Wahrscheinlichkeitsrechnung und Statistik für Biologen
Lineare Regression

Dirk Metzler

31. Mai 2021

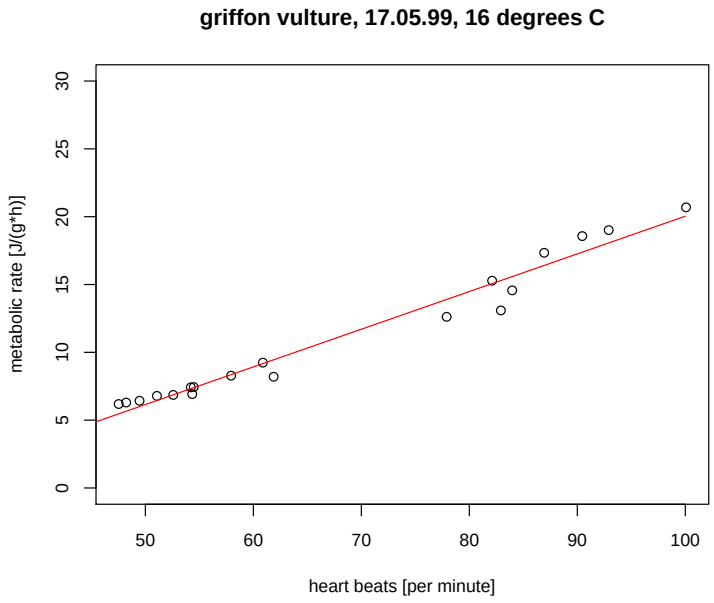
Inhaltsverzeichnis

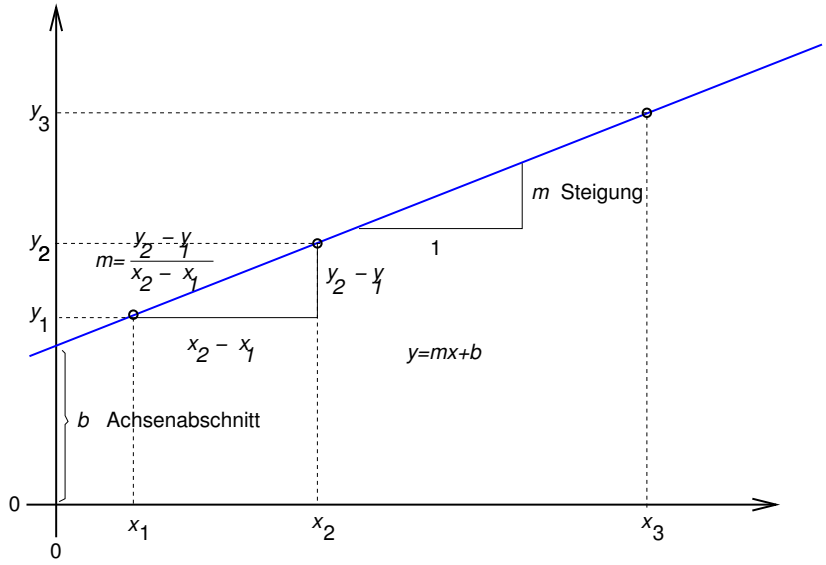
1 Lineare Regression: wozu und wie?

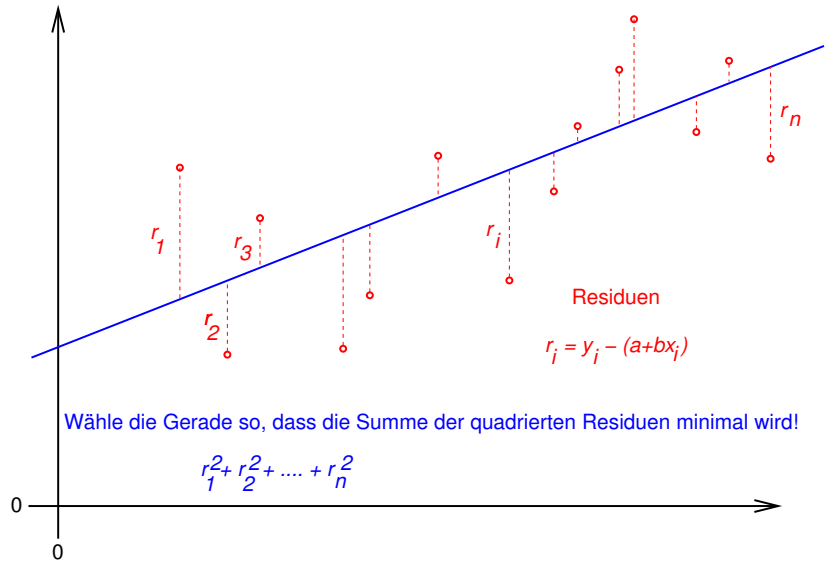
Literatur

[1] 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?


griffon vulture, 17.05.99, 16 degrees C
metabolic rate [J/(g*h)]
heart beats [per minute]
0
5
10
15
20
25
30
50
60
70
80
90
100


y_3
y_2
y_1
$m = \frac{y_2 - y_1}{x_2 - x_1}$
m Steigung
1
$y_2 - y_1$
$x_2 - x_1$
$y=mx+b$
b Achsenabschnitt
0
0
x_1
x_2
x_3



Definiere die Regressiosgerade

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

durch die Minimierung der Summe der quadrierten Residuen:

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

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 und alle dieselbe Varianz σ^2 haben.

gegebene Daten:

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

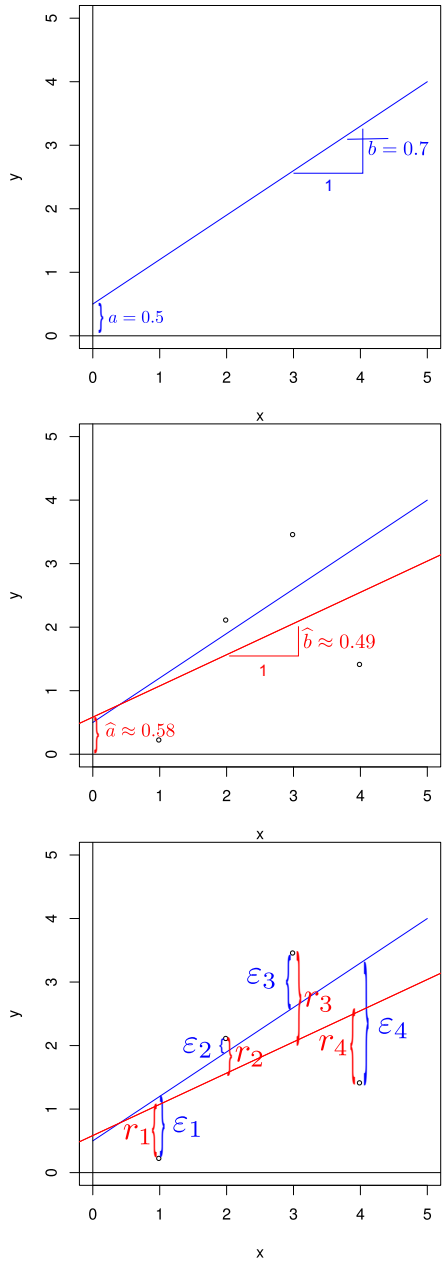
Modell: es gibt Zahlen a, b, σ^2, so dass

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

Dabei sind $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ unabhängig $\sim \mathcal{N}(0, \sigma^2)$.[1.5ex] $\Rightarrow y_1, y_2, \ldots, y_n$ sind unabhängig $y_i \sim \mathcal{N}(a + b \cdot x_i, \sigma^2)$.[1.5ex] a, b, σ^2 sind unbekannt, aber **nicht zufällig**.

