Multivariate Statistics in Ecology and Quantitative Genetics

Linear Regression and Linear Models

Dirk Metzler & Noémie Becker

http://evol.bio.lmu.de/_statgen

23. Juni 2014

Contents

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

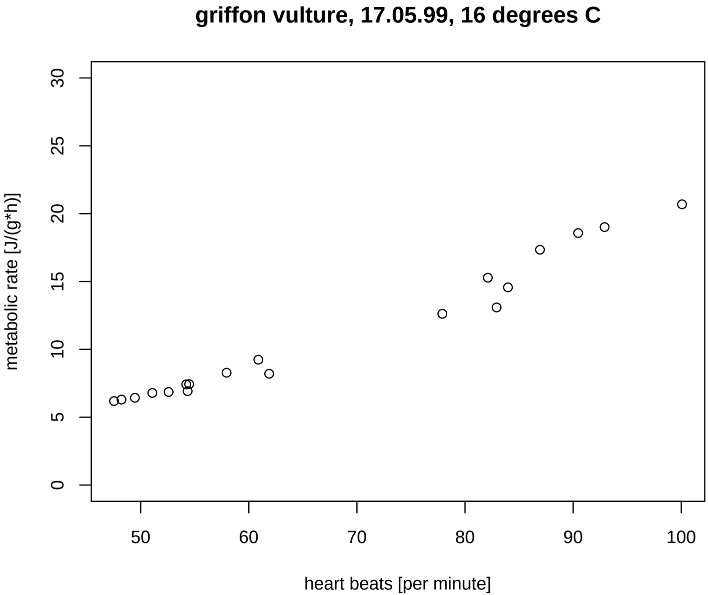
- Data from Goethe-University, Group of Prof. Prinzinger
- Developed telemetric system for measuring heart beats of flying birds
- Important for ecological questions: metabolic rate.

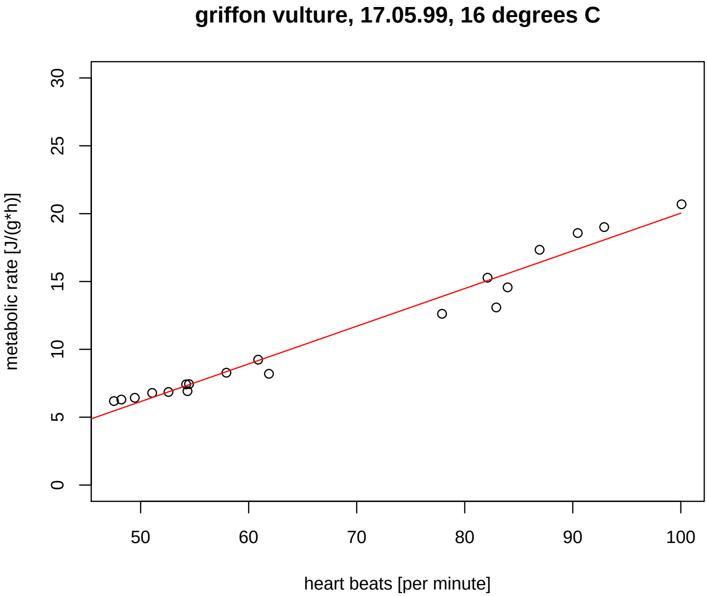
📄 Prinzinger, R., E. Karl, R. Bögel, Ch. Walzer (1999): Energy metabolism, body temperature, and cardiac work in the Griffon vulture Gyps vulvus - telemetric investigations in the laboratory and in the field.
Zoology **102**, Suppl. II: 15

- 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?

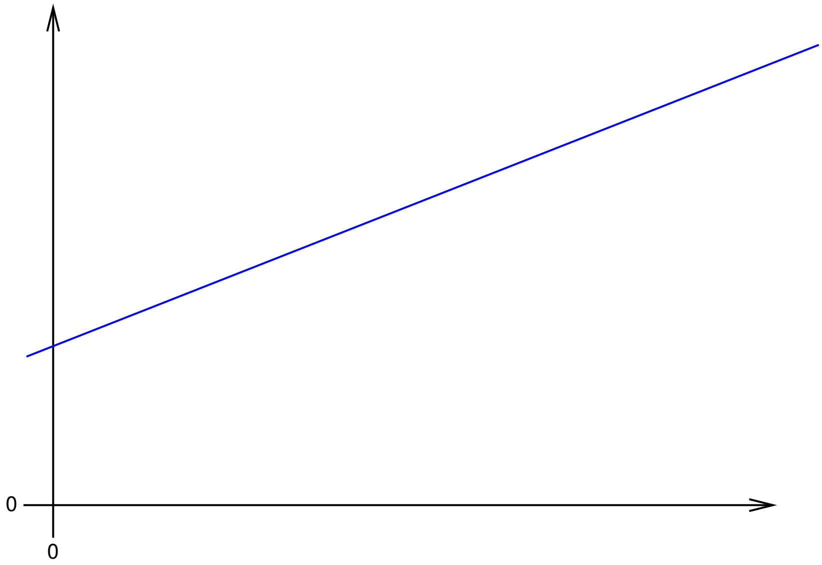

griffon vulture, 17.05.99, 16 degrees C
metabolic rate [J/(g*h)]
0
5
10
15
20
25
30
50
60
70
80
90
100
heart beats [per minute]

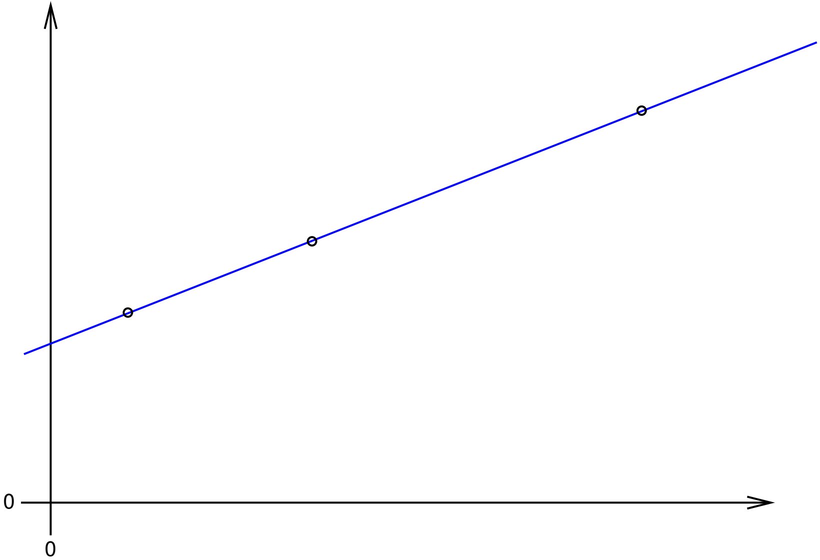

griffon vulture, 17.05.99, 16 degrees C
metabolic rate [J/(g*h)]
0
5
10
15
20
25
30
heart beats [per minute]
50
60
70
80
90
100

