Statistics for EES

Linear regression and linear models

Dirk Metzler

April 21, 2026

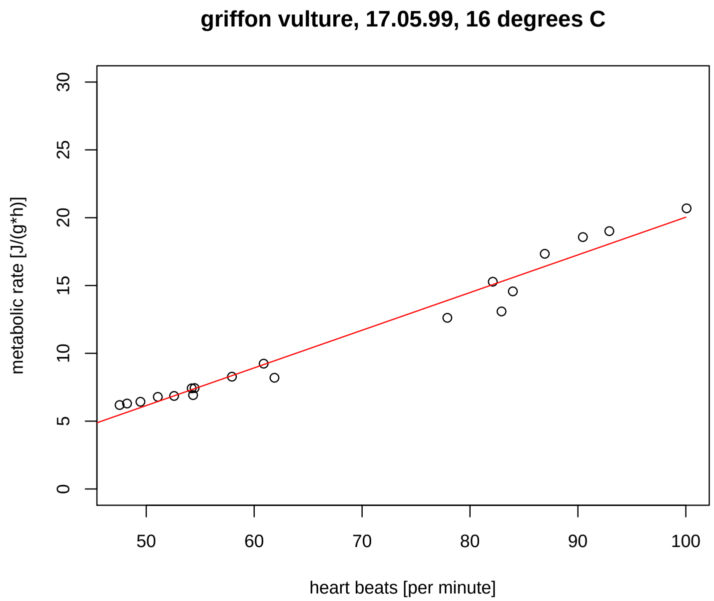
Contents

1 Linear regression

References

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

- 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?

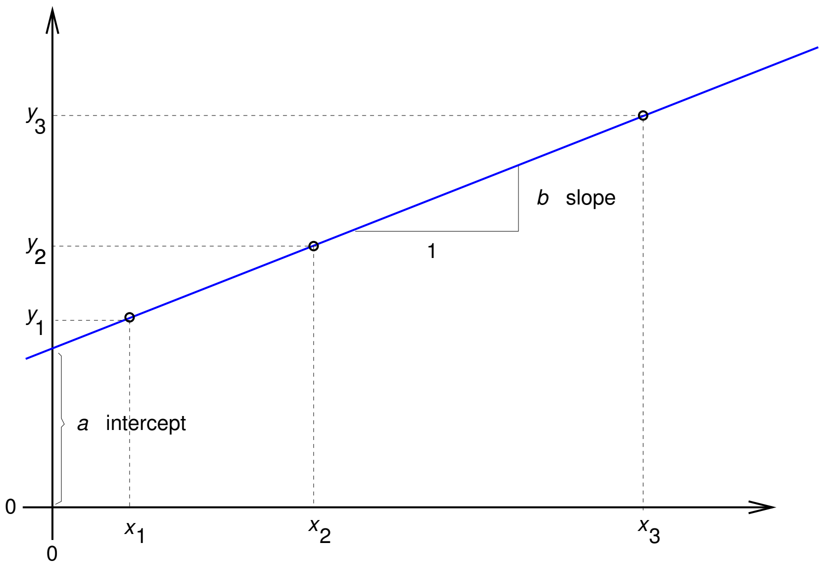


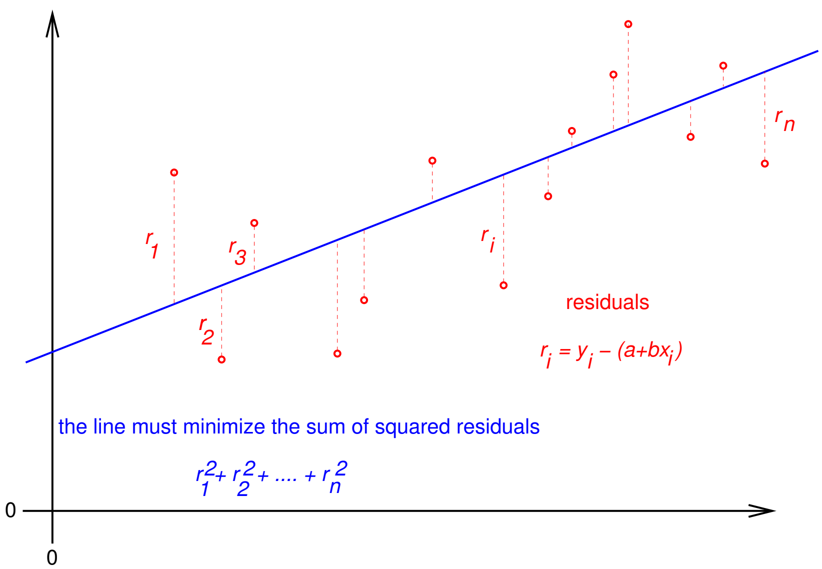

```
vulture
              day heartbpm metabol minTemp maxTemp medtemp
1   01.04./02.04.    70.28   11.51      -6       2    -2.0
2   01.04./02.04.    66.13   11.07      -6       2    -2.0
3   01.04./02.04.    58.32   10.56      -6       2    -2.0
4   01.04./02.04.    58.63   10.62      -6       2  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





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.

givend 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 $\sim \mathcal{N}(0, \sigma^2)$.[1.5ex] $\Rightarrow y_1, y_2, \ldots, y_n$ are independent $y_i \sim \mathcal{N}(a+b\cdot x_i, \sigma^2)$.[1.5ex] a, b, σ^2 are unknown, but **not random**.

We estimate a and b by computing

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

Theorem 1. *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 $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.

Sketch of the proof of the theorem

Let $g(a, b) = \sum_i (y_i - (a + b\cdot x_i))^2$. We optimize g, by setting the derivatives of g

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

to 0 and obtain

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

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

gives us

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

and the theorem follows by solving this for $\hat{a}$ and $\hat{b}$. □

Regression and Correlation

For the bias-corrected (that is, computed with $n-1$) standard deviations s_x and s_y and the bias-corrected sample covarinance

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

we obtain for the estimated slope of the regression line:

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

For the sample correlation $\operatorname{cor}(x, y) = \operatorname{cov}(x, y)/(s_x \cdot s_y)$ we obtain

$$\widehat{b} = \frac{\operatorname{cov}(x, y)}{s_x^2} = \frac{\operatorname{cor}(x, y) \cdot s_x \cdot s_y}{s_x^2} = \operatorname{cor}(x, y) \cdot \frac{s_y}{s_x}.$$

In particular, $\widehat{b}$ is equal to the correlation if and only if $s_x = s_y$.

Model:

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

[1.5ex] How to compute the significance of a relationship between the *explanatory trait* X and the *target variable* Y? [1.5ex] In other words: How can we test the null hypothesis $b = 0$? [1.5ex] We have estimated b by $\hat{b} \neq 0$. Could the true b be 0? [1.5ex] How large is the standard error of $\hat{b}$?

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

not random: a, b, x_i, σ^2 random: ε, y_i

$$\operatorname{var}(y_i) = \operatorname{var}(a + b \cdot x_i + \varepsilon) = \operatorname{var}(\varepsilon) = \sigma^2$$

and $y_1, y_2, \ldots, y_n$ are stochastically independent.

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

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

In fact $\hat{b}$ is normally distributed with mean b and

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

Problem: We do not know σ^2. We estimate σ^2 by considering the residual variance:

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

Note that we divide by $n-2$. The reason for this is that two model parameters a and b have been estimated, which means that two degrees of freedom got lost.

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

Estimate σ^2 by

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

Then

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

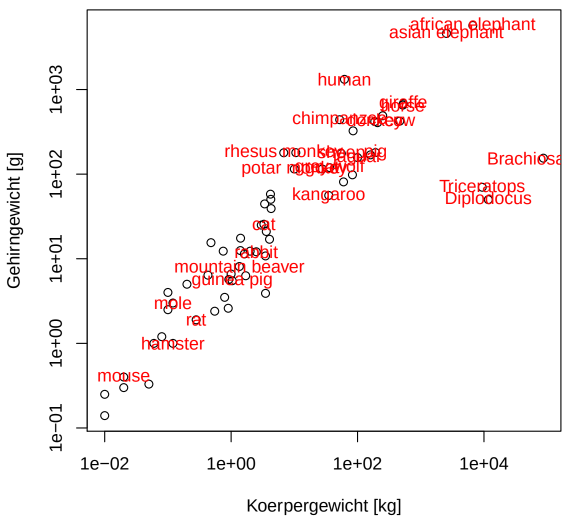
is Student-t-distributed with $n-2$ degrees of freedom and we can apply the t-test to test the null hypothesis $b = 0$.

2 log-scaling the data

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

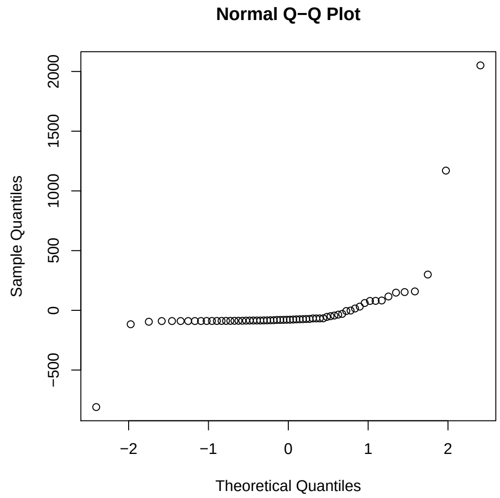
```
> data
  weight.kg. brain.weight.g          species  extinct
1     6654.00        5712.00 african elephant   no
2        1.00           6.60                    no
3        3.39          44.50                    no
4        0.92           5.70                    no
5     2547.00        4603.00   asian 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
.          .               .            .
.          .               .            .
.          .               .            .

