Multivariate Statistics in Ecology and Quantitative Genetics

3. Linear Regression and Linear Models

Dirk Metzler & Martin Hutzenthaler

http://evol.bio.lmu.de/StatGen.html

19. Mai 2010

Contents

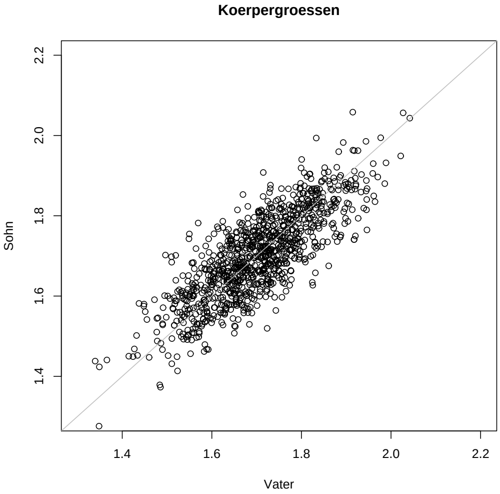
Origin of the word “Regression”

Sir Francis Galton (1822–1911): Regression toward the mean.

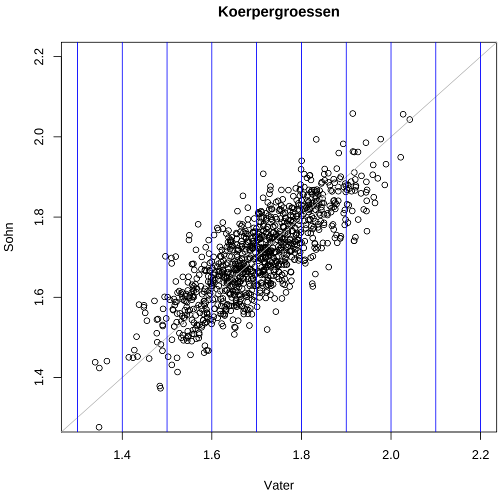
Origin of the word “Regression”

Sir Francis Galton (1822–1911): Regression toward the mean.

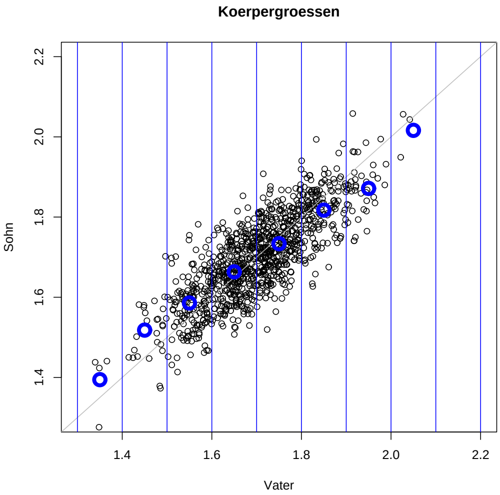
Tall fathers tend to have sons that are slightly smaller than the fathers. Sons of small fathers are on average larger than their fathers.

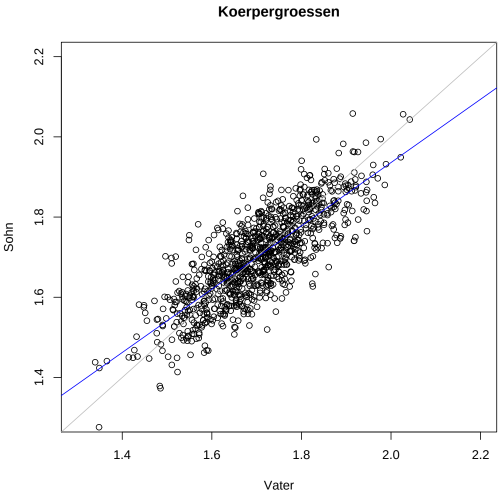

Koerpergroessen
Sohn
Vater
1.4
1.6
1.8
2.0
2.2
1.4
1.6
1.8
2.0
2.2

Regression toward the mean

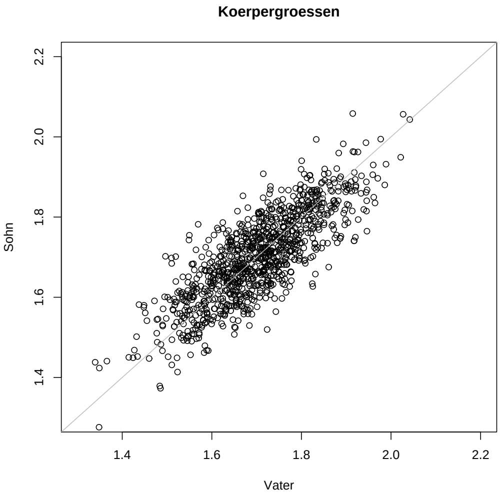


Regression toward the mean

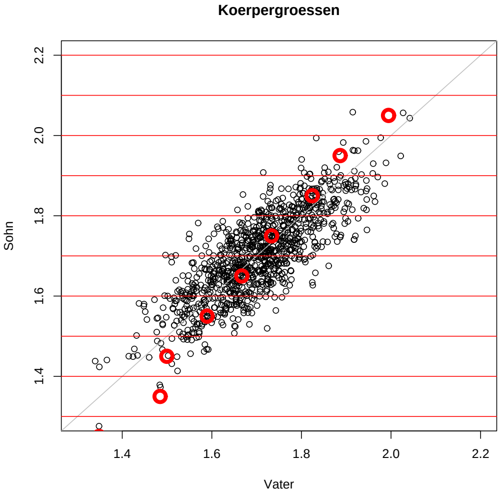



Koerpergroessen
Sohn
Vater
1.4
1.6
1.8
2.0
2.2
1.4
1.6
1.8
2.0
2.2

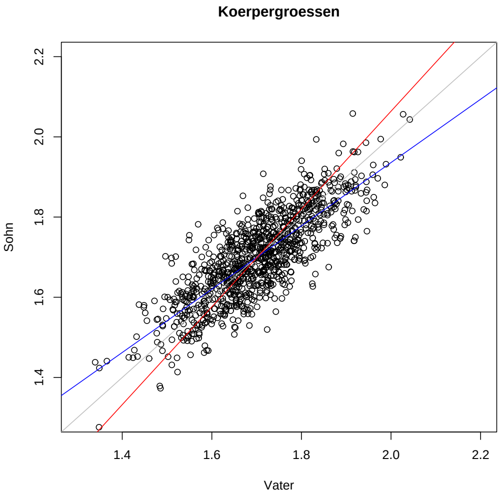
Regression toward the mean

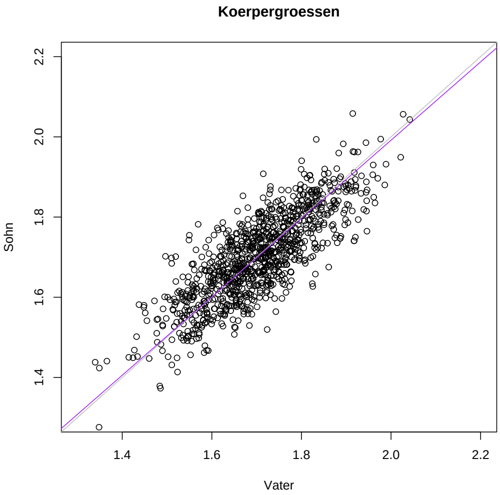


Regression toward the mean



Regression toward the mean




Koerpergroessen
Sohn
Vater
1.4
1.6
1.8
2.0
2.2
1.4
1.6
1.8
2.0
2.2

Similar effects

- In sports: The champion of the season will tend to fail the high expectations in the next year.

Similar effects

- In sports: The champion of the season will tend to fail the high expectations in the next year.
- In school: If the worst 10% of the students get extra lessons and are not the worst 10% in the next year, then this does not proof that the extra lessons are useful.

Contents

- Regression toward the mean
- Univariate linear regression: how and why?
- t-test for linear regression
- Examples with transformed variables
  - log-scaling brain sizes and body weights
  - root of numbers of inhabitants and deaths
- Multivariate Regression
  - Example: species richness on sandy beaches
  - Example: Success of different therapies
  - Example: Daphnia
- Cross validation and AIC

photo (c) by Jörg Hempel

Griffon Vulture
Gypus fulvus
German:
Gänsegeier

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

- Data from Goethe-University, Group of Prof. Prinzinger
- Developed telemetric system for measuring heart beats of flying birds

📄 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

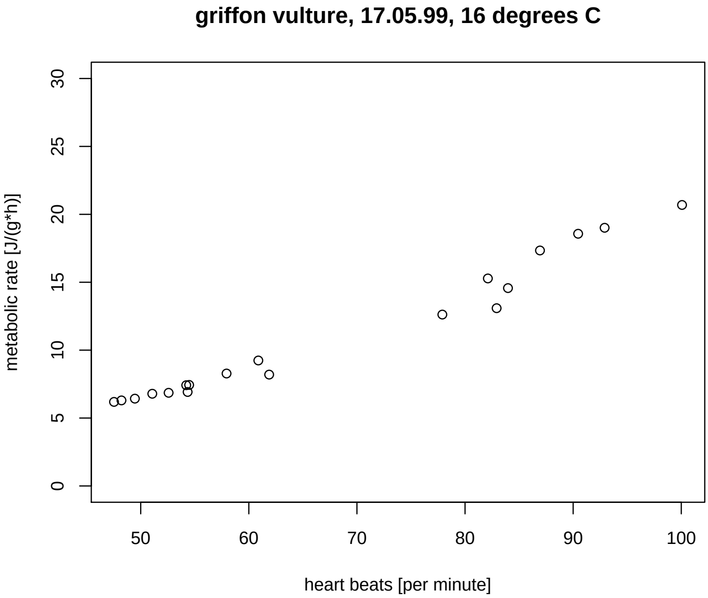
- Data from Goethe-University, Group of Prof. Prinzinger
- Developed telemetric system for measuring heart beats of flying birds
- Important for ecological questions: metabolic rate.

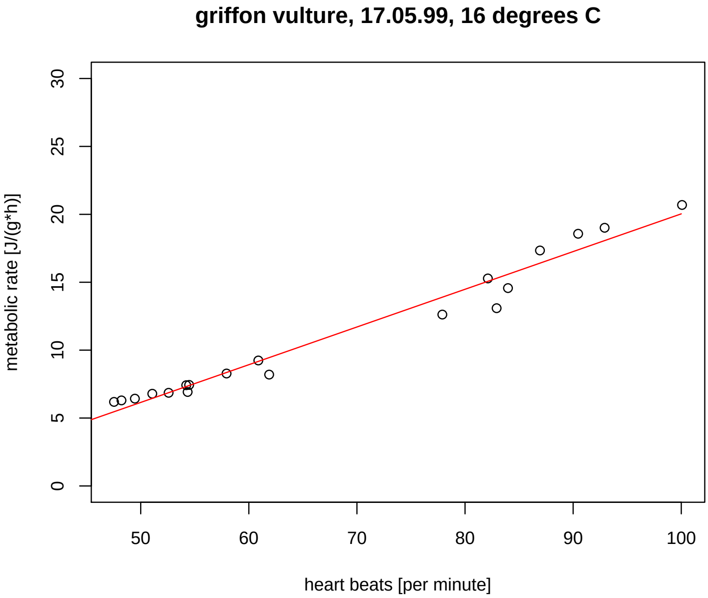
Prinzinger, R., E. Karl, R. Bögel, Ch. Walzer (1999): Energy metabolism, body temperature, and cardiac work in the Griffon vulture Gyps vulvus - telemetric investigations in the laboratory and in the field.
Zoology **102**, Suppl. II: 15

- Data from Goethe-University, Group of Prof. Prinzinger
- Developed telemetric system for measuring heart beats of flying birds
- Important for ecological questions: metabolic rate.
- metabolic rate can only be measured in the lab

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

- Data from Goethe-University, Group of Prof. Prinzinger
- Developed telemetric system for measuring heart beats of flying birds
- Important for ecological questions: metabolic rate.
- metabolic rate can only be measured in the lab
- can we infer metabolic rate from heart beat frequency?