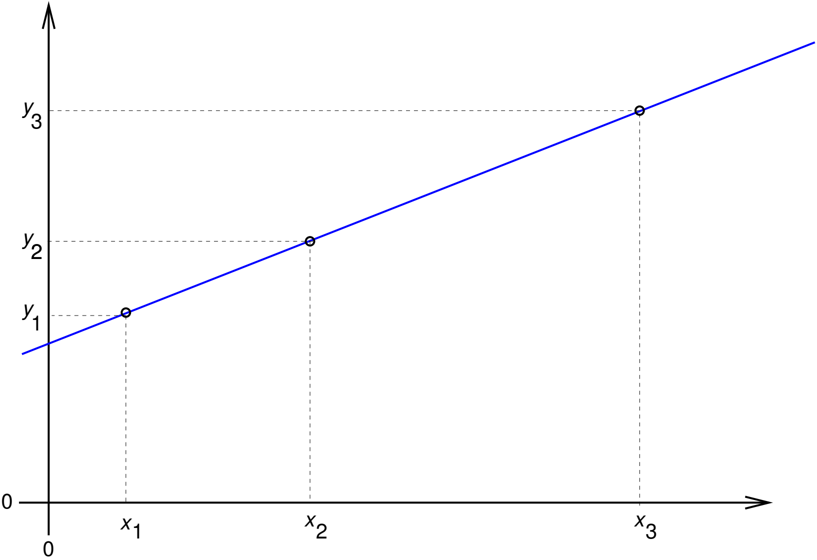
```
vulture
              day heartbpm metabol minTemp maxTemp medtemp
1    01.04./02.04.    70.28   11.51      -6       2    -2.0
2    01.04./02.04.    66.13   11.07      -6       2    -2.0
3    01.04./02.04.    58.32   10.56      -6       2    -2.0
4    01.04./02.04.    58.63   10.62      -6      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

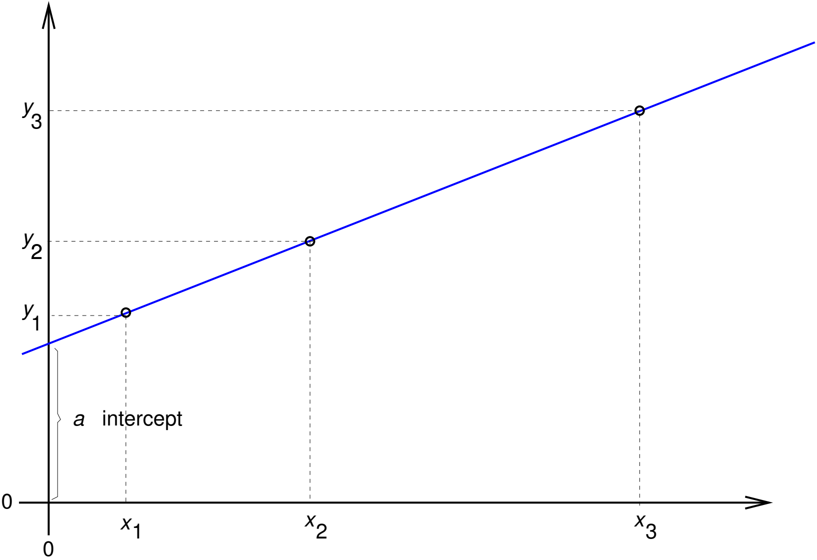
```
> model <- lm(metabol~heartbpm,data=vulture,
              subset=day=="17.05.")
> summary(model)
Call:
lm(formula = metabol ~ heartbpm, data = vulture, subset = day
    "17.05.")
Residuals:
    Min      1Q  Median      3Q     Max
-2.2026 -0.2555  0.1005  0.6393  1.1834
Coefficients:
           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