Unterschied zwischen ε_i und r_i

```
> g <- function(x) { 0.5 + 0.7*x }
> plot(g,xlab="x",ylab="y",col="blue",
+     xlim=c(0,5),ylim=c(0,5))
> abline(h=0)
> abline(v=0)

> x <- 1:4
> eps <- rnorm(4,mean=0,sd=0.8)
> y <- g(x)+eps
> points(x,y)

> abh <- lm(y~x)$coef
> abh
(Intercept)           x
  0.5835982   0.4912893
> abline(a=abh[1],b=abh[2],col="red")
```



Wir schätzen a und b, indem wir

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

Theorem 1. *Man kann $\hat{a}$ und $\hat{b}$ berechnen durch*

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

und

$$\hat{a} = \bar{y} - \hat{b} \cdot \bar{x}.$$

Bitte merken: Die Gerade $y = \hat{a} + \hat{b} \cdot x$ geht genau durch den Schwerpunkt der Punktwolke $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.

Beweisskizze zum Theorem

Sei $g(a,b) = \sum_i (y_i - (a + b \cdot x_i))^2$. Wir optimieren g, indem wir die Ableitungen von g

$$\begin{aligned} \frac{\partial g(a,b)}{\partial a} &= \sum_i 2 \cdot (y_i - (a + bx_i)) \cdot (-1) \\ \frac{\partial g(a,b)}{\partial b} &= \sum_i 2 \cdot (y_i - (a + bx_i)) \cdot (-x_i) \end{aligned}$$

nach beiden Variablen auf 0 setzen, und erhalten:

$$\begin{aligned} 0 &= \sum_i (y_i - (\hat{a} + \hat{b}x_i)) \cdot (-1) \\ 0 &= \sum_i (y_i - (\hat{a} + \hat{b}x_i)) \cdot (-x_i) \end{aligned}$$

$$\begin{aligned} 0 &= \sum_i (y_i - (\hat{a} + \hat{b}x_i)) \\ 0 &= \sum_i (y_i - (\hat{a} + \hat{b}x_i)) \cdot x_i \end{aligned}$$

kann man ausmultiplizieren zu

$$\begin{aligned} 0 &= \left(\sum_i y_i\right) - n \cdot \hat{a} - \hat{b} \cdot \left(\sum_i x_i\right) \\ 0 &= \left(\sum_i y_i x_i\right) - \hat{a} \cdot \left(\sum_i x_i\right) - \hat{b} \cdot \left(\sum_i x_i^2\right) \end{aligned}$$

und das Theorem folgt durch Auflösen nach $\hat{a}$ und $\hat{b}$. □

Regression und Korrelation

Sind s_x und s_y die korrigierten Standardabweichungen der x- bzw. y-Werte, und ist

$$\mathrm{cov}(x,y) = \frac{1}{n-1} \sum_i (x_i - \overline{x}) \cdot (y_i - \overline{y})$$

die korrigierte Kovarianz, so erhalten wir für die geschätzte Steigung der Korrelationsgerade:

$$\widehat{b} = \frac{\sum_i (x_i - \overline{x}) \cdot (y_i - \overline{y})}{\sum_i (x_i - \overline{x})^2} = \frac{\frac{1}{n-1} \sum_i (x_i - \overline{x}) \cdot (y_i - \overline{y})}{\frac{1}{n-1} \sum_i (x_i - \overline{x})^2} = \frac{\mathrm{cov}(x,y)}{s_x^2} = \mathrm{cor}(x,y) \cdot \frac{s_y}{s_x}.$$

$\widehat{b}$ stimmt also mit der Korrelation $\mathrm{cor}(x,y) = \frac{\mathrm{COV}(x,y)}{s_x \cdot s_y}$ dann (und nur dann!) überein, wenn $s_x = s_y$.

`vulture`

	day	heartbpm	metabol	minTemp	maxTemp	medtemp
1	01.04./02.04.	70.28	11.51	-6	2	-2.0
2	01.04./02.04.	66.13	11.07	-6	2	-2.0
3	01.04./02.04.	58.32	10.56	-6	2	-2.0

```
4   01.04./02.04.    58.63   10.62      -6       2    -2.0
5   01.04./02.04.    58.05    9.52      -6       2    -2.0
6   01.04./02.04.    66.37    7.19      -6       2    -2.0
7   01.04./02.04.    62.43    8.78      -6       2    -2.0
8   01.04./02.04.    65.83    8.24      -6       2    -2.0
9   01.04./02.04.    47.90    7.47      -6       2    -2.0
10  01.04./02.04.    51.29    7.83      -6       2    -2.0
11  01.04./02.04.    57.20    9.18      -6       2    -2.0
.     .               .       .          .       .       .
.     .               .       .          .       .       .
.     .               .       .          .       .       .
```

(14 different days)

```
> model <- lm(metabol~heartbpm,data=vulture,
              subset=day=="17.05.")
> summary(model)
Call:
lm(formula = metabol ~ heartbpm, data = vulture,
    subset = day == "17.05.")
Residuals:
    Min      1Q  Median      3Q     Max
-2.2026 -0.2555  0.1005  0.6393  1.1834
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -7.73522    0.84543  -9.149 5.60e-08 ***
heartbpm     0.27771    0.01207  23.016 2.98e-14 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
Residual standard error: 0.912 on 17 degrees of freedom
Multiple R-squared: 0.9689,Adjusted R-squared: 0.9671
F-statistic: 529.7 on 1 and 17 DF,  p-value: 2.979e-14
```

Optimierung der Gelegegröße

Beispiel:

Der Erbsensamenkäfer

Callosobruchus maculatus

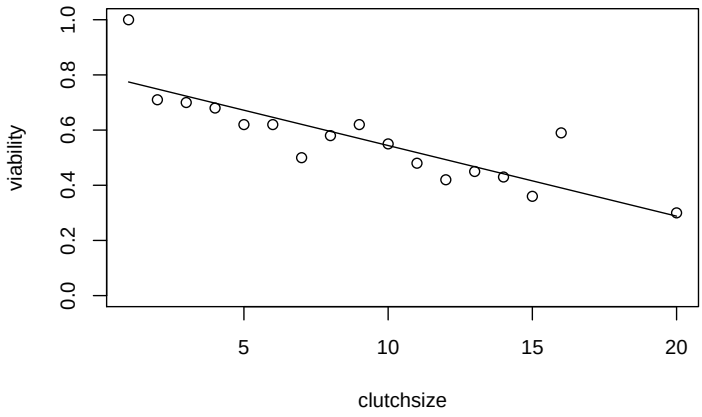
englisch: *Cowpea weevil* oder auch *bruchid beetle*