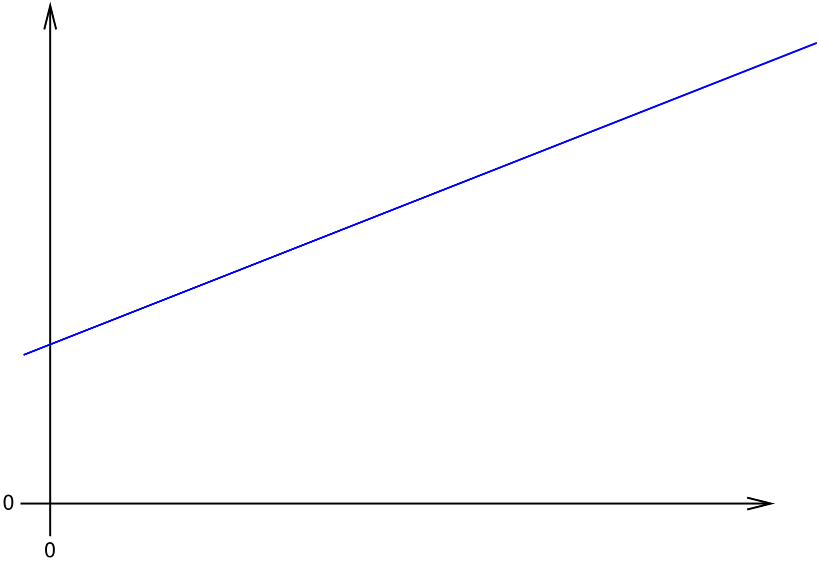

griffon vulture, 17.05.99, 16 degrees C
metabolic rate [J/(g*h)]
0
5
10
15
20
25
30
heart beats [per minute]
50
60
70
80
90
100

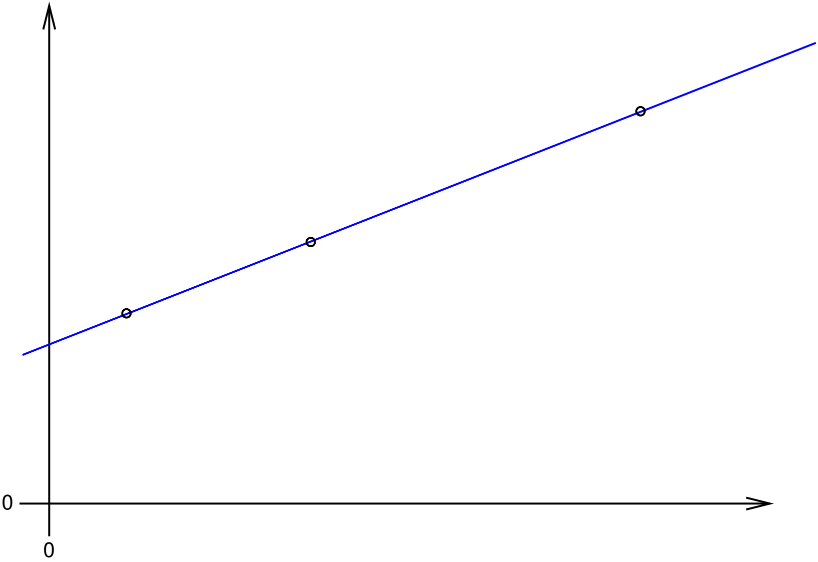

griffon vulture, 17.05.99, 16 degrees C
metabolic rate [J/(g*h)]
0
5
10
15
20
25
30
50
60
70
80
90
100
heart beats [per minute]

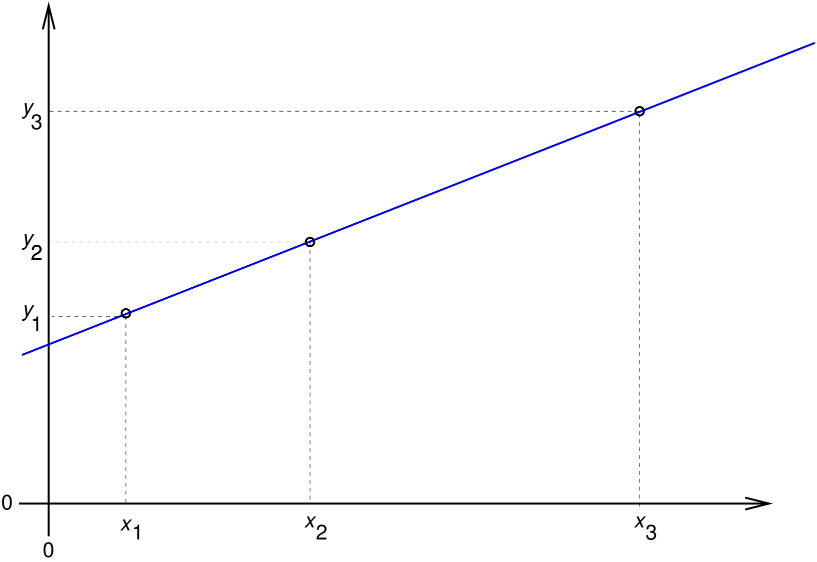
```
vulture
                 day heartbpm metabol minTemp maxTemp medtemp
1    01.04./02.04.    70.28   11.51      -6       2    -2.0
2    01.04./02.04.    66.13   11.07      -6       2    -2.0
3    01.04./02.04.    58.32   10.56      -6       2    -2.0
4    01.04./02.04.    58.63   10.62      -6       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
.       .                .       .          .       .       .
.       .                .       .          .       .       .
.       .                .       .          .       .       .
```

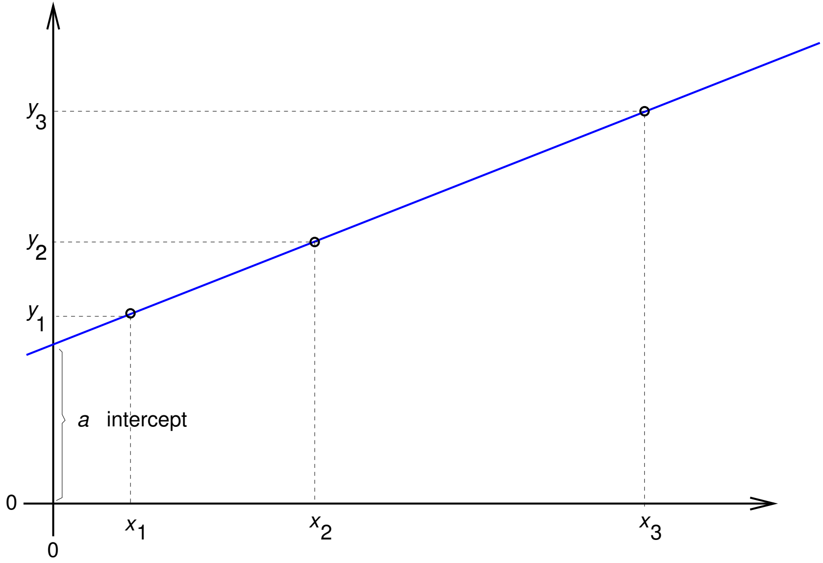
(14 different days)

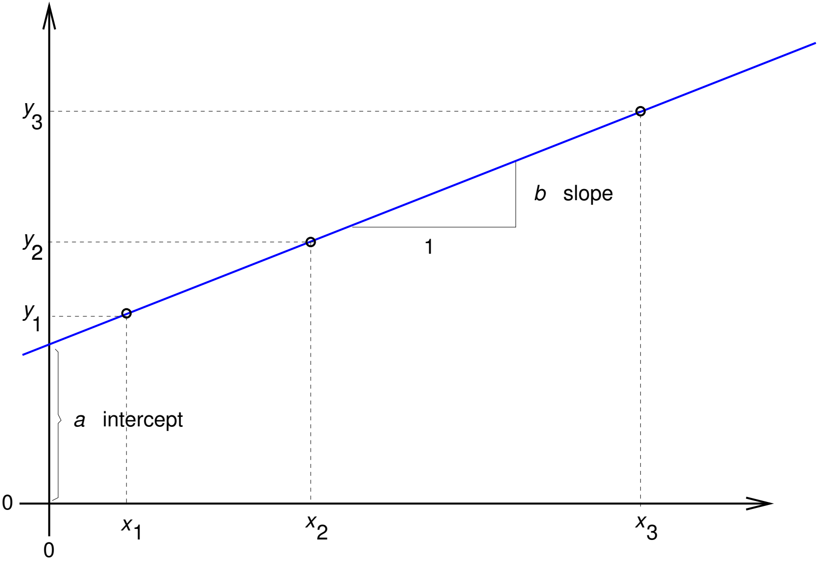
```
> model <- lm(metabol~heartbpm,data=vulture,
              subset=day=="17.05.")
> summary(model)
Call:
lm(formula = metabol ~ heartbpm, data = vulture, subset = day
    "17.05.")
Residuals:
    Min      1Q  Median      3Q     Max
-2.2026 -0.2555  0.1005  0.6393  1.1834
Coefficients:
            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
```