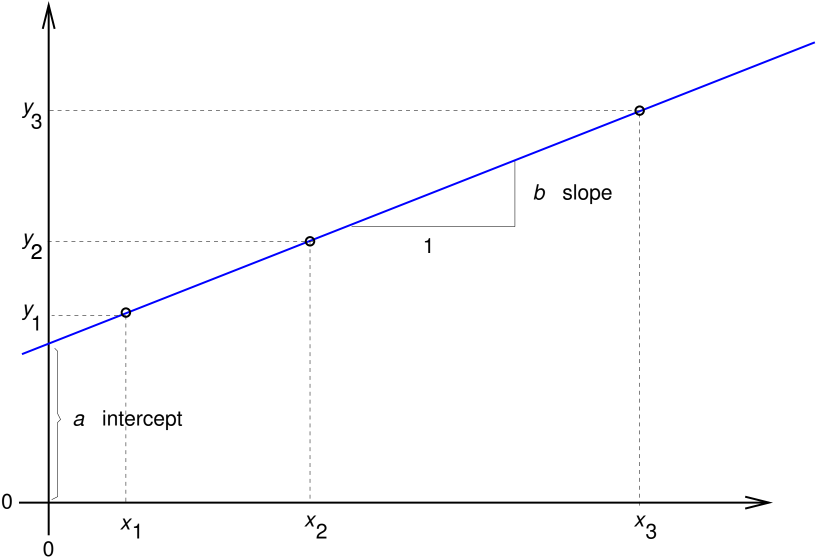

0
0

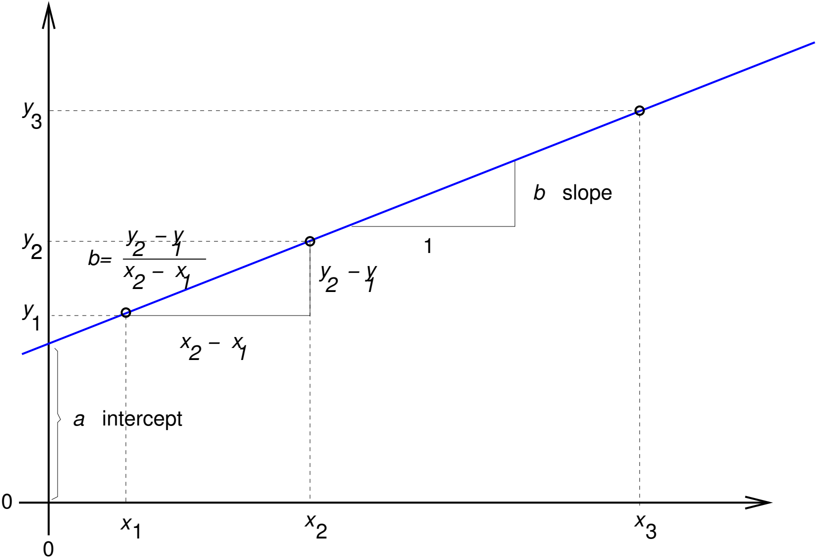

0
0

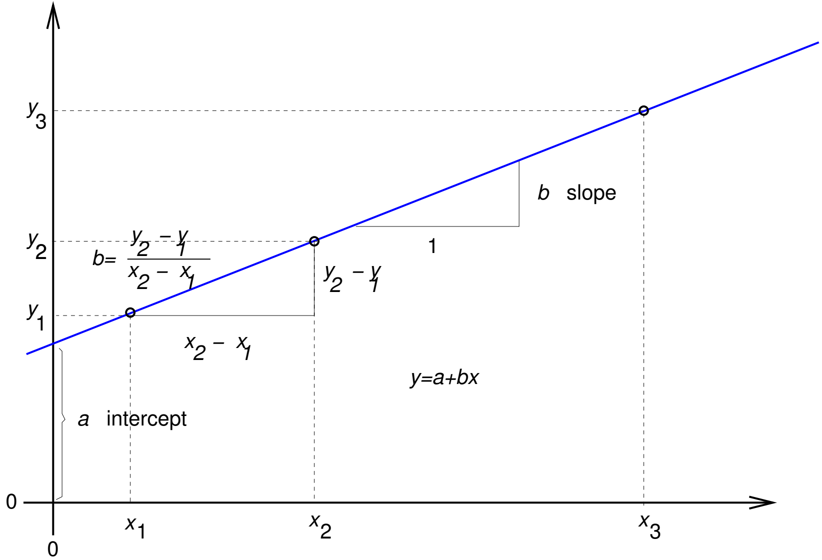

y_3
y_2
y_1
0
0
x_1
x_2
x_3

Univariate linear regression

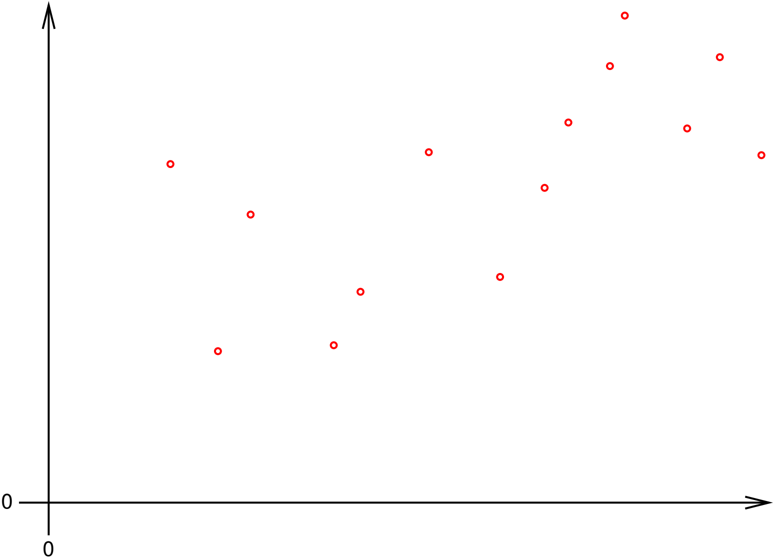



y_3
y_2
y_1
b slope
1
a intercept
0
x_1
x_2
x_3
0

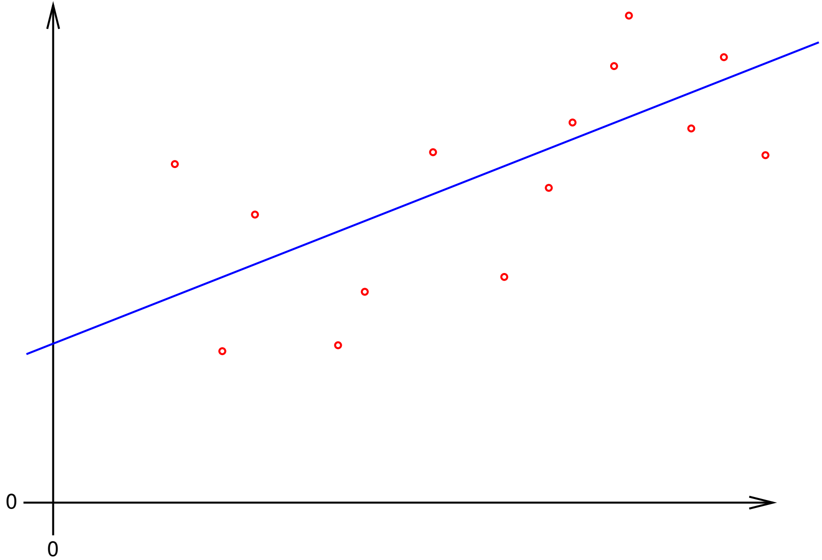

y_3
y_2
y_1
$b = \frac{y_2 - y_1}{x_2 - x_1}$
b slope
1
$y_2 - y_1$
$x_2 - x_1$
a intercept
0
0
x_1
x_2
x_3

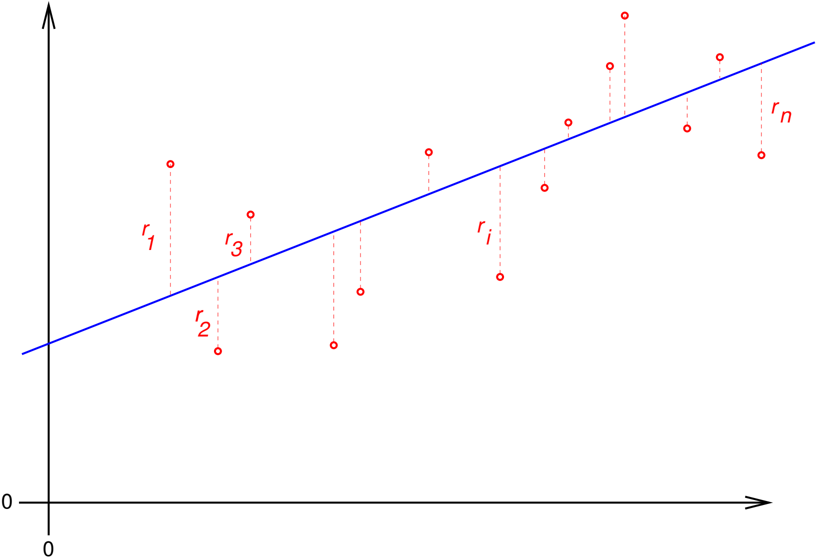

y_3
y_2
y_1
0
0
x_1
x_2
x_3
$b=\frac{y_2-y_1}{x_2-x_1}$
y_2-y_1
x_2-x_1
b slope
1
$y=a+bx$
a intercept

