und betreffen die Gene mit den Bezeichnungen „M81830_at“ bzw. „U58048_at“ von 27 Leukämiepatienten der Tumorklasse 0 (vgl. http://www.bioconductor.org/.

```
      -0.532, 0.211,  0.192,  0.473, -0.188, -0.066, -0.702,  0.922,
      -0.382, -0.076, -0.250,  0.276,  0.764
Y:    0.564,  0.295,  0.817,  0.530,  0.388,  0.051,  0.908, 0.604,
      0.377,  0.717,  0.626,  0.165,  0.519,  0.530, 0.389,  0.495,
      0.872,  0.471,  0.656,  0.500,  0.014, 0.893,  0.158,  0.613,
      0.236,  0.756,  0.702
```

Lösung mit R:[5]

```
> # Beispiel 6.2 (Schätzung der POroduktmomentkorrelation)
> x <- c(0.194, -0.011,  0.270, -0.248, -0.391,  0.005, -0.027,  0.363,
+       -0.195, -0.123, -0.056, -0.138, -0.436,  0.002, -0.532,  0.211,
+        0.192,  0.473, -0.188, -0.066, -0.702,  0.922, -0.382, -0.076,
+       -0.250,  0.276,  0.764)
> y <- c(0.564, 0.295, 0.817, 0.530, 0.388, 0.051, 0.908, 0.604, 0.377,
+        0.717, 0.626, 0.165, 0.519, 0.530, 0.389, 0.495, 0.872, 0.471,
+        0.656, 0.500, 0.014, 0.893, 0.158, 0.613, 0.236, 0.756, 0.702)
> options(digits=4)
> n <- length(x)
> mw_x <- mean(x); mw_y <- mean(y)
> s_x <- sd(x); s_y <- sd(y)
> print(cbind(n, mw_x, s_x)); print(cbind(n, mw_y, s_y))
      n      mw_x    s_x
[1,] 27 -0.005519 0.3704
      n   mw_y    s_y
[1,] 27 0.5128 0.2479
> # bivariate Statistiken
> s_xy <- cov(x, y) # Kovarianz
> r_xy <- cor(x, y, method="pearson") # Produktmoment(=Pearson)korrelation
> print(cbind(s_xy, r_xy))
        s_xy   r_xy
[1,] 0.05362 0.5839
> # Konfidenzintervall und Abhängigkeitsprüfung mit cor.test()
> cor.test(x, y, method="pearson",alternative="two.sided", conf.level=0.95)

        Pearson's product-moment correlation

data:  x and y
t = 3.596, df = 25, p-value = 0.001386
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.2620 0.7889
sample estimates:
   cor
0.5839

> # manuelle Berechnung (Konfidenzintervall, P-Wert)
> sxy <- sum((x-mw_x)*(y-mw_y))/(n-1); rxy <- sxy/s_x/s_y
> muz <- 0.5*log((1+rxy)/(1-rxy))+rxy/2/(n-1); sz <- 1/sqrt(n-3)
> print(cbind(sxy, rxy, muz, sz))
        sxy    rxy    muz     sz
[1,] 0.05362 0.5839 0.6796 0.2041
> alpha <- 0.05; zq <- qnorm(1-alpha/2)
> xx <- muz-rxy/(n-1); zu <- xx-zq*sz; zo <- xx+zq*sz
> rhou <- (exp(2*zu)-1)/(exp(2*zu)+1) # untere Grenze für rho
> rhoo <- (exp(2*zo)-1)/(exp(2*zo)+1) # obere Grenze für rho
> print(cbind(zu, zo, rhou, rhoo))
        zu    zo   rhou   rhoo
[1,] 0.257 1.057 0.2515 0.7846
```

[5] Die in der R-Prozedur *cor.test* berechneten Grenzen des Konfidenzintervalls für ρ weichen von den manuell bestimmten Grenzen ab. Die Abweichung ist dadurch bedingt, dass in R bei der Bestimmung von z_u bzw. z_o der Term $r_{XY}/[2(n-1)]$ vernachlässigt wird, was bei hinreichend großem n vertretbar ist.

6.3 Einfache lineare Regression

Lernziel 6.6:
Die Parameter der Regression von Y auf X im Modell A mit zweidimensional-normalverteilten Variablen schätzen und die Abhängigkeitsprüfung durchführen können.

Ablaufschema:

- Beobachtungsdaten und Modell:
  Beobachtung der Variablen X und Y an n Untersuchungseinheiten ergibt n Wertepaare $(x_1, y_1), (x_2, y_2), \ldots (x_n, y_n)$;
  Aus der Definitionsgleichungen

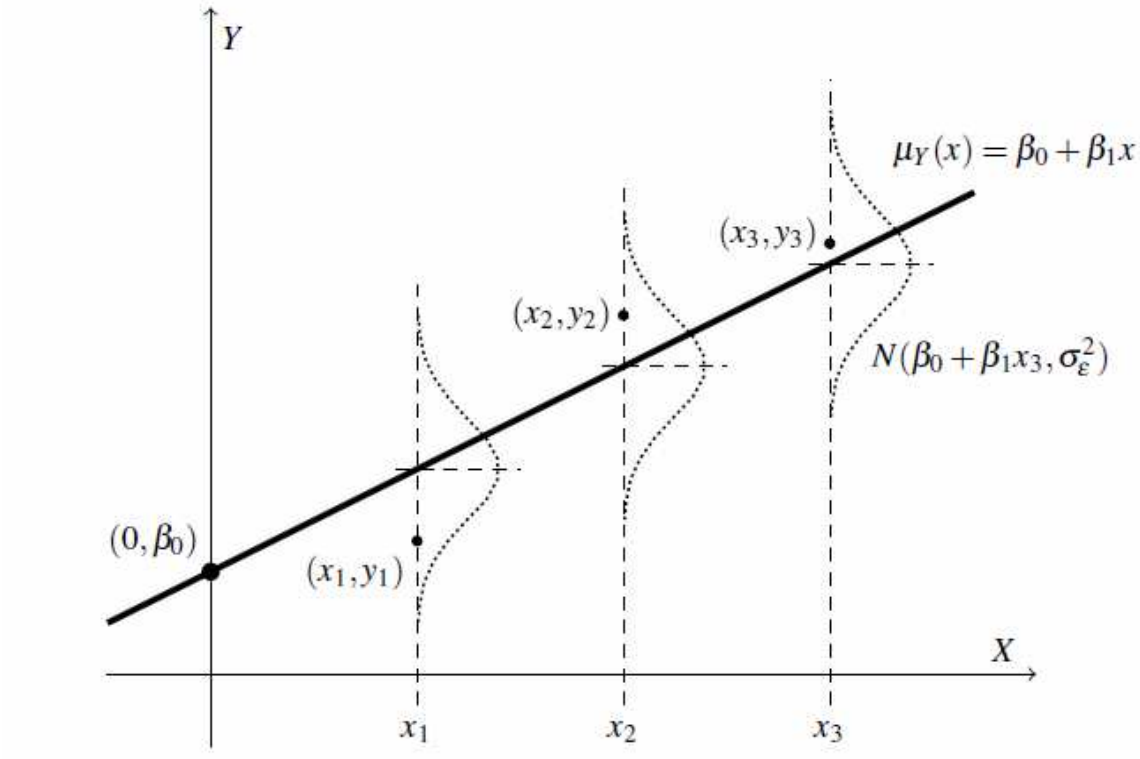
  $$X = \sigma_X Z_1 + \mu_X, \quad Y = \sigma_Y \rho Z_1 + \sigma_Y \sqrt{1-\rho^2}\, Z_2 + \mu_Y$$

  der zweidimensionalen Normalverteilung mit den Parametern μ_X, σ_X, μ_Y, σ_Y und ρ folgt für die Abhängigkeit der Variablen Y von X das lineare Modell:

  $$Y = \beta_0 + \beta_1 X + \varepsilon \quad \text{mit} \quad \beta_1 = \rho_{XY} \frac{\sigma_Y}{\sigma_X}, \quad \beta_0 = \mu_Y - \beta_1 \mu_X \quad \text{und}$$

  $$\varepsilon = \sigma_Y \sqrt{1-\rho^2}\, Z_2 \sim N(0, \sigma_\varepsilon^2), \quad \sigma_\varepsilon^2 = \sigma_Y^2 \left(1-\rho^2\right)$$

  Für jeden festen Wert x von X ist Y normalverteilt mit dem Mittelwert $\mu_Y(x)=\beta_0 + \beta_1 x$ und der von x unabhängigen Varianz σ_ε^2.