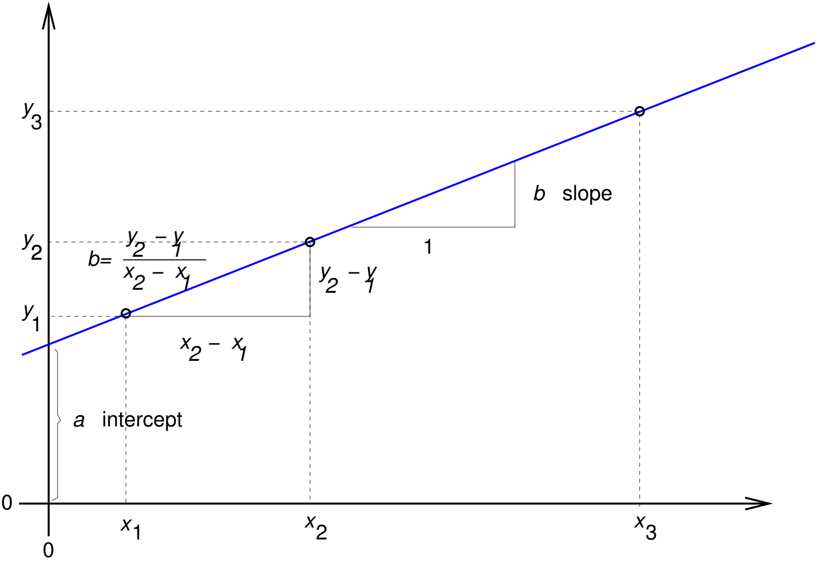

0
0

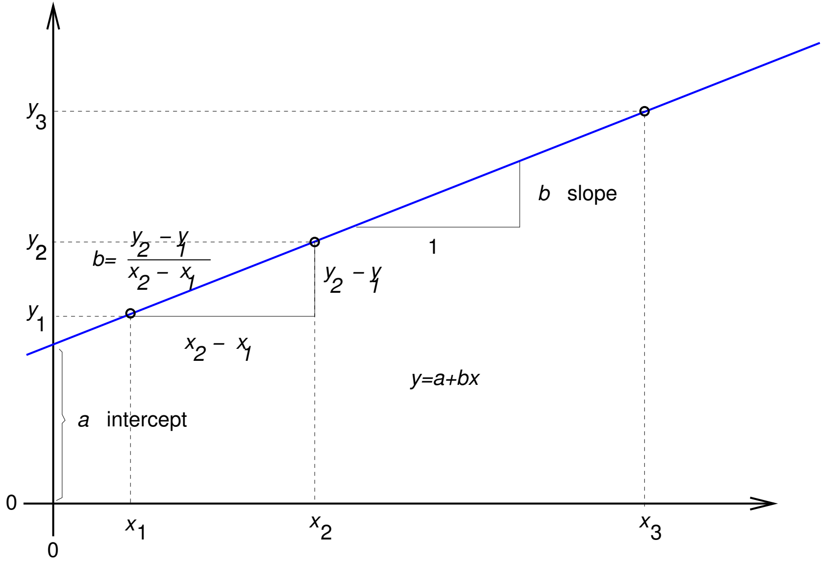

0
0

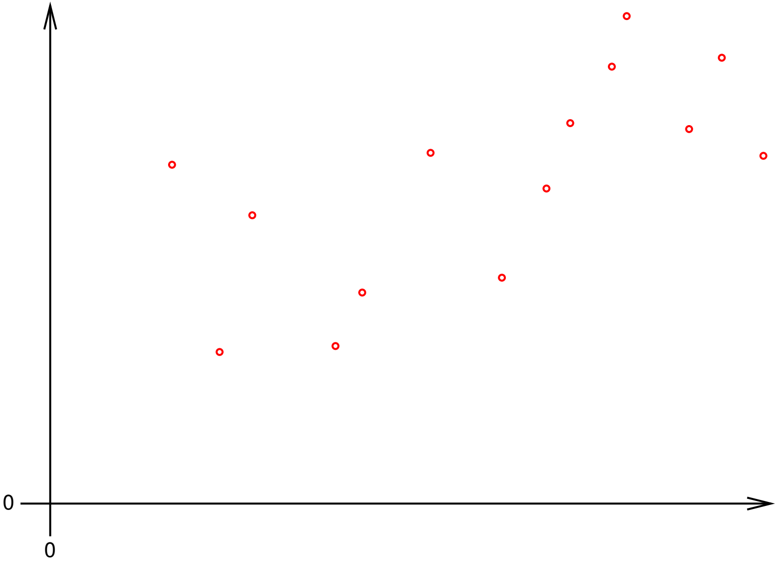

y_3
y_2
y_1
0
x_1
x_2
x_3
0

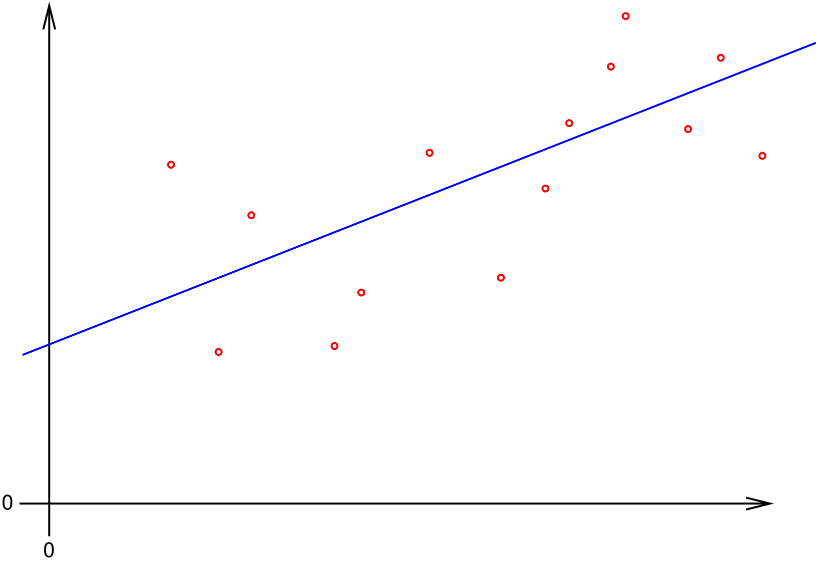

y_3
y_2
y_1
a intercept
0
x_1
x_2
x_3
0

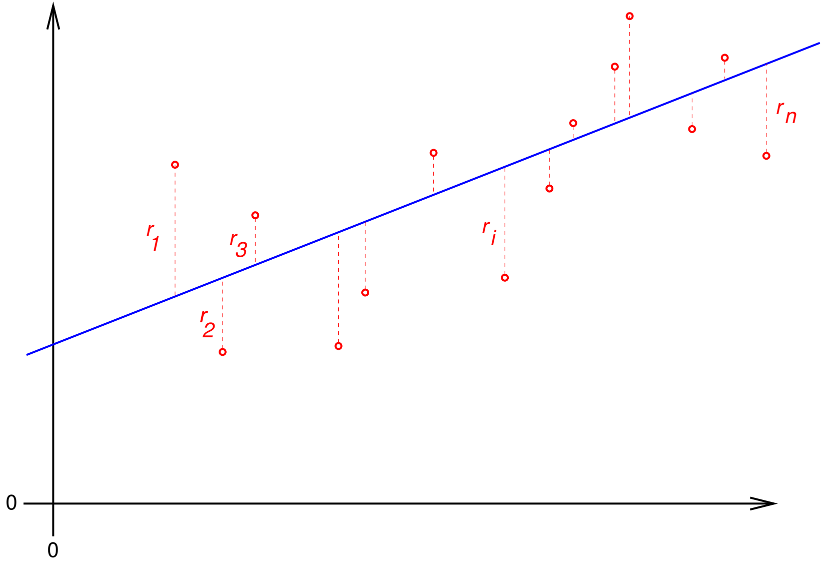

y_3
b slope
1
y_2
y_1
a intercept
0
x_1
x_2
x_3
0

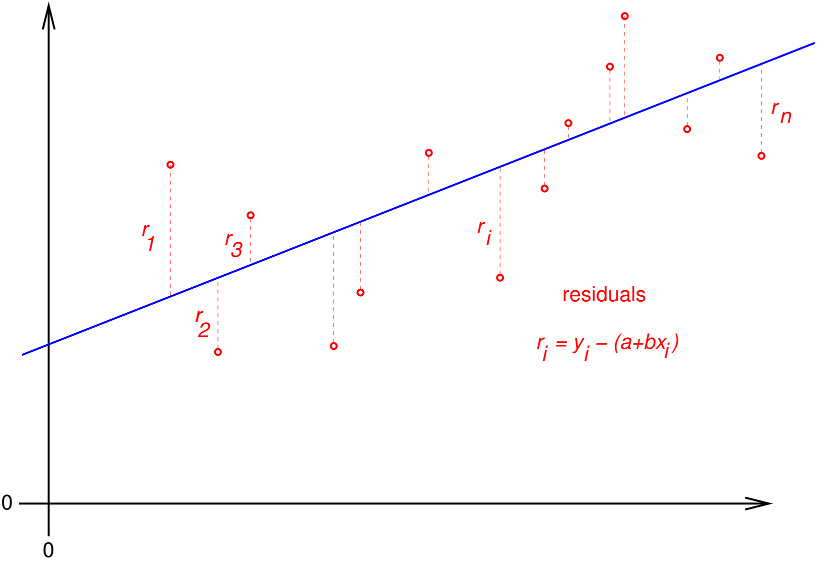

y_3
y_2
y_1
0
$b = \frac{y_2 - y_1}{x_2 - x_1}$
b slope
1
$y_2 - y_1$
$x_2 - x_1$
a intercept
x_1
x_2
x_3
0


y_3
y_2
y_1
$b = \frac{y_2 - y_1}{x_2 - x_1}$
b slope
1
$y_2 - y_1$
$x_2 - x_1$
$y=a+bx$
a intercept
0
x_1
x_2
x_3
0

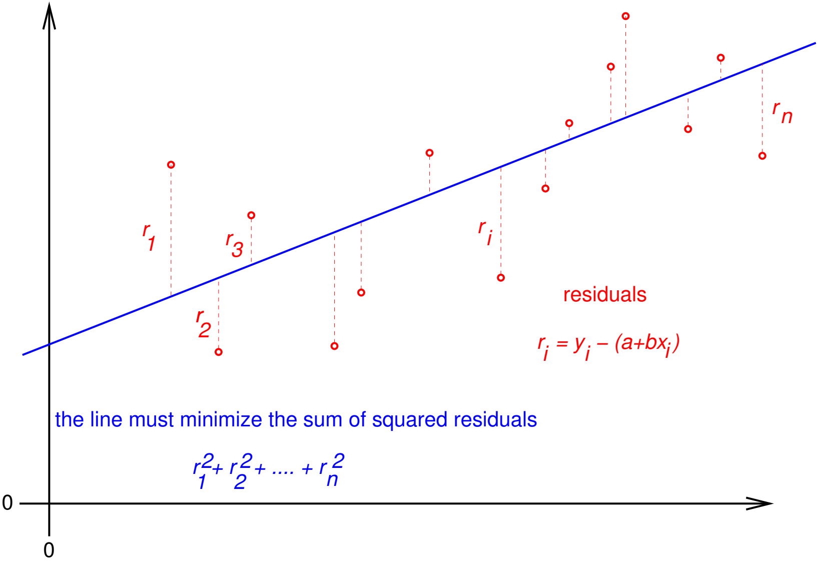

0
0


0
0


r_1
r_2
r_3
r_i
r_n
0
0


r_1
r_2
r_3
r_i
r_n
residuals
$r_i = y_i - (a+bx_i)$
0
0

Univariate linear regression: how and why?



residuals

$r_i = y_i - (a+bx_i)$

the line must minimize the sum of squared residuals

$r_1^2 + r_2^2 + \dots + r_n^2$

define the regression line

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

by minimizing the sum of squared residuals:

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

this is based on the model assumption that values a, b exist, such that, for all data points (x_i, y_i) we have

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

whereas all ε_i are independent and normally distributed with the same variance σ^2.

given data:

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

given data:

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

Model: there are values a, b, σ^2 such that

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

given data:

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

Model: there are values *a*, *b*, σ^2 such that

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

$\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$ are independent $\sim \mathcal{N}(0, \sigma^2)$.

given data:

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

Model: there are values a, b, σ^2 such that

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

$\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are independent $\sim \mathcal{N}(0, \sigma^2)$.

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

given data:

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

Model: there are values *a*, *b*, σ^2 such that

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

$\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$ are independent $\sim \mathcal{N}(0, \sigma^2)$.

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

a, b, σ^2 are unknown, but **not random**.

We estimate *a* and *b* by computing

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

We estimate a and b by computing

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

Theorem

Compute â and b̂ by

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

and

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

We estimate a and b by computing

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

Theorem

Compute â and $\hat{b}$ by

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

and

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

Please keep in mind:
The line $y = \hat{a} + \hat{b} \cdot x$ goes through the center of gravity of the cloud of points $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.

```
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
```

Optimizing clutch sizes

Example:*Cowpea weevil* (also *bruchid beetle*)
Callosobruchus maculatus
German: Erbsensamenkäfer

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.

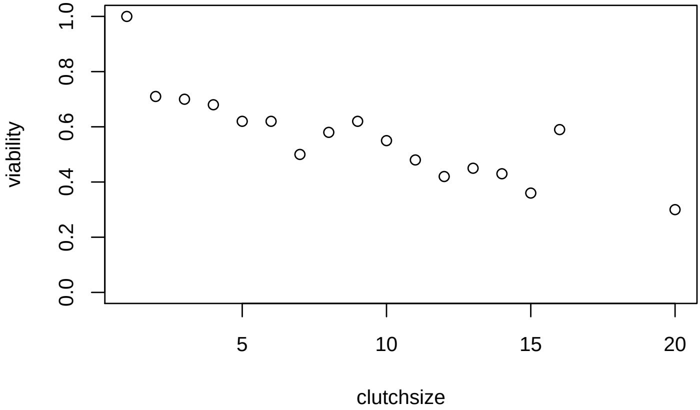
How does survival probability depnend on clutch size?

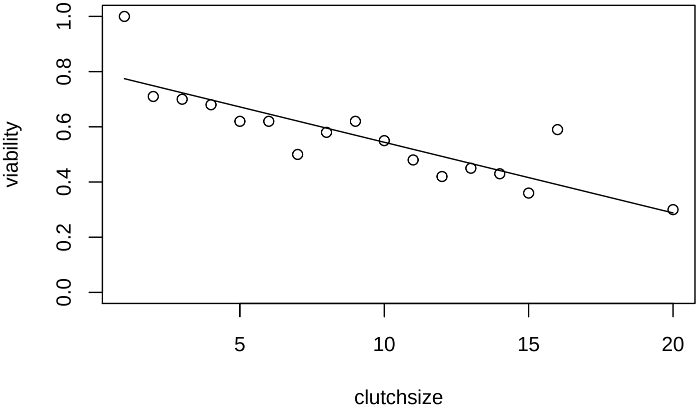
Optimizing clutch sizes

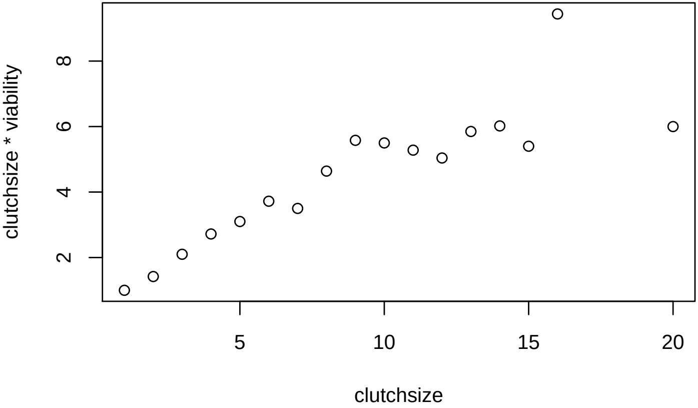
Example:*Cowpea weevil* (also *bruchid beetle*)
Callosobruchus maculatus
German: Erbsensamenkäfer

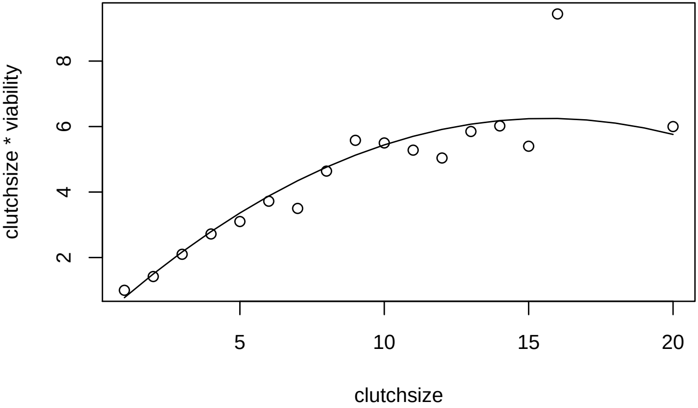
Wilson, K. (1994) Evolution of clutch size in insects. II. A test of static optimality models using the beetle Callosobruchus maculatus (Coleoptera: Bruchidae).
Journal of Evolutionary Biology **7:** 365–386.