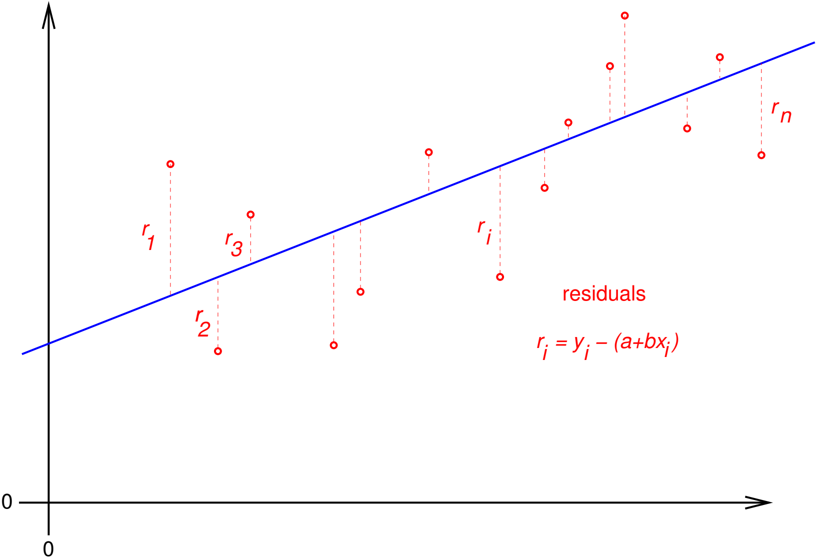
Univariate linear regression



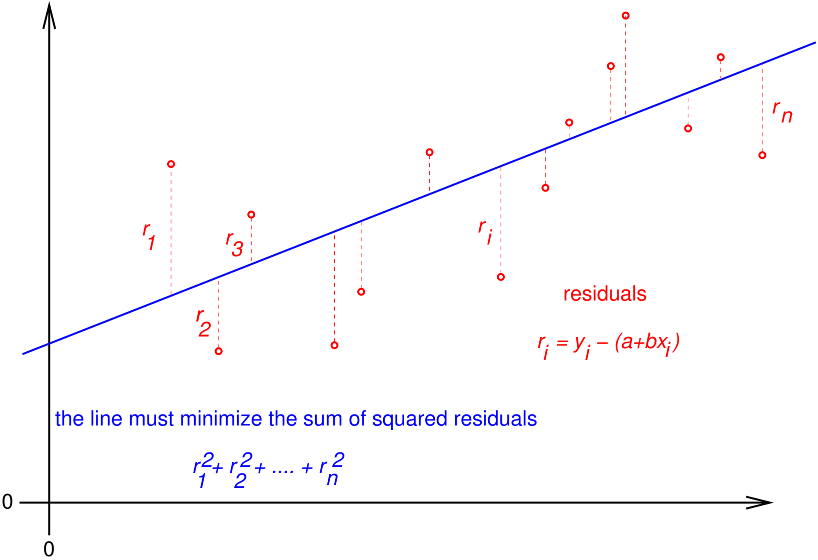
Univariate linear regression




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



residuals

$r_i = y_i - (a+bx_i)$

the line must minimize the sum of squared residuals

$r_1^2 + r_2^2 + ... + r_n^2$

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

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

$\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ 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

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

$\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ 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

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

$\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$ 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 $\hat{a}$ *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

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

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 $\hat{a}$ 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 $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$.

```
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
```

Contents

t-test for $\hat{b}$

Estimate σ^2 by

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

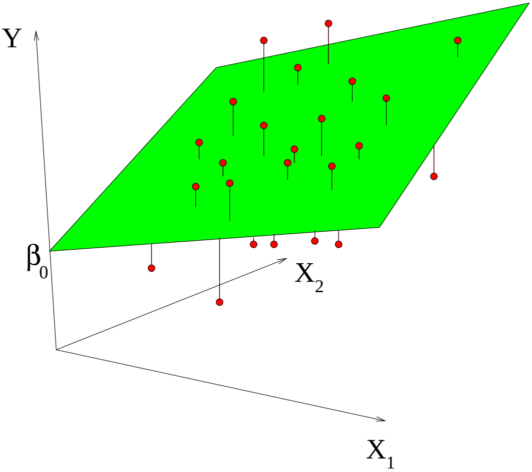
Then,

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

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

Contents

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:

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

Multivariate Regression

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

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

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:

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

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:

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

Model:

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

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

Contents

- 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
```

```
> modell0 <- lm(richness ~ angle2+NAP+grainsize+humus,
+                  data = rikz)
> summary(modell0)
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:

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

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

e.g. b_6 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 `week`$= k$ and 0 otherwise.

e.g. b_6 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”.

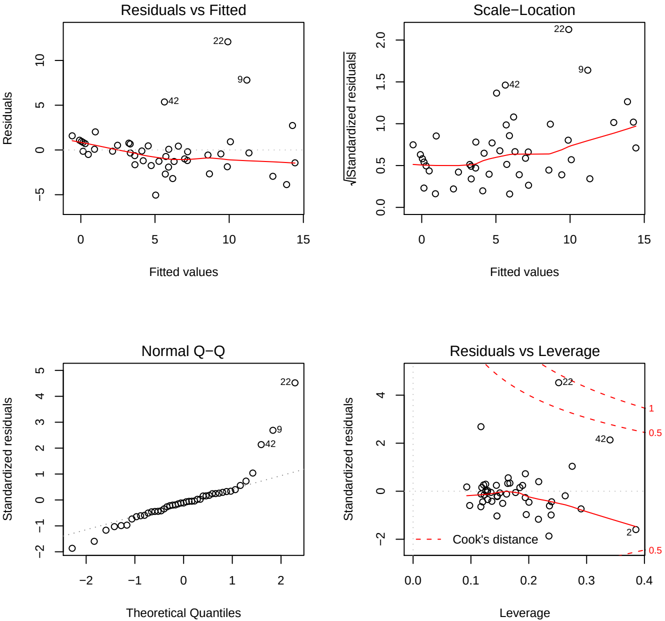
```
> modell0 <- lm(richness ~ angle2+NAP+grainsize+humus,
+                  data = rikz)
> modell <- lm(richness ~ angle2+NAP+grainsize+humus
+                          +factor(week), data = rikz)
> anova(modell0, modell)
Analysis of Variance Table

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

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 whether 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(model1B)
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(model1B,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(modell1B)
[...]
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
```

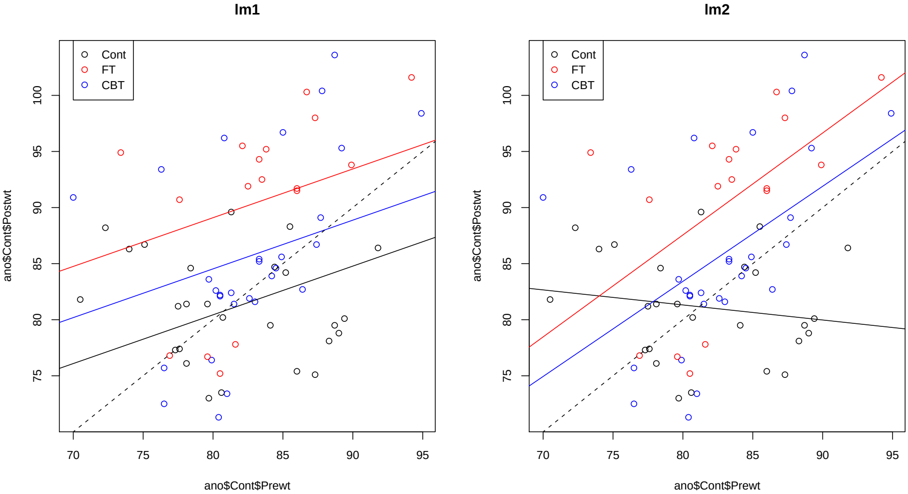
Contents

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 Preweight.