Die Funktion $x \mapsto \mu_Y(x)$ heißt lineare Regressionsfunktion (von Y auf X).[6]

- <u>Parameterschätzung und Abhängigkeitsprüfung:</u>
  - Schätzwerte für die Modellparameter β_1, β_0, das (von *X* abhängige) Zielgrößenmittel $\hat{y}(x)$ und die Varianz σ_ε^2:

$$b_1 = \hat{\beta}_1 = r_{XY}\frac{s_Y}{s_X}, \quad b_0 = \hat{\beta}_0 = \bar{y} - b_1\bar{x},$$

$$\hat{y}(x) = b_0 + b_1 x = \bar{y} + b_1(x - \bar{x}),$$

$$MQE = \hat{\sigma}_\varepsilon^2 = \frac{1}{n-2}\sum_{i=1}^{n} e_i^2 = \frac{1}{n-2}\sum_{i=1}^{n}\left(y_i - \hat{y}(x_i)\right)^2$$

    Die Funktion $x \mapsto \hat{\mu}(x)$ heißt empirische Regressionsfunktion, ihr Graph empirische Regressionsgerade.
  - (1-α)-Konfidenzintervall für den Geradenanstieg β_1:

$$b_1 \pm t_{n-2,1-\alpha/2} SE(b_1) = b_1 \pm t_{n-2,1-\alpha/2}\sqrt{\frac{MQE}{(n-1)s_X^2}}$$

    Offensichtlich hängt die Zielgröße *Y* im Rahmen des einfach linearen Regressionsmodells von der Einflussgröße *X* ab, wenn der Geradenanstieg $\beta_1 \neq 0$ ist. Bei einem vorgegebenen Irrtumsrisiko α lautet die Entscheidung auf $\beta_1 \neq 0$, wenn das (1- α)-Konfidenzintervall für β_1 die null nicht enthält. Gleichwertig mit der Prüfung H_0: $\beta_1 = 0$ gegen H_1: $\beta_1 \neq 0$ ist die Prüfung auf Abhängigkeit mit dem Korrelationskoeffizienten, d.h. die Prüfung der Hypothesen H_0: $\rho_{XY} = 0$ vs. H_1: $\rho_{XY} \neq 0$.
  - (1-α)-Konfidenzintervall für das Zielgrößenmittel $\mu_{Y(x)}$ an der Stelle *x*:

$$\hat{y}(x) \pm t_{n-2,1-\alpha/2} SE(\hat{y}) = \hat{y}(x) \pm t_{n-2,1-\alpha/2}\sqrt{MQE\left(\frac{1}{n} + \frac{(x-\bar{x})^2}{(n-1)s_X^2}\right)}$$

- <u>Anpassungsgüte:</u>
  Es empfiehlt sich, nach Schätzung der Regressionsparameter die Regressionsgerade gemeinsam mit den Datenpunkten in ein Streudiagramm einzuzeichnen. Auf diese Weise gewinnt man eine Vorstellung, wie "gut" die Punkteverteilung durch die

[6] Die unabhängige Variable X wird auch Einflussgröße oder Regressor, die abhängige Variable Y auch Zielgröße oder Regressand genannt.

Regressionsgerade wiedergegeben wird. Eine Kennzahl füre die Anpassungsgüte ist das Bestimmtheitsmaß:

$$B = r_{XY}^2 = \left(\frac{s_{XY}}{s_X s_Y} \right)^2 = \frac{SQY - SQE}{SQY},$$

$$SQY = (n-1)s_Y^2, \quad SQE = \sum_{i=1}^{n} e_i^2 = SQY(1 - r_{XY}^2)$$

Eigenschaften von B:

- Es gilt: $0 \leq B \leq 1$.
- *B* ist der Anteil der durch *X* erklärten Variation von *Y.*

Beispiel 6.4:

In einer Studie wurden u.a. die Serumkonzentrationen X und Y der Na- bzw. Cl-Ionen (in mmol/l) von n=15 Probanden bestimmt. Die Messwerte sind:

```
X:   135.0, 147.0, 148.5, 130.0, 139.0, 129.0, 142.0, 146.0,
     131.0, 143.5, 138.5, 145.0, 143.0, 153.0, 149.0
Y:    99.0, 106.5, 105.5,  94.0,  98.0,  92.0,  97.0, 106.0,
     102.5,  98.5, 105.0, 103.0, 101.0, 107.0, 104.0
```

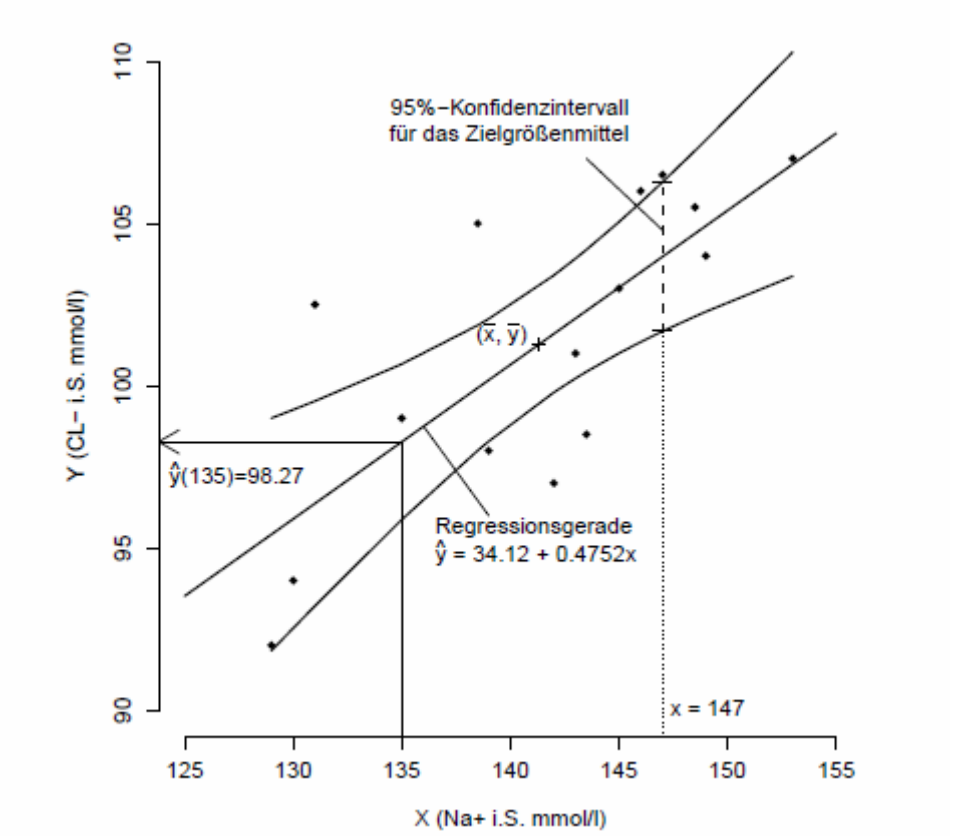
Man bestimme unter der Voraussetzung einer zweidimensional-normalverteilten Grundgesamtheit
a) die Parameter der Regressionsgeraden (von Y auf X),
b) die Summe SQE der Quadrate der Residuen, das mittlere Residuenquadrat MQE und das Bestimmtheitsmaß B sowie
c) 95%-Konfidenzintervalle für den Anstieg der Regressionsgeraden und die Zielgrößenmittelwerte.