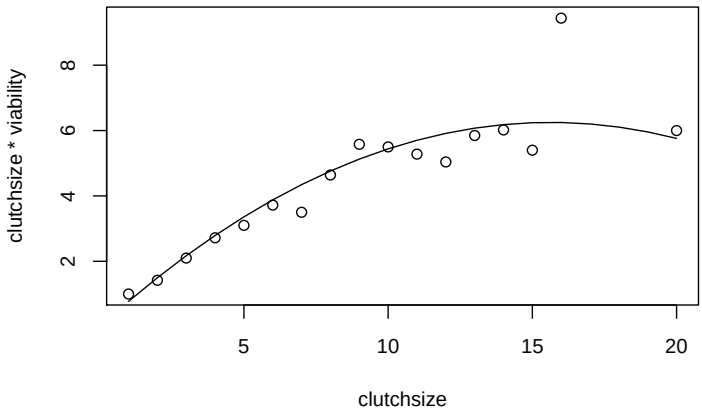
Literatur

[Wil94] Wilson, K. (1994) Evolution of clutch size in insects. II. A test of static optimality models using the beetle Callosobruchus maculatus (Coleoptera: Bruchidae). *Journal of Evolutionary Biology* **7:** 365–386.

Wie hängt die Überlebenswahrscheinlichkeit von der Gelegegröße ab?

Mit welcher Gelegegröße wird die erwartete Anzahl Nachkommen optimiert?





2 t-Test fuer lineare Zusammenhänge

Beispiel: Rothirsch (*Cervus elaphus*)

Theorie: Hirschkühe können das Geschlecht ihrer Nachkommen beeinflussen.

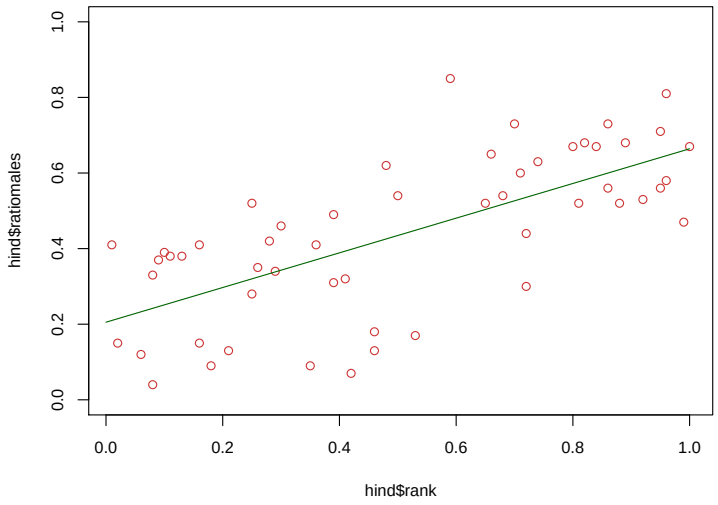
Unter dem Gesichtspunkt evolutionär stabiler Strategien ist zu erwarten, dass schwache Tiere eher zu weiblichem und starke Tiere eher zu männlichem Nachwuchs tendieren.

Literatur

[CAG86] 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.

```
> hind
   rank ratiomales
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
```

ACHTUNG: Simulierte Daten, die sich an den Daten aus der Originalpublikation lediglich orientieren.




```
> mod <- lm(ratiomales~rank,data=hind)
> summary(mod)
Call:
lm(formula = ratiomales ~ rank, data = hind)
Residuals:
     Min       1Q   Median       3Q      Max
-0.32798 -0.09396  0.02408  0.11275  0.37403

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.20529    0.04011   5.119 4.54e-06 ***
rank         0.45877    0.06732   6.814 9.78e-09 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.154 on 52 degrees of freedom
Multiple R-squared: 0.4717,Adjusted R-squared: 0.4616
```

```
F-statistic: 46.44 on 1 and 52 DF,  p-value: 9.78e-09
```

Modell:

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

[1.5ex] Wie berechnet man die Signifikanz eines Zusammenhangs zwischen dem *erklärenden Merkmal* X und der *Zielgröße* Y? [1.5ex] Anders formuliert: Mit welchem Test können wir der Nullhypothese $b = 0$ zu Leibe rücken? [1.5ex] Wir haben b durch $\hat{b}$ geschätzt (und gehen jetzt mal von $\hat{b} \neq 0$ aus). Könnte das wahre b auch 0 sein? [1.5ex] Wie groß ist der Standardfehler unserer Schätzung $\hat{b}$?

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

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.

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

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

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

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

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

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

Zu beachten ist, 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.

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

Schätze σ^2 durch

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

Dann ist

$$\frac{\hat{b} - b}{s \Big/ \sqrt{\sum_i (x_i - \bar{x})^2}}$$

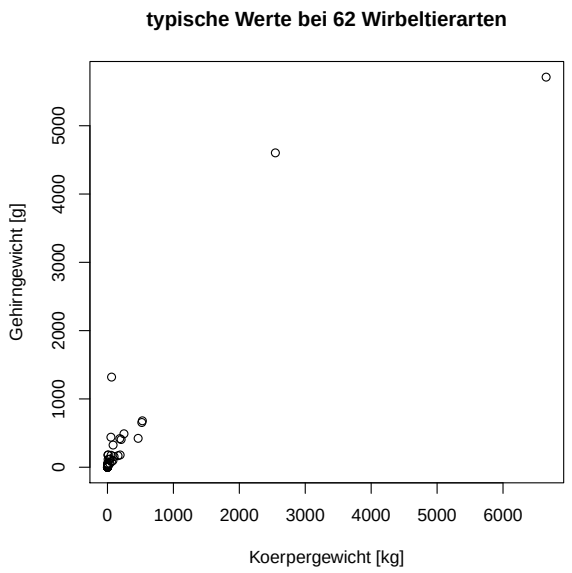
Student-t-verteilt mit $n - 2$ Freiheitsgraden und wir können den t-Test anwenden, um die Nullhypothese $b = 0$ zu testen.