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
75
80
85
90
95
100
70
75
80
85
90
95

```
> 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 complex model fits significantly better than the nested model.

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

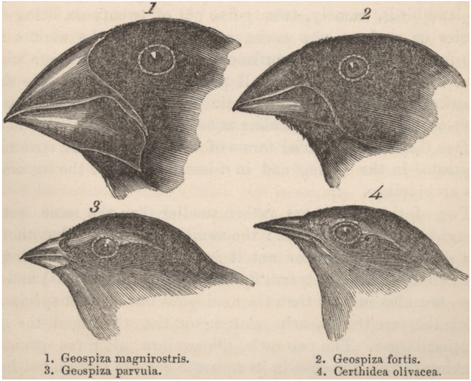
interpretation: The role of the weight before the treatment depends on the type of the treatment.

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

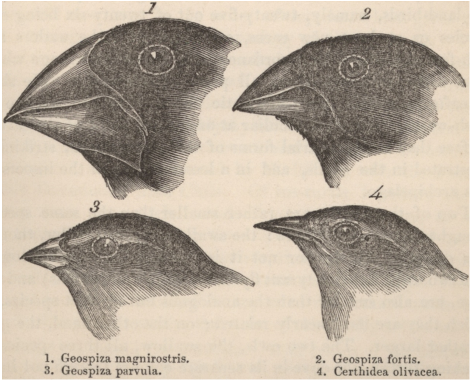
interpretation: The role of the weight before the treatment depends on the type of the treatment.
or: The difference between effects of the treatments depends on the weight before the treetment.

Contents

Contents

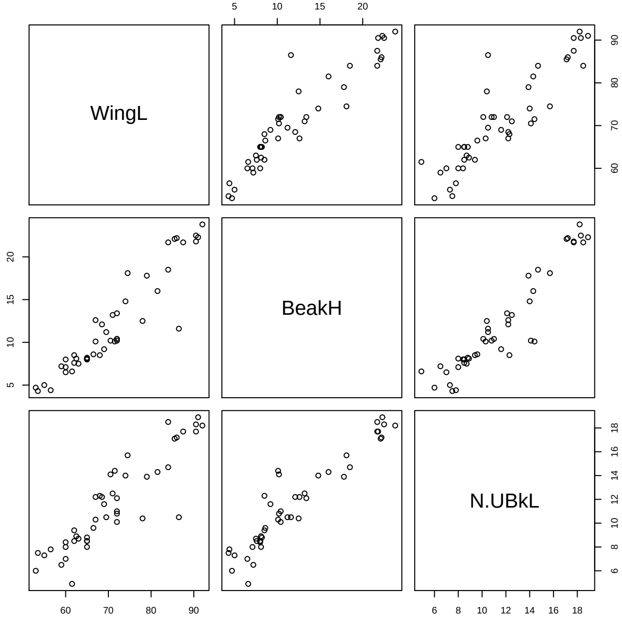


You find a beak of a Darwin finch. The beak is 14 mm long and 10 mm high. How accurately can you predict the winglength of the bird?



You find a beak of a Darwin finch. The beak is 14 mm long and 10 mm high. How accurately can you predict the winglength of the bird?

Your “training data” are the winglengths (WingL), beak heights (BeakH) and beak lengths (N.UBkL) of 46 Darwin finches.


WingL
BeakH
N.UBkL
5
10
15
20
90
80
70
60
20
15
10
5
18
16
14
12
10
8
6
60
70
80
90
6
8
10
12
14
16
18

```
> modH <- lm(WingL~BeakH)
> summary(modH)

Call:
lm(formula = WingL ~ BeakH)

Residuals:
    Min      1Q  Median      3Q     Max
-7.1882 -2.5327 -0.2796  1.8325 16.2702

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 49.78083    1.33103   37.40   <2e-16 ***
BeakH        1.76284    0.09961   17.70   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 3.868 on 44 degrees of freedom
Multiple R-squared: 0.8768, Adjusted R-squared: 0.874
F-statistic: 313.2 on 1 and 44 DF,  p-value: < 2.2e-16

> predict(modH,newdata=data.frame(BeakH=10))
       1
67.40924
```

```
> modL <- lm(WingL~N.UBkL)
> summary(modL)

Call:
lm(formula = WingL ~ N.UBkL)

Residuals:
    Min      1Q  Median      3Q     Max 
-7.1321 -3.3974  0.4737  2.2966 18.2299 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  41.5371     2.2884   18.15   <2e-16 ***
N.UBkL        2.5460     0.1875   13.58   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 4.838 on 44 degrees of freedom
Multiple R-squared: 0.8074, Adjusted R-squared: 0.803 
F-statistic: 184.4 on 1 and 44 DF,  p-value: < 2.2e-16

> predict(modL,newdata=data.frame(N.UBkL=14))
       1 
77.18117 
```

```
> modHL <- lm(WingL~BeakH+N.UBkL)
> summary(modHL)

Call:
lm(formula = WingL ~ BeakH + N.UBkL)

Residuals:
    Min      1Q  Median      3Q     Max
-7.3185 -2.5022 -0.2752  1.5352 16.5893

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  48.1740     2.2572  21.343  < 2e-16 ***
BeakH         1.5133     0.2999   5.047 8.69e-06 ***
N.UBkL        0.3984     0.4513   0.883    0.382
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 3.878 on 43 degrees of freedom
Multiple R-squared: 0.879, Adjusted R-squared: 0.8734
F-statistic: 156.2 on 2 and 43 DF,  p-value: < 2.2e-16

> predict(modHL,newdata=data.frame(BeakH=10,N.UBkL=14))
       1
68.88373
```

Which of the three predictions 67.4mm, 77.2mm und 68.9mm for the winglength is most reliable?

Which of the three predictions 67.4mm, 77.2mm und 68.9mm for the winglength is most reliable?

In the model modHL (with beak length and height) the influence of beak length is not significant.

Which of the three predictions 67.4mm, 77.2mm und 68.9mm for the winglength is most reliable?

In the model modHL (with beak length and height) the influence of beak length is not significant.

We can not draw conclusion from non-significance. Beak length could still improve the prediction.

Which of the three predictions 67.4mm, 77.2mm und 68.9mm for the winglength is most reliable?

In the model modHL (with beak length and height) the influence of beak length is not significant.

We can not draw conclusion from non-significance. Beak length could still improve the prediction.

Is it always good to use as much data as possible?

Which of the three predictions 67.4mm, 77.2mm und 68.9mm for the winglength is most reliable?

In the model modHL (with beak length and height) the influence of beak length is not significant.

We can not draw conclusion from non-significance. Beak length could still improve the prediction.

Is it always good to use as much data as possible?

This could lead to “overfitting”: If too many parameters are available, the model will learn all the little details of the data including random fluctions. It will learn just memorize the trainig data. This may corrupt the model’s predictions for new data.

We could judge the models by the standard deviation of the ε_i, which we estimate by the standard deviation of the residuals r_i.

We could judge the models by the standard deviation of the ε_i, which we estimate by the standard deviation of the residuals r_i.

We must account for the different number d of model parameters, because we lose one degree of freedom for each estimated parameter:

$$\widehat{\sigma}_\epsilon = \sqrt{\frac{1}{n-d}\sum_i r_i^2} = \sigma_r \cdot \sqrt{\frac{n-1}{n-d}}.$$

We could judge the models by the standard deviation of the ε_i, which we estimate by the standard deviation of the residuals r_i.

We must account for the different number d of model parameters, because we lose one degree of freedom for each estimated parameter:

$$\widehat{\sigma}_\epsilon = \sqrt{\frac{1}{n-d}\sum_i r_i^2} = \sigma_r \cdot \sqrt{\frac{n-1}{n-d}}.$$

These values are reported in R by the command “summary”:
modH:

```
Residual standard error: 3.868 on 44 degrees of freedom
```

modL:

```
Residual standard error: 4.838 on 44 degrees of freedom
```

modHL:

```
Residual standard error: 3.878 on 43 degrees of freedom
```

Another possibility to judge the prediction error of a model is *cross validation* (aka *Jackknife*).

Another possibility to judge the prediction error of a model is *cross validation* (aka *Jackknife*).

The idea is: Remove one of the 46 birds from the dataset and fit the model to the other 45. How well can the model predict the winglength of the omitted bird?

Another possibility to judge the prediction error of a model is *cross validation* (aka *Jackknife*).

The idea is: Remove one of the 46 birds from the dataset and fit the model to the other 45. How well can the model predict the winglength of the omitted bird?

Repeat this for all 46 birds.

Another possibility to judge the prediction error of a model is *cross validation* (aka *Jackknife*).

The idea is: Remove one of the 46 birds from the dataset and fit the model to the other 45. How well can the model predict the winglength of the omitted bird?

Repeat this for all 46 birds.

We have to decide how we measure the error. How to judge a model with many medium errors compared to a model with rare large errors? We use (the square root of) the sum of squared errors.