Lösung mit R:

```
> # Beispiel 6.3
> # Dateneingabe, univariate Statistiken
> x <- c(135.0, 147.0, 148.5, 130.0, 139.0,
+         129.0, 142.0, 146.0, 131.0, 143.5,
+          138.5, 145.0, 143.0, 153.0, 149.0)
> y <- c(99.0, 106.5, 105.5, 94.0, 98.0,
+         92.0, 97.0, 106.0, 102.5, 98.5,
+         105.0, 103.0, 101.0, 107.0, 104.0)
> options(digits=4)
> xy <- data.frame(x, y)
> # Ordnen des Datensatzes nach aufsteigender Größe von x
> iv <- order(xy$x)
> xy <- xy[iv,]; x <- xy$x; y <- xy$y
> # a) lineare Regression von Y auf X
> modyx <- lm(formula = y ~ x, data=xy)
> summary(modyx) # Abhängigkeitsprüfung, Regressionsparameter

Call:
lm(formula = y ~ x, data = xy)
```

```
Residuals:
   Min     1Q Median     3Q    Max
-4.599 -2.035 -0.025  1.656  6.128

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   34.120     16.498    2.07   0.0591 .
x              0.475      0.117    4.08   0.0013 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.22 on 13 degrees of freedom
Multiple R-squared:  0.561,     Adjusted R-squared:  0.527
F-statistic: 16.6 on 1 and 13 DF,  p-value: 0.00131

> paryx <- coefficients(modyx); paryx # Regressionsparameter
(Intercept)           x
    34.1195      0.4752
```

Grafik: Streudiagramm, Regressionsgerade, 95-Konfidenzband




```
> # b) Bestimmung von SQE, MQE, B
> n <- length(x); vary <- var(y); rxy <- cor(x, y)
> SQE <- (n-1)*vary*(1-rxy^2); MQE <- SQE/(n-2); B <- rxy^2
> print(cbind(n, vary, rxy, SQE, MQE, B))
      n  vary   rxy   SQE   MQE      B
[1,] 15 21.92 0.749 134.8 10.37 0.5609
> # c) 95%-Konfidenzintervalle für den Anstieg und die Zielvariablenmittel
> confint(modyx, level=0.95) # 95%-Konfidenzintervalle fuer Parameter
               2.5 %  97.5 %
(Intercept) -1.5213 69.7603
x            0.2233  0.7271
> tabpredict <- predict(modyx, xy , level=0.95, interval="confidence")
```

```
> data.frame(xy, tabpredict)
       x     y    fit    lwr    upr
6  129.0  92.0  95.42  91.84  99.00
4  130.0  94.0  95.90  92.53  99.26
9  131.0 102.5  96.37  93.22  99.53
1  135.0  99.0  98.27  95.88 100.67
11 138.5 105.0  99.94  98.01 101.87
5  139.0  98.0 100.17  98.29 102.06
7  142.0  97.0 101.60  99.79 103.40
13 143.0 101.0 102.07 100.23 103.92
10 143.5  98.5 102.31 100.43 104.19
12 145.0 103.0 103.02 101.00 105.05
8  146.0 106.0 103.50 101.35 105.65
2  147.0 106.5 103.98 101.68 106.27
3  148.5 105.5 104.69 102.14 107.24
15 149.0 104.0 104.93 102.28 107.57
14 153.0 107.0 106.83 103.38 110.28
```

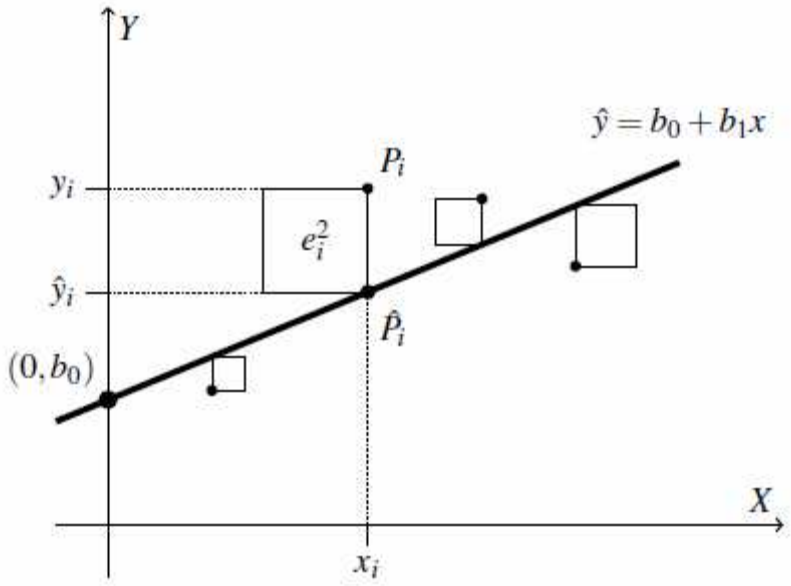
Lernziel 6.7:
Die Parameter der Regression von Y auf X im Modell B (mit zufallsgestörter linearer Regressionsfunktion) schätzen und die Abhängigkeitsprüfung durchführen können.

Ablaufschema:

- Beobachtungsdaten: wie beim Modell A

- Modell (Modell B):
  $Y(x) = \mu_Y(x) + \varepsilon$ mit
  $\mu_Y(x) = \beta_0 + \beta_1 x, \quad \varepsilon \sim N(0, \sigma_\varepsilon^2)$

- Parameterschätzung und Abhängigkeitsprüfung:
  - Prinzip (Kleinste Quadrat – Schätzung):



  - Formeln:

$$Q(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right)^2 = \min!$$

$$\rightarrow \hat{\beta}_1 = b_1 = \frac{s_{XY}}{s_X^2} = r_{XY} \frac{s_Y}{s_X}, \quad \hat{\beta}_0 = b_0 = \bar{y} - b_1 \bar{x},$$

$$SQE = Q(b_0, b_1) = (n-1) s_Y^2 (1 - r_{XY}^2), \quad MQE = \frac{SQE}{n-2}$$

- Konfidenzintervalle und Abhängigkeitsprüfung: wie bei Modell A

Beispiel 6.5:

Um herauszufinden, wie die Entwicklungsdauer Y des Bachflohkrebses *Gammarus fossarum* von der Wassertemperatur X abhängt, wurde ein Laboratoriumsexperiment mit vorgegebenen Temperaturwerten durchgeführt. Die Versuchsergebnisse sind:

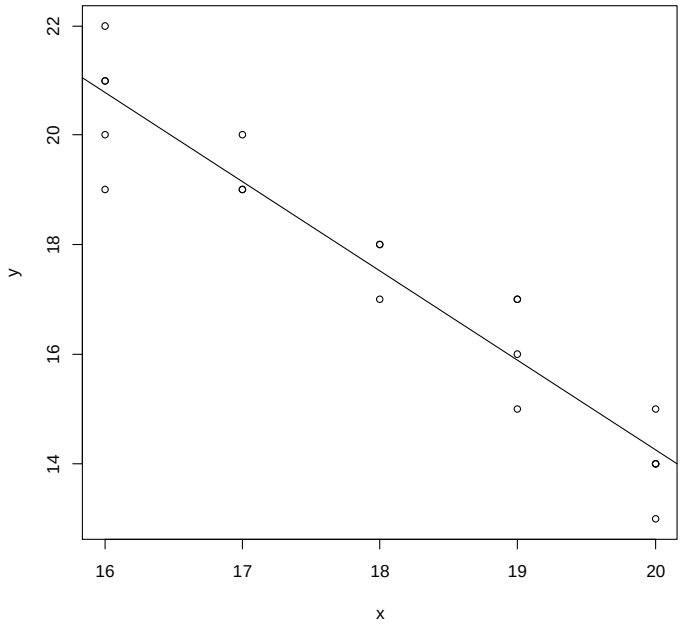
lfd. Nr.	X	Y	lfd. Nr.	X	Y	lfd. Nr.	X	Y	lfd. Nr.	X	Y
1	16	22	6	17	19	11	18	17	16	20	14
2	16	20	7	17	20	12	19	17	17	20	14
3	16	19	8	17	19	13	19	15	18	20	14
4	16	21	9	18	18	14	19	16	19	20	15
5	16	21	10	18	18	15	19	17	20	20	13

Es soll im Rahmen einer Regressionsanalyse auf 5%igem Signifikanzniveau geprüft werden, ob die mittlere Entwicklungsdauer linear von der Temperatur abhängt. Ferner sind die Regressionsparameter zu schätzen, die Regressionsgerade mit dem Streudiagramm darzustellen und für den Anstieg ein 95%iges Konfidenzintervall anzugeben.