How does survival probability depnend on clutch size?
Which clutch size optimizes the expected number of surviving offspring?


viability
0.0
0.2
0.4
0.6
0.8
1.0
5
10
15
20
clutchsize


viability
0.0
0.2
0.4
0.6
0.8
1.0
5
10
15
20
clutchsize


clutchsize * viability
2
4
6
8
5
10
15
20
clutchsize


clutchsize * viability
2
4
6
8
5
10
15
20
clutchsize

Contents

Regression toward the mean

Univariate linear regression: how and why?

t-test for linear regression

Examples with transformed variables

- log-scaling brain sizes and body weights
- root of numbers of inhabitants and deaths

Multivariate Regression

- Example: species richness on sandy beaches
- Example: Success of different therapies
- Example: Daphnia

Cross validation and AIC

Example: red deer (*Cervus elaphus*)

theory: femals can influence the sex of their offspring

Example: red deer (*Cervus elaphus*)

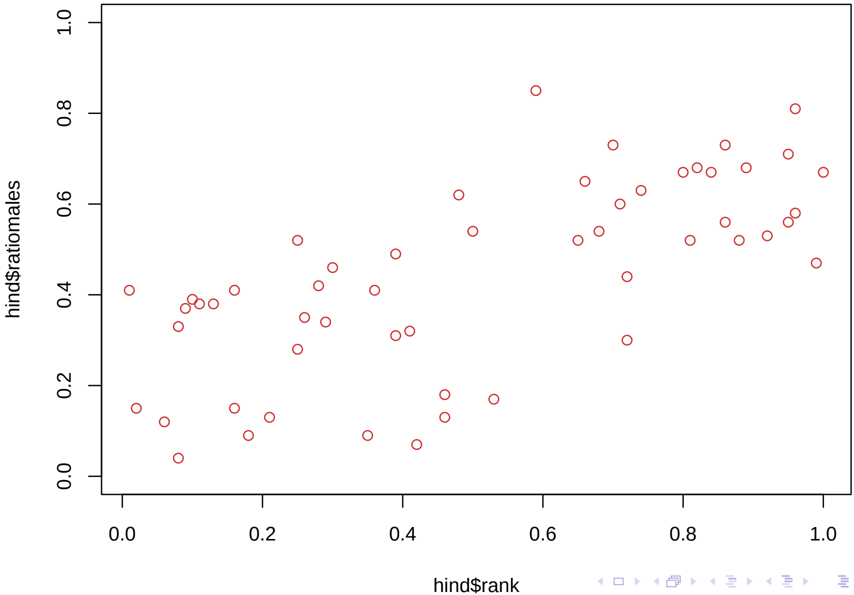
theory: femals can influence the sex of their offspring

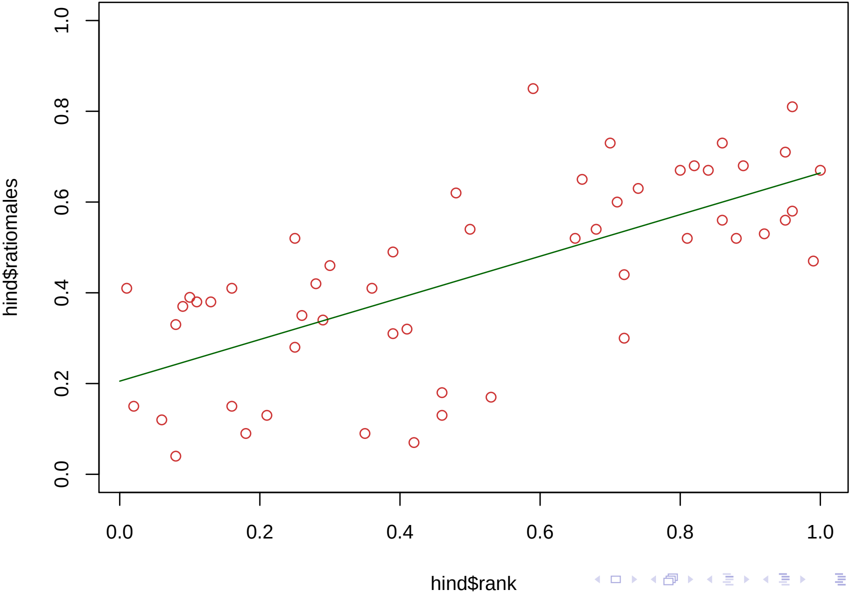
Evolutionary stable strategy: weak animals may tend to have female offspring, strong animals may tend to have male offspring.

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
```

CAUTION: Simulated data, inspired by original paper


hind$ratiomales
0.0
0.2
0.4
0.6
0.8
1.0
hind$rank
0.0
0.2
0.4
0.6
0.8
1.0


hind$ratiomales
hind$rank
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0

```
> 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
```

Model:

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

Model:

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

How to compute the significance of a relationship between the *explanatory trait* X and the *target variable* Y?

Model:

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

How to compute the significance of a relationship between the *explanatory trait* X and the *target variable* Y?

In other words: How can we test the null hypothesis $b = 0$?

Model:

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

How to compute the significance of a relationship between the *explanatory trait X* and the *target variable Y*?

In other words: How can we test the null hypothesis $b = 0$?

We have estimated b by $\hat{b} \neq 0$. Could the true b be 0?

Model:

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

How to compute the significance of a relationship between the *explanatory trait* X and the *target variable* Y?

In other words: How can we test the null hypothesis $b = 0$?

We have estimated b by $\hat{b} \neq 0$. Could the true b be 0?

How large is the standard error of $\hat{b}$?

t-test for $\hat{b}$

Estimate σ^2 by

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

Then,

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

is t-distributed with $n-2$ degrees of freedom. Thus, we can apply a t-test to test the null-hypothesis $b = 0$.

Contents

Regression toward the mean

Univariate linear regression: how and why?

t-test for linear regression

Examples with transformed variables
- log-scaling brain sizes and body weights
- root of numbers of inhabitants and deaths

Multivariate Regression
- Example: species richness on sandy beaches
- Example: Success of different therapies
- Example: Daphnia

Cross validation and AIC

Contents

Regression toward the mean

Univariate linear regression: how and why?

t-test for linear regression

Examples with transformed variables

- log-scaling brain sizes and body weights
- root of numbers of inhabitants and deaths

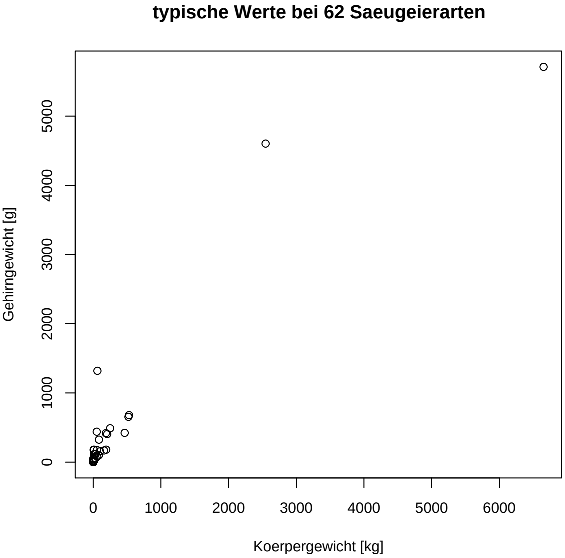
Multivariate Regression

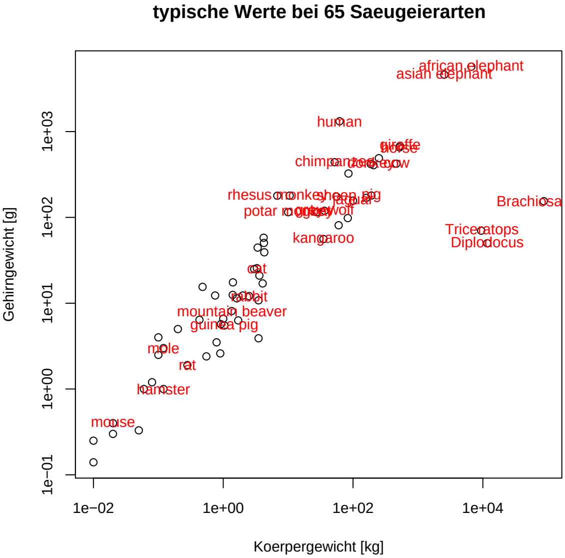
- Example: species richness on sandy beaches
- Example: Success of different therapies
- Example: Daphnia

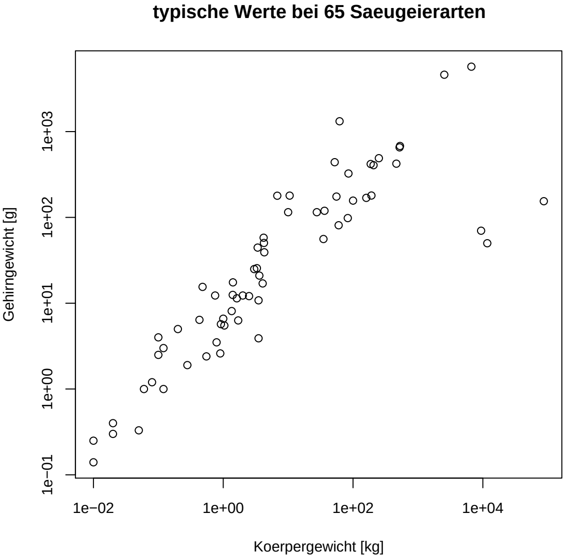
Cross validation and AIC

Data example: typical body weight [kg] and and brain weight [g] of 62 mammals species (and 3 dinosaurs)

```
> 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
10       52.16         440.00       chimpanzee  no
11        0.43           6.40
.         .               .           .
.         .               .           .
.         .               .           .
```

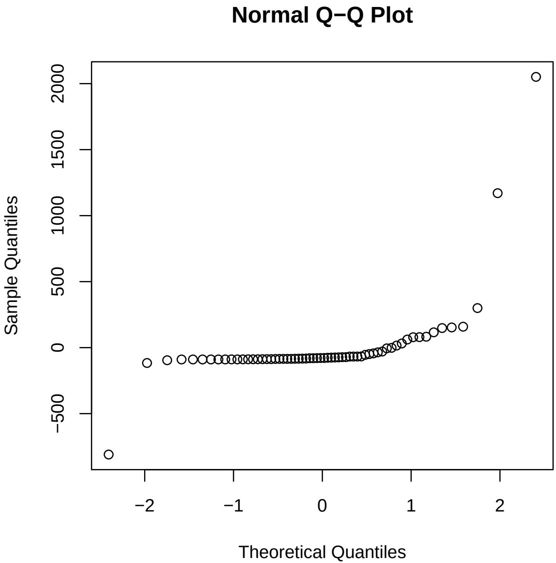

typische Werte bei 62 Saeugeierarten
Gehirngewicht [g]
0
1000
2000
3000
4000
5000
0
1000
2000
3000
4000
5000
6000
Koerpergewicht [kg]


typische Werte bei 65 Saeugeierarten
african elephant
asian elephant
human
giraffe
horse
chimpanzee
donkey
cow
rhesus monkey
sheep
pig
jaguar
Brachiosa
potar monkey
goat
grey wolf
Triceratops
Diplodocus
kangaroo
cat
rabbit
mountain beaver
guinea pig
mole
rat
hamster
mouse
Gehirngewicht [g]
1e−01
1e+00
1e+01
1e+02
1e+03
1e−02
1e+00
1e+02
1e+04
Koerpergewicht [kg]

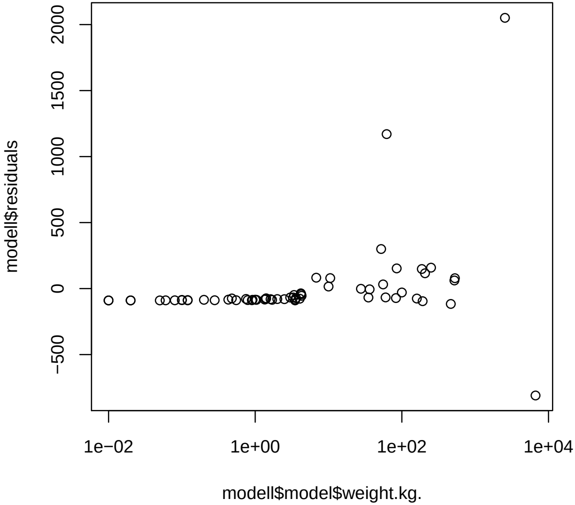

typische Werte bei 65 Saeugeierarten
Gehirngewicht [g]
1e−01
1e+00
1e+01
1e+02
1e+03
1e−02
1e+00
1e+02
1e+04
Koerpergewicht [kg]

```
> 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
```

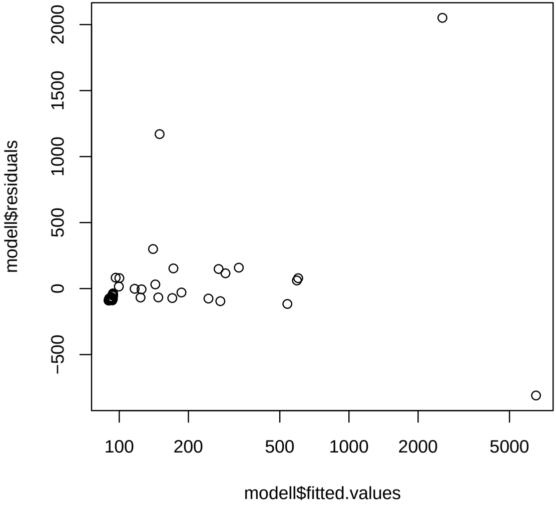
```
qqnorm(modell$residuals)
```



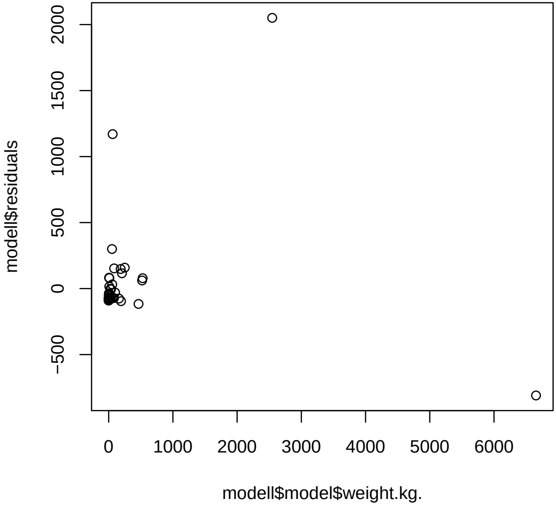
```
plot(modell$fitted.values,modell$residuals)
```



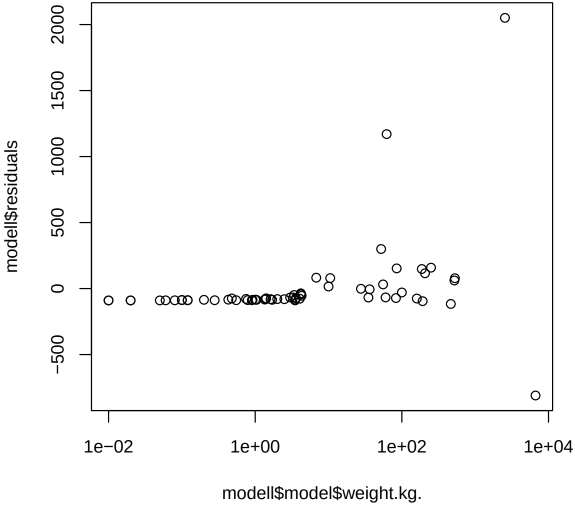
```
plot(modell$fitted.values,modell$residuals,log='x')
```