```
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
$\sigma(\text{Residuals}) \cdot \sqrt{\frac{n-1}{n-d}}$	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
$\sigma(\text{Residuals})\cdot\sqrt{\frac{n-1}{n-d}}$	3.86	4.84	3.87
cross validation.	3.96	4.97	3.977
AIC	259.0	279.5	260.1

	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

	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:

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

	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
$\sigma(\text{Residuals})\cdot\sqrt{\frac{n-1}{n-d}}$	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:

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

For $n \geq 8$ holds $\log(n) > 2$ and BIC penalizes every additional parameter harder than AIC. (As always, log is the natural logarithm.)

Low values of AIC and BIC favor the model.

Low values of AIC and BIC favor the model. (At least in R. There may be programs that show AIC and BIC with inverse sign)

Low values of AIC and BIC favor the model. (At least in R. There may be programs that show AIC and BIC with inverse sign)

AIC is based on the idea to approximate the prediction error (which is exact under certain conditions).

Low values of AIC and BIC favor the model. (At least in R. There may be programs that show AIC and BIC with inverse sign)

AIC is based on the idea to approximate the prediction error (which is exact under certain conditions).

BIC approximates (up to a constant) the log of the posterior probability of the model, where all models are a priori assumed to be equally probable.

	height	length	height and length
σ(Residuals)	3.83	4.78	3.79
$d =$ (Number of parameters)	2	2	3
$\sigma(\text{Residuals}) \cdot \sqrt{\frac{n-1}{n-d}}$	3.87	4.84	3.88
cross validation.	26.56	33.34	26.68
AIC	259.0	279.5	260.1
BIC	264.4	285.0	267.4

It seems best to use only the beak height.

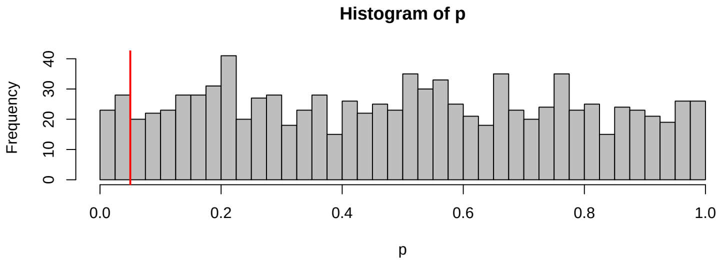
Problem with extensive model selection

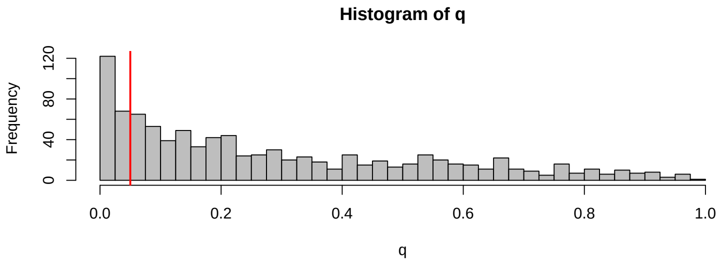
If you have optimized the model e.g. by AIC and than compute p-values with the same data, you find too much significance.

Problem with extensive model selection

If you have optimized the model e.g. by AIC and than compute *p*-values with the same data, you find too much significance.We explore this with a little simulation:

```
A <- as.factor(rep(c("a","b","c"),c(40,40,40)))
B <- as.factor(rep(rep(c("w","x","y","z"),c(10,10,10,10)),3))
C <- as.factor(rep(c("p","q","r"),40))
D <- as.factor(rep(rep(c("m","n"),c(5,5)),12))
X <- rnorm(120,10,2)

library(MASS)

p <-  numeric()
q <-  numeric()
for(i in 1:1000) {
  X <- rnorm(120,10,2)
  p[i] <- anova(lm(X~1),lm(X~A*B*C*D))$"Pr(>F)"[[2]]
  q[i] <- anova(lm(X~1),stepAIC(lm(X~A*B*C*D)))$"Pr(>F)"[[2]]
}
```


Histogram of p
Frequency
0
10
20
30
40
0.0
0.2
0.4
0.6
0.8
1.0
p


Histogram of q
Frequency
0
40
80
120
0.0
0.2
0.4
0.6
0.8
1.0
q

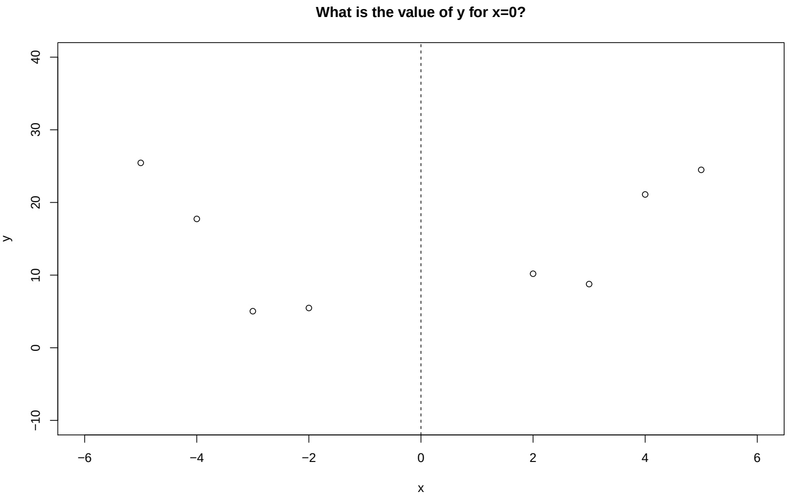
Save model selection and checking if you have lots of data

1. Divide the data randomly into 3 subsets A, B, C, where A may contain half of the data, and B and C a quarter each.
2. Fit each candidate model to the data subset A.
3. Assess the accuracy of these fitted models with data subset B. Let M be the best model in this contest.
4. Assess the accuracy of M again and also its *p*-values, this time with dataset C.