Lösung mit R:

```
> # Beispiel 6.5 (Zufallsgestörte lineare Abhängigkeit)
> # Dateneingabe, univariate Statistiken
> x <- c(rep(16, 5), rep(17, 3), rep(18, 3), rep(19, 4), rep(20, 5))
> y <- c(22, 20, 19, 21, 21, 19, 20, 19, 18, 18,
+        17, 17, 15, 16, 17, 14, 14, 14, 15, 13)
> options(digits=4)
> n <- length(x); mwx <- mean(x); mwy <- mean(y)
> sx <- sd(x); sy <- sd(y)
> print(cbind(n, mwx, sx)); print(cbind(n, mwy, sy))
      n   mwx    sx
[1,] 20 18.05 1.572
      n   mwy    sy
[1,] 20 17.45 2.685
```

```
> # bivariate Statistiken
> s_xy <- cov(x, y) # Kovarianz
> r_xy <- cor(x, y, method="pearson") # Produktmoment(=Pearson)korrelation
> print(cbind(s_xy, r_xy))
       s_xy    r_xy
[1,] -4.024 -0.9534
> # Abhängigkeitsprüfung
> cor.test(x, y, method="pearson", alternative="two.sided",
conf.level=0.95)

        Pearson's product-moment correlation

data:  x and y
t = -13.41, df = 18, p-value = 8.295e-11
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.9817 -0.8837
sample estimates:
    cor
-0.9534

> # Schätzung der Regressionsparameter, Fehlervarianz
> b1 <- s_xy/sx^2; b0 <- mwy-b1*mwx
> print(cbind(b1, b0))
         b1    b0
[1,] -1.628 46.84
> # Abhängigkeitsprüfung (Berechnung des P-Wertes)
> tgs <- r_xy*sqrt(n-2)/sqrt(1-r_xy^2)
> P <- 2*pt(-abs(tgs), n-2); q <- qt(0.975, n-2)
> print(cbind(tgs, P, q))
        tgs         P     q
[1,] -13.41 8.295e-11 2.101
> # Loesung mit Funktion lm
> xy <- data.frame(x, y)
> modyx <- lm(formula = y ~ x, data=xy)
> summary(modyx)

Call:
lm(formula = y ~ x, data = xy)

Residuals:
    Min      1Q  Median      3Q     Max
-1.7881 -0.3389 -0.0314  0.5327  1.2119

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   46.841      2.200    21.3  3.3e-14 ***
x             -1.628      0.121   -13.4  8.3e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.832 on 18 degrees of freedom
Multiple R-squared:  0.909,     Adjusted R-squared:  0.904
F-statistic:  180 on 1 and 18 DF,  p-value: 8.29e-11

> confint(modyx, level=0.95) # 95%-Konfidenzintervalle fuer Parameter
              2.5 % 97.5 %
(Intercept) 42.219 51.463
x           -1.883 -1.373
> # Streudiagramm mit Regressionsgeraden
> plot(x, y); abline(modyx)
```



Gleichung der Regressionsgeraden: $\hat{y} = -1.628x + 46.841$

Lernziel 6.8:
Linearisierende Transformationen anwenden können, um nichtlineare Abhängigkeiten (allometrische, exponentielle bzw. gebrochen lineare) mit Hilfe von linearen Regressionsmodellen erfassen zu können.

Linearisierende Transformationen:
Nichtlineare Regressionsfunktion $\mu_{Y'}(X')$ (Zielvariable *Y'*, Einflussvariable *X'*) → lineare Regressionsfunktion

Aus der Geradengleichung $y = \beta_0 + \beta_1 x$ durch logarithmische bzw. reziproke Skalentransformationen ableitbare nichtlineare Funktionstypen:

Transformation	Nichtlineare Gleichung	Funktionstyp
$x = \ln x', y = \ln y'$	$y' = \beta_0' x'^{\beta_1}, \beta_0' = e^{\beta_0}$	Allometrische Funktion
$x = x', y = \ln y'$	$y' = \beta_0' e^{\beta_1 x'}, \beta_0' = e^{\beta_0}$	Exponentialfunktion
$x = x', y = 1/y'$	$y' = 1/(\beta_0 + b_1 x')$	Gebrochene lineare Funktion
$x = 1/x', y = 1/y'$	$y' = x'/(\beta_0 x' + \beta_1)$	Gebrochene lineare Funktion

Beispiel 6.6:
Die folgende Tabelle enthält Angaben über die Länge X' (in mm) und Masse Y' (in mg) von 15 Exemplaren des Bachflohkrebses

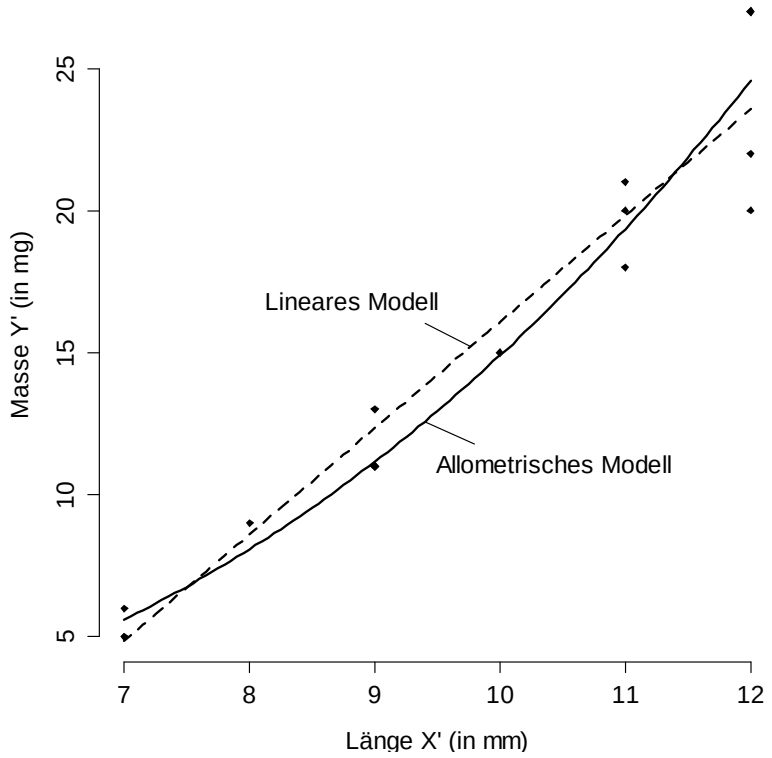
Gammarus fossarum. Es soll die Abhängigkeit der Masse von der Länge durch ein geeignetes Regressionsmodell dargestellt werden.

X'	Y'	$X = \ln X'$	$Y = \ln Y'$	X'	Y'	$X = \ln X'$	$Y = \ln Y'$	X'	Y'	$X = \ln X'$	$Y = \ln Y'$
7	5	1.946	1.609	9	11	2.197	2.398	11	21	2.398	3.045
7	5	1.946	1.609	9	13	2.197	2.565	12	20	2.485	2.996
7	6	1.946	1.792	10	15	2.303	2.708	12	22	2.485	3.091
8	9	2.079	2.197	11	18	2.398	2.890	12	27	2.485	3.296
9	11	2.197	2.398	11	20	2.398	2.996	12	27	2.485	3.296

Lösung mit R:

```
> # Beispiel 6.6
> xs <- c(rep(7, 3), 8, rep(9, 3), 10, rep(11, 3), rep(12, 4))
> ys <- c(5, 5, 6, 9, 11, 11, 13, 15, 18, 20, 21, 20, 22, 27, 27)
> options(digits=4)
> # Grafikparameter
> par(pin=c(6, 4), mai=c(0.8, 0.9, 0.2, 0.1))
> par(cex.axis=1.3, cex.lab=1.3)
> # Streudiagramm mit Originalvariablen
> plot(xs, ys, type="p", col="black", xlab="Länge X' (in mm)",
+     ylab="Masse Y' (in mg)", pch=18, frame.plot=F,
+     xlim=c(7, 12), ylim=c(5, 27), lwd=2)
> # log/log-Transformation
> x <- log(xs); y <- log(ys)
> daten <- data.frame(xs, ys, x, y); daten
   xs ys     x     y
1   7  5 1.946 1.609
2   7  5 1.946 1.609
3   7  6 1.946 1.792
4   8  9 2.079 2.197
5   9 11 2.197 2.398
6   9 11 2.197 2.398
7   9 13 2.197 2.565
8  10 15 2.303 2.708
9  11 18 2.398 2.890
10 11 20 2.398 2.996
11 11 21 2.398 3.045
12 12 20 2.485 2.996
13 12 22 2.485 3.091
14 12 27 2.485 3.296
15 12 27 2.485 3.296
> # univariate Statistiken
> n <- length(x); mwx <- mean(x); mwy <- mean(y); sx <- sd(x); sy <- sd(y)
> print(cbind(n, mwx, sx, mwy, sy))
      n   mwx     sx   mwy     sy
[1,] 15 2.263 0.2073 2.592 0.5779
> # lineare Regression mit logaithmierten Variablen
> modyx <- lm(y ~ x); summary(modyx)

Call:
lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max
-0.20537 -0.09106  0.00704  0.08841  0.15294

Coefficients:
```