```
plot(modell$model$weight.kg.,modell$residuals)
```



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



We see that the residuals’ varaince depends on the fitted values (or the body weight): “heteroscadiscity”

We see that the residuals' varaince depends on the fitted values (or the body weight): “heteroscadiscity”
The model assumes *homoscedascity*, i.e. the random deviations must be (almost) independent of the explaining traits (body weight) and the fitted values.

We see that the residuals' varaince depends on the fitted values (or the body weight): "heteroscadiscity"
The model assumes *homoscedascity*, i.e. the random deviations must be (almost) independent of the explaining traits (body weight) and the fitted values.
variance-stabilizing transformation:
can be rescale body- and brain size to make deviations independent of variables

Actually not so surprising: An elephant's brain of typically 5 kg can easily be 500 g lighter or heavier from individual to individual. This can not happen for a mouse brain of typically 5 g. The latter will rather also vary by 10%, i.e. 0.5 g. Thus, the variance is not additive but rather multiplicative:

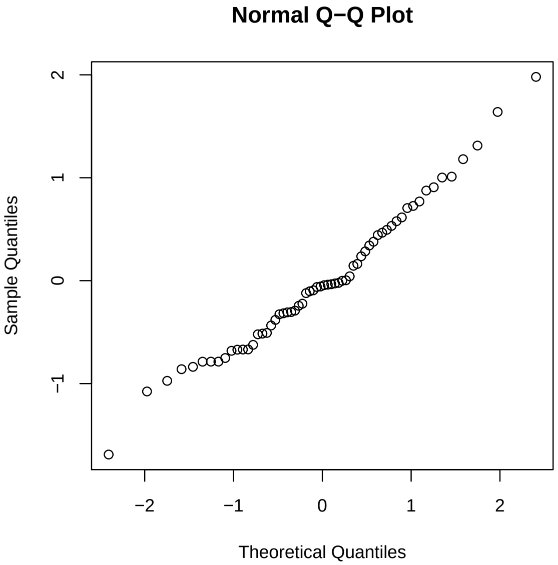
$$\text{brain mass} = (\text{expected brain mass}) \cdot \text{random}$$

We can convert this into something with additive randomness by taking the log:

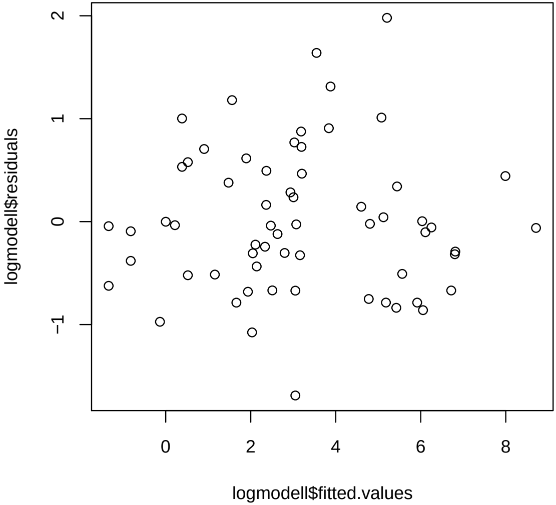
$$\log(\text{brain mass}) = \log(\text{expected brain mass}) + \log(\text{random})$$

```
> logmodell <- lm(log(brain.weight.g)~log(weight.kg.),subset=
> summary(logmodell)

Call:
lm(formula = log(brain.weight.g) ~ log(weight.kg.), subset = 
    "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
```

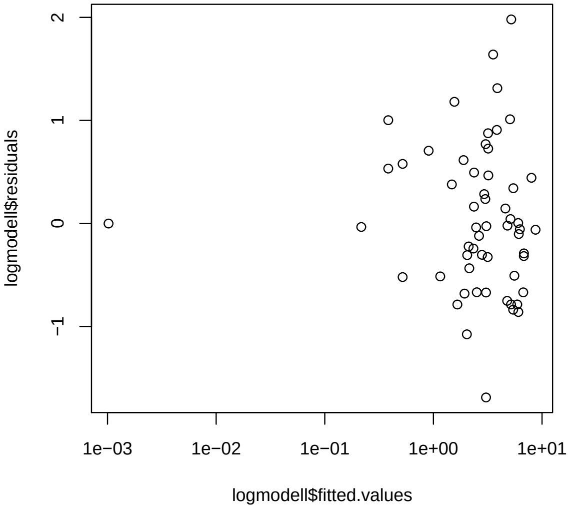
```
qqnorm(modell$residuals)
```



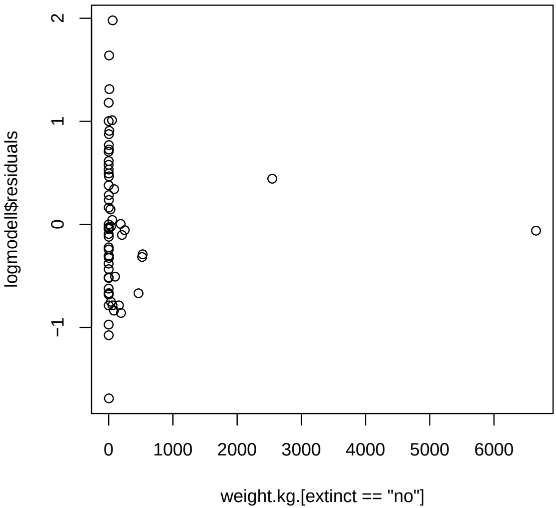
```
plot(logmodell$fitted.values,logmodell$residuals)
```



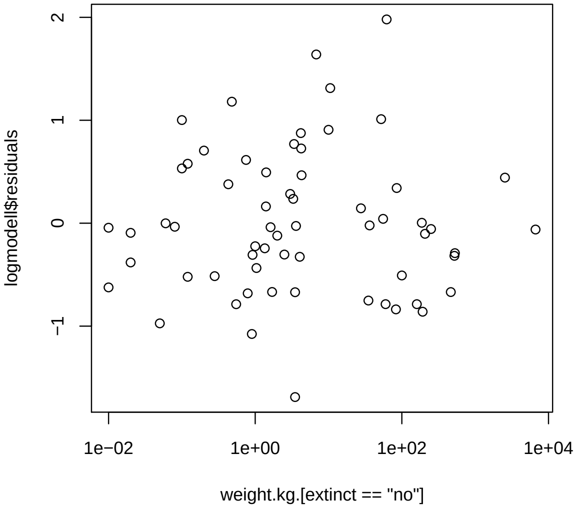
```
plot(logmodell$fitted.values,logmodell$residuals,log='x'
)
```



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



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



Contents

Regression toward the mean

Univariate linear regression: how and why?

t-test for linear regression

Examples with transformed variables
- log-scaling brain sizes and body weights
- root of numbers of inhabitants and deaths

Multivariate Regression
- Example: species richness on sandy beaches
- Example: Success of different therapies
- Example: Daphnia

Cross validation and AIC

Data: For 301 US-american (Counties) number of white female inhabitants from 1960 and number of deaths by breast cancer in this group between 1950 and 1960. (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
```

Is the average number of deaths proportional to population size, i.e.

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

or does the cancer risk depend on the size of the county, such that a different model fits better? e.g.

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

with $a \neq 0$.

```
> 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
```

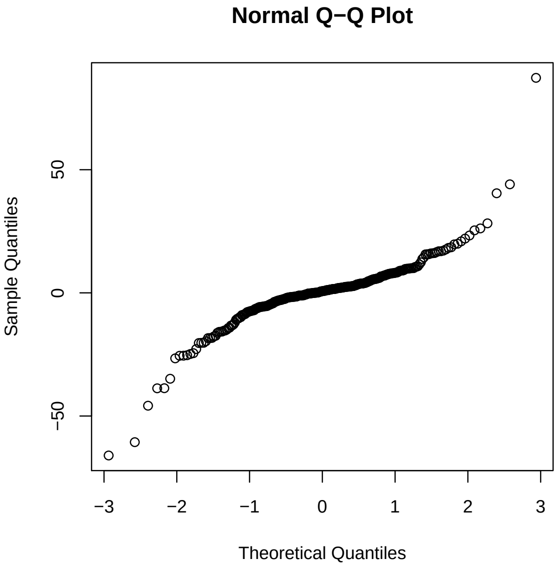
The intercept is estimated to -0.526, but not significantly different from 0.

The intercept is estimated to -0.526, but not significantly different from 0.

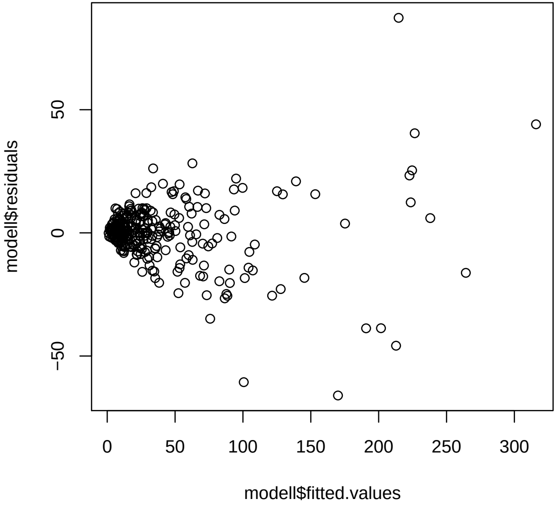
Thus we cannot reject the null hypothesis that the county size has no influence on the cancer risk.

The intercept is estimated to -0.526, but not significantly different from 0.
Thus we cannot reject the null hypothesis that the county size has no influence on the cancer risk.
But.. does the model fit?

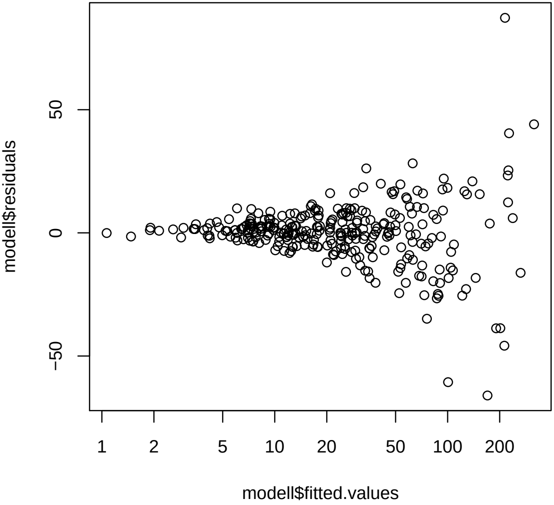
```
qqnorm(modell$residuals)
```



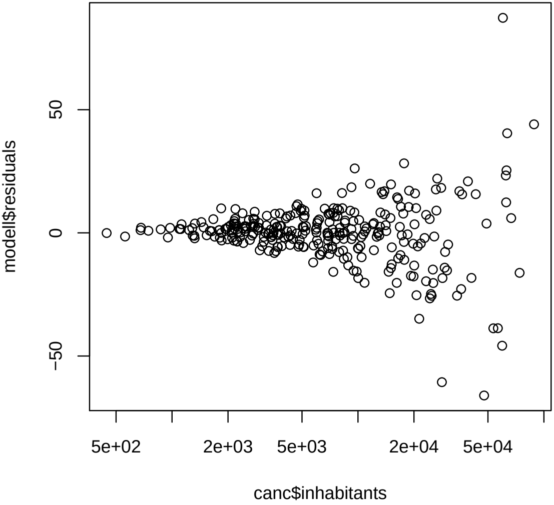
```
plot(modell$fitted.values,modell$residuals)
```



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



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



The variance of the residuals depends on the fitted values.
Heteroscedasticity

The variance of the residuals depends on the fitted values.
Heteroscedasticity
The linear model assumgs *Homoscedasticity*.