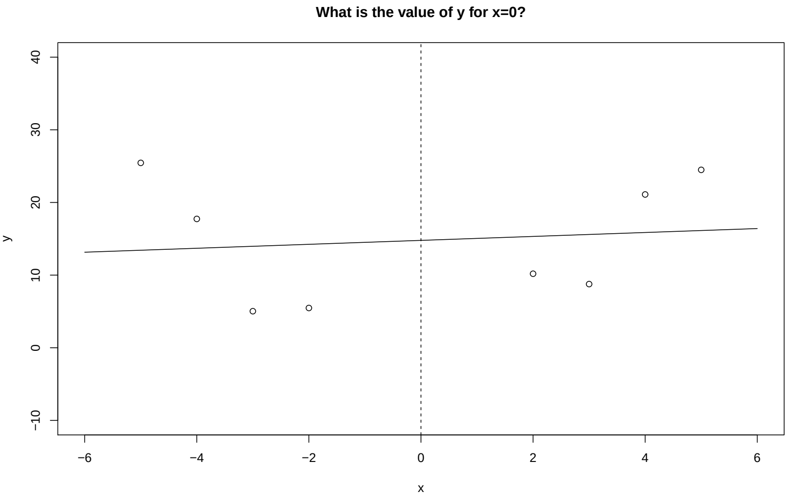
Graphical methods are also very important in model fitting, especially applied to residuals. Plot resudials against variables. If this uncovers dependencies, they should be added to the model.

Contents

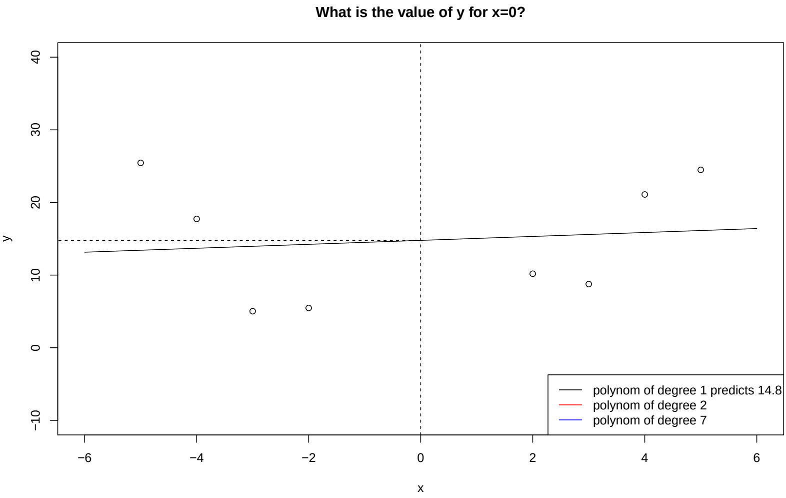
Overfitting



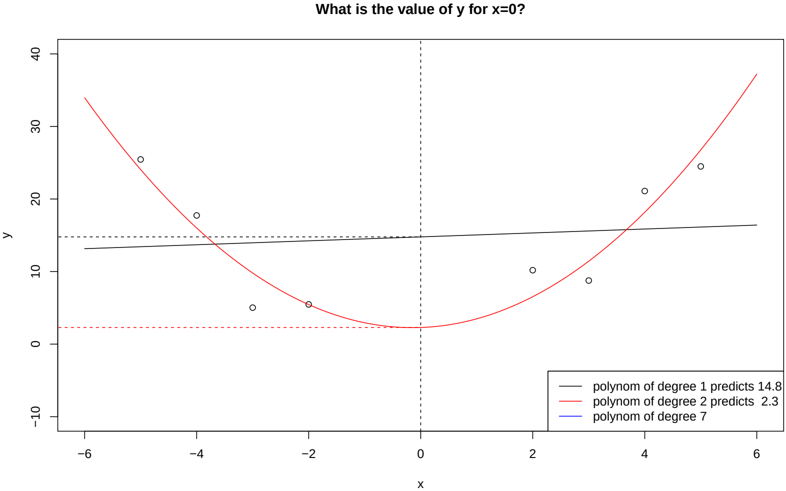
Overfitting



Overfitting

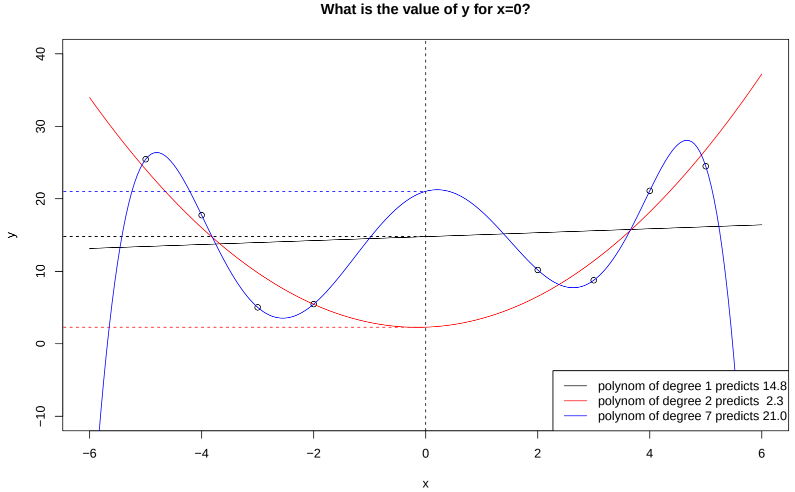


Overfitting



`lm(y~poly(x,2))` *is the same as* `lm(y~x+I(x^2))`

Overfitting



`lm(y~poly(x,7))`

Contents

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
```

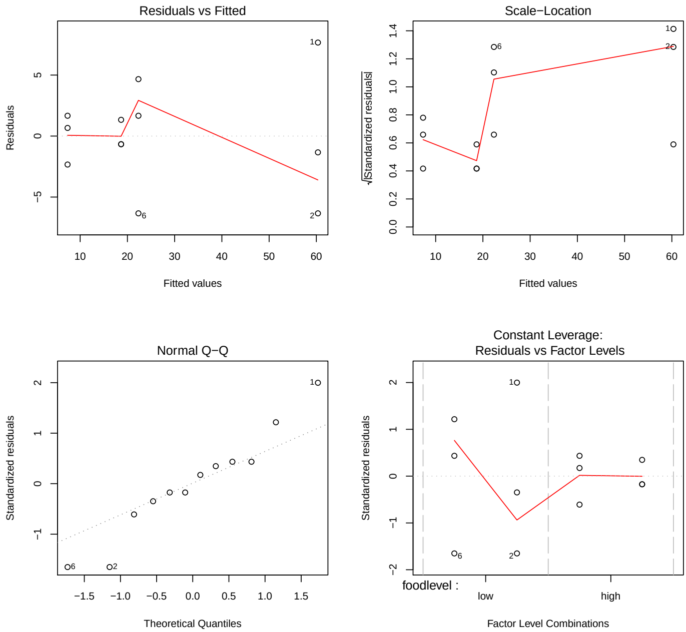
```
> mod1 <- lm(counts~foodlevel+species,data=daph)
> mod2 <- lm(counts~foodlevel*species,data=daph)
> anova(mod1,mod2)
Analysis of Variance Table

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.