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   -3.615      0.316   -11.4  3.7e-08 ***
x              2.743      0.139    19.7  4.5e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.108 on 13 degrees of freedom
Multiple R-squared:  0.968,     Adjusted R-squared:  0.965
F-statistic:  389 on 1 and 13 DF,  p-value: 4.54e-11

> paryx <- coefficients(modyx); b0 <- paryx[[1]]; b1 <- paryx[[2]]
> b0s <- exp(b0); b0s; # Rücktransformation
[1] 0.02692
> curve(b0s*x^b1, lty=1, lwd=2, ad=T)
> segments(9.4, b0s*9.4^b1, 9.8, b0s*9.4^b1-0.8)
> text(9.4, b0s*9.4^b1-1.5, expression("Allometrisches Modell"),
> +  pos=4, cex=1.3)
> # lineare Regression mit Originalvariablen
> modysxs <- lm(ys ~ xs); summary(modysxs)

Call:
lm(formula = ys ~ xs)

Residuals:
   Min     1Q Median     3Q    Max
-3.562 -1.341  0.140  0.899  3.438

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -21.323      2.708   -7.87  2.7e-06 ***
xs             3.740      0.271   13.78  3.9e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.97 on 13 degrees of freedom
Multiple R-squared:  0.936,     Adjusted R-squared:  0.931
F-statistic:  190 on 1 and 13 DF,  p-value: 3.92e-09

> parysxs <- coefficients(modysxs)
> bb0 <- parysxs[[1]]; bb1 <- parysxs[[2]]
> print(cbind(bb0, bb1))
       bb0  bb1
[1,] -21.32 3.74
> curve(bb0+bb1*x, lty=2, lwd=2, ad=T)
> segments(9.77, bb0+9.77*bb1, 9.4, bb0+9.77*bb1+0.8)
> text(9.6, bb0+9.77*bb1+1.6, expression("Lineares Modell"),
> +  pos=2, cex=1.3)
```



Regressionsfunktion (allometrisches Modell mit Orignialvariablen):

$y' = 0.02692 x'^{2.743}$

Lernziel 6.9:
Regressionsgeraden durch den Nullpunkt bestimmen können.

Ablaufschema:

- Beobachtungsdaten: wie beim Modell A

- Modell (Modell C):
  Wenn von der Regressionsgeraden auf Grund sachlogischer Überlegungen verlangt wird, dass sie durch einen festen Punkt $P=(x_0,y_0)$ der Merkmalsebene verläuft. Ohne Beschränkung der Allgemeinheit kann P im Nullpunkt des Koordinatensystems liegend angenommen, also $x_0=y_0=0$ vorausgesetzt werden. Zur Erfüllung der Forderung nach einer durch den Nullpunkt verlaufenden Regressionsgeraden macht man den Modellansatz:

  $$Y(x) = \mu_Y(x) + \varepsilon \quad \text{mit}$$
  $$\mu_Y(x) = \beta_1 x, \quad \varepsilon \approx N(0, \sigma_\varepsilon^2)$$

- Parameterschätzung und Abhängigkeitsprüfung:

- Schätzwerte für die Modellparameter β_1 und σ_ε^2:

$$\hat{\beta}_1 = b_1 = \sum_{i=1}^{n} x_i y_i \Big/ \sum_{i=1}^{n} x_i^2 ,$$

$$MQE = \frac{SQE}{n-1} \quad \text{mit} \quad SQE = \sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} x_i y_i \right)^2 \Big/ \sum_{i=1}^{n} x_i^2$$

- (1-α)-Konfidenzintervall für den Anstieg:

$$b_1 \pm t_{n-1,1-\alpha/2} SE(b_1) = b_1 \pm t_{n-1,1-\alpha/2} \sqrt{\frac{MQE}{\sum_{i=1}^{n} x_i^2}}$$

H_0: $\beta_1 = 0$ auf dem Testniveau α ablehnen, wenn das (1-α)-Konfidenzintervall für β_1 den Wert 0 nicht enthält.

Beispiel 6.7:

Es sei C die Plasmakonzentration eines Wirkstoffes und c_0 der Anfangswert. Die Abnahme der auf den Anfangswert bezogenen Konzentration $Y'=C/c_0$ in Abhängigkeit von der Zeit X (in h) ist durch folgende Daten dokumentiert:

```
X :  1,    2,    3,    4,    5,     6,     7,     8
Y‘:  0.72, 0.29, 0.16, 0.11, 0.075, 0.046, 0.025, 0.014
```

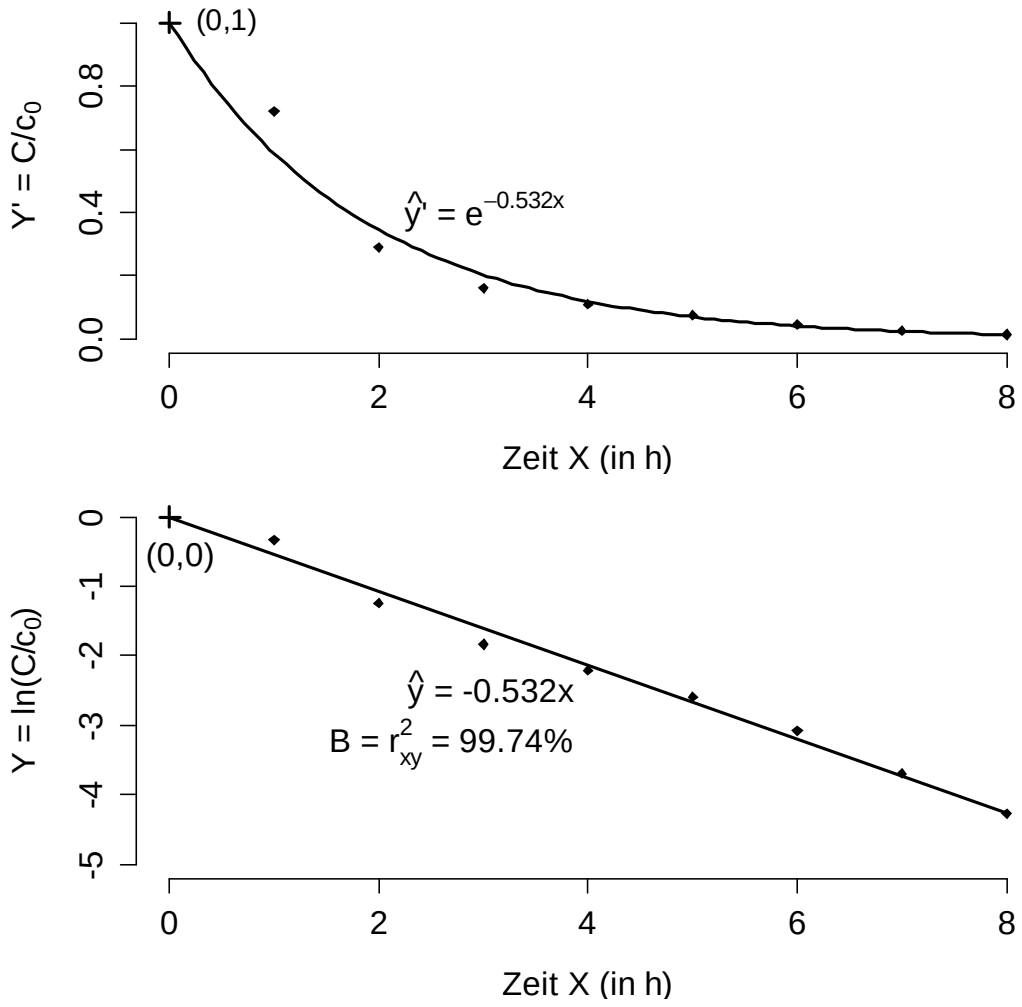
Offensichtlich muss Y'(0)=1 gelten. Unter der (auch durch das Streudiagramm nahegelegten) Annahme, dass Y' im Mittel nach dem Exponentialgesetz $\mu_{Y'}(x) = e\exp(\beta_1 x)$ abnimmt, bestimme man einen Schätzwert (samt 95%igem Konfidenzintervall) für β_1.

Lösung mit R:

```
> # Beispiel 6.7 (Regressionsgerade durch den Nullpunkt)
> x <- seq(from=1, to=8, by=1)
> ys <- c(0.72, 0.29, 0.16, 0.11, 0.075, 0.046, 0.025, 0.014)
> options(digits=4)
> # Grafikparameter
> par(mfrow=c(2, 1))
> par(pin=c(6, 4), mai=c(0.8, 0.9, 0.2, 0.1))
> par(cex.axis=1.3, cex.lab=1.3)
> n <- length(x)
> # Streudiagramm mit Originalvariablen
> plot(x, ys, type="p", col="black", xlab="Zeit X (in h)",
+     ylab=expression("Y' = C/"*c[0]), pch=18, frame.plot=F,
```

```
+     xlim=c(0, 8), ylim=c(0, 1.1), lwd=2)
> text(2.1, 0.4, expression(hat(y)*"' = "*e^{-0.532*x}), pos=4, cex=1.3)
> points(0, 1, pch=3, lwd=2, cex=1.3)
> text(0.1,1, expression("(0,1)"), pos=4, cex=1.2)
> #
> # log-Transformation
> y <- log(ys); xy <- x*y
> daten <- data.frame(x, ys, y, xy)
> # Schätzung des Anstiegs
> sumxy <- sum(xy); sumx2 <- sum(x^2); sumy2 <- sum(y^2)
> b1 <- sumxy/sumx2
> print(cbind(b1, sumx2, sumy2, sumxy), digits=6)
            b1 sumx2   sumy2    sumxy
[1,] -0.532004   204 57.8907 -108.529
> curve(exp(b1*x), lty=1, lwd=2, ad=T)
> # Bestimmung von SQE und MQE
> SQE <- sumy2-sumxy^2/sumx2; MQE <- SQE/(n-1)
> print(cbind(SQE, MQE), digits=5)
        SQE      MQE
[1,] 0.1529 0.021843
> # Bestimmtheitsmaß
> B <- 1-SQE/sumy2; B
[1] 0.9974
> # 95%-Konfidenzintervall für den Anstieg
> q <- qt(0.975, n-1); seb1 <- sqrt(MQE/sumx2)
> ug <- b1-q*seb1; og <- b1+q*seb1
> print(cbind(q, b1, seb1, ug, og))
         q     b1    seb1      ug      og
[1,] 2.365 -0.532 0.01035 -0.5565 -0.5075
> # Lösung mit R-Funktion lm()
> mod <- lm(y ~ 0+x)
> ergebnis <- summary(mod); ergebnis

Call:
lm(formula = y ~ 0 + x)

Residuals:
    Min      1Q  Median      3Q     Max
-0.2366 -0.1029  0.0112  0.0805  0.2035

Coefficients:
  Estimate Std. Error t value Pr(>|t|)
x  -0.5320     0.0103   -51.4  2.8e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.148 on 7 degrees of freedom
Multiple R-squared:  0.997,     Adjusted R-squared:  0.997
F-statistic: 2.64e+03 on 1 and 7 DF,  p-value: 2.76e-10

> B <- ergebnis$r.squared; B
[1] 0.9974
> parx <- coefficients(mod); b1 <- parx[[1]]; b1
[1] -0.532
> confint(mod)
    2.5 %  97.5 %
x -0.5565 -0.5075
> # Streudiagramm mit logarithmiertem Y
> plot(x, y, type="p", col="black", xlab="Zeit X (in h)",
+     ylab=expression("Y = ln(C/"*c[0]*")"), pch=18, frame.plot=F,
+     xlim=c(0, 8), ylim=c(-5, 0), lwd=2)
> segments(0, 0, 8, b1*8, lty=1, lwd=2)
```

```
> text(4, -2.5, expression(hat(y)*" = -0.532x"), pos=2, cex=1.3)
> text(4, -3.3, expression("B = "*r[xy]^2*" = 99.74%"), pos=2, cex=1.3)
> points(0, 0, pch=3, lwd=2, cex=1.3)
> text(0.1,-0.15, expression("(0,0)"), pos=1, cex=1.3)
```



Lernziel 6.10:
Probenmesswerte mit Hilfe von linearen Kalibrationsfunktionen schätzen können.

Ablaufschema:

- Bestimmung der linearen Kalibrationsfunktion:
  Die Kalibrationsfunktion bestimmt man in der Regel so, dass man zu vorgegebenen Kalibrierproben (Werte x_i von X) die entsprechenden Werte y_i der Hilfsgröße Y misst und eine lineare Regression von Y auf X durchführt (Modell B). Schätzwerte für die Modellparameter β_1, β_0 und σ_E^2:

$$\hat{\beta}_1 = b_1 = \frac{s_{XY}}{s_X^2} = r_{XY}\frac{s_Y}{s_X}, \quad \hat{\beta}_0 = b_0 = \bar{y} - b_1\bar{x},$$

$$\hat{\sigma}_E^2 = MQE = \frac{SQE}{n-2} \quad \text{mit} \quad SQE = (n-1)s_Y^2(1-r_{XY}^2)$$

Gleichung der Kalibrationsfunktion: $\hat{y} = f(x, b_0, b_1) = b_0 + b_1 x$
Voraussetzung: Anstieg b_1 weicht auf dem vorgegebenen Testniveau α signifikant von Null ab, d.h. :

$$TG = \frac{r_{XY}\sqrt{n-2}}{\sqrt{1-r_{XY}^2}} = \sqrt{\frac{b_1^2(n-1)s_X^2}{MQE}} > t_{n-2,1-\alpha/2}$$

- <u>Rückschluss von Y auf X:</u>
  Bei bekannten Regressionsparametern β_1 und β_0 sowie bekanntem Erwartungswert η von *Y* ergibt sich der gesuchte X-Wert ξ einfach aus der Regressionsgleichung: $\xi = (\eta - \beta_0)/\beta_1$. Im Allgemeinen kennt man weder die Regressionsparameter β_1 und β_0 noch den Erwartungswert η. Naheliegend ist nun folgende Vorgangsweise: Wir bilden den Mittelwert $\bar{y}'$ aus *m* zum selben ξ gemessenen *Y*-Werten (im Extremfall kann *m*=1 sein), setzen $\bar{y}'$ an Stelle von $\hat{y}$ in die Regressionsgleichung $\hat{y} = \bar{y} + b_1(x - \bar{x})$ ein und lösen nach *x* auf. Die so erhaltene Größe – wir bezeichnen sie mit $\hat{x}$ - nehmen wir als Schätzfunktion für *x*. Es ist also $\hat{x} = \bar{x} + (\bar{y}' - \bar{y})/b_1$.
- Berechnung eines Konfidenzintervalls für ξ:
  Unter der Voraussetzung $g = t_{n-2,1-\alpha/2}^2 / TG^2 < 0.1$ erhält man das approximative (1-α)-Konfidenzintervall für den gesuchten X-Wert:

$$UG = \hat{x} - t_{n-2,1-\alpha/2}s_{\hat{x}} \quad \text{und} \quad OG = \hat{x} + t_{n-2,1-\alpha/2}s_{\hat{x}}$$

$$s_{\hat{x}} = \frac{\sqrt{MQE}}{|b_1|}\sqrt{\left(\frac{1}{m} + \frac{1}{n} + \frac{(\bar{y}' - \bar{y})^2}{b_1^2(n-1)s_X^2}\right)}$$

Man beachte, dass die Genauigkeit der Schätzung von der Anzahl *n* der Kalibrierproben und vom Umfang *m* der *Y*-Stichprobe abhängt. Für ein optimales Design der Kalibrationsfunktion wird man ferner darauf achten, dass $(\bar{y}' - \bar{y})$ möglichst klein und s_X^2 möglichst groß ist.

Beispiel 6.8:

Zur Messung von Fe-Konzentrationen sollen die Peakhöhen von Atomabsorptionsspektrallinien herangezogen werden. Zwecks Kalibration des Messverfahrens wurden die Peakhöhen (Variable Y, in cm) in Abhängigkeit von einigen vorgegebenen Massenwerten (Variable X, in ng) bestimmt. Wir berechnen
a) die lineare Kalibrationsfunktion im Rahmen einer linearen Regression von Y auf X und schätzen
b) die Masse einer neuen Probe auf Grund einer gemessenen Peakhöhe von 0.055cm (α=5%).