The variance of the residuals depends on the fitted values.
Heteroscedasticity
The linear model assumgs *Homoscedasticity*.
Variance Stabilizing Transformation:
How can we rescale the population size such that we obtain homoscedastic data?

Where does the variance come from?

Where does the variance come from?
If n is the number of white female inhabitants and p the individual probability to die by breast cancer within 10 years, then np is the expected number of deaths and the variance is

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

(Maybe approximate binomial by Poisson). Standard deviation: $\sqrt{n \cdot p}$.

Where does the variance come from?
If n is the number of white female inhabitants and p the individual probability to die by breast cancer within 10 years, then np is the expected number of deaths and the variance is

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

(Maybe approximate binomial by Poisson). Standard deviation: $\sqrt{n \cdot p}$.
In this case we can approximately stabilize variance by taking the root on both sides of the equation.

Explanation:

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

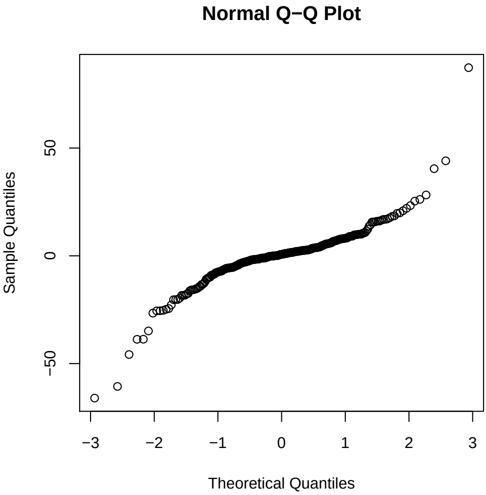
SD is not exactly proportional to $\sqrt{x}$, but at least $2 \cdot b \cdot \sqrt{x} \cdot \varepsilon$ has SD prop. to $\sqrt{x}$, namely $2 \cdot b \cdot \sqrt{x} \cdot \sigma$. The Term ε^2 is the σ^2-fold of a χ^2_1-distributed random variable and has SD=$\sigma^2 \cdot \sqrt{2}$. If σ is small compared to $b \cdot \sqrt{x}$, the approximation

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

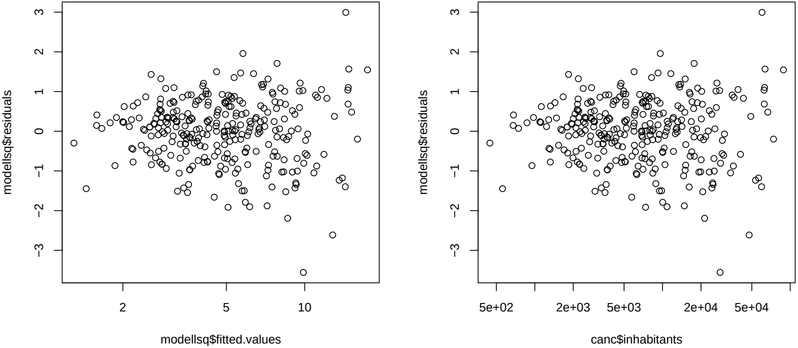
is reasonable and the SD of y is approximately proportional to $\sqrt{x}$.

```
> 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')
```



The qqnorm plot is not perfect by at least the variance is stabilized.

The qqnorm plot is not perfect by at least the variance is stabilized.
The result remains the same: No significant relation between county size and breast cancer death risk.

Contents

Regression toward the mean

Univariate linear regression: how and why?

t-test for linear regression

Examples with transformed variables

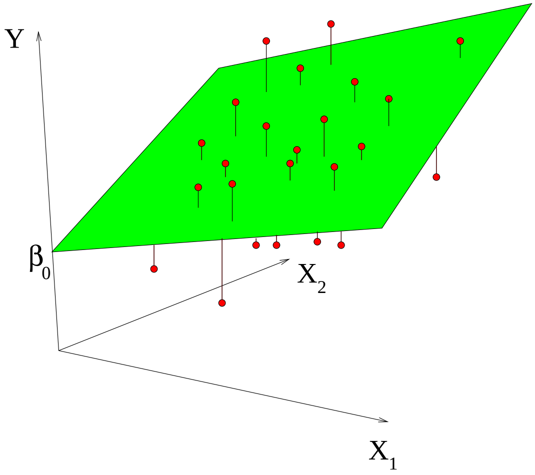
- log-scaling brain sizes and body weights
- root of numbers of inhabitants and deaths

Multivariate Regression

- Example: species richness on sandy beaches
- Example: Success of different therapies
- Example: Daphnia

Cross validation and AIC

Multivariate Regression



Multivariate Regression

Problem: Predict Y from X_1, X_2,. . . ,X_m.

Multivariate Regression

Problem: Predict Y from X_1, X_2,...,X_m.

Observations:

$$\begin{array}{ccc}
Y_1 & , & X_{11}, X_{21}, \ldots, X_{m1} \\
Y_2 & , & X_{12}, X_{22}, \ldots, X_{m2} \\
\vdots & & \vdots \\
Y_n & , & X_{1n}, X_{2n}, \ldots, X_{mn}
\end{array}$$

Multivariate Regression

Problem: Predict Y from X_1, X_2,...,X_m.

Observations:

$$\begin{array}{ccl} Y_1 & , & X_{11}, X_{21}, \ldots, X_{m1} \\ Y_2 & , & X_{12}, X_{22}, \ldots, X_{m2} \\ \vdots & & \vdots \\ Y_n & , & X_{1n}, X_{2n}, \ldots, X_{mn} \end{array}$$

Model: $Y = a + b_1 \cdot X_1 + b_2 \cdot X_2 + \cdots + b_m \cdot X_m + \varepsilon$

Multivariate Regression

Problem: Predict Y from X_1, X_2,. . . ,X_m.

Observations:

$$\begin{array}{ccl}
Y_1 & , & X_{11}, X_{21}, \ldots, X_{m1} \\
Y_2 & , & X_{12}, X_{22}, \ldots, X_{m2} \\
\vdots & & \vdots \\
Y_n & , & X_{1n}, X_{2n}, \ldots, X_{mn}
\end{array}$$

Model: $Y = a + b_1 \cdot X_1 + b_2 \cdot X_2 + \cdots + b_m \cdot X_m + \varepsilon$

Equation system to determine a, b_1, b_2, . . . , b_m:

$$\begin{array}{ccccccccccccc}
Y_1 & = & a & + & b_1 \cdot X_{11} & + & b_2 \cdot X_{21} & + & \ldots & + & b_m \cdot X_{m1} & + & \varepsilon_1 \\
Y_2 & = & a & + & b_1 \cdot X_{12} & + & b_2 \cdot X_{22} & + & \ldots & + & b_m \cdot X_{m2} & + & \varepsilon_2 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\
Y_n & = & a & + & b_1 \cdot X_{1n} & + & b_n \cdot X_{2n} & + & \ldots & + & b_m \cdot X_{mn} & + & \varepsilon_n
\end{array}$$

Model:

$$\begin{array}{ccccccccccccc}
Y_1 & = & a & + & b_1 \cdot X_{11} & + & b_2 \cdot X_{21} & + & \ldots & + & b_m \cdot X_{m1} & + & \varepsilon_1 \\
Y_2 & = & a & + & b_1 \cdot X_{12} & + & b_2 \cdot X_{22} & + & \ldots & + & b_m \cdot X_{m2} & + & \varepsilon_2 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\
Y_n & = & a & + & b_1 \cdot X_{1n} & + & b_n \cdot X_{2n} & + & \ldots & + & b_m \cdot X_{mn} & + & \varepsilon_n
\end{array}$$

target variable Y
explanatory variables $X_1, X_2, \ldots, X_m$
parameter to be estimated $a, b_1, \ldots, b_m$
independent normally distributed pertubations $\varepsilon_1, \ldots, \varepsilon_m$ with unknown variance σ^2.

Contents

Regression toward the mean

Univariate linear regression: how and why?

t-test for linear regression

Examples with transformed variables
- log-scaling brain sizes and body weights
- root of numbers of inhabitants and deaths

Multivariate Regression
- Example: species richness on sandy beaches
- Example: Success of different therapies
- Example: Daphnia

Cross validation and AIC

- Which factors influence the species richness on sandy beaches?
- Data from the dutch National Institute for Coastal and Marine Management Rijkswaterstaat/RIKZ
- see also

  Zuur, Ieno, Smith (2007) *Analysing Ecological Data.* Springer

```
  richness angle2  NAP  grainsize  humus week
1       11     96  0.045     222.5  0.05  1
2       10     96 -1.036     200.0  0.30  1
3       13     96 -1.336     194.5  0.10  1
4       11     96  0.616     221.0  0.15  1
.        .      .    .         .       .  .
.        .      .    .         .       .  .
21       3     21  1.117     251.5  0.00  4
22      22     21 -0.503     265.0  0.00  4
23       6     21  0.729     275.5  0.10  4
.        .      .    .         .       .  .
.        .      .    .         .       .  .
43       3     96 -0.002     223.0  0.00  3
44       0     96  2.255     186.0  0.05  3
45       2     96  0.865     189.5  0.00  3
```

Meaning of the Variables

richness	Number of species that were found in a plot.
angle2	slope of the beach a the plot
NAP	altitude of the plot compared to the mean sea level.
grainsize	average diameter of sand grains
humus	fraction of organic material
week	in which of 4 was this plot probed.

(many more variables in original data set)

Model 0:

$$\begin{aligned}\text{richness} &= a + b_1 \cdot \text{angle2} + b_2 \cdot \text{NAP} + b_3 \cdot \text{grainsize} + \\ &\quad + b_4 \cdot \text{humus} + \varepsilon\end{aligned}$$

Model 0:

$$\begin{aligned} \text{richness} &= a + b_1 \cdot \text{angle2} + b_2 \cdot \text{NAP} + b_3 \cdot \text{grainsize} + \\ &\quad + b_4 \cdot \text{humus} + \varepsilon \end{aligned}$$

in R notation:

```
richness ∼ angle2 + NAP + grainsize + humus
```

```
> model10 <- lm(richness ~ angle2+NAP+grainsize+humus,
+                 data = rikz)
> summary(model10)
Call:
lm(formula = richness ~ angle2 + NAP + grainsize + humus, data
Residuals:
    Min      1Q  Median      3Q     Max
-4.6851 -2.1935 -0.4218  1.6753 13.2957
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 18.35322    5.71888   3.209  0.00262 **
angle2      -0.02277    0.02995  -0.760  0.45144
NAP         -2.90451    0.59068  -4.917 1.54e-05 ***
grainsize   -0.04012    0.01532  -2.619  0.01239 *
humus       11.77641    9.71057   1.213  0.23234
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
Residual standard error: 3.644 on 40 degrees of freedom
Multiple R-squared: 0.5178, Adjusted R-squared: 0.4696
```

- e.g. -2.90451 is the estimator for b_2, the coefficient of `NAP`

- e.g. -2.90451 is the estimator for b_2, the coefficient of `NAP`
- The *p* value `Pr(>|t|)` refers to the null hypothesis that the true parameter value may be 0, i.e. the (potentially) explanatory variable (e.g. `NAP`) has actually no effect on the target variable (the species richness).

- e.g. -2.90451 is the estimator for b_2, the coefficient of `NAP`
- The p value `Pr(>|t|)` refers to the null hypothesis that the true parameter value may be 0, i.e. the (potentially) explanatory variable (e.g. `NAP`) has actually no effect on the target variable (the species richness).
- `NAP` is judged to be highly significant, `grainsize` also.

- e.g. -2.90451 is the estimator for b_2, the coefficient of `NAP`
- The p value `Pr(>|t|)` refers to the null hypothesis that the true parameter value may be 0, i.e. the (potentially) explanatory variable (e.g. `NAP`) has actually no effect on the target variable (the species richness).
- `NAP` is judged to be highly significant, `grainsize` also.
- Is there a significant week effect?

- e.g. -2.90451 is the estimator for b_2, the coefficient of `NAP`
- The p value `Pr(>|t|)` refers to the null hypothesis that the true parameter value may be 0, i.e. the (potentially) explanatory variable (e.g. `NAP`) has actually no effect on the target variable (the species richness).
- `NAP` is judged to be highly significant, `grainsize` also.
- Is there a significant week effect?
- Not the number 1,2,3,4 of the week should be multiplied with a coefficient. Instead, the numbers are taken as a non-numerical factor, i.e. each of the weeks 2,3,4 get a parameter that describes how much the species richness is increased compared to week 1.

- e.g. -2.90451 is the estimator for b_2, the coefficient of `NAP`
- The p value `Pr(>|t|)` refers to the null hypothesis that the true parameter value may be 0, i.e. the (potentially) explanatory variable (e.g. `NAP`) has actually no effect on the target variable (the species richness).
- `NAP` is judged to be highly significant, `grainsize` also.
- Is there a significant week effect?
- Not the number 1,2,3,4 of the week should be multiplied with a coefficient. Instead, the numbers are taken as a non-numerical factor, i.e. each of the weeks 2,3,4 get a parameter that describes how much the species richness is increased compared to week 1.
- In R this is done by changing `week` into a `factor`.

