# Linear regression

## Statistical modelling

Gilles Guillot

gigu@dtu.dk

September 17, 2013

# Example

Concentration of DDT (a toxic chemical) in 15 pike fish as a function of fish age....

See file `pike_data.txt` in `data` folder.

# Various questions

- Does concentration increase with age? How much?
  *Parameter estimation (aka inference)*
- How much confidence should we place in the answer?
  *Testing significance, model checking*
- What is the average concentration of a 3.5 or 8 year old fish?
  *Prediction*

# Empirical covariance and correlation I

**Definition: Empirical covariance**

$$Cov(x, y) = \frac{1}{n} \sum_i (x_i - \bar{x})(y_i - \bar{y})$$

Tends to be large when $x_i$ and $y_i$ are large simultaneously. Hence quantifies how much $x$ and $y$ "co-vary together".
The covariance is scale-dependent: $Cov(ax, y) = aCov(x, y)$

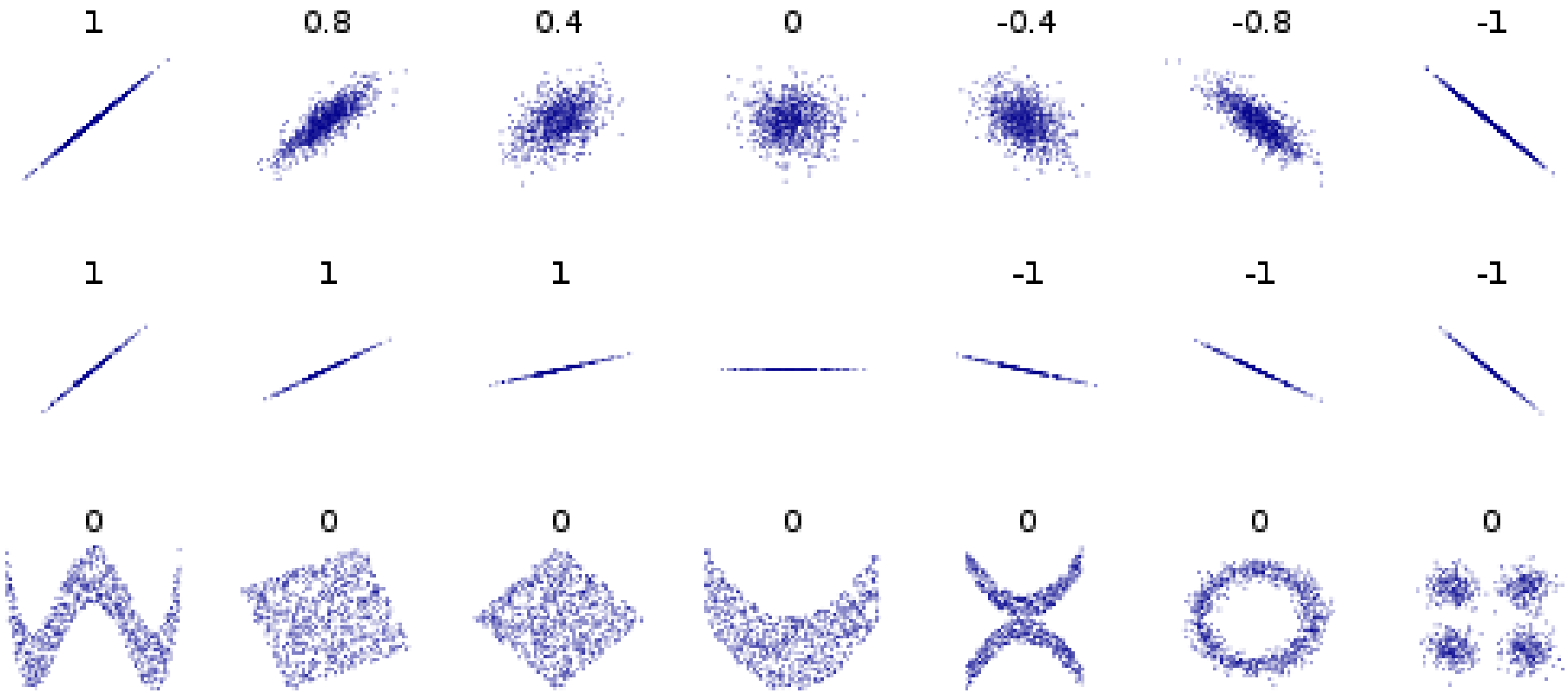
# Empirical covariance and correlation II

**Definition: Empirical correlation (aka Pearson's correlation coef.)**

$$Cor(x, y) = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2} \sqrt{\sum_i (y_i - \bar{y})^2}}$$

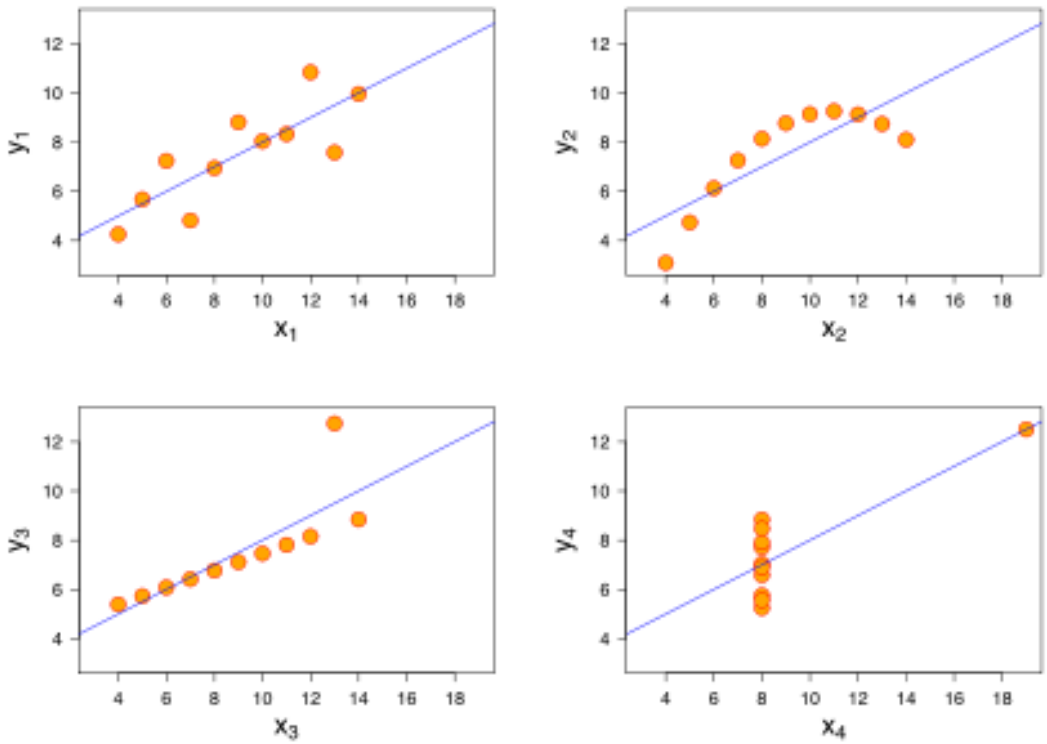
Cor. = Cov. rescaled by standard deviations

# Correlation coefficient: interpretation and pitfalls



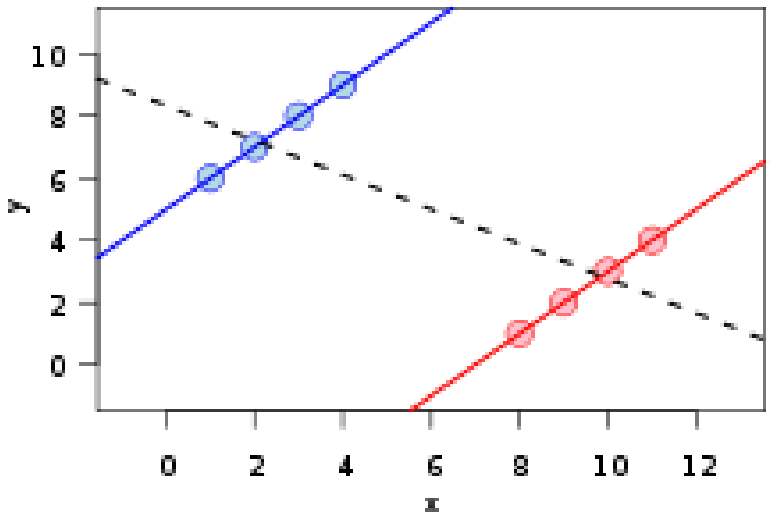

Correlation for various data patterns (reprinetd from wikipedia)

# The correlation coefficient $\rho$ does not tell the whole story




All four sets have identical $\rho \approx 0.816$, but vary considerably when graphed. Data plotted above are synthetic data made up by F. Anscombe in 1973 to illustrate the pitfalls associated with $\rho$.

# Structure effect (aka Simpson's paradox)




Quick interpretation of $\rho$ values can lead to erroneaous conclusions.
See also Simpson's effect on wikipedia for further detail and examples.

# The simple regression model

Notation: $x_1, ..., x_n$ ages of the various individiduals, $y_1, ..., y_n$ DDT concentrations

## The simple linear model

- $Y_i = ax_i + b + \varepsilon_i$
- the $x_i$s are deterministic variables
  (a somehow arbitrary modelling choice)
- $a$ and $b$ are unknown deterministic coeffficients
- the $\varepsilon_i$'s are independent realisations of a $\mathcal{N}(0, \sigma^2)$ variable
  (made for convienience, consistency with data has to be checked, see later)
- There are three parameters in this problem: $(a, b, \sigma) = \theta$

# Least square error estimation

- For arbitrary values $a$ and $b$, $e_i = y_i - ax_i - b$ measures the error made by the linear model on obs. $i$
- A good model should yield low errors on all observations

## Definition: the Least Square estimator

The unknown parameters $(a, b)$ are estimated as $\widehat{(a, b)}_{LS}$ that jointly minimize $\sum_i (y_i - ax_i - b)^2$.

In math style:

$$\widehat{(a, b)}_{LS} = Argmin_{a,b} \sum_i (y_i - ax_i - b)^2$$

# Connection with lecture on *Statistical Estimation* (lecture 1)

- The vector $\widehat{(a,b)}_{LS}$ is an *estimate* of $(a,b)$
- The procedure Data $\longrightarrow \widehat{(a,b)}_{LS}$, i.e. the generic process associating an estimate to a dataset is an *estimator*
- This procedure or function is deterministic in the sense that the same dataset will yield the same estimate
- In the framework of this course, Data $= (x_i, y_i)_{i=1,\ldots,n}$ with $Y_i = ax_i + b + \varepsilon_i$ and where $\varepsilon_i$ is a random variable Data are random therefore $\widehat{(a,b)}_{LS}$ should be seen as random.

# Remarks and questions on the Least Square principle

- What if we attempt to minimize $\sum |y_i - ax_i - b|$ ?
- What if we swap $x$ and $y$?
- What if the data points are approximately located on a circle?
- Why the squared error?

# Explicit expression of $\widehat{(a,b)}_{LS}$

$$\widehat{(a,b)}_{LS} = Argmin_{a,b} \sum_i (y_i - ax_i - b)^2$$

- $SSE(a,b) = \sum_i (y_i - ax_i - b)^2$ is a second order polynomial in $a$ and $b$
- Solving $\frac{\partial}{\partial a} SSE(a,b) = 0$ and $\frac{\partial}{\partial b} SSE(a,b) = 0$ yields:

Expression of MSE estimates:

$$\hat{a}_{LS} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}$$

and

$$\hat{b}_{LS} = \bar{y} - \hat{a}_{LS}\bar{x}$$

# Computational detail I

Zero-ing the partial derivatives in $a$ and $b$ we get

$$\frac{\partial}{\partial a} SSE(a,b) = -2 \sum_i x_i (y_i - a x_i - b) = 0 \quad (1)$$

$$\frac{\partial}{\partial b} SSE(a,b) = -2 \sum_i (y_i - a x_i - b) = 0 \quad (2)$$

Hence

$$a \sum_i x_i^2 + b \sum x_i = \sum_i x_i y_i \quad (3)$$

$$a \sum_i x_i + nb = \sum_i y_i \quad (4)$$

We have a linear system in $a$ and $b$.

# Computational detail II

A subsitution yields:

$$\hat{a}_{LS} = \frac{n \sum_i x_i y_i - \sum x_i \sum y_i}{n \sum x_i^2 - (\sum_i x_i)^2} \tag{5}$$

and

$$\hat{b}_{LS} = \frac{1/n \sum_i x_i y_i - \bar{x}\bar{y}}{1/n \sum x_i^2 - \bar{x}^2} \tag{6}$$

□

# Estimating the variance $\sigma^2$ of the residuals I

The LSE does not provide an estimate of $\sigma^2$.

A reasonable idea could be to define $\hat{\varepsilon}_i = y_i - \hat{a}_{LS}x + \hat{b}_{LS}$ and estimate $\sigma^2$ as $\frac{1}{n}\sum_i \hat{\varepsilon}_i^2$.


This would lead to a biased estimator


# Estimating the variance $\sigma^2$ of the residuals II

Unbiased estimate of $\sigma^2$:

We define

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_i \hat{\varepsilon}_i^2$$

with $\hat{\varepsilon}_i = y_i - \hat{a}_{LS}x_i + \hat{b}_{LS}$

It is an unbiased estimator of $\sigma^2$, i.e. $E[\hat{\sigma}^2] = \sigma^2$

# Connection to maximum likelihood estimation I

Remember: in the Linear Model, the $x_i$'s are considered deterministic.

And our model says: $Y_i = ax_i + b + \varepsilon_i$ and $(\varepsilon_i)_{i=1,\dots,n} \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, \sigma^2)$.
(i.i.d stands for independent and identically distributed)

The two lines above can be re-written equivalently as

$$(Y_i)_{i=1,\dots,n} \overset{\text{indep.}}{\sim} \mathcal{N}(ax_i + b, \sigma^2)$$

# Connection to maximum likelihood estimation II

The density of probability of $Y_i$ is Gaussian with mean $\mu_i = ax_i + b$ and variance $\sigma^2$:

$$f_{Y_i}(y) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{y - ax_i - b}{\sigma}\right)^2\right]$$

The likelihood in this problem is

$$L(y_1, ..., y_n; a, b, \sigma) = \prod_{i=1}^{n} f_{Y_i}(y_i) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{y_i - ax_i - b}{\sigma}\right)^2\right]$$

# Connection to maximum likelihood estimation III

The log-likelihood (up to an additive constant) is:

$$l(a,b,\sigma) = \ln L(a,b,\sigma) = -n\ln\sigma - 1/2\sum_i \left(\frac{y_i - ax_i - b}{\sigma}\right)^2$$

The MLE can be estimated by zero-ing $\frac{\partial}{\partial a}l(a,b,\sigma)$ , $\frac{\partial}{\partial b}l(a,b,\sigma)$ and $\frac{\partial}{\partial \sigma}l(a,b,\sigma)$:

$$\frac{\partial}{\partial a}l(a,b,\sigma) = \frac{1}{\sigma^2}\sum_i x_i(y_i - ax_i - b) = 0 \quad (7)$$

$$\frac{\partial}{\partial b}l(a,b,\sigma) = \frac{1}{\sigma^2}\sum_i (y_i - ax_i - b) = 0 \quad (8)$$

$$\frac{\partial}{\partial \sigma}l(a,b,\sigma) = -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_i (y_i - ax_i - b)^2 = 0 \quad (9)$$

# Connection to maximum likelihood estimation IV

In Eq. (7-8), we recognize the expressions in the LS estimator.

Hence

$\hat{a}_{ML} = \hat{a}_{LS}$ and $\hat{b}_{ML} = \hat{b}_{LS}$

Plugging $\hat{a}_{ML}$ and $\hat{b}_{ML}$ in Eq. (9) yields:

$$\hat{\sigma}^2_{ML} = \frac{1}{n}\sum_i (y_i - \hat{a}_{ML}x_i - \hat{b}_{ML})^2$$

Which is biased...

# Geometric interpretation of linear regression

- $\boldsymbol{x} = (x_1