```
X:   1.409, 3.013, 5.508, 8.100, 10.303
Y:   0.027, 0.040, 0.065, 0.084,  0.102
```

Lösung mit R:

```
> # Beispiel 6.8 (Lineare Kalibration)
> masse <- c(1.409,3.013, 5.508, 8.100, 10.303)
> peak <- c(0.027, 0.040, 0.065, 0.084, 0.102)
> options(digits=4)
> # a) Abhängigkeitsprüfung und Parameterschätzung:
> x <- masse; y <- peak; n <- length(x)
> mwx <- mean(x); mwy <- mean(y)
> sx <- sd(x); sy <- sd(y)
> print(cbind(n, mwx, sx, mwy, sy))
     n   mwx    sx    mwy      sy
[1,] 5 5.667 3.627 0.0636 0.03078
> s_xy <- cov(x, y) # Kovarianz
> r_xy <- cor(x, y, method="pearson"); B <- r_xy^2
> print(cbind(s_xy, r_xy, B))
       s_xy   r_xy      B
[1,] 0.1115 0.9987 0.9974
> # Schätzung der Regressionsparameter
> b1 <- s_xy/sx^2; b0 <- mwy-b1*mwx
> print(cbind(b1, b0))
           b1      b0
[1,] 0.008476 0.01557
> SQE <- (n-1)*sy^2*(1-r_xy^2); MQE <- SQE/(n-2);
> SQY <- (n-1)*sy^2; SQR <- (n-1)*sy^2*r_xy^2
> print(cbind(SQY, SQE, MQE, SQR))
          SQY       SQE       MQE      SQR
[1,] 0.003789 9.879e-06 3.293e-06 0.003779
> # Abhängigkeitsprüfung mit t-Test
> tgs <- r_xy*sqrt(n-2)/sqrt(1-r_xy^2)
> P <- 2*pt(-abs(tgs), n-2); q <- qt(0.975, n-2)
> print(cbind(tgs, P, q))
       tgs         P     q
[1,] 33.88 5.654e-05 3.182
>
```