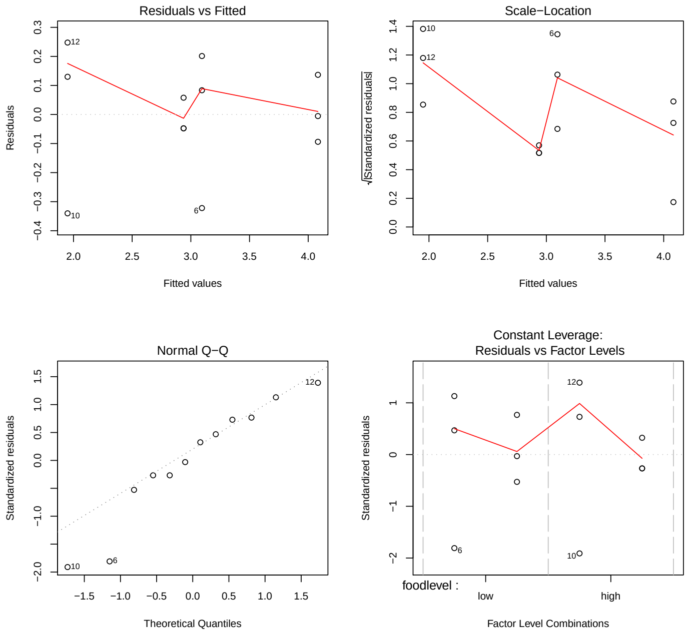
Result: the more complex model, in which different species react differently to low food level, fits significantly better.

But does it fit well enough...?


Residuals vs Fitted
Residuals
5
0
−5
10
20
30
40
50
60
Fitted values
1
6
2
Scale−Location
√|Standardized residuals|
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
10
20
30
40
50
60
Fitted values
1
2
6
Normal Q−Q
Standardized residuals
2
1
0
−1
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
Theoretical Quantiles
1
6
2
Constant Leverage:
Residuals vs Factor Levels
Standardized residuals
2
1
0
−1
−2
foodlevel :
low
high
Factor Level Combinations
1
6
2

```
> mod3 <- lm(log(counts)~foodlevel+species,data=daph)
> mod4 <- lm(log(counts)~foodlevel*species,data=daph)
> anova(mod3,mod4)
Analysis of Variance Table

Model 1: log(counts) ~ foodlevel + species
Model 2: log(counts) ~ foodlevel * species
  Res.Df     RSS Df Sum of Sq      F Pr(>F)
1      9 0.38041
2      8 0.37856  1 0.0018545 0.0392  0.848
```

```
> summary(mod3)

Call:
lm(formula = log(counts) ~ foodlevel + species, data = daph)

Residuals:
     Min       1Q   Median       3Q      Max
-0.34017 -0.05915  0.02622  0.13153  0.24762

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.0946     0.1028  30.104 2.41e-10 ***
foodlevellow  -1.1450     0.1187  -9.646 4.83e-06 ***
speciesmagna   0.9883     0.1187   8.326 1.61e-05 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.2056 on 9 degrees of freedom
Multiple R-squared: 0.9475, Adjusted R-squared: 0.9358
F-statistic: 81.19 on 2 and 9 DF,  p-value: 1.743e-06
```


Residuals vs Fitted
Residuals
Fitted values
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.4
2.0
2.5
3.0
3.5
4.0
12
10
6
Scale−Location
√|Standardized residuals|
Fitted values
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
2.0
2.5
3.0
3.5
4.0
10
12
6
Normal Q−Q
Standardized residuals
Theoretical Quantiles
1.5
1.0
0.5
0.0
−1.0
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
12
10
6
Constant Leverage:
Residuals vs Factor Levels
Standardized residuals
Factor Level Combinations
1
0
−1
−2
foodlevel :
low
high
12
6
10

The qqplot looks better now but not really good.

The qqplot looks better now but not really good.

The reason is perhaps that the values of the target variable `counts` were small integers such that the normal distribution assumption is dubious.

The qqplot looks better now but not really good.

The reason is perhaps that the values of the target variable `counts` were small integers such that the normal distribution assumption is dubious.

Instead of the normal linear model we can fit a log transformed generalized linear model of type Poisson. We will see this in a few days.

The qqplot looks better now but not really good.

The reason is perhaps that the values of the target variable `counts` were small integers such that the normal distribution assumption is dubious.

Instead of the normal linear model we can fit a log transformed generalized linear model of type Poisson. We will see this in a few days.

For now we only compare the models with normality assumptions.

```
> AIC(mod1,mod2,mod3,mod4)
     df        AIC
mod1  4 91.0188246
mod2  5 76.3268216
mod3  4  0.6376449
mod4  5  2.5790019
```

The log-linear models clearly have better AIC values than the linear models with untransformed data. But one should not compare AIC values between models with different (or differently scaled) target variable.

```
> AIC(mod1,mod2,mod3,mod4)
     df        AIC
mod1  4 91.0188246
mod2  5 76.3268216
mod3  4  0.6376449
mod4  5  2.5790019
```

The log-linear models clearly have better AIC values than the linear models with untransformed data. But one should not compare AIC values between models with different (or differently scaled) target variable.

The interaction in model mod4 is not only non-significant, the model mod3 without interaction also has the better AIC values.

So we favor mod3:

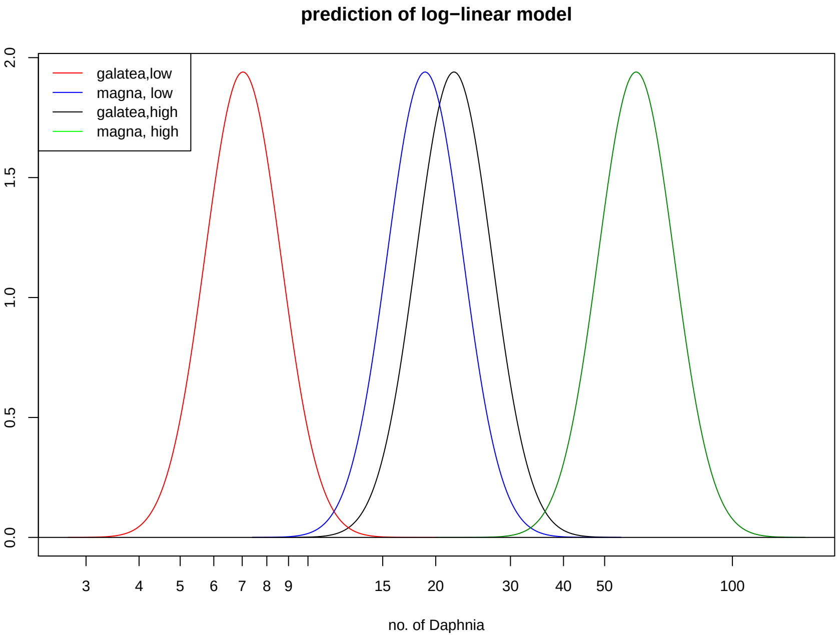
$$\log(\texttt{counts}) = 3.09 - 1.14 \cdot I_{\text{low food}} + 0.99 \cdot I_{\text{magna}} + \varepsilon$$

So we favor mod3:

$$\log(\texttt{counts}) = 3.09 - 1.14 \cdot I_{\text{low food}} + 0.99 \cdot I_{\text{magna}} + \varepsilon$$

By applying the *e* function we obtain:

$$\texttt{counts} = 21.98 \cdot 0.32^{I_{\text{low food}}} \cdot 2.69^{I_{\text{magna}}} \cdot e^{\varepsilon}$$


prediction of log−linear model
galatea,low
magna, low
galatea,high
magna, high
0.0
0.5
1.0
1.5
2.0
3
4
5
6
7
8
9
15
20
30
40
50
100
no. of Daphnia

But is it reasonable at all to assume normal distribution when the data are counts 0,1,2,. . . ?

But is it reasonable at all to assume normal distribution when the data are counts 0,1,2,. . . ?

We will come back to this dataset when we discuss GLMs.