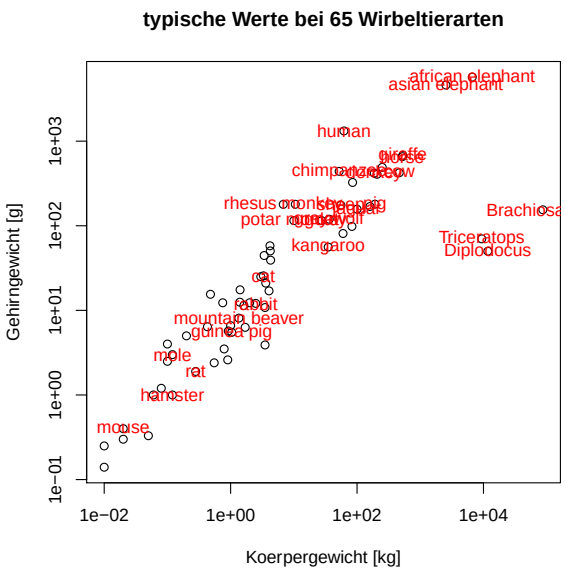
3 Skalierung der Daten

3.1 Beispiel: Körper- und Gehirngewicht

Daten: Typisches Körpergewicht [kg] und Gehirngewicht [g] von 62 Säugetierarten (und 3 Dinosaurierarten)

```
> data
   weight.kg. brain.weight.g          species  extinct
1      6654.00        5712.00 african elephant  no
2         1.00           6.60                   no
3         3.39          44.50                   no
4         0.92           5.70                   no
5      2547.00        4603.00   asian elephant  no
6        10.55         179.50                   no
7         0.02           0.30                   no
8       160.00         169.00                   no
9         3.30          25.60              cat  no
.          .              .           .
.          .              .           .
.          .              .           .

64     9400.00          70.00       Triceratops yes
65    87000.00         154.50     Brachiosaurus yes
```






```
> modell <- lm(brain.weight.g~weight.kg.,subset=extinct=="no")
> summary(modell)
Call:
lm(formula = brain.weight.g ~ weight.kg., subset = extinct ==
    "no")
Residuals:
    Min      1Q  Median      3Q     Max
-809.95  -87.43  -78.55  -31.17 2051.05
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 89.91213   43.58134   2.063   0.0434 *
```

```
weight.kg.   0.96664    0.04769  20.269   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
Residual standard error: 334.8 on 60 degrees of freedom
Multiple R-squared: 0.8726,Adjusted R-squared: 0.8704
F-statistic: 410.8 on 1 and 60 DF,  p-value: < 2.2e-16
```

Wie gut passt das Modell $Y_i = a + b \cdot X_i + \varepsilon_i$?

Sind die Residuen $r_i = Y_i - (\widehat{a} + \widehat{b} \cdot X_i)$ einigermaßen normalverteilt?

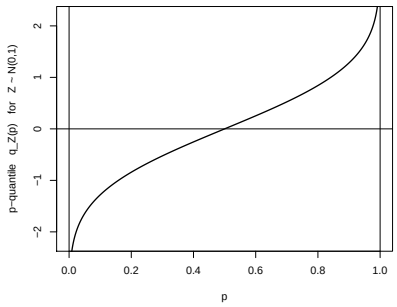
Graphische Methode: vergleiche die theoretischen Quantile der Standardnormalverteilung $\mathcal{N}(0,1)$ mit denen der Residuen.

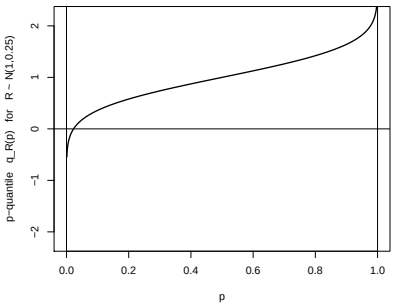
Zur Erinnerung: Ist $p \in [0,1]$ so ist das p-Quantile $q_Z(p)$ einer Zufallsvariablen Z definiert durch

$$\Pr(Z \leq q_Z(p)) = p.$$

Die **Quantilfunktion** q_Z ist also die Umkehrfunktion der **Verteilungsfunktion** $x \mapsto \Pr(Z \leq x)$.

Quantilfunktion q_Z für $Z \sim \mathcal{N}(0,1)$ Quantilfunktion q_R für $R \sim \mathcal{N}(1, 0.25)$





Welchen Graphen erhalten wir, wenn wir $q_Z(p)$ für alle $p \in [0,1]$ direkt gegen $q_R(p)$ auftragen?

Ist $Z \sim \mathcal{N}(0,1)$ und $R = a + b \cdot Z$, so ist $R \sim \mathcal{N}(a, b^2)$.

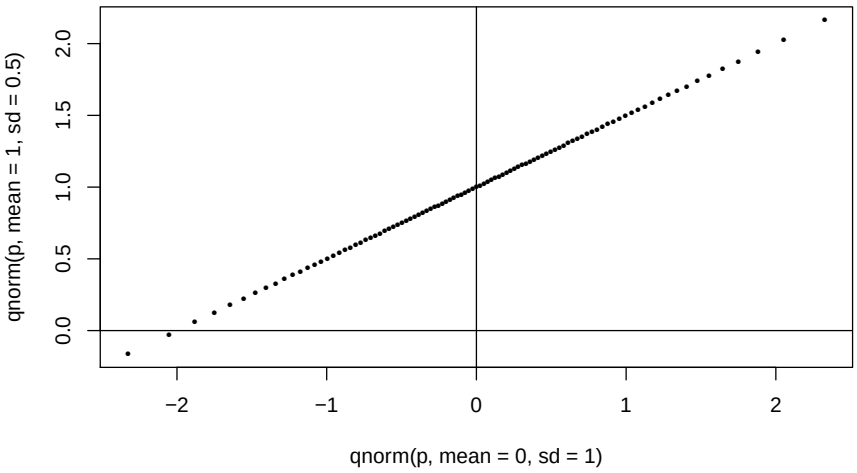
Also gilt für alle $p \in [0,1]$

$$\begin{aligned} p &= \Pr(Z < q_Z(p)) = \Pr(a + b \cdot Z < a + b \cdot q_Z(p)) \\ &= \Pr(R < a + b \cdot q_Z(p)), \end{aligned}$$

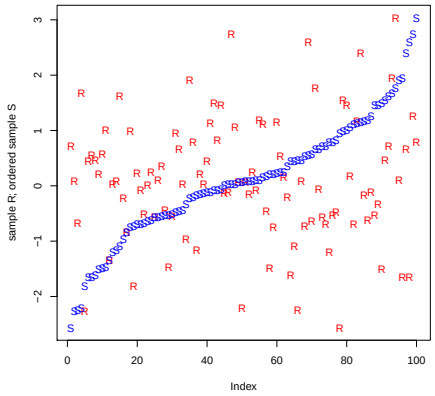
und somit ist $a + b \cdot q_Z(p)$ genau das p-Quantil $q_R(p)$ von R.

Trägt mal also z.B. für alle p die Punkte $(q_Z(p), q_R(p))$ in ein Koordinatensystem ein, so erhält man eine Gerade mit Achsenabschnitt a und Steigung b.

```
p <- seq(from=0,to=1,by=0.01)
plot(qnorm(p,mean=0,sd=1),qnorm(p,mean=1,sd=0.5),
                                    pch=16,cex=0.5)
abline(v=0,h=0)
```



Erhält man $S_1, S_2, \ldots, S_n$, indem man eine Stichprobe $R_1, R_2, \ldots, R_n$ der Größe ordnet, so dass $S_1 \leq S_2 \leq \cdots \leq S_n$, dann entspricht S_i ungefähr dem $\frac{i}{n+1}$-Quantil der Verteilung, aus der die Stichprobe stammt.



Wir bezeichnen S_i im folgenden als
empirisches $\frac{i}{n+1}$-Quantil der Stichprobe.