Model 0:

$$\begin{aligned}\text{richness} &= a + b_1 \cdot \text{angle2} + b_2 \cdot \text{NAP} + b_3 \cdot \text{grainsize} + \\ &\quad + b_4 \cdot \text{humus} + \\ &\quad b_5 \cdot I_{\texttt{week}=2} + b_6 \cdot I_{\texttt{week}=3} + b_7 \cdot I_{\texttt{week}=4} + \varepsilon\end{aligned}$$

$I_{\texttt{week}=k}$ is a so-called indicator variable which is 1 if $\texttt{week}= k$ and 0 otherwise.

Model 0:

$$\begin{aligned}\text{richness} \quad = \quad & a + b_1 \cdot \text{angle2} + b_2 \cdot \text{NAP} + b_3 \cdot \text{grainsize} + \\ & + b_4 \cdot \text{humus} + \\ & b_5 \cdot I_{\texttt{week}=2} + b_6 \cdot I_{\texttt{week}=3} + b_7 \cdot I_{\texttt{week}=4} + \varepsilon\end{aligned}$$

$I_{\texttt{week}=k}$ is a so-called indicator variable which is 1 if $\texttt{week} = k$ and 0 otherwise.

e.g. b_7 describes, by how much the species richness in an average plot probed in week 3 is increased compared to week 1.

Model 0:

$$\begin{aligned} \text{richness} &= a + b_1 \cdot \text{angle2} + b_2 \cdot \text{NAP} + b_3 \cdot \text{grainsize} + \\ &\quad + b_4 \cdot \text{humus} + \\ &\quad b_5 \cdot I_{\texttt{week}=2} + b_6 \cdot I_{\texttt{week}=3} + b_7 \cdot I_{\texttt{week}=4} + \varepsilon \end{aligned}$$

$I_{\texttt{week}=k}$ is a so-called indicator variable which is 1 if $\texttt{week}= k$ and 0 otherwise.

e.g. b_7 describes, by how much the species richness in an average plot probed in week 3 is increased compared to week 1.

in R notation:

```
richness ∼ angle2 + NAP + grainsize + humus +
factor(week)
```

```
> modell <- lm(richness ~ angle2+NAP+grainsize+humus
+                   +factor(week), data = rikz)
> summary(modell)
 .
 .
 .

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    9.298448   7.967002   1.167 0.250629
angle2         0.016760   0.042934   0.390 0.698496
NAP           -2.274093   0.529411  -4.296 0.000121 ***
grainsize      0.002249   0.021066   0.107 0.915570
humus          0.519686   8.703910   0.060 0.952710
factor(week)2 -7.065098   1.761492  -4.011 0.000282 ***
factor(week)3 -5.719055   1.827616  -3.129 0.003411 **
factor(week)4 -1.481816   2.720089  -0.545 0.589182
---
```

- Obviously, in weeks 2 and 3 significantly less species were found than in week 1, which is our reference point here.

- Obviously, in weeks 2 and 3 significantly less species were found than in week 1, which is our reference point here.
- The estimated `Intercept` is thus the expected species richness in week 1 in a plot where all other parameters take the value 0.

- Obviously, in weeks 2 and 3 significantly less species were found than in week 1, which is our reference point here.
- The estimated `Intercept` is thus the expected species richness in week 1 in a plot where all other parameters take the value 0.
- An alternative representation without `Intercept` takes 0 as reference point.

```
> modell.alternativ <- lm(richness ~ angle2+NAP+
+             grainsize+humus+factor(week)-1, data = rikz)
> summary(modell.alternativ)
  .
  .
  .
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
angle2         0.016760   0.042934   0.390 0.698496
NAP           -2.274093   0.529411  -4.296 0.000121 ***
grainsize      0.002249   0.021066   0.107 0.915570
humus          0.519686   8.703910   0.060 0.952710
factor(week)1  9.298448   7.967002   1.167 0.250629
factor(week)2  2.233349   8.158816   0.274 0.785811
factor(week)3  3.579393   8.530193   0.420 0.677194
factor(week)4  7.816632   6.522282   1.198 0.238362
```

the *p* values refer to the question whether the four intercepts for the different weeks are significantly different from 0.
The four *p* values refer to the null hypotheses that the additive parameter of a week is 0.

How do we test whether there is a difference between the weeks?

How do we test whether there is a difference between the weeks?

We saw before that weeks 2 and 3 are significantly different from week 1.

How do we test whether there is a difference between the weeks?

We saw before that weeks 2 and 3 are significantly different from week 1. However, the *p* value refers to the situation of single testing.

How do we test whether there is a difference between the weeks?

We saw before that weeks 2 and 3 are significantly different from week 1. However, the *p* value refers to the situation of single testing.

If we perform pairwise test for the weeks, we end up with $\binom{4}{2} = 6$ tests.

How do we test whether there is a difference between the weeks?

We saw before that weeks 2 and 3 are significantly different from week 1. However, the *p* value refers to the situation of single testing.

If we perform pairwise test for the weeks, we end up with $\binom{4}{2} = 6$ tests.

Bonferroni correction: Multiply each *p* value with the number of tests performed, in our case 6.

Bonferroni correction

Problem: If you perform many tests, some of them will reject the null hypothesis even if the null hypothesis is true.

Bonferroni correction

Problem: If you perform many tests, some of them will reject the null hypothesis even if the null hypothesis is true.

Example: If you perform 20 tests where the null hypothesis is actually true, then on average 1 test will falsly reject the null hypothesis on the 5% level.

Bonferroni correction

Problem: If you perform many tests, some of them will reject the null hypothesis even if the null hypothesis is true.

Example: If you perform 20 tests where the null hypothesis is actually true, then on average 1 test will falsly reject the null hypothesis on the 5% level.

Bonferroni correction: Multiply all *p* values with the number of tests performed. Reject the null hypotheses where the result is still smaller than the significance level.

Bonferroni correction

Problem: If you perform many tests, some of them will reject the null hypothesis even if the null hypothesis is true.

Example: If you perform 20 tests where the null hypothesis is actually true, then on average 1 test will falsly reject the null hypothesis on the 5% level.

Bonferroni correction: Multiply all p values with the number of tests performed. Reject the null hypotheses where the result is still smaller than the significance level.

Disadvantage: Conservative: Often, the null hypothies cannot be rejected even it is not true (type-2-error).

Alternative: Test whether there is a week effect by using an analysis of variance (anova) to compare a model with week effect to a model without week effect.

Alternative: Test whether there is a week effect by using an analysis of variance (anova) to compare a model with week effect to a model without week effect.

Only works for nested models, i.e. the simpler model can be obtained by restricting some parameters of the richer model to certain values or equations. In our case: “all week summands are equal”.

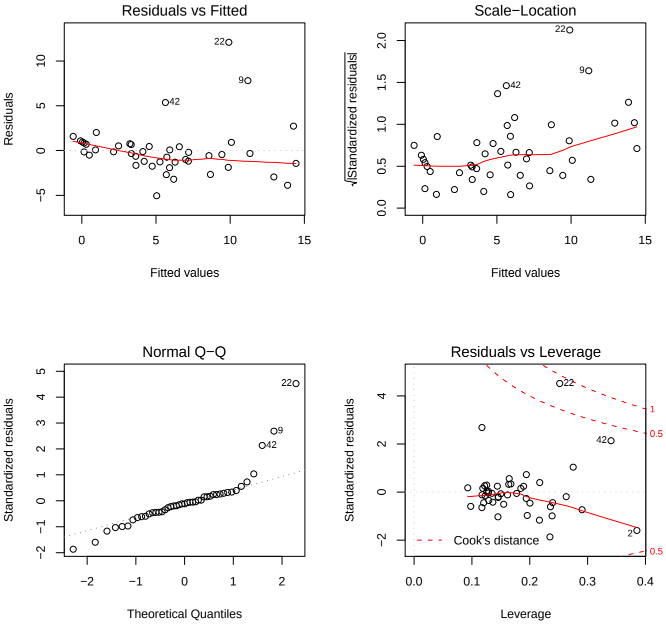
```
> modell0 <- lm(richness ~ angle2+NAP+grainsize+humus,
+                   data = rikz)
> modell <- lm(richness ~ angle2+NAP+grainsize+humus
+                          +factor(week), data = rikz)
> anova(modell0, modell)
Analysis of Variance Table

Model 1: richness ~ angle2 + NAP + grainsize + humus
Model 2: richness ~ angle2 + NAP + grainsize + humus + factor
  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1     40 531.17
2     37 353.66  3    177.51 6.1902 0.00162 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

We reject the null hypothesis that the weeks have no effect with a p-value of 0.00162.

We reject the null hypothesis that the weeks have no effect with a *p*-value of 0.00162.

But wait! We can only do that if the more complex model fits well to the data. We check this graphically.

```
plot(modell)
```


Residuals vs Fitted
Residuals
Fitted values
10
5
0
−5
0
5
10
15
22
9
42
Scale−Location
√|Standardized residuals|
Fitted values
2.0
1.5
1.0
0.5
0.0
0
5
10
15
22
9
42
Normal Q−Q
Standardized residuals
Theoretical Quantiles
5
4
3
2
1
0
−1
−2
−2
−1
0
1
2
22
9
42
Residuals vs Leverage
Standardized residuals
Leverage
4
2
0
−2
0.0
0.1
0.2
0.3
0.4
22
42
2
1
0.5
0.5
Cook's distance

Probes 22, 42, and 9 are considered as outliers.

Probes 22, 42, and 9 are considered as outliers.

Can we explain this by taking more parameters into account or are these real outliers, which are atypical and must be analysed separately.

Is there an interaction between NAP and angle2?

Is there an interaction between NAP and angle2?

$$\begin{aligned}\text{richness} &= a + b_1 \cdot \text{angle2} + b_2 \cdot \text{NAP} + b_3 \cdot \text{grainsize} + \\ &\quad + b_4 \cdot \text{humus} + \\ &\quad + b_5 \cdot I_{\texttt{week}=2} + b_6 \cdot I_{\texttt{week}=3} + b_7 \cdot I_{\texttt{week}=4} \\ &\quad b_8 \cdot \text{angle2} \cdot \text{NAP} + \varepsilon\end{aligned}$$

in R notation:

```
richness ~ angle2 + NAP + angle2:NAP+grainsize + humus
+ factor(week)
```

Is there an interaction between NAP and angle2?

$$\begin{aligned}\text{richness} \quad = \quad & a + b_1 \cdot \text{angle2} + b_2 \cdot \text{NAP} + b_3 \cdot \text{grainsize} + \\ & + b_4 \cdot \text{humus} + \\ & + b_5 \cdot I_{\texttt{week}=2} + b_6 \cdot I_{\texttt{week}=3} + b_7 \cdot I_{\texttt{week}=4} \\ & b_8 \cdot \text{angle2} \cdot \text{NAP} + \varepsilon\end{aligned}$$

in R notation:

```
richness ∼ angle2 + NAP + angle2:NAP+grainsize + humus
+ factor(week)
```

short-cut:

```
richness ∼ angle2*NAP+grainsize + humus + factor(week)
```

```
> modell3 <- lm(richness ~ angle2*NAP+grainsize+humus
+                  +factor(week), data = rikz)
> summary(modell3)
[...]
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   10.438985   8.148756   1.281 0.208366
angle2         0.007846   0.044714   0.175 0.861697
NAP           -3.011876   1.099885  -2.738 0.009539 **
grainsize      0.001109   0.021236   0.052 0.958658
humus          0.387333   8.754526   0.044 0.964955
factor(week)2 -7.444863   1.839364  -4.048 0.000262 ***
factor(week)3 -6.052928   1.888789  -3.205 0.002831 **
factor(week)4 -1.854893   2.778334  -0.668 0.508629
angle2:NAP     0.013255   0.017292   0.767 0.448337
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Different types of ANOVA tables

If you apply the R command `anova` to a single model, the variables are added consecutively in the same order as in the command. Each *p* value refers to the test wether the model gets significantly better by adding the variable to only those that are listed above the variable. In contrast to this, the *p* values that are given by `summary` or by `dropterm` from the MASS library always compare the model to a model where only the corresponding variable is set to 0 and all other variables can take any values. The *p* values given by `anova` thus depend on the order in which the variables are given in the command. This is not the case for `summary` and `dropterm`. The same options exist in other software packages, sometimes under the names “type I analysis” and “type II analysis”.