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

typische Werte bei 65 Wirbeltierarten

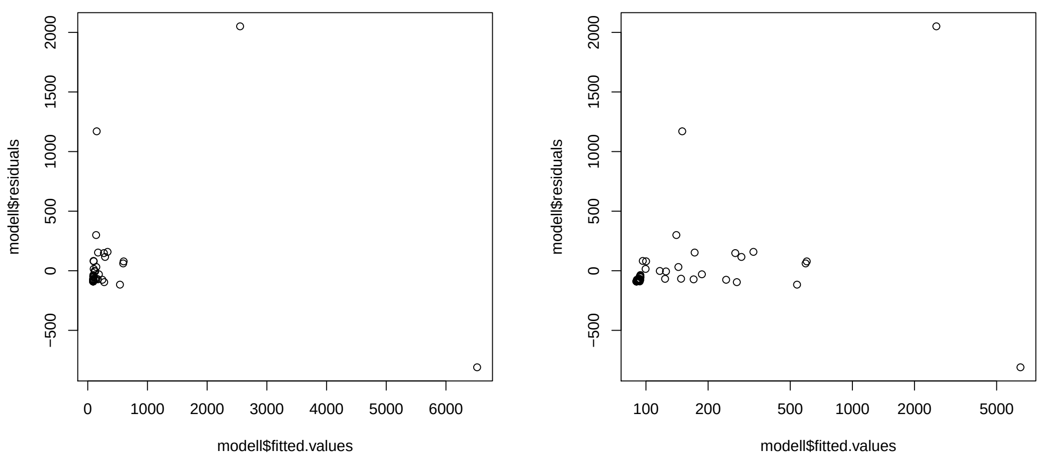



```
> modell <- lm(brain.weight.g~weight.kg.,subset=extinct=="no")
> summary(modell)
Call:
lm(formula = brain.weight.g ~ weight.kg., subset = extinct ==
    "no")
Residuals:
    Min       1Q  Median       3Q      Max
-809.95  -87.43  -78.55  -31.17 2051.05
Coefficients:
            Estimate Std. Error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

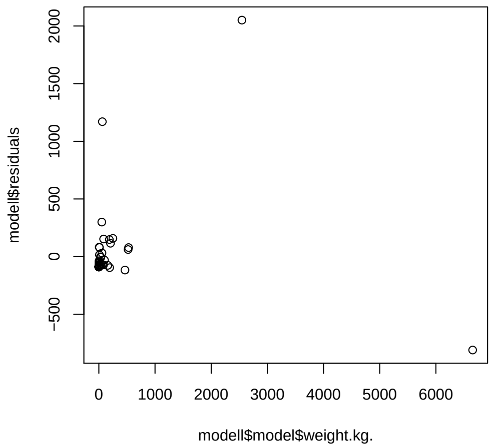
```
qqnorm(modell$residuals)
```

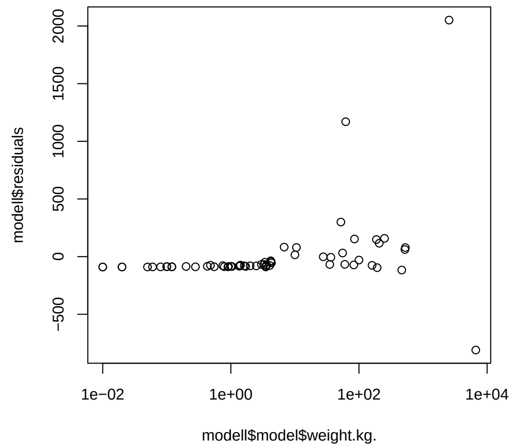



```
plot(modell$fitted.values,modell$residuals)
plot(modell$fitted.values,modell$residuals,log='x')
```




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





We see that the residuals' variance depends on the fitted values (or the body weight): "heteroscedasticity"

The model assumes *homoscedasticity*, 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

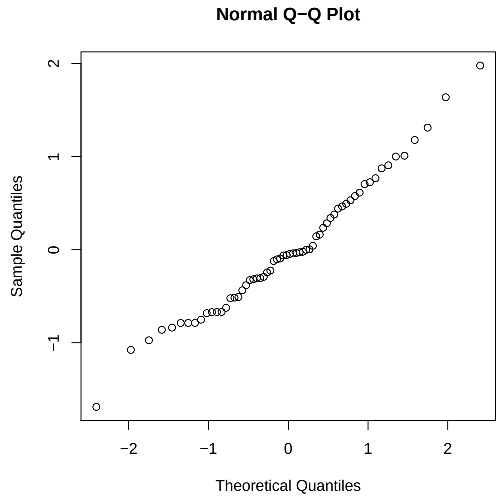
Actually not so surprising: An elephant's brain of typically 5 kg can easily be 500 g lighter or heavier from individual to individual. This can not happen for a mouse brain of typically 5 g. The latter will rather also vary by 10%, i.e. 0.5 g. Thus, the variance is not additive but rather multiplicative:

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

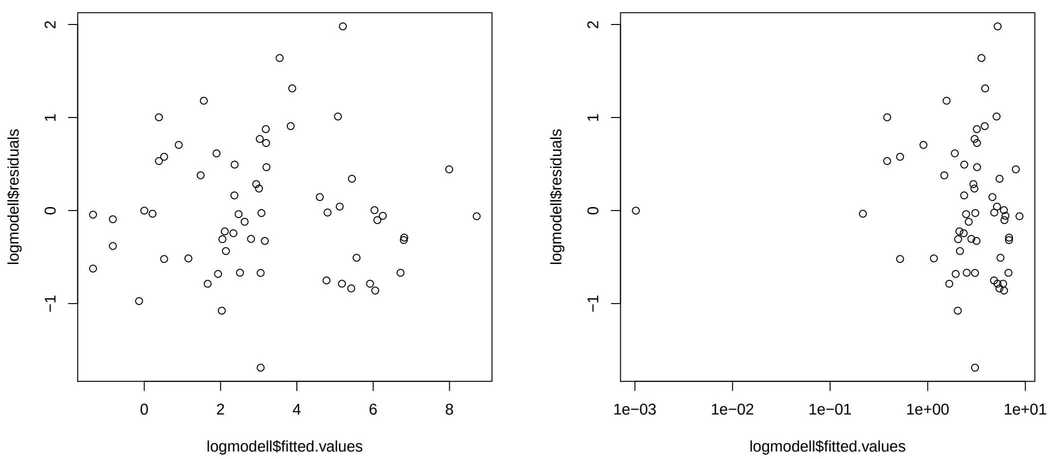
We can convert this into something with additive randomness by taking the log:

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

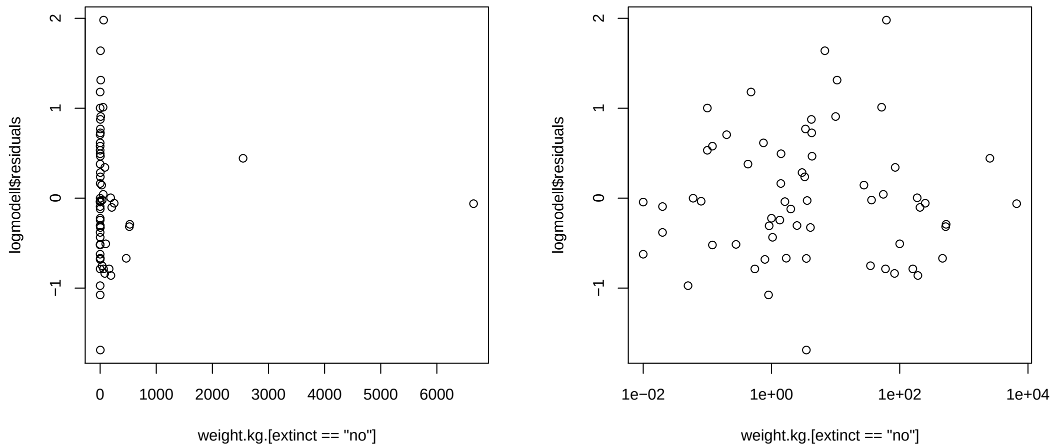
```
> logmodell <- lm(log(brain.weight.g)~log(weight.kg.),subset=extinct=="no")
> summary(logmodell)

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

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      2.11067    0.09794   21.55   <2e-16 ***
log(weight.kg.)  0.74985    0.02888   25.97   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 0.7052 on 60 degrees of freedom
Multiple R-squared: 0.9183, Adjusted R-squared: 0.9169
F-statistic: 674.3 on 1 and 60 DF,  p-value: < 2.2e-16

qqnorm(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' )
```



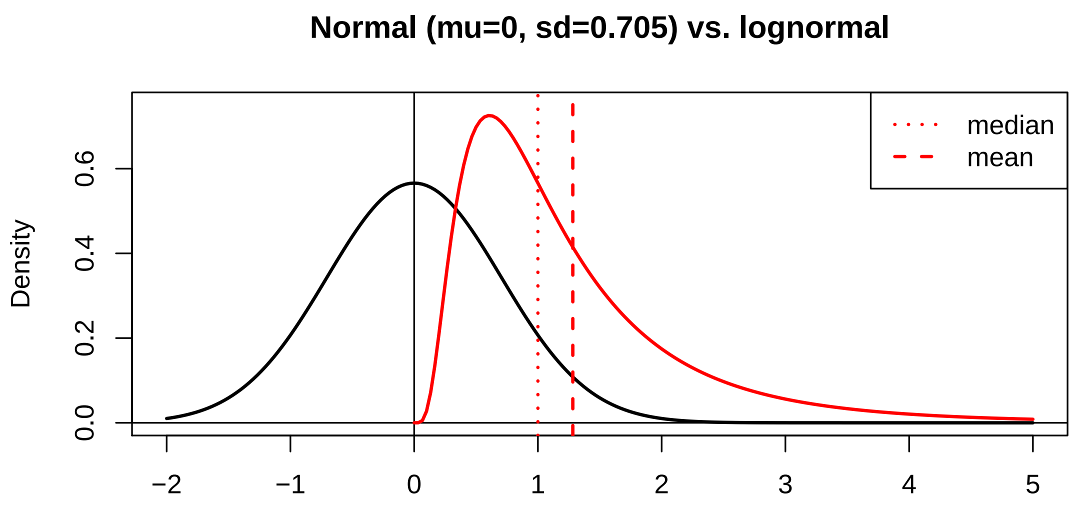
What does this model predict?