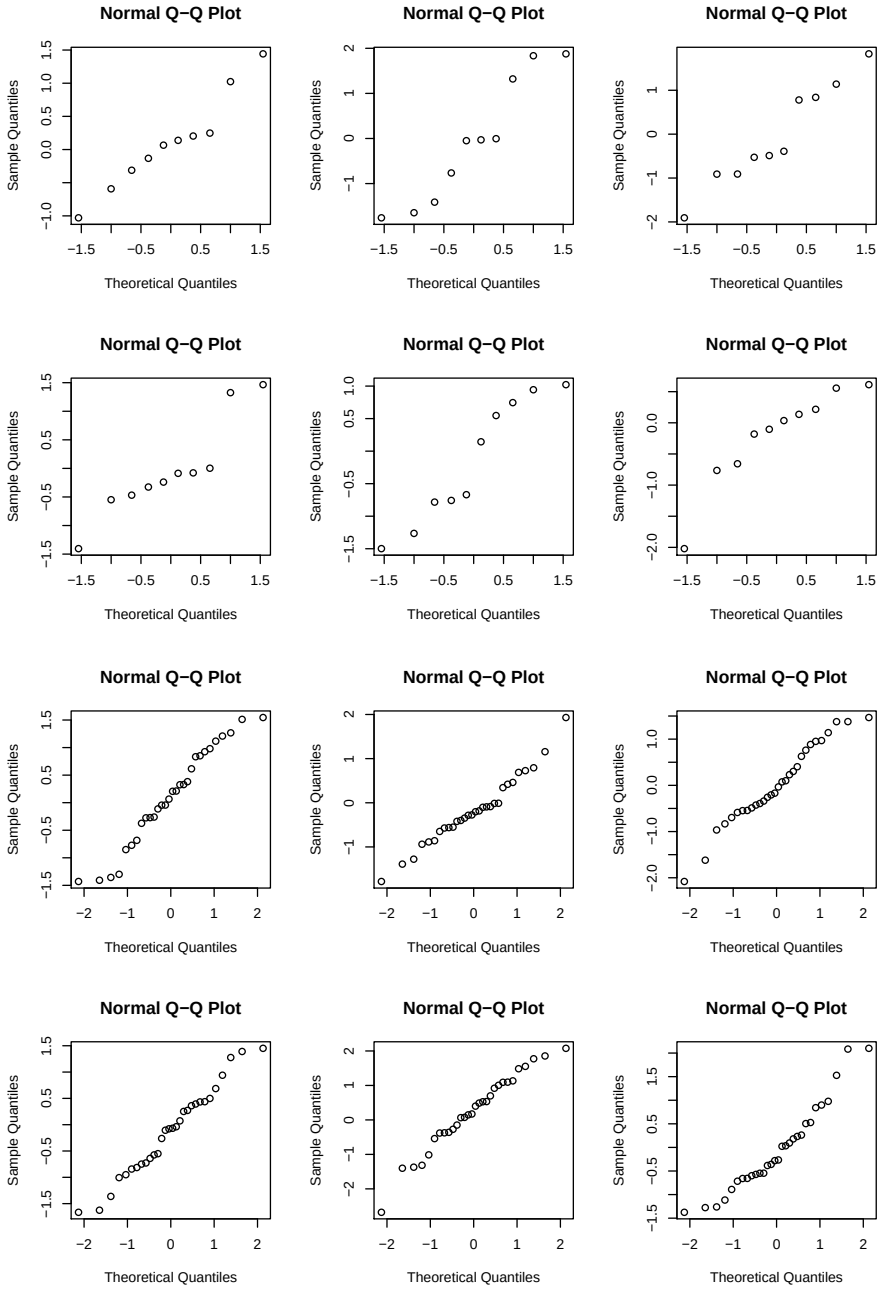
Seien $R_1, R_2, \ldots, R_n$ unabhängig gemäß einer Normalverteilung erzeugt.

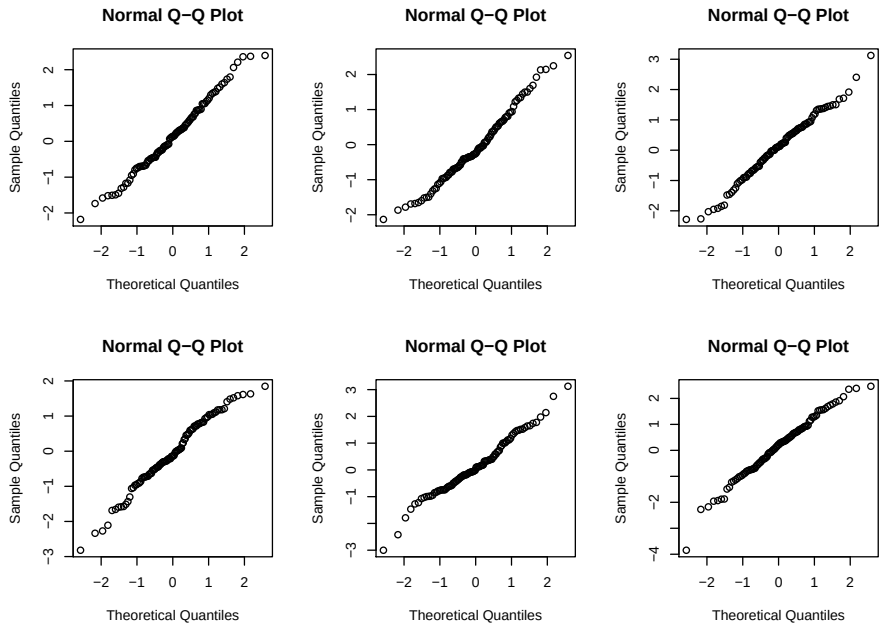
Trägt man statt der theoretischen Quantile die empirischen Quantile von $R_1, \ldots, R_n$ gegen die Quantile einer Normalverteilung auf, so liegen die Werte nicht genau auf einer Geraden, aber in der Nähe einer Geraden.

Es sollten keine *systematischen* Abweichungen von einer imaginären Geraden erkennbar sein.

Die nächsten Seiten zeigen qqnorm-Plots von simulierten, unabhängig standardnormalverteilten Daten

mit n=10, n=30 und $n = 100$.





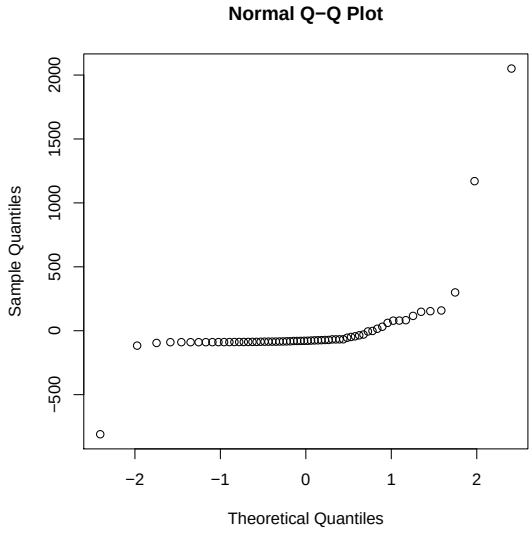
Nun zurück zur Frage ob das Modell

$$[\text{brain.weight.g}]_i = a + b \cdot [\text{weight.kg}]_i + \varepsilon_i$$

passt.