The same model is specified twice:

```
> modellA <- lm(richness ~ angle2+NAP+humus
+          +factor(week)+grainsize,data = rikz)
> modellB <- lm(richness ~ angle2+grainsize
+          +NAP+humus+factor(week), data = rikz)
```

Look at the *p*-valus of `grainsize`

```
> anova(modellA)
Analysis of Variance Table

Response: richness
             Df Sum Sq Mean Sq F value    Pr(>F)
angle2        1 124.86  124.86 13.0631 0.0008911 ***
NAP           1 319.32  319.32 33.4071 1.247e-06 ***
humus         1  35.18   35.18  3.6804 0.0627983 .
factor(week)  3 268.51   89.50  9.3638 9.723e-05 ***
grainsize     1   0.11    0.11  0.0114 0.9155704
Residuals    37 353.66    9.56
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

```
> anova(modellB)
Analysis of Variance Table

Response: richness
            Df Sum Sq Mean Sq F value    Pr(>F)
angle2       1 124.86  124.86 13.0631 0.00089 ***
grainsize    1  35.97   35.97  3.7636 0.06003 .
NAP          1 390.11  390.11 40.8127 1.8e-07 ***
humus        1  19.53   19.53  2.0433 0.16127
factor(week) 3 177.51   59.17  6.1902 0.00162 **
Residuals   37 353.66    9.56
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

```
> library(MASS)
> dropterm(modellA,test="F")
Single term deletions

Model:
richness ~ angle2 + NAP + humus + factor(week) + grainsize
           Df Sum of Sq    RSS    AIC  F Value    Pr(F)
<none>                  353.66 108.78
angle2      1      1.46 355.12 106.96     0.15 0.6984
NAP         1    176.37 530.03 124.98    18.45 0.0001 ***
humus       1      0.03 353.70 106.78 0.003565 0.9527
factor(week)3    177.51 531.17 121.08     6.19 0.0016 **
grainsize   1      0.11 353.77 106.79     0.01 0.9155
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

```
> dropterm(modellB,test="F")
Single term deletions

Model:
richness ~ angle2 + grainsize + NAP + humus + factor(week
           Df Sum of Sq    RSS    AIC  F Value     Pr(F)
<none>                  353.66 108.78
angle2      1      1.46 355.12 106.96     0.15 0.6984
grainsize   1      0.11 353.77 106.79     0.01 0.9155
NAP         1    176.37 530.03 124.98    18.45 0.0001 ***
humus       1      0.03 353.70 106.78 0.003565 0.9527
factor(week)3    177.51 531.17 121.08     6.19 0.0016 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

```
> summary(modellA)
[...]
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    9.298448   7.967002   1.167 0.2506
angle2         0.016760   0.042934   0.390 0.6984
NAP           -2.274093   0.529411  -4.296 0.0001 ***
humus          0.519686   8.703910   0.060 0.9527
factor(week)2 -7.065098   1.761492  -4.011 0.0002 ***
factor(week)3 -5.719055   1.827616  -3.129 0.0034 **
factor(week)4 -1.481816   2.720089  -0.545 0.5891
grainsize      0.002249   0.021066   0.107 0.9155
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

```
> summary(model1B)
[...]
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    9.298448   7.967002   1.167 0.2506
angle2         0.016760   0.042934   0.390 0.6984
grainsize      0.002249   0.021066   0.107 0.9155
NAP           -2.274093   0.529411  -4.296 0.0001  ***
humus          0.519686   8.703910   0.060 0.9527
factor(week)2 -7.065098   1.761492  -4.011 0.0002  ***
factor(week)3 -5.719055   1.827616  -3.129 0.0034  **
factor(week)4 -1.481816   2.720089  -0.545 0.5891
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Contents

Regression toward the mean

Univariate linear regression: how and why?

t-test for linear regression

Examples with transformed variables

- log-scaling brain sizes and body weights
- root of numbers of inhabitants and deaths

Multivariate Regression

- Example: species richness on sandy beaches
- Example: Success of different therapies
- Example: Daphnia

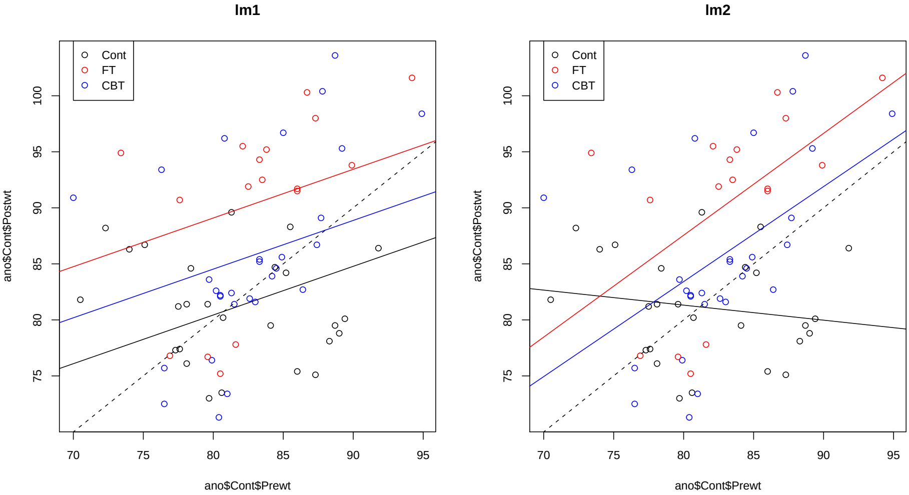
Cross validation and AIC

For young anorexia patients the effect of family therapy (FT) and cognitive behavioral therapy (CBT) is compared to a control group (Cont) by comparing the weight before (Prewt) and after (Postwt) the treatment (Treat).

Hand, D. J., Daly, F., McConway, K., Lunn, D. and Ostrowski, E. eds (1993) *A Handbook of Small Data Sets.* Chapman & Hall

Model lm1 There is a linear relation with the pre-weight. Each treatment changes the weight by a value that depends on the treatment but not on the treatment.

Model lm2 Interaction between Treatment und Preweight: The effect of the pre-weight depends on the kind of treatment.


lm1
lm2
Cont
FT
CBT
ano$Cont$Postwt
ano$Cont$Prewt
70
75
80
85
90
95
100

```
> lm1 <- lm(Postwt~Prewt+Treat,anorexia)
> lm2 <- lm(Postwt~Prewt*Treat,anorexia)
> anova(lm1,lm2)
Analysis of Variance Table

Model 1: Postwt ~ Prewt + Treat
Model 2: Postwt ~ Prewt * Treat
  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1     68 3311.3
2     66 2844.8  2     466.5 5.4112 0.006666 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

result: the more camplex model fits significantly better than the nested model.

result: the more camplex model fits significantly better than the nested model.

interpretation: The role of the weight before the treatment depends on the type of the treatment.

result: the more camplex model fits significantly better than the nested model.

interpretation: The role of the weight before the treatment depends on the type of the treatment.
or: The difference between effects of the treatments depends on the weight before the treetment.

Contents

Regression toward the mean

Univariate linear regression: how and why?

t-test for linear regression

Examples with transformed variables
- log-scaling brain sizes and body weights
- root of numbers of inhabitants and deaths

Multivariate Regression
- Example: species richness on sandy beaches
- Example: Success of different therapies
- Example: Daphnia

Cross validation and AIC

Question: Is there a difference between Daphnia magna and Daphnia galeata in their reaction on food supply?

Question: Is there a difference between Daphnia magna and Daphnia galeata in their reaction on food supply?

Data from Justina Wolinska's ecology course for Bachelor students.

```
> daph <- read.table("daphnia_justina.csv",h=T)
> daph
   counts foodlevel  species
1       68   high     magna
2       54   high     magna
3       59   high     magna
4       24   high   galeata
5       27   high   galeata
6       16   high   galeata
7       20    low     magna
8       18    low     magna
9       18    low     magna
10       5    low   galeata
11       8    low   galeata
12       9    low   galeata
```

```
> mod1 <- lm(counts~foodlevel+species,data=daph)
> mod2 <- lm(counts~foodlevel*species,data=daph)
> anova(mod1,mod2)
Analysis of Variance Table

Model 1: counts ~ foodlevel + species
Model 2: counts ~ foodlevel * species
  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1      9 710.00
2      8 176.67  1    533.33 24.151 0.001172 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

```
> summary(mod2)
[...]
Coefficients:
                      Estimate  Std.Error t.value Pr(>|t|)
(Intercept)               22.33  2.713   8.232 3.55e-05 ***
countslow                -15.00  3.837  -3.909  0.00449 **
foodlevelmagna            38.00  3.837   9.904 9.12e-06 ***
countslow:foodlevelmagna -26.67  5.426  -4.914  0.00117 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 4.699 on 8 degrees of freedom
Multiple R-squared: 0.9643, Adjusted R-squared: 0.9509
F-statistic: 71.95 on 3 and 8 DF,  p-value: 3.956e-06
```

result: the more complex model, in which different species react differently to low food level, fits significantly better.

result: the more complex model, in which different species react differently to low food level, fits significantly better.

But can we really assume normal distribution on numbers like 5, 8, 9...?

result: the more complex model, in which different species react differently to low food level, fits significantly better.

But can we really assume normal distribution on numbers like 5, 8, 9...?

We will come back to this in the Lecture about GLMs.

Contents

Regression toward the mean

Univariate linear regression: how and why?

t-test for linear regression

Examples with transformed variables

- log-scaling brain sizes and body weights
- root of numbers of inhabitants and deaths

Multivariate Regression

- Example: species richness on sandy beaches
- Example: Success of different therapies
- Example: Daphnia

Cross validation and AIC

How to predict the winglength of a Darwin finch by its beak size?

How to predict the winglength of a Darwin finch by its beak size? Shall we take beak height, beak length or both into account?

How to predict the winglength of a Darwin finch by its beak size? Shall we take beak height, beak length or both into account? Residual variance should be small....

How to predict the winglength of a Darwin finch by its beak size? Shall we take beak height, beak length or both into account? Residual variance should be small....

Leave-one-out cross validation: If you leave out one bird and fit the model to the others, how well can this model predict the wing span?

How to predict the winglength of a Darwin finch by its beak size? Shall we take beak height, beak length or both into account? Residual variance should be small....

Leave-one-out cross validation: If you leave out one bird and fit the model to the others, how well can this model predict the wing span?

```
prederrorHL <- numeric()
for (i in 1:46) {
  selection <- rep(TRUE,46)
  selection[i] <- FALSE
  modHL.R <- lm(WingL~N.UBkL+BeakH,data=finchdata,
                                   subset=selection)
  prederrorHL[i]=WingL[i]-predict(modHL.R,finchdata[i,])
}
```

	Height	Length	Height and Length
σ(Residuals)	3.83	4.78	3.79

	Height	Length	Height and Length
σ(Residuals)	3.83	4.78	3.79
$d =$ (Number Parameters)	2	2	3

	Height	Length	Height and Length
σ(Residuals)	3.83	4.78	3.79
$d =$ (Number Parameters)	2	2	3
σ(Residuals)$\cdot\sqrt{\frac{n-1}{n-d}}$	3.86	4.84	3.87

	Height	Length	Height and Length
σ(Residuals)	3.83	4.78	3.79
$d =$ (Number Parameters)	2	2	3
σ(Residuals)$\cdot\sqrt{\frac{n-1}{n-d}}$	3.86	4.84	3.87
cross validation.	3.96	4.97	3.977

	Height	Length	Height and Length
σ(Residuals)	3.83	4.78	3.79
$d =$ (Number Parameters)	2	2	3
σ(Residuals)$\cdot\sqrt{\frac{n-1}{n-d}}$	3.86	4.84	3.87
cross validation.	3.96	4.97	3.977
AIC	259.0	279.5	260.1

	Height	Length	Height and Length
σ(Residuals)	3.83	4.78	3.79
$d =$ (Number Parameters)	2	2	3
σ(Residuals)$\cdot\sqrt{\frac{n-1}{n-d}}$	3.86	4.84	3.87
cross validation.	3.96	4.97	3.977
AIC	259.0	279.5	260.1
BIC	264.4	285.0	267.4

	Height	Length	Height and Length
σ(Residuals)	3.83	4.78	3.79
$d =$ (Number Parameters)	2	2	3
σ(Residuals)$\cdot\sqrt{\frac{n-1}{n-d}}$	3.86	4.84	3.87
cross validation.	3.96	4.97	3.977
AIC	259.0	279.5	260.1
BIC	264.4	285.0	267.4

Akaike's Information Criterion:

$$\mathrm{AIC} = -2 \cdot \log L + 2 \cdot (\mathrm{NumberofParameters})$$

	Height	Length	Height and Length
σ(Residuals)	3.83	4.78	3.79
$d =$ (Number Parameters)	2	2	3
$\sigma(\text{Residuals})\cdot\sqrt{\frac{n-1}{n-d}}$	3.86	4.84	3.87
cross validation.	3.96	4.97	3.977
AIC	259.0	279.5	260.1
BIC	264.4	285.0	267.4

Akaike's Information Criterion:

$$\mathrm{AIC} = -2 \cdot \log L + 2 \cdot (\mathrm{NumberofParameters})$$

Bayesian Information Criterion:

$$\mathrm{BIC} = -2 \cdot \log L + \log(n) \cdot (\mathrm{NumberofParameters})$$

	Height	Length	Height and Length
σ(Residuals)	3.83	4.78	3.79
$d =$ (Number Parameters)	2	2	3
σ(Residuals)$\cdot\sqrt{\frac{n-1}{n-d}}$	3.86	4.84	3.87
cross validation.	3.96	4.97	3.977
AIC	259.0	279.5	260.1
BIC	264.4	285.0	267.4

Akaike's Information Criterion:

$$\mathrm{AIC} = -2 \cdot \log L + 2 \cdot (\mathrm{NumberofParameters})$$

Bayesian Information Criterion:

$$\mathrm{BIC} = -2 \cdot \log L + \log(n) \cdot (\mathrm{NumberofParameters})$$