```
                  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
```

x body weight in kg [`weight.kg.`]

Y brain mass in gram [`brain.weight.`]

$$
\begin{aligned}
\log(Y) &\approx 2.11 + 0.75 \cdot \log(x) + \varepsilon \\
Y = e^{\log(Y)} &= e^{2.11 + 0.75 \cdot \log(x) + \varepsilon} \\
&= e^{2.11} \cdot e^{0.75 \cdot \log(x)} \cdot e^{\varepsilon} \approx 8.25 \cdot \left(e^{\log(x)}\right)^{0.75} \cdot e^{\varepsilon} \\
&= 8.25 \cdot x^{3/4} \cdot e^{\varepsilon}
\end{aligned}
$$



$$\varepsilon \sim \mathcal{N}\left(0, 0.705^2\right)$$

If $Z \sim \mathcal{N}(\mu, \sigma^2)$, then e^Z is log-normally distributed and $\mathbb{E}\left(e^Z\right) = e^{\mu+\sigma^2/2}$. Therefore,

$$\begin{aligned} \mathbb{E}e^{\varepsilon} &= e^{0+0.705^2/2} \approx 1.28 \\ \mathbb{E}Y &\approx 8.25 \cdot x^{3/4} \cdot 1.28 = 10.56 \cdot x^{3/4}. \end{aligned}$$

3 Checking model assumptions

Is the model appropriate for the data?, e.g

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

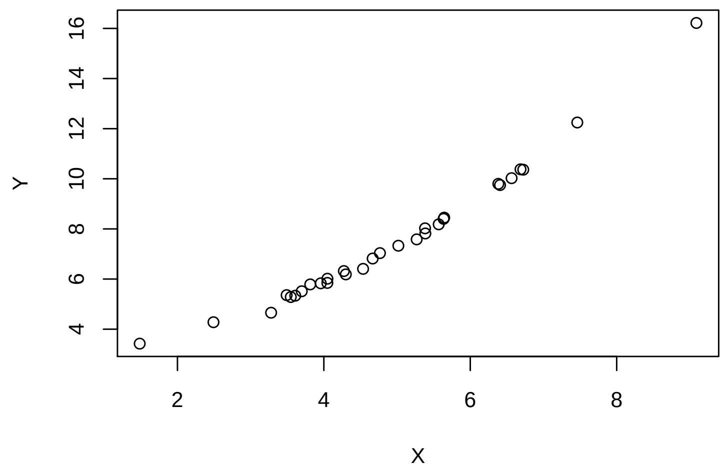
If the model fits, the residuals

$$r_i = y_i - \left(\widehat{a} + \widehat{b} \cdot x_i\right)$$

approximate the $\varepsilon_i = y_i - (a + b \cdot x_i)$
and therefore must

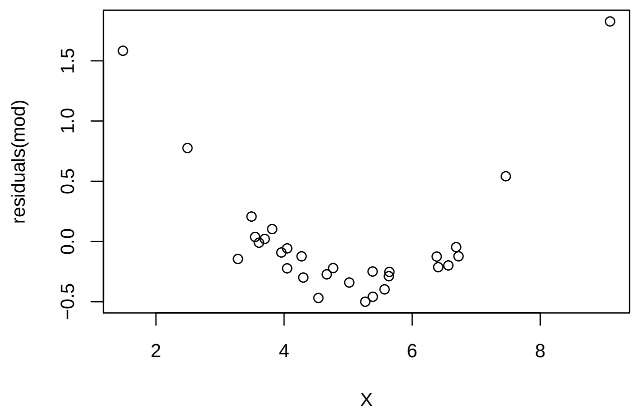
- look normally distributed and
- must not have obvious dependencies with X or $\widehat{a} + \widehat{b} \cdot X$.

Example: is the relation between X and Y sufficiently well described by the linear equation $Y_i = a + b \cdot X_i + \varepsilon_i$? [-0.5cm]




```
> mod <- lm(Y ~ X)
> summary(mod)

Call:
lm(formula = Y ~ X)

Residuals:
     Min       1Q   Median       3Q      Max
-0.49984 -0.26727 -0.13472  0.01344  1.82718

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.61118    0.33295  -1.836    0.077 .
X            1.65055    0.06472  25.505   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5473 on 28 degrees of freedom
Multiple R-squared: 0.9587, Adjusted R-squared: 0.9573
F-statistic: 650.5 on 1 and 28 DF,  p-value: < 2.2e-16
```

```
> plot(X,residuals(mod))
```
[-0.5cm]



Obviously, the residuals tend to be larger for very large and very small values of X than for mean values of X. That should not be!

Idea: Instead fit a section of a parabola instead of aline to (x_i, y_i), i.e. a model of the form

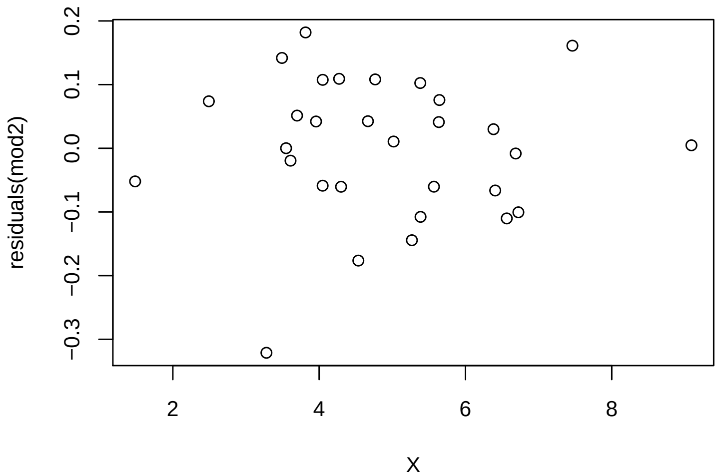
$$Y_i = a + b \cdot X_i + c \cdot X_i^2 + \varepsilon_i.$$

Is this still a linear model? Yes: Let $Z = X^2$, then Y is linear in X and Z.

In R:

```
> Z <- X^2
> mod2 <- mod <- lm(Y ~ X+Z)


> summary(mod2)

Call:
lm(formula = Y ~ X + Z)

Residuals:
      Min        1Q    Median        3Q       Max
-0.321122 -0.060329  0.007706  0.075337  0.181965

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.933154   0.158825  18.468   <2e-16 ***
X           0.150857   0.061921   2.436   0.0217 *
Z           0.144156   0.005809  24.817   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1142 on 27 degrees of freedom
Multiple R-squared: 0.9983, Adjusted R-squared: 0.9981
F-statistic:  7776 on 2 and 27 DF,  p-value: < 2.2e-16
```

For this model there is no obvious dependence between X and the residuals:
`plot(X,residuals(mod2))` [-5mm]



Is the assumption of normality in the model $Y_i = a + b \cdot X_i + \varepsilon_i$ in accordance with the data?

Are the residuals $r_i = Y_i - (\widehat{a} + \widehat{b} \cdot X_i)$ more or less normally distributed?

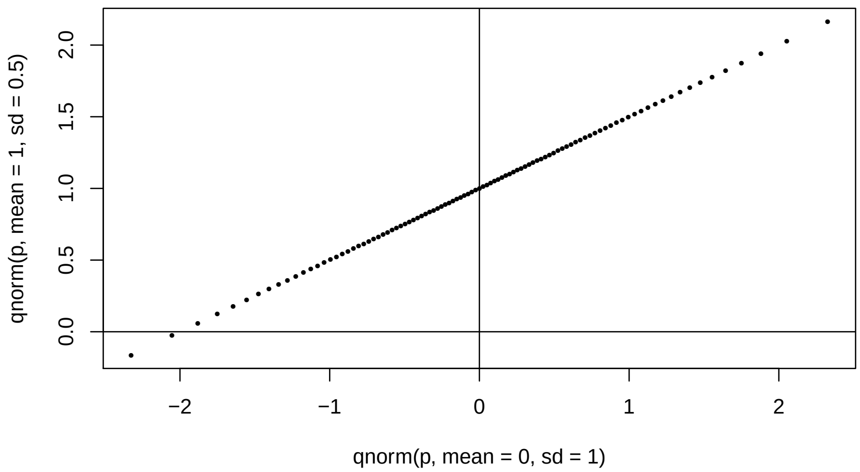
Graphical Methods: compare the theoretical quantiles of the standard normal distribution $\mathcal{N}(0, 1)$ with those of the residuals.

Background: If we plot the quantiles of $\mathcal{N}(\mu, \sigma^2)$ against those of $\mathcal{N}(0, 1)$, we obtain a line $y(x) = \mu + \sigma \cdot x$. (Reason: If X is standard-normally distributed and $Y = a + b \cdot X$, then Y is normally distributed with mean a and variance b^2.)

Before we fit the model with `lm()` we first have to check whether the model assumptions are fulfilled.
~~Before we fit the model with `lm()` we first have to check whether the model assumptions are fulfilled.~~

To check the assumptions underlying a linear model we need the residuals. To compute the residuals we first have to fit the model (in R with `lm()`). After that we can check the model assumptions and decide whether we stay with this model or still have to modify it.