Sind die Residuen einigermaßen normalverteilt?

```
qqnorm(modell$residuals)
```

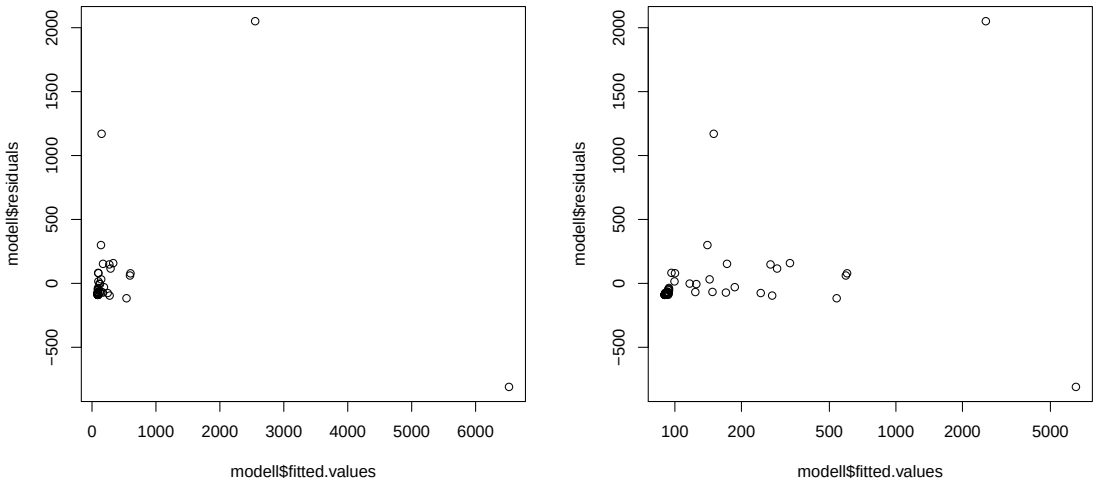


Die Residuen sind offensichtlich nicht normalverteilt. Zumindest gibt es extreme Ausreißer.

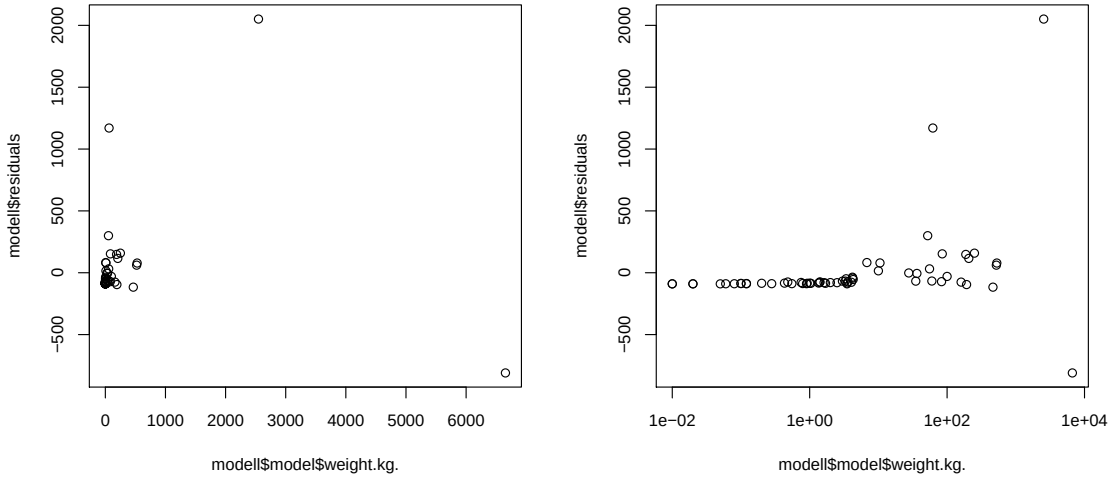
Ein weiteres Kriterium ist, dass die Residuen und ihre Varianz von anderen Größen, einschließlich der vorhergesagten Werte $\widehat{a} + \widehat{b} \cdot X_i$ unabhängig sind.

Auch das überprüfen wir graphisch.

```
plot(modell$fitted.values,modell$residuals)
plot(modell$fitted.values,modell$residuals,log='x')
```




```
plot(modell$model$weight.kg.,modell$residuals)
plot(modell$model$weight.kg.,modell$residuals,log='x' )
```



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?

Eigentlich ist es ja offensichtlich: 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

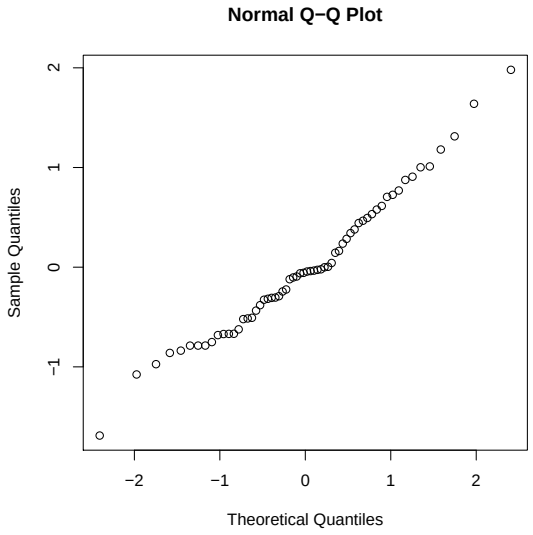
ist, wird es nicht um 500 g variieren können, sondern vielleicht ebenfalls um 10%, also ±0.5 g. Die Varianz ist hier also nicht additiv, sondern multiplikatiiv:

$$\text{Hirnmasse} = (\text{erwartete Hirnmasse}) \cdot \text{Zufall}$$

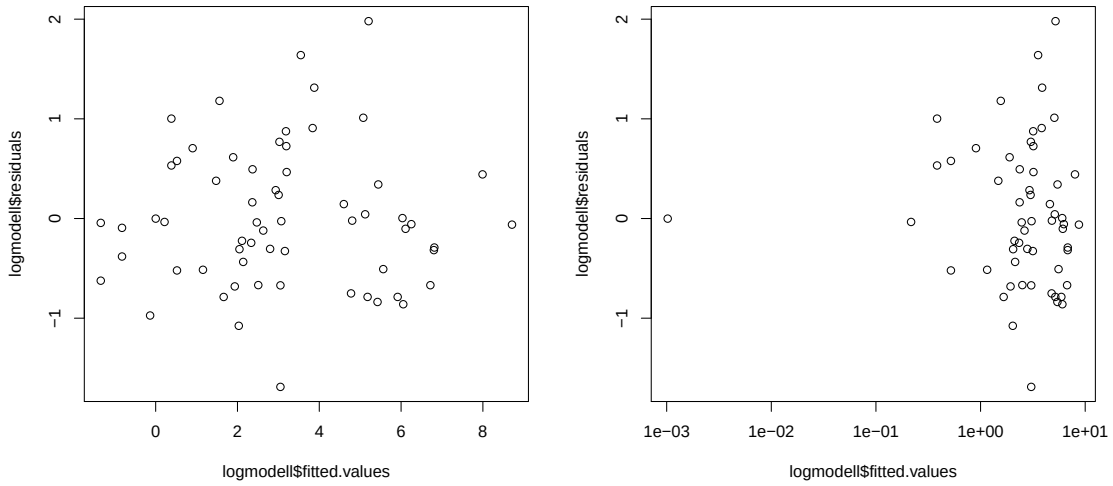
Das können wir aber in etwas mit additivem Zufallterm umwandeln, indem wir auf beiden Seiten den (natürlichen) Logarithmus ziehen:

$$\log(\text{Hirnmasse}) = \log(\text{erwartete Hirnmasse}) + \log(\text{Zufall})$$