```
> # Lösung mit R-Funktion lm()
> daten <- data.frame(masse, peak)
> modell <- lm(formula=peak~masse, data=daten)
> ergebnis <- summary(modell)
> ergebnis$coefficients
             Estimate Std. Error t value  Pr(>|t|)
(Intercept) 0.015573  0.0016335   9.533 2.448e-03
masse       0.008476  0.0002502  33.877 5.654e-05
> b0 <- ergebnis$coefficients[1,1]; b0
[1] 0.01557
> b1 <- ergebnis$coefficients[2,1]; b1
[1] 0.008476
> rxy <- cor(masse, peak); B <- rxy^2
> print(cbind(rxy, B))
         rxy      B
[1,] 0.9987 0.9974
>
> # b) Schätzung der Probenmasse zur gegebenen Peakhöhe
> peakhoehe <- 0.055; alpha <- 0.05
> masse_erwartet <- (peakhoehe-b0)/b1
> sigma <- ergebnis$sigma
> mw_peak <- mean(peak)
> var_masse <- var(masse)
> SE_masse_erwartet <- sigma/abs(b1)*sqrt(1+1/5+
+   (peakhoehe-mw_peak)^2/b1^2/4/var_masse)
> t_quantil <- qt(1-alpha/2, 3)
> UG <- masse_erwartet - t_quantil*SE_masse_erwartet
> OG <- masse_erwartet + t_quantil*SE_masse_erwartet
> print(cbind(peakhoehe, masse_erwartet, SE_masse_erwartet, UG, OG))
     peakhoehe masse_erwartet SE_masse_erwartet    UG    OG
[1,]     0.055          4.652            0.2364 3.899 5.404
> # Überprüfung der Voraussetzung
> tgs <- cor(masse, peak)*sqrt(n-2)/sqrt(1-cor(masse, peak)^2); tgs
[1] 33.88
> g <- t_quantil^2/tgs^2; g # muss < 0.1 sein!
[1] 0.008825
```

6.4 Übungsbeispiele

1. An bestimmten von sechs verschiedenen Grasarten stammenden Chromosomen wurden die Gesamtlänge L sowie die Teillänge H des C-Band Heterochromatins gemessen (Angaben in µm; aus H.M. Thomas, Heredity, 46: 263-267, 1981). Man berechne und interpretiere die Produktmomentkorrelation r_{lh}. (0.78, Teil-Ganzheitskorr.)

L	77.00	79.00	72.50	65.50	56.50	57.25
H	6.00	5.00	5.00	3.00	2.75	4.25

2. An 15 Pflanzen (*Biscutella laevigata*) wurden u.a. die Sprosshöhe X und die Länge Y des untersten Stengelblattes gemessen (Angaben in mm).

X	Y	X	Y	X	Y
298	39	380	50	232	70
345	47	92	33	90	14
183	18	380	70	200	28
340	29	195	20	350	45
350	45	265	52	620	40

   a) Man berechne die Produktmomentkorrelation.
   b) Was ergibt sich, wenn man das Wertepaar X=620, Y=40 als ausreißerverdächtig weglässt?
   c) Man zeige an Hand der Stichprobe (ohne das letzte Wertepaar), dass die Korrelationskoeffizienten signifikant von null abweichen (α=5%).

   (r_{xy}=0.439; ohne letztes Wertepaar: r_{xy}= 0.605, r_s=0.689; sign. ungleich null)

3. Die folgenden Häufigkeiten sind einer auf F. Galton zurückgehenden Studie über die Augenfarben von Ehepartnern entnommen. Wenn man lediglich zwischen "heller" und "dunkler" Augenfarbe unterscheidet, haben von 774 beobachteten Ehepaaren 309 die Kombination "hell/hell" (d.h., Ehemann und Ehefrau haben eine helle Augenfarbe), 214 die Kombination "hell/dunkel", 132 die Kombination "dunkel/hell" und 119 die Kombination "dunkel/dunkel".
   a) Welche Häufigkeiten sind zu erwarten, wenn man annimmt, dass die Augenfarbe keinen Einfluss bei der Partnerwahl hat?
   b) Man prüfe, ob zwischen den Augenfarben der Ehepartner eine Abhängigkeit besteht (α=5%).

   (erw. Häufigk.: 297.99, 225.01, 143.01, 107.99; k. sign. Abh.)

4. In einer Studie wurde untersucht, ob zwischen der Mortalität X in der Perinatalperiode und der Rauchergewohnheit (Raucher/Nichtraucher) Y während der Schwangerschaft ein Zusammenhang besteht. Folgende Daten stehen zur Verfügung: In der Kategorie "Raucher" gab es 246 Todesfälle (von insgesamt 8406), in der Kategorie "Nichtraucher" 264 von insgesamt 19874. Man zeige auf dem 5%-Niveau, dass X und Y nicht unabhängig variiert, und beurteile das Mortalitätsrisiko mit dem odds-ratio. (1.22)

5. Die Wirksamkeit einer Behandlung wurde einerseits durch den Probanden und andererseits durch den Prüfarzt beurteilt. Man beschreibe den Zusammenhang zwischen den Beurteilungen mit einem geeigneten Korrelationsmaß. Wie groß

sind die bei einer angenommenen Unabhängigkeit zu erwartenden absoluten Häufigkeiten? Was ergibt die Abhängigkeitsprüfung? (α = 5%)? (Abhängigkeit)

	Arzt		
Proband	sehr gut	gut	mäßig
sehr gut	36	10	4
gut	6	16	8
mäßig	5	8	12

6. In einer Geburtenstation wurden 120 Mütter nach ihren Rauchergewohnheiten befragt und nach dem Zigarettenkonsum in 3 Klassen eingeteilt. Unter den Müttern waren 50 "Nichtraucher", 39 Mütter "mittlere Raucher" und 31 "starke Raucher". Die Mütter der Kategorie "Nichtraucher" brachten 28 Mädchen und 22 Knaben zur Welt, in der Kategorie "mittlere Raucher" gab es 21 Mädchen- und 18 Knabengeburten und in der Kategorie "starke Raucher" gab es 16 Mädchen- und 15 Knabengeburten. Man prüfe, ob das Geschlecht vom Zigarettenkonsum abhängt (α=5%). (k. sign. Abh.)

7. Um den Zusammenhang zwischen dem Puppariengewicht und dem Alter von Tsetsefliegenweibchen (*Glossina p. palpalis*) bei der Puparienablage zu beschreiben, wurden 550 Puparien untersucht. Das Alter wurde in 4, das Gewicht in 5 Klassen eingeteilt (Angaben in Tagen bzw. Milligramm). Man untersuche, ob das Gewicht vom Alter abhängt (α=5%). (Abhängigkeit)

	Alter			
Gewicht	bis 20	21 bis 40	41 bis 60	über 60
bis 23	5	6	6	10
24 bis 27	23	28	39	35
28 bis 31	34	61	60	41
32 bis 35	19	55	42	21
über 35	5	26	16	5

8. Man beschreibe die Abhängigkeit der Variablen Y von der Variablen X durch ein lineares Regressionsmodell. Besteht überhaupt eine signifikante Abhängigkeit (α=5%)? Wie groß ist die zu erwartende Änderung Δ von Y, wenn X um 100 Einheiten zunimmt? Mittels einer Regression von X auf Y berechne man zusätzlich auch die zu erwartende Änderung Δ' von X bei Variation von Y um Δ Einheiten. (b_1=0.1058 sign. ungleich null, b_0=12.04; Δ= 10.58, Δ'=36.62)

X	Y	X	Y	X	Y
298	39	380	50	232	70
345	47	92	33	90	14
183	18	380	70	200	28
340	29	195	20	350	45
350	45	265	52		

9. Man beschreibe die Abnahme der Säuglingssterblichkeit Y (Anzahl der gestorbenen Säuglinge auf 1000 Lebendgeborene) in Österreich von 1977 bis

1987 durch ein lineares Regressionsmodell. Wie groß ist die durchschnittliche Abnahme der Säuglingssterblichkeit pro Jahr innerhalb des angegebenen Beobachtungszeitraumes? Gibt es eine signifikante Änderung der Säuglingssterblichkeit mit der Zeit (α=5%)? (b_1=-0.64 sign. ungleich null, b_0=-39.67)

X	77	78	79	80	81	82	83	84	85	86	87
Y	16.8	15.0	14.7	14.3	12.7	12.8	11.9	11.4	11.2	10.3	9.8

10. Die nachfolgende Tabelle enthält die über das Jahr gemittelten Wassertemperaturen (in ^oC) der Donau. Man prüfe im Rahmen einer linearen Regression, ob sich im Beobachtungszeitraum die Temperatur signifikant verändert hat (α=5%). (b_1=0.0588 n. sign. ungleich null)

Jahr	Temp.	Jahr	Temp.	Jahr	Temp.
80	9.4	86	10.7	92	11.5
81	10.6	87	9.6	93	10.6
82	10.5	88	10.6	94	11.5
83	10.0	89	10.4	95	9.9
84	9.9	90	10.9		
85	10.1	91	10.2		

11. Die Wirkung eines Präparates A auf den (systolischen) Blutdruck wird durch Blutdruckmessungen vor und nach Gabe von A ermittelt. Ergänzend zu diesen Zielvariablen wird das Gewicht (in kg) als Kovariable mit erfasst. Man prüfe, ob der Behandlungseffekt (= Differenz der Blutdruckwerte vor und nach Gabe des Präparates) vom Körpergewicht linear abhängt (α=5%). (Abhängigkeitsprüfung n. sign.)

Gewicht	Blutdruck/vor	Blutdruck/nach
67	170	148
68	190	155
78	175	137
94	189	143
89	180	145
82	178	140

12. In einer Stichprobe von 10 Frauen wurden der Blutdruck Y (mm Hg) und das Alter X registriert. Kann man mit einem linearen Regressionsmodell vom Alter auf den Blutdruck schließen (α=5%)? (b_1=0.555 sign. ≠0, b_0=102.1)

Proband	Alter	Blutdruck	Proband	Alter	Blutdruck
1	36	115	6	31	120
2	57	122	7	49	135
3	61	139	8	27	118
4	42	127	9	35	125
5	46	125	10	58	140

13. Von einem Gebiet der Schweiz liegen aus 10 Wintern (Dezember bis März) die in der folgenden Tabelle angeführten Werte der Schneehöhe X (in cm) und der Lawinenabgänge Y vor. Man stelle die Abhängigkeit der Anzahl der Lawinenabgänge von der Schneehöhe durch ein lineares Regressionsmodell dar. (α=5%)

X	80	300	590	170	302	515	609	843	221	616
Y	31	44	78	65	75	38	51	104	37	91

14. Der Energieumsatz E (in kJ pro kg Körpergewicht und Stunde) wurde in Abhängigkeit von der Laufgeschwindigkeit v (in m/s) gemessen. Man stelle die Abhängigkeit des Energieumsatzes von der Laufgeschwindigkeit durch ein geeignetes Regressionsmodell dar und prüfe, ob im Rahmen des Modells überhaupt ein signifikanter Einfluss der Geschwindigkeit auf den Energieumsatz besteht (α=5%). ($E = 0.514v^{3.3}$, b_1=3.3 sign. ≠0)

v	3.1	4.2	5.0	5.4	6.6
E	27.6	50.6	62.7	147.1	356.3

15. Der durch die folgenden Daten belegte Zusammenhang zwischen der Länge L und der Fluggeschwindigkeit V von Tieren ist offensichtlich nichtlinear (aus T.A. McMahon und J.T. Bonner, Form und Leben, Heidelberg, Spektrum d. Wissenschaft, 1985). Wie man sich an Hand eines Streudiagramms klar machen kann, erreicht man mit einer doppelt-logarithmischen Transformation eine Linearisierung. Man beschreibe die Abhängigkeit der Fluggeschwindigkeit von der Länge durch eine geeignete Regressionsfunktion. Welcher Streuungsanteil von V ist durch L erklärbar? ($V = 469.7 \cdot L^{0.3612}$, b_1=0.3612 sign. ≠0 bei α = 5%)

Art	L in cm	V in cm/s
Fruchtfliege	0.2	190
Pferdebremse	1.3	660
Rubinkehlkolibri	8.1	1120
Wasserjungfer	8.5	1000
Gr. braune Fledermaus	11.0	690
Grasmücke	11.0	1200
Gewöhnl. Mauersegler	17.0	2550
Fliegender Fisch	34.0	1560
Regenbrachvogel	41.0	2320
Spießente	56.0	2280
Bewik-Schwan	120.0	1880
Rosapelikan	160.0	2280

16. Für die Wandermuschel *Dreissena polymorpha pallas* wurden (nach 5 Altersklassen aufgegliedert) Gewichts- und Längenmaße bestimmt und die in der nachstehenden Tabelle angegebenen Klassenmittelwerte L bzw. G berechnet. Man stelle die Abhängigkeit des Gewichts G von der Länge L durch eine allometrische Funktion dar und beurteile die Güte der Anpassung mit Hilfe des Bestimmtheitsmaßes. Vgl. Schulz, N.: Die Wandermuschel im Keutschacher See. Carinthia II, 170/90, 549 (1980). ($G=0.000134\ L^{2.976}$; 99.9%)

Länge L/mm	7.56	11.92	16.40	24.83	29.03
Gewicht G/g	0.055	0.213	0.564	1.894	3.012

17. Die folgende Tabelle enthält die altersspezifischen Lebensraten L_i (Anteil der Individuen, die das Alter a_i erleben), die an einer Kohorte von ursprünglich 142 Individuen eines Rankenfüßers (*Balanus glandula*) festgestellt wurden (aus Ch.J. Krebs, Ecology, New York, Harper & Row, 1985; die Variable a_i zählt das Lebensalter in Jahren). Für die Abnahme der Lebensrate mit dem Alter versuche man den exponentiellen Ansatz $L=\exp(ß_1 a)$, der insbesondere auch der Forderung $L_0=1$ genügt. Mittels einer einfach-logarithmischen Transformation erhält man daraus eine lineare Funktion, deren Parameter b_1 zu bestimmen ist. Man bestimme ein 95%-Konfidenzintervall für den Parameter $ß_1$.
(L = exp(-0.557a), [-0.604, -0.511])

a_i	1	2	3	4	5	6	7	8
L_i	0.437	0.239	0.141	0.109	0.077	0.046	0.014	0.014