```
p <- seq(from=0.01,to=0.99,by=0.01)
plot(qnorm(p,mean=0,sd=1),qnorm(p,mean=1,sd=0.5),
                                       pch=16,cex=0.5)
abline(v=0,h=0)
```



If we plot the *empirical* quantiles of a sample from a normal distribution against the theoretical quantiles of a standard normal distribution, the values are not precisely on the line but are scattered around a line.

If no *systematic* deviations from an imaginary line are recognizable: Normal distribution assumption is acceptable

If *systematic* deviations from an imaginary line are obvious: Assumption of normality may be problematic. It may be necessary to rescale variables or to take additional explanatory variables into account.

Some of what you should be able to explain

- Model assumptions underlying linear regression
  - Equation
  - What is random, what is fixed?
- approach: minimize sum of squared residuals
- optimal solution for slope and intercept
- slope vs. correlation
- t-test for the slope (standard error, test statistic and df)
- scaling the data: when, why, how?
- qqnorm plots
  - theory
  - how to use them to judge model assumptions

4 Applications in Quantitative Genetics

Quantitative Genetics

- natural selection needs phenotypic variation to operate

- many traits are influenced by few major and many minor genes
- Q.G. has been successfully applied in animal an plant breeding
- application to evolutionary and ecological processes not trivial
- no exact knowledge of genetic mechanisms, rather statistical approach
- QTL analysis to search for genomic regions that influence a trait

Aims for now

- use formulas for Var and Cov to understand how
  - natural variation and
  - correlation of a trait with fitness
  - heritability of the trait

  influence the effect of selection
- based on the theoretical considerations how to predict effect of selection based on data?
- Results will be summarized in
  - Robertson-Price identity
  - breeder's equation

Recommended Books

References

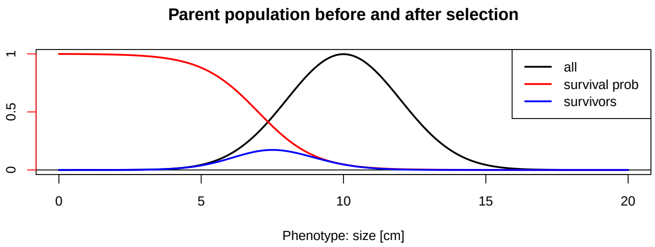
[FM96] T.S. Falconer, T.F.C. Mackay (1996) *Introduction to Quantitative Genetics* (4. ed.) Pearson Education Ltd., UK

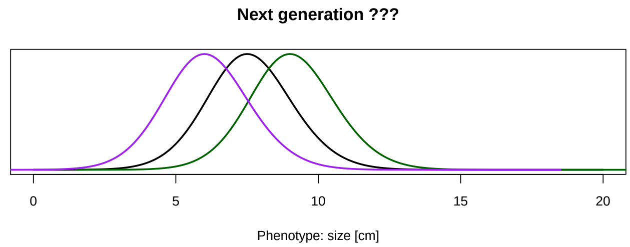
[LW98] M. Lynch, B. Walsh (1998) *Genetics and Analysis of Quantitative Traits* Sinauer Associates, Inc., Sunderland, MA, USA

[F19] F19] J. Felsenstein (2019+) *Theoretical Evolutionary Genetics* https://felsenst.github.io/pgbook/pgbook.html

[BB+07] N.H. Barton, D.E.G. Briggs, J.A. Eisen, D.B. Goldstein, N.H. Patel (2007) *Evolution* Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, USA

Selection on quantitative trait

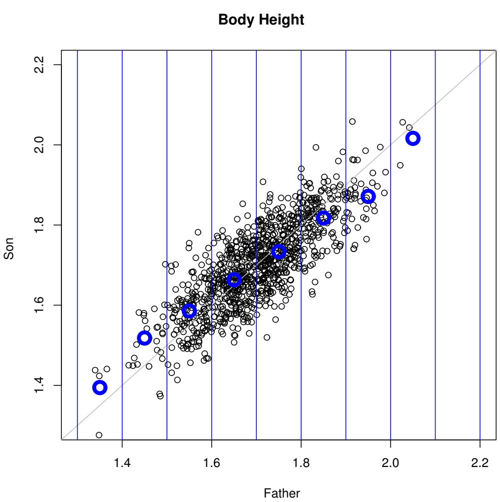


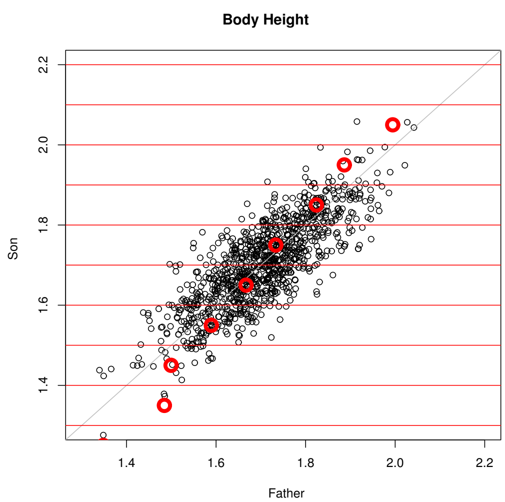


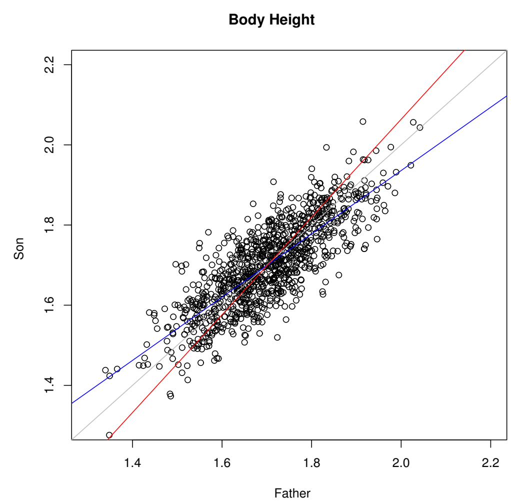
Origin of the word "Regression"

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.

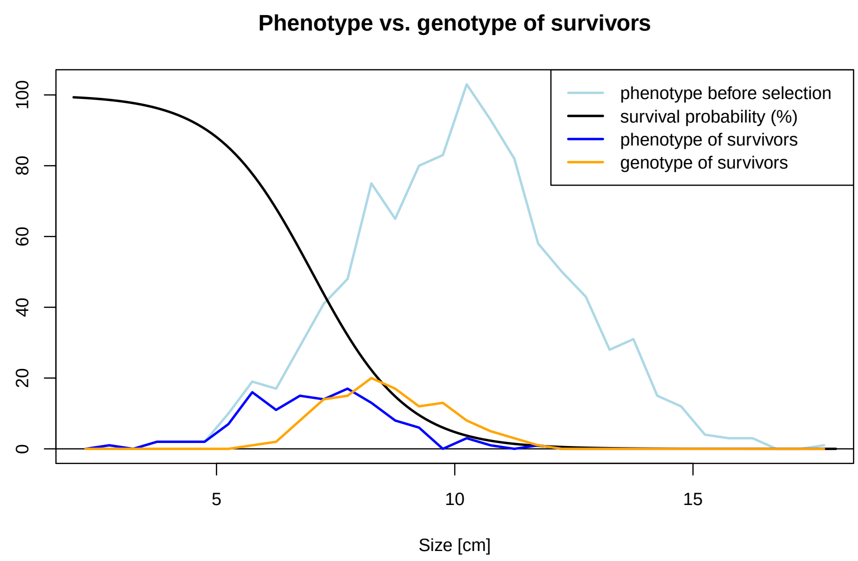




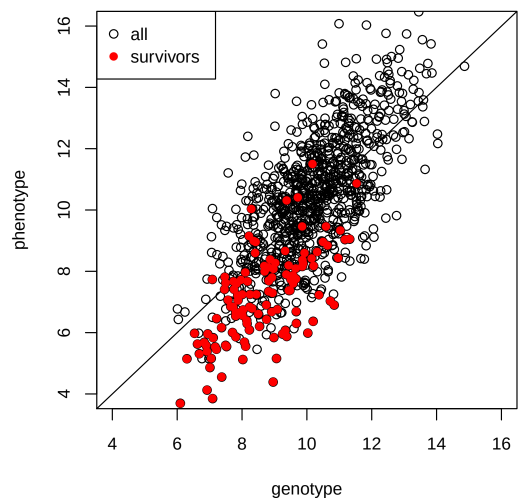


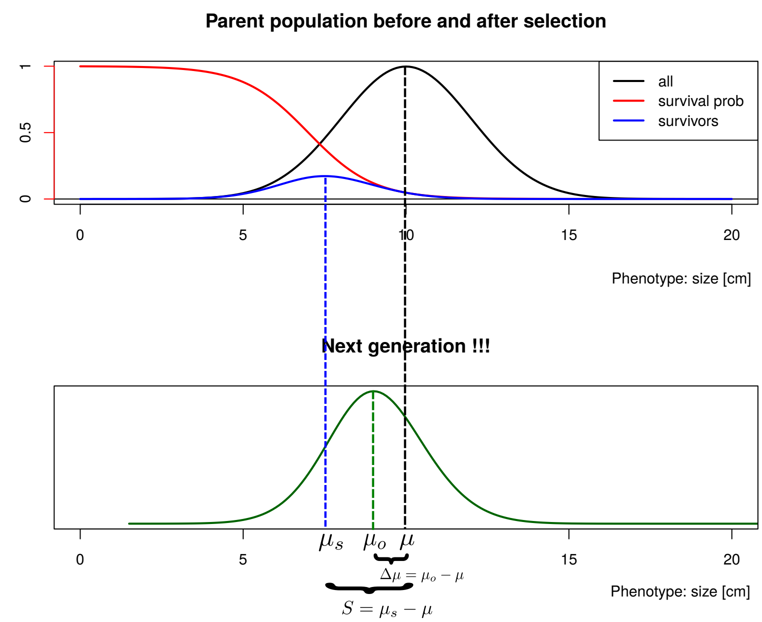
Similar effects

- 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.




```
genotype <- rnorm(1000,10,1.5)
environment <- rnorm(1000,0,1.5)
phenotype <- genotype + environment

hist(phenotype,col="lightblue",breaks=4:36/2)

survival.prob <- function(x) {
  1-1/(1+exp(-x+7))
}

lines(20:180/10,survival.prob(20:180/10)*100,lwd=2)

survivors <- rbinom(1000,size=1,prob=(survival.prob(phenotype)))

hist(phenotype[survivors==1],add=TRUE,col="blue",breaks=4:36/2)
hist(genotype[survivors==1],add=TRUE,col="orange",breaks=4:36/2)
```





Classical estimated of heritabilities after Falconer (1981) *Introduction to quantitative genetics*

Species	*Trait*	*Heritability*
humans	stature	0.65
	serum immunoglobulin	0.45
cattle	body weight	0.65
	milk yield	0.35
poultry	body weight	0.40
	egg production	0.10

We will now derive two equations:

(Robertson-)Price-equation: How selection shifts the mean phenotype (in the same generation)

breeders' equation:

- Predict change from one generation to the next
- Account for selection and heritability
- use a measure of heritability that can be estimated from parent-offspring comparisons

μ mean phenotype before selection

μ_s mean phenotype after selection but before reproduction

$S = \mu_s - \mu$ **directional selection differential**

μ_o mean phenotype in offspring generation

$\Delta\mu = \mu_o - \mu$

$W(z)$ **individual fitness**: probability that individual with phenotype z will survive to reproduce

$p(z)$ density of phenotype z before selection

$\overline{W} = \int W(z) \cdot p(z)\ dz$ mean individual fitness

$w(z) = W(z)/\overline{W}$ relative individual fitness

$p_s(z) = w(z)p(z)$ density of phenotype z after selection but before reproduction (density in a stochastic sense, i.e. integrates to 1)

Let Z be the phenotype of an individual drawn randomly from the parent population before selection.

$$\begin{aligned} \mu &= \mathbb{E}Z \qquad \mathbb{E}(w(Z)) = 1 \\ \mu_S &= \int_z z \cdot p_S(z)\ dz = \int_z z \cdot w(z) \cdot p(z)\ dz = \mathbb{E}(Z \cdot w(Z)) \end{aligned}$$

$$\Rightarrow S = \mu_s - \mu = \mathbb{E}(Z \cdot w(Z)) - \mathbb{E}(Z) \cdot \mathbb{E}(w(Z)) = \mathrm{Cov}(Z, w(Z))$$

Thus, we obtain:

Theorem 2 (Robertson-Price identity; Robertson 1966; Price 1970/72)**.**

$$S = Cov(Z, w(Z))$$

Assume we can partition the phenotypic value Z into a genotypic value G and an environmental (or random) deviation E:

$$Z = G + E$$

Then,

$$\mathrm{Cov}(Z, G) = \mathrm{Cov}(G + E, G) = \mathrm{Var}(G) + \mathrm{Cov}(E, G)$$

and

$$\mathrm{Cor}(Z, G) = \frac{\mathrm{Var}(G) + \mathrm{Cov}(G, E)}{\sigma_G \cdot \sigma_Z}.$$

In the special case of $\mathrm{Cov}(G, E) = 0$, we obtain for the genetic contribution of the phenotypic variance

$$\mathrm{Cor}^2(G, Z) = \frac{\mathrm{Var}(G)}{\mathrm{Var}(Z)}.$$

(Note that if $\mathrm{Cov}(G, E) = 0$, then $\mathrm{Var}(Z)$=$\mathrm{Var}(G)$+$\mathrm{Var}(E)$)

Note that if E is really due to environmental effects, $\mathrm{Cov}(G, E)$ may not be 0 if the population is genetically and spatially structured (and for many other possible reasons).

In any case,

$$\frac{\mathrm{Var}(G)}{\mathrm{Var}(Z)} =: H^2$$

is called **heritability in the broad sense**.

Problem: $\mathrm{Var}(G)$ and thus also H^2 are parameters that are hard to estimate.

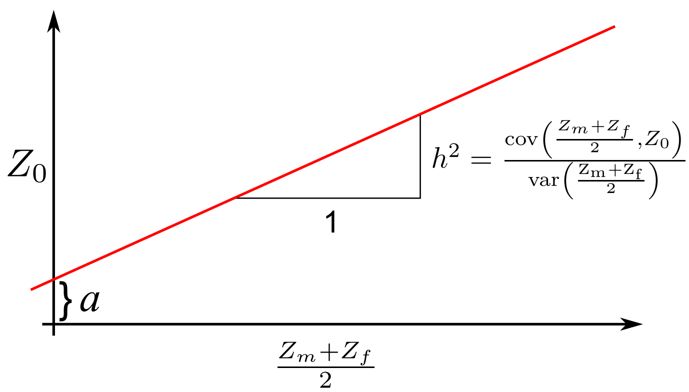
narrow-sense heritability

Let Z_m, Z_f, Z_o be a the phenotype sampled from a triplet of mother, father and an offspring, sampled from the population. The narrow-sense heritability h^2 is defined by

$$h^2 := \frac{\mathrm{Cov}\left(\frac{Z_m+Z_f}{2}, Z_o\right)}{\mathrm{Var}\left(\frac{Z_m+Z_f}{2}\right)}.$$

It is the slope of the regression line to predict Z_0 from the mid-parental phenotype $\frac{Z_m+Z_f}{2}$ and can be estimated from a sample of many parent-offspring triples.

Remember: The line that predicts Y from X has slope $\mathrm{Cov}(X, Y)/\mathrm{Var}(X)$.



If there was selection, then:

$$\mu_o = \mathbb{E}Z_o = \mathbb{E}\left(a + h^2 \cdot \frac{Z_m + Z_f}{2}\right) = a + h^2 \cdot \mathbb{E}\left(\frac{Z_m + Z_f}{2}\right) = a + h^2 \cdot \mu_S$$

If the values $\widetilde{Z}_m, \widetilde{Z}_f, \widetilde{Z}_o$ stem from a population with no selection, we assume that the mean phenotype is the same in the two generations:

$$\mu = \mathbb{E}\widetilde{Z}_o = a + h^2 \cdot \mathbb{E}\left(\frac{\widetilde{Z}_m + \widetilde{Z}_f}{2}\right) = a + h^2 \cdot \mu$$

This implies: $\Delta\mu = \mu_o - \mu = (a + h^2 \cdot \mu_S) - (a + h^2 \cdot \mu) = h^2 \cdot (\mu_S - \mu) = h^2 \cdot S$

Theorem 3 (breeders' equation)**.**

$$\Delta\mu = h^2 S$$

Equivalent definition of h^2

Assume that Z_m and Z_f are independent and have the same distribution as Z. Then follows

$$\mathrm{Var}\left(\frac{Z_m+Z_f}{2}\right) \;=\; \frac{1}{4}\mathrm{Var}\left(Z_m+Z_f\right) \;=\; \frac{1}{4}\left(\mathrm{Var}\left(Z_m\right)+\mathrm{Var}\left(Z_f\right)\right) \;=\; \frac{1}{2}\mathrm{Var}\left(Z\right),$$

and

$$\mathrm{Cov}\left(\frac{Z_m+Z_f}{2}, Z_0\right) \;=\; \frac{1}{2}\mathrm{Cov}\left(Z_m+Z_f, Z_0\right) \;=\; \frac{\mathrm{Cov}\left(Z_m, Z_0\right)+\mathrm{Cov}\left(Z_f, Z_0\right)}{2}.$$

And thus

$$h^2 = \frac{\mathrm{Cov}\left(\frac{Z_m+Z_f}{2}, Z_0\right)}{\mathrm{Var}\left(\frac{Z_m+Z_f}{2}\right)} = \frac{\mathrm{Cov}\left(Z_m, Z_0\right)+\mathrm{Cov}\left(Z_f, Z_0\right)}{\mathrm{Var}\left(Z\right)}$$

Equivalent definition of h^2 under certain assumptions

Let G_m and G_f be the phenotypic effects of the genes transmitted by the mother and the father to the offspring.

If mating is so random and if there are no correlations (between parental genotypes and environmental effects etc.), and if genetic effects are **additive**, we obtain

$$\mathrm{Cov}\left(\frac{Z_m+Z_f}{2}, Z_o\right) = \mathrm{Cov}\left(\frac{G_m+G_f}{2}, G_m+G_f\right) = \frac{\mathrm{Var}G_m+\mathrm{Var}G_f}{2},$$

and thus

$$h^2 \;=\; \frac{\mathrm{Cov}\left(\frac{Z_m+Z_f}{2}, Z_o\right)}{\frac{1}{2}\mathrm{Var}\left(Z\right)} \;=\; \frac{\mathrm{Var}G_m+\mathrm{Var}G_f}{\mathrm{Var}\left(Z\right)}$$

How to define h^2 if genetic effects are not additive

Let A and B be the alleles at one locus and let z_{AA}, z_{AB} and z_{BB} be the average phenotypes of individuals with genotypes AA, AB and BB.

What if $z_{AB} \neq (z_{AA}+z_{BB})/2$?

Then, decompose the genetic effects for each $(u,v) \in \{(A,A),(A,B),(B,B)\}$ as follows:

$$z_{uv} = \mu + G(u) + G(v) + D(u,v)$$

by setting these components as follows, where p is the population frequency of A.

$$\begin{aligned} \mu &= p^2 \cdot z_{AA} + 2p(1-p)\cdot z_{AB} + (1-p)^2 \cdot z_{BB} \\ G(A) &= p\cdot z_{AA} + (1-p)\cdot z_{AB} - \mu \\ G(B) &= p\cdot z_{AB} + (1-p)\cdot z_{BB} - \mu \\ D(u,v) &= z_{uv} - \mu - G(u) - G(v) \end{aligned}$$

If U and V are sampled independently from $\{A,B\}$ according to the population allele frequencies p and $1-p$, we obtain that $G(U)$, $G(V)$ and $D(U,V)$ are random variables with the following properties:

- their expected values $\mathbb{E}G(U)$, $\mathbb{E}G(V)$ and $\mathbb{E}D(U,V)$ are 0.
- $G(U)$, $G(V)$ and $D(U,V)$ are uncorrelated

see appendix in handout or e.g. Felsenstein (2019+) *Theoretical Evolutionary Genetics* https://felsenst.github.io/pgbook/pgbook.html

$G(A)$ **and** $G(B)$ are called *additive effects*. (I find this a bit misleading – what is additive here? Actutally they are *average effects*.)

$D(u,v)$ is called *dominance deviation*.

Note that the separation between *additive effects* and *dominance deviation* depends on population allele frequencies.

Assume now we have n unlinked loci with additive effects $G_1(.)$, $G_2(.)$, ..., $G_n(.)$ and dominance deviations $D_1(.,.)$, $D_2(.,.)$, ..., $D_n(.,.)$, and the effects are **additive among the loci**, that is, no epistasis. (Otherwise: how to separate additive from non-additive locus interactions, see e.g. Falconer, Mackay (1996) *Introduction to Quantitative Genetics.* 4th ed.)

Then, the phenotypic variance is the sum of

the so-called additive variance $V_A = \mathrm{Var}\left(G_1(U_1) + G_1(V_1) + G_2(U_2) + G_2(V_2) + \cdots + G_n(U_n) + G_n(V_n)\right)$,

the so-called dominance variance $V_D = \mathrm{Var}\left(D_1(U_1, V_1) + D_2(U_2, V_2) + \cdots + D_n(U_n, V_n)\right)$ and

and the environmental variance V_E.

We can then define narrow-sense heritability as the **fraction of phenotypic variation that is due to additive genetic effects**

$$h^2 = \frac{V_A}{V_A + V_D + V_E},$$

and this is still $\frac{\mathrm{Cov}((Z_m+Z_f)/2, Z_O)}{\mathrm{Var}((Z_m+Z_f)/2)}$, see handout appendix or e.g. Felsenstein (2019+) for details.

Example

References

[1] Galen (1996) Rates of floral evolution: adaptation to bumblebee pollination in an alpine wildflower, *Polemonium viscosum* *Evolution* 50(1): 120–125

- long-term experiment, trait is corolla flare
- S was measured as
  - 7% when estimated from number of seeds
  - 17% when estimated from number of surviving offspring after 6 years
- $h^2 \approx 1$
- Change of trait 9% in one generation

Some of the things you should be able to explain

- Robertson-Price identity
- Breeder's equation
- Why is the narrow-sense heritability and not the broad-sense heritability used in the breeder's equation
- Why additive genetic effects depend on population allele frequencies.
- Connection of narrow-sense heritability and additive genetics effects
  - definition of additive genetics effects depends on population allele frequencies
  - additivity between loci still required

Appendix: Some proofs regarding narrow-sense heritability

We consider the case of a single diploid locus with two alleles A and B and phenotypes z_{AA}, $z_{AB} = z_{BA}$ and z_{BB}. Note that μ as defined above in the context of narrow-sense heritability can also be written (or even defined) as $\mu = \mathbb{E}(z_{UV})$, where here and in the following U and V are random alleles that were independently draw according to the allele frequencies (in the context of Price's equation after selection; also p above should refer to this). Also that the "additive effect" $G(A)$ can also be written with expected values as

$$G(A) = \mathbb{E}(z_{AU}) - \mu.$$

To check and better understand some of the above statements on narrow-sense heritability, we will used the concept of *conditional expectations* and conditional covariances. For a random variables X and Y with a finite or countable state spaces, we can define

$$\mathbb{E}(X|Y = y) = \sum_x x \cdot \Pr(X = x|Y = y),$$

and $\mathbb{E}(X|Y)$ is then the random variable that depends on Y: it takes the value $\mathbb{E}(X|Y = y)$ if $Y = y$. With this we obtain for our random alleles U and V (which are independently drawn according to the allele probabilities):

$$G(U) = \mathbb{E}(z_{UV}|U) - \mu = \mathbb{E}(z_{UV} - \mu \mid U)$$

A useful formula is the law of total expectation

$$\mathbb{E}(\mathbb{E}(X|Y)) = \mathbb{E}X.$$

At least for random variables with finite or countable statespace it is easy to check:

$$\mathbb{E}(\mathbb{E}(X|Y)) = \sum_y \Pr(Y = y) \sum_x x \Pr(X = x|Y = y) = \sum_x x \sum_y \Pr(X = x, Y = y) = \sum_x x \Pr(X = x) = \mathbb{E}X.$$

With the law of total expectation we obtain

$$\mathbb{E}(G(U)) = \mathbb{E}(\mathbb{E}(z_{UV}|U)) - \mu = \mathbb{E}(z_{UV}) - \mu = 0.$$

The *conditional covariance* is defined as follows:

$$\mathsf{Cov}(X, Y|Z) = \mathbb{E}(X \cdot Y|Z) - \mathbb{E}(X|Z) \cdot \mathbb{E}(Y|Z)$$

Below we will use the *covariance decomposition formula*

$$\mathsf{Cov}(X, Y) = \mathbb{E}(\mathsf{Cov}(X, Y|Z)) + \mathsf{Cov}(\mathbb{E}(X|Z), \mathbb{E}(Y|Z)),$$

which we can derive from the law of total expectation as follows:

$$\begin{aligned}
\mathsf{Cov}(X,Y) &= \mathbb{E}(X\cdot Y)-\mathbb{E}X\cdot\mathbb{E}Y\\
&= \mathbb{E}(\mathbb{E}(X\cdot Y|Z))-\mathbb{E}(\mathbb{E}(X|Z))\cdot\mathbb{E}(\mathbb{E}(Y|Z))\\
&= \mathbb{E}(\mathbb{E}(X\cdot Y|Z))-\mathbb{E}(\mathbb{E}(X|Z)\cdot\mathbb{E}(Y|Z))+\mathbb{E}(\mathbb{E}(X|Z)\cdot\mathbb{E}(Y|Z))-\mathbb{E}(\mathbb{E}(X|Z))\cdot\mathbb{E}(\mathbb{E}(Y|Z))\\
&= \mathbb{E}(\mathbb{E}(X\cdot Y|Z)-\mathbb{E}(X|Z)\cdot\mathbb{E}(Y|Z))+\mathsf{Cov}(\mathbb{E}(X|Z),\mathbb{E}(Y|Z))\\
&= \mathbb{E}(\mathsf{Cov}(X,Y|Z))+\mathsf{Cov}(\mathbb{E}(X|Z),\mathbb{E}(Y|Z))
\end{aligned}$$

We apply the covariance decomposition formula to show that $G(U)$ and $D(U,V)$ are uncorrelated:

$$\begin{aligned}
\mathsf{Cov}(G(U),D(U,V)) &= \mathsf{Cov}(G(U),z_{UV}-G(U)-G(V)-\mu)\\
&= \mathsf{Cov}(G(U),z_{UV}-G(U)) \qquad \text{(as } G(V) \text{ is independent of } G(V) \text{ and } \mu \text{ is constant))}\\
&= \mathsf{Cov}(G(U),z_{UV})-\mathsf{Cov}(G(U),G(U))\\
&= \mathbb{E}(\mathsf{Cov}(G(U),z_{UV}|V))+\mathsf{Cov}(\mathbb{E}(G(U)|V),\mathbb{E}(z_{UV}|V))-\mathsf{Cov}(G(U),G(U))
\end{aligned}$$

Note that the second summand is 0 because $\mathbb{E}(G(U)|V)$ is constant as $G(U)$ is independent of V. Further, from the independence of $G(U)$ of V and the bilinearity of the covariance follows that

$$\mathbb{E}(\mathsf{Cov}(G(U),z_{UV}|V))=\mathsf{Cov}(G(U),\mathbb{E}(z_{UV}|V))=\mathsf{Cov}(G(U),G(U)+\mu)=\mathsf{Cov}(G(U),G(U)),$$

from which follows that

$$\mathsf{Cov}(G(U),D(U,V))=\mathsf{Cov}(G(U),G(U))-\mathsf{Cov}(G(U),G(U))=0.$$

Okay, the argument with the bilinearity was a bit hand-waving, so here are the details:

$$\begin{aligned}
\mathbb{E}(\mathsf{Cov}(G(U),z_{UV}|V)) &= p\cdot\mathsf{Cov}(G(U),z_{UA})+(1-p)\cdot\mathsf{Cov}(G(U),z_{UB})\\
&= \mathsf{Cov}(G(U),p\cdot z_{UA}+(1-p)z_{UB})\\
&= \mathsf{Cov}(G(U),G(U)+\mu)\\
&= \mathsf{Cov}(G(U),G(U))
\end{aligned}$$

In the same way it follows that $G(V)$ and $D(U,V)$ are uncorrelated, and $G(U)$ are $G(V)$ independent anyway.

If U, V and W are independently sampled according to allele frequencies, $D(U,V)$ and $D(U,W)$, that is, the dominance component of a parent an offspring, are uncorrelated. For this, we first show that $\mathbb{E}(Z_{UV}\cdot Z_{UW})=\mathbb{E}(G(U)^2)+\mu$ and that $\mathbb{E}(Z_{UV}\cdot G(U))=\mathbb{E}(G(U)^2)$. We use that Z_{UV} and Z_{UW} are conditionally independent given U.

$$\begin{aligned}
\mathbb{E}(Z_{UV}\cdot Z_{UW}) &= \mathbb{E}(\mathbb{E}(Z_{UV}Z_{UW}|U))\\
&= \mathbb{E}(\mathbb{E}(Z_{UV}|U)\cdot\mathbb{E}(Z_{UW}|U))\\
&= \mathbb{E}((G(U)+\mu)\cdot(G(U)+\mu))\\
&= \mathbb{E}(G(U)^2)+2\mu\cdot\mathbb{E}(G(U))+\mu^2\\
&= \mathbb{E}(G(U)^2)+\mu^2
\end{aligned}$$

Next we use that $G(U)$ and Z_{UV} are conditionally independent given U.

$$\begin{aligned}
\mathbb{E}(G(U)\cdot z_{UV}) &= \mathbb{E}(\mathbb{E}(G(U)\cdot z_{UV}|U))\\
&= \mathbb{E}(\mathbb{E}(G(U)|U)\cdot\mathbb{E}(z_{UV}|U))\\
&= \mathbb{E}(G(U)\cdot(G(U)+\mu))\\
&= \mathbb{E}(G(U)^2)+\mathbb{E}(G(U))\cdot\mu))\\
&= \mathbb{E}(G(U)^2)
\end{aligned}$$

Now we can show that the dominance effects in a random pair consiting of one parent and one of its offspring are uncorrelated:

$$\begin{aligned}
\mathsf{Cov}(D(U,V),D(U,W)) &= \mathbb{E}(D(U,V)\cdot D(U,W)) \\
&= \mathbb{E}((z_{UV}-G(U)-G(V)-\mu)\cdot(z_{UW}-G(U)-G(W)-\mu)) \\
&= \mathbb{E}(z_{UV}\cdot z_{UW}-z_{UV}\cdot G(U)-\mu\cdot z_{UW}-\mu\cdot z_{UV}-z_{UW}\cdot G(U)+G(U)^2+\mu^2) \\
&\quad +\mathbb{E}(G(V))\cdot\mathbb{E}(z_{UW}-G(U)-G(W)-\mu) \\
&\quad +\mathbb{E}(G(W))\cdot\mathbb{E}(z_{UV}-G(U)-G(V)-\mu) \\
&= \mathbb{E}(z_{UV}\cdot z_{UW})-\mathbb{E}(z_{UV}\cdot G(U))-\mu\cdot\mathbb{E}(z_{UW})-\mu\cdot\mathbb{E}(z_{UV})-\mathbb{E}(z_{UW}\cdot G(U))+\mathbb{E}(G(U)^2)+\mu^2 \\
&= \mathbb{E}(z_{UV}\cdot z_{UW})-2\mathbb{E}(z_{UV}\cdot G(U))+\mathbb{E}(G(U)^2)-\mu^2 \\
&= \mathbb{E}(G(U)^2)+\mu^2-2\mathbb{E}(G(U)^2)+\mathbb{E}(G(U)^2)-\mu^2 \\
&= 0
\end{aligned}$$

Now, finally, for the proof that $\frac{V_A}{V_A+V_D+V_E}=\frac{\mathrm{Cov}((Z_m+Z_f)/2,Z_O)}{\mathrm{Var}((Z_m+Z_f)/2)}$:
For the denominator

$$\begin{aligned}
\mathsf{Var}((Z_m+Z_f)/2) &= \frac{1}{4}(\mathsf{Var}(Z_m)+\mathsf{Var}(Z_f)) \\
&= \frac{1}{2}\mathsf{Var}(Z_m) \\
&= \frac{1}{2}(V_A+V_D+V_E)
\end{aligned}$$

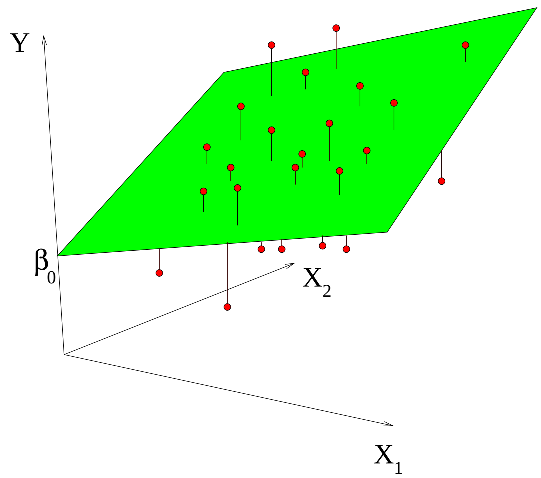
follows from the fact that the "additive" and dominance component are independent of each other and the assumption that the environmental component is independent of the genetic components. (A bit unclear is still whether the factor of 2 should be in the definition of narrow-sense heritability.)

Now for the numerator:

$$\begin{aligned}
\mathsf{Cov}((Z_m+Z_f)/2,Z_O) &= \frac{1}{2}\cdot(\mathsf{Cov}(Z_m,Z_O)+\mathsf{Cov}(Z_f,Z_O)) \\
&= \mathsf{Cov}(Z_m,Z_O) \\
&= \mathsf{Cov}(\mu+G(U)+G(V)+D(U,V),\mu+G(U)+G(W)+D(U,W)) \\
&\qquad \text{(as parent ans offspring share one allele)}U \\
&= \mathsf{Cov}(G(U)+D(U,V),G(U)+G(W)+D(U,W)) \\
&\qquad \text{(removed what is independent of the other side)} \\
&= \mathsf{Cov}(G(U),G(U))+\mathsf{Cov}(D(U,V),D(U,W)) \\
&\qquad \text{(because } G(U) \text{ is uncorrelated to } D(U,V) \text{ and } D(U,W)\text{)} \\
&= \mathsf{Cov}(G(U),G(U)) \qquad \text{(because } D(U,V) \text{ and } D(U,W) \text{ are uncorrelated)} \\
&= \frac{1}{2}V(A)
\end{aligned}$$

5 Multiple Regression

Multiple Regression



Multiple Regression

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

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

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

Example: species richness on sandy beaches

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

References

[ZIS07] 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:

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

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 = rikz)
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
F-statistic: 10.74 on 4 and 40 DF,  p-value: 5.237e-06
```

- 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 `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.

- 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?

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.

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.

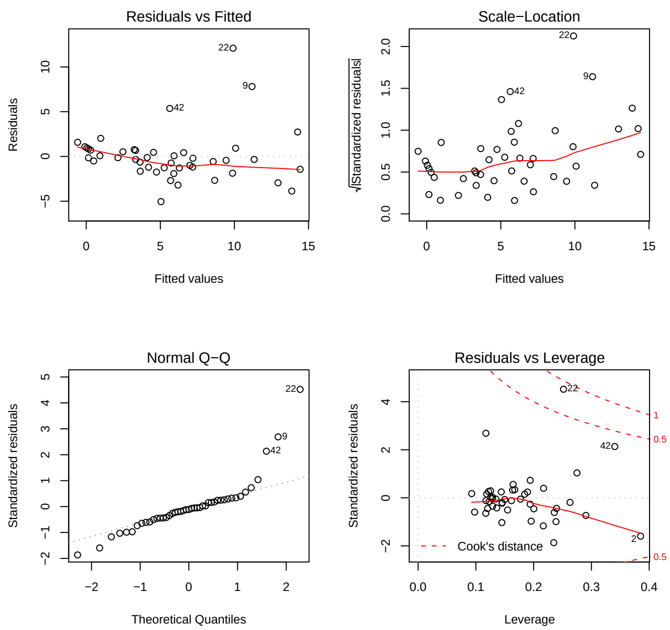
Only works for nested models, i.e. the simpler model can be obtained by restricting some parameters of the richer model to certain values or equations. In our case: "all week summands are equal".

```
> 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(week)
  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.

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

```
plot(modell)
```



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?

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

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 `drop1` 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".

Example: Success of different therapies

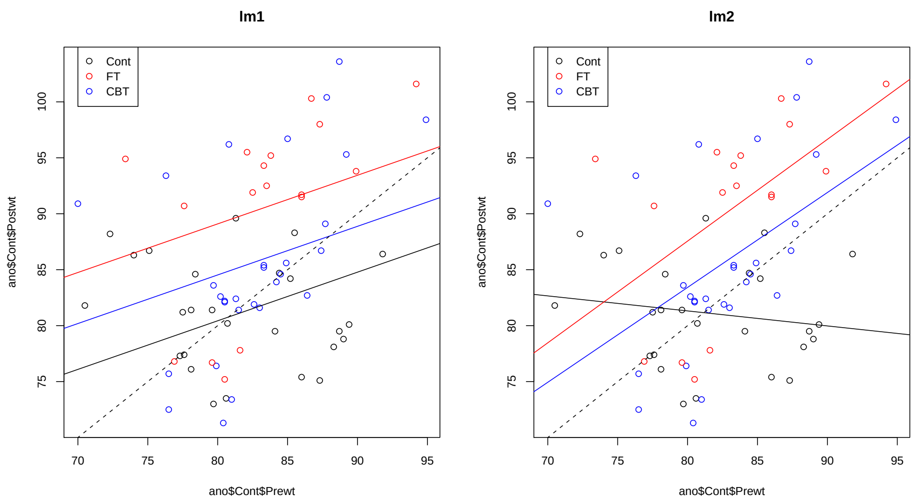
For young anorexia patients the effect of family therapy (FT) and cognitive behavioral therapy (CBT) is compared to a control group (Cont) by comparing the weight before (Prewt) and after (Postwt) the treatment (Treat).

References

[HD+93] 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 <- 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
```

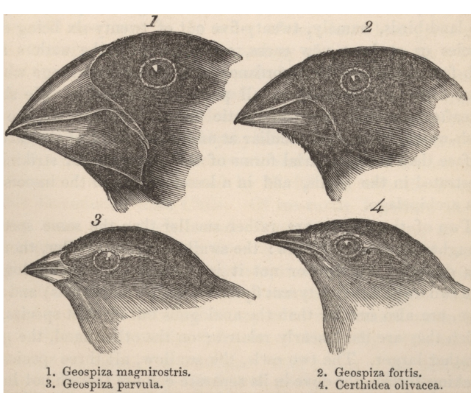
result: the more complex model fits significantly better than the nested model.

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.

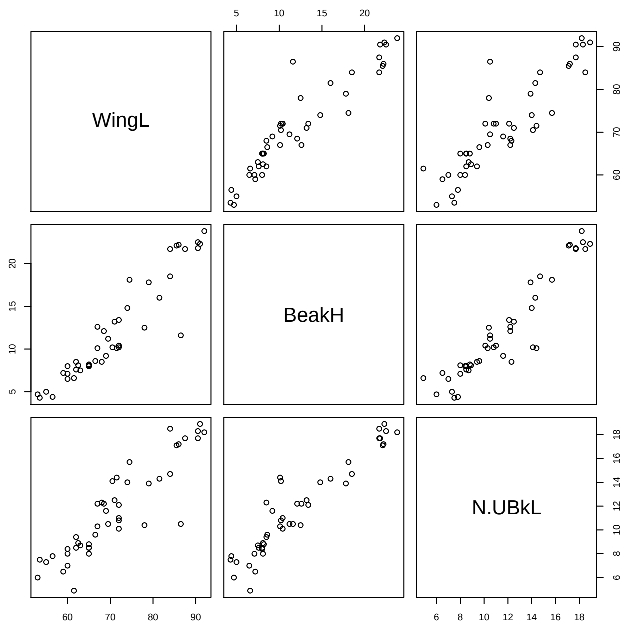
6 Cross validation and AIC

Example: Beak sizes and winglengths in Darwin finches



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.



Shall we account only for beak heights, only for beak lengths or for both?

```
> 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?

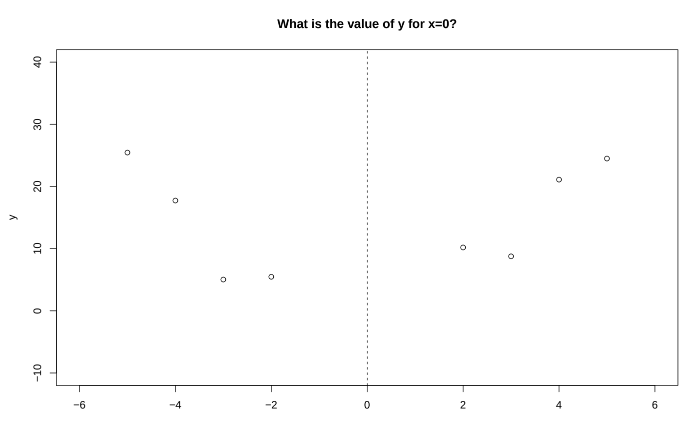
In the model modHL (with beak length and height) the influence of beak length is not significant.

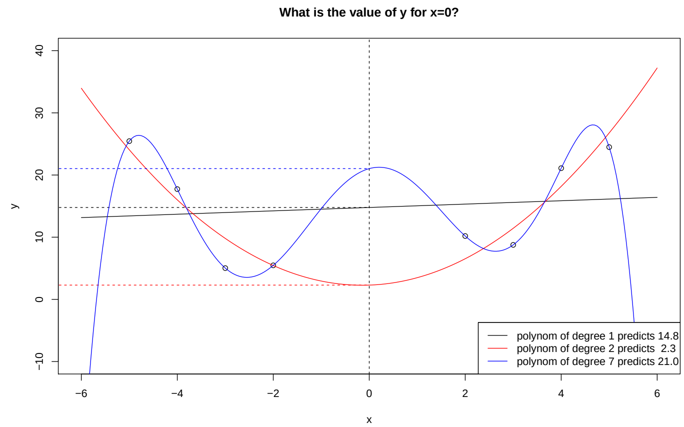
We can not draw conclusion from non-significance. Beak length could still improve the prediction.

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.

Overfitting





`lm(y~poly(x,2))` *is the same as* `lm(y~x+I(x^2))` `lm(y~poly(x,7)))`

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*).

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
$d =$ (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:

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

Bayesian Information Criterion:

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

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. (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.

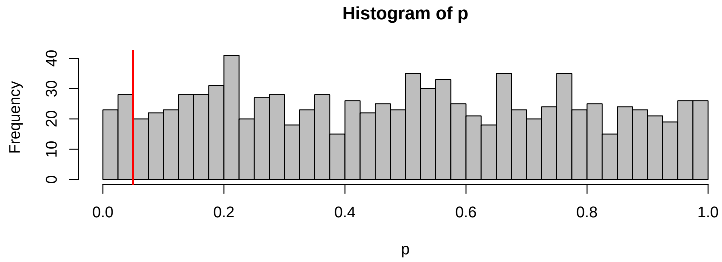
	height	length	height and length
σ(Residuals)	3.83	4.78	3.79
$d =$ (Number of parameters)	2	2	3
$\sigma(\text{Residuals})\cdot\sqrt{\frac{n-1}{n-d}}$	3.87	4.84	3.88
cross validation.	26.56	33.34	26.68
AIC	259.0	279.5	260.1
BIC	264.4	285.0	267.4

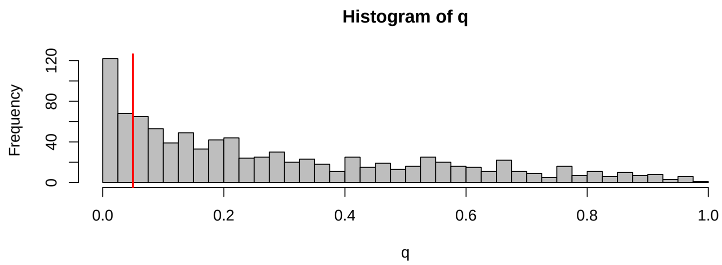
It seems best to use only the beak height.

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.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]]
}
```





Safe 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 residuals against variables. If this uncovers dependencies, they should be added to the model.

Example: Daphnia

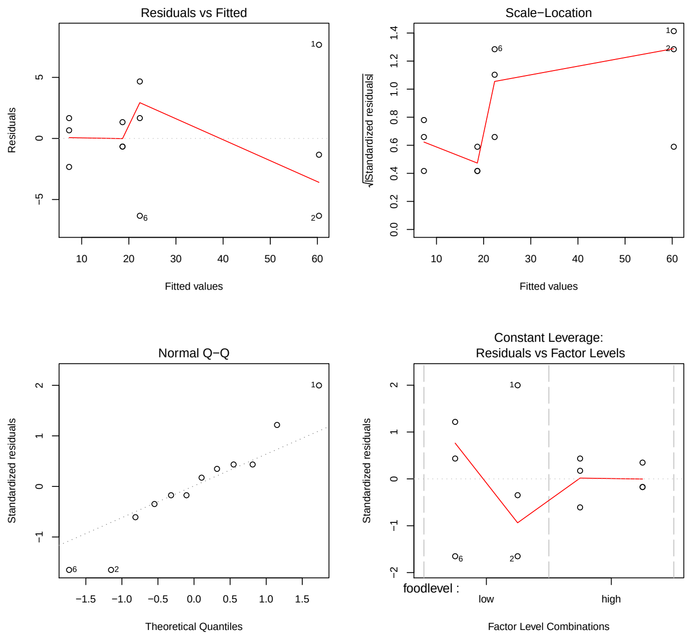
Question: Is there a difference between *Daphnia magna* and *Daphnia galeata* in their reaction on food supply?

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 ***
foodlevellow            -15.00  3.837  -3.909  0.00449 **
speciesmagna             38.00  3.837   9.904 9.12e-06 ***
foodlevellow:speciesmagna -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
```

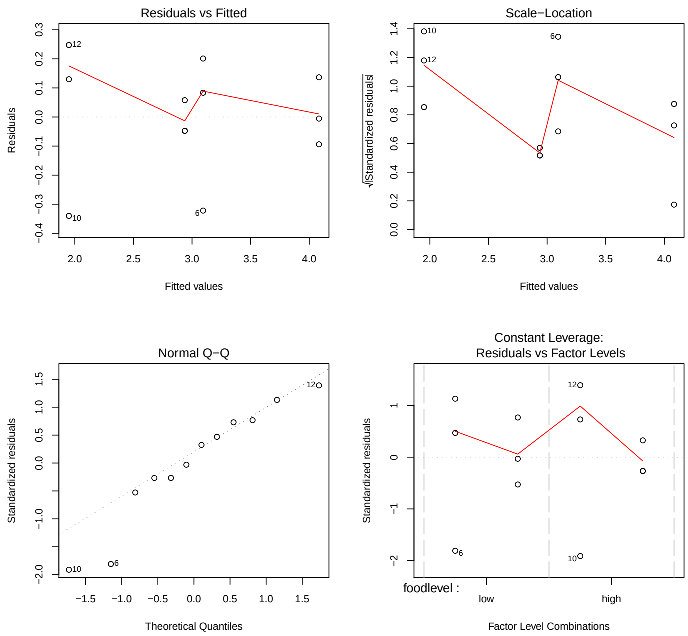
Result: the more complex model, in which different species react differently to low food level, fits significantly better.

But does it fit well enough...?




```
> 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
```



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.

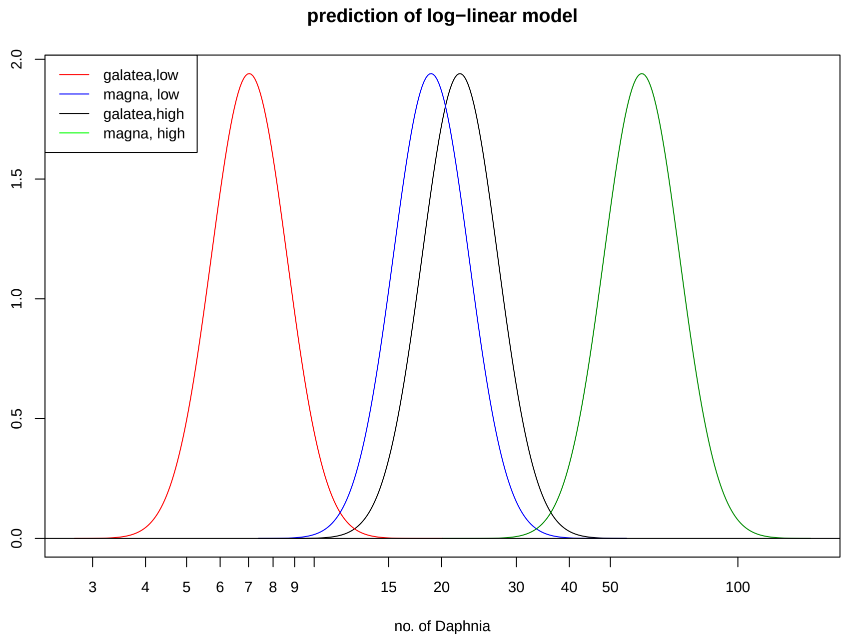
The interaction in model mod4 is not only non-significant, the model mod3 without interaction also has the better AIC values.

So we favor mod3:

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

By applying the e function we obtain:

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



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.

7 Extensions of linear models

Extensions of linear models

multiple linear model: models as we discussed, with more than one explanatory variable

multivariate linear model: the response variable y_i is multi-dimensional. That is, y consists of two or more columns that my be correlated

General linear model: the errors ε_i can be correlated. They still have $\mathbb{E}(\varepsilon_i) = 0$ but even the assumption of normality can be dropped.

Generalized linear model (GLM): The response variable y_i are not normally distributed; possible distributions are Poisson, binomial (e.g. logistic regression) or gamma. There may be no ε_i.

Linear mixed models: The coefficients of one or more factor variables (that typically have many classes) are assumed to be normally distributed.

Some of the things you should be able to explain

- interpretation of interaction terms
- how to specify all assumtions of multiple linear models ...
  - ...in precise mathematical terms
  - ...in R notation
  - and how to translate these notations into each other

- graphical methods to check model assumptions
- meaning of Anova p-values in different kinds of R output
- overfitting and how to avoid it
- cross validation, AIC, BIC and how to apply them
- connection betweed standard deviation of residuals and of ϵ_i
- Items listed on page 16