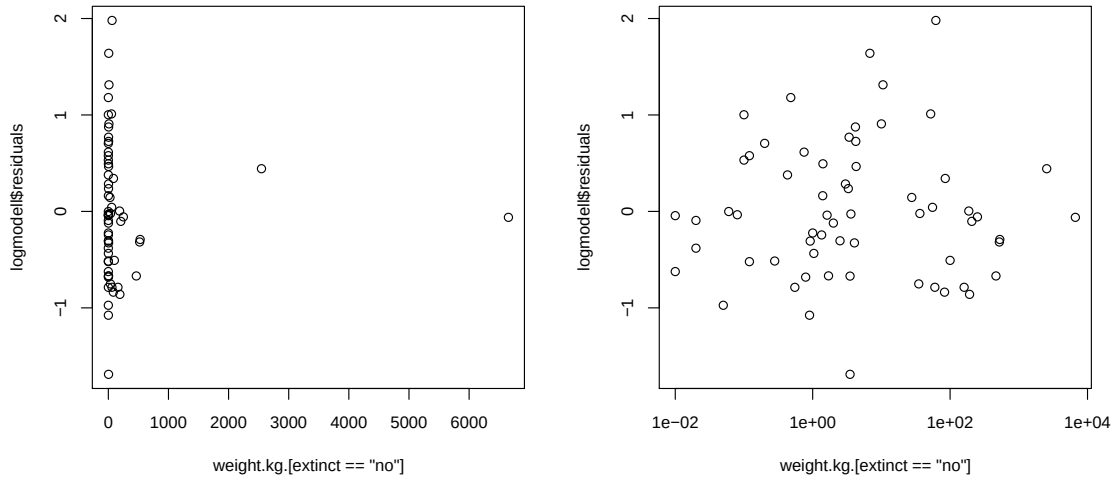
```
> logmodell <- lm(log(brain.weight.g)~log(weight.kg.),subset=extinct=="no")
> summary(logmodell)

Call:
lm(formula = log(brain.weight.g) ~ log(weight.kg.), subset = extinct ==
    "no")
Residuals:
     Min       1Q   Median       3Q      Max
-1.68908 -0.51262 -0.05016  0.46023  1.97997

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      2.11067    0.09794   21.55   <2e-16 ***
log(weight.kg.)  0.74985    0.02888   25.97   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
Residual standard error: 0.7052 on 60 degrees of freedom
Multiple R-squared: 0.9183,Adjusted R-squared: 0.9169
F-statistic: 674.3 on 1 and 60 DF,  p-value: < 2.2e-16

qqnorm(logmodell$residuals)
```




```
plot(logmodell$fitted.values,logmodell$residuals)
plot(logmodell$fitted.values,logmodell$residuals,log='x' )
```




```
plot(weight.kg.[extinct=='no'],logmodell$residuals)
plot(weight.kg.[extinct='no'],logmodell$residuals,log='x' )
```



3.2 Beispiel: Mortalität und Einwohnerzahl

Daten: Für 301 US-amerikanische Landkreise (Counties) die Anzahl weißer Einwohnerinnen von 1960 und die Anzahl der Brustkrebstoten aus dieser Gruppe zwischen 1950 und 1960. (Aus Rice (2007) Mathematical Statistics and Data Analysis.)

```
> canc
    deaths inhabitants
1        1         445
2        0         559
3        3         677
4        4         681
```

```
5        3         746
6        4         869
.        .          .
.        .          .
.        .          .

300    248       74005
301    360       88456
```

Fragestellung: Ist die durchschnittliche Anzahl an Todesopfer proportional zur Einwohnerzahl, d.h.

$$\mathbb{E}\text{deaths} = b \cdot \text{inhabitants}$$

oder hängt das Krebsrisiko von der Größe des Bezirks ab (evtl. wegen Urbanität), so dass ein anderes Modell besser passt? z.B.

$$\mathbb{E}\text{deaths} = a + b \cdot \text{inhabitants}$$

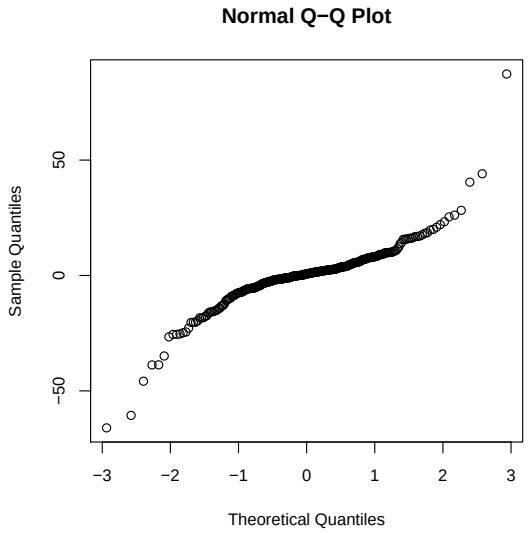
mit $a \neq 0$.

```
> modell <- lm(deaths~inhabitants,data=canc)
> summary(modell)
Call:
lm(formula = deaths ~ inhabitants, data = canc)
Residuals:
     Min       1Q   Median       3Q      Max
-66.0215  -4.1279   0.6769   5.2357  87.2989
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -5.261e-01  9.692e-01  -0.543    0.588
inhabitants  3.578e-03  5.446e-05  65.686   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
Residual standard error: 13 on 299 degrees of freedom
Multiple R-squared: 0.9352,Adjusted R-squared: 0.935
F-statistic:  4315 on 1 and 299 DF,  p-value: < 2.2e-16
```

Die additive Konstante ("Achsenabschnitt") wir auf -0.526 geschätzt, ist aber nicht signifikant von 0 verschieden.

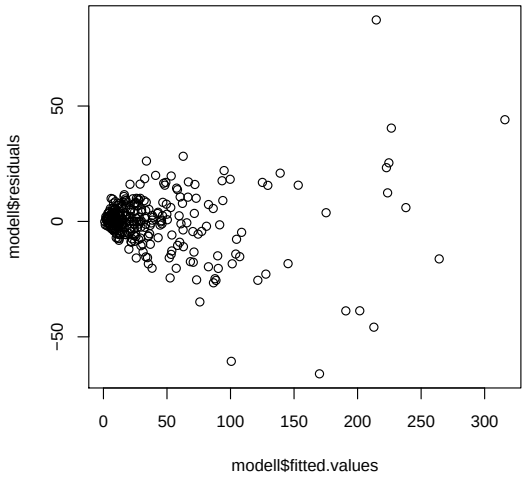
Man kann also nicht die Nullhypothese verwerfen, dass das Krebsrisiko von der Größe des Bezirks unabhängig ist.

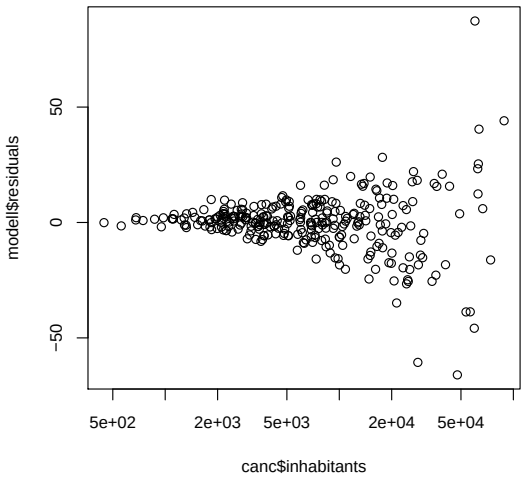
Aber passt das Modell eigentlich?




```
qqnorm(modell$residuals)
```

```
plot(modell$fitted.values,modell$residuals)
plot(modell$fitted.values,modell$residuals,log='x')
```






```
plot(canc$inhabitants,modell$residuals,log='x')
```

Wir sehen, dass die Varianz der Residuen von den angepassten Werten bzw. der Einwohnerzahl abhängt. Man sagt, es liegt *Heteroskedastizität* vor.

Das Modell geht aber von *Homoskedastizität* aus, d.h. die Residuenvarianz soll von den erklärenden Merkmalen (der Einwohnerzahl) und den angepassten Werten (annähernd) unabhängig sein.

Varianzstabilisierende Transformation: Wie können wir die Einwohnerzahl und die Anzahl der Todesfälle umskalieren, um Homoskedastizität zu erreichen?

Woher kommt die Abhängigkeit der Varianz von der Einwohnerzahl?

Ist n die Anzahl der (weißen) Einwohnerinnen und p die Wahrscheinlichkeit, innerhalb von 10 Jahren an Brustkrebs zu sterben, so ist np die erwartete Anzahl solcher Todesfälle, und die Varianz ist

$$n \cdot p \cdot (1-p) \approx n \cdot p$$

(Approximation der Binomial- durch die Poissonverteilung bietet sich an). Die Standardabweichung ist also $\sqrt{n \cdot p}$.

In einem solchen Fall kann man die Varianz annähernd stabilisieren, indem man sowohl das erklärende Merkmal als auch die Zielgröße durch Wurzelziehen stabilisiert.

Für alle, die es etwas genauer wissen wollen:

$$\begin{aligned} \sqrt{y} &= b \cdot \sqrt{x} + \varepsilon \\ \Rightarrow \quad y &= (b \cdot \sqrt{x} + \varepsilon)^2 \\ &= b^2 \cdot x + 2 \cdot b \cdot \sqrt{x} \cdot \varepsilon + \varepsilon^2 \end{aligned}$$

Die Standardabweichung ist hier nicht genau proportional zu $\sqrt{x}$, aber zumindest hat der Teil $2 \cdot b \cdot \sqrt{x} \cdot \varepsilon$ hat eine zu $\sqrt{x}$ proportionale Standardabweichung, nämlich $2 \cdot b \cdot \sqrt{x} \cdot \sigma$. Der Term ε^2 ist hingegen das σ^2-fache einer χ^2_1-verteilten Zufallsvariablen und hat Standardabweichung $\sigma^2 \cdot \sqrt{2}$. Zumindest wenn σ sehr klein ist im Vergleich zu $b \cdot \sqrt{x}$, gilt also die Approximation

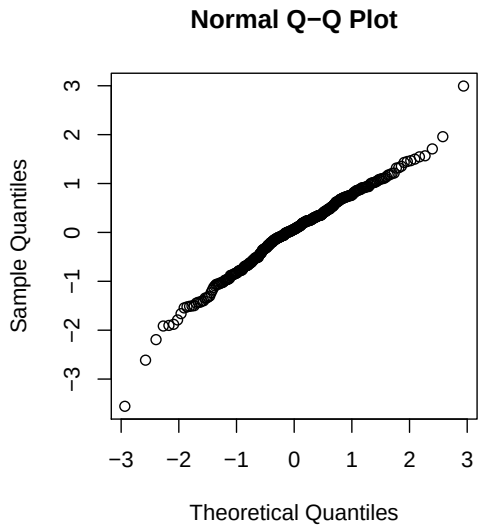
$$y \approx b^2 \cdot x + 2 \cdot b \cdot \sqrt{x} \cdot \varepsilon$$

und die Standardabweichung von y ist ungefähr proportional zu $\sqrt{x}$.

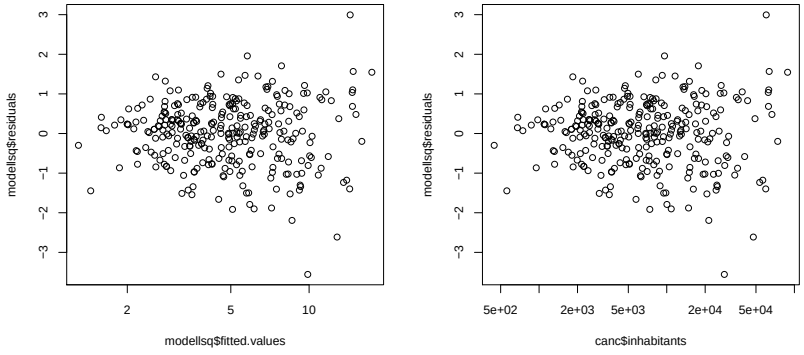
```
> modellsq <- lm(sqrt(deaths)~sqrt(inhabitants),data=canc)
> summary(modellsq)
```

```
Call:
lm(formula = sqrt(deaths) ~ sqrt(inhabitants), data = canc)
Residuals:
     Min       1Q   Median       3Q      Max
-3.55639 -0.51900  0.06204  0.54277  2.99434
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       0.0664320  0.0974338   0.682    0.496
sqrt(inhabitants) 0.0583722  0.0009171  63.651   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
Residual standard error: 0.8217 on 299 degrees of freedom
Multiple R-squared: 0.9313,Adjusted R-squared: 0.931
F-statistic:  4051 on 1 and 299 DF,  p-value: < 2.2e-16
```

```
qqnorm(modell$residuals)
```




```
plot(modellsq$fitted.values,modellsq$residuals,log='x')
  plot(canc$inhabitants,modellsq$residuals,log='x')
```



Der QQ-Plot mit der Normalverteilung sieht nicht perfekt aus, aber immerhin ist die Varianz stabilisiert.

Das Ergebnis bleibt aber dasselbe: Der "Intercept" ist nicht signifikant von 0 verschieden. Also kann der erwartete Wert von $\sqrt{y}$ proportional zu $\sqrt{x}$ sein, was bedeutet, dass auch der erwartete Wert von y ungefähr (!) proportional zu x ist. Also gibt es nach wie vor keinen deutlichen Hinweis auf eine Abhängigkeit zwischen der Einwohnerzahl und der Todesrate.

Schlusswort

Manchmal ist die Suche nach einer geeigneten varianzstabilisierenden Transformation mit viel Herumprobieren verbunden. In vielen Fällen passt die log-Transformation einigermaßen gut.

Was Sie u.a. erklären können sollten

- Modellannahmen der linearen Regression
  - Gleichung
  - Was ist zufällig und was nicht?
- Ansatz: kleinste Summe der quadrierten Residuen
- Optimale Lösung für Steigung und Achsenabschnitt
- Zusammenhang zwischen Steigung und Korrelation
- t-Test für lineare Zusammenhänge (Standardfehler, Teststatistik, Freiheitsgrade)
- Skalierung der Daten: wann, wie und wieso
- qqnorm-Plots
  - Theorie
  - Anwendung zur Beurteilung, ob das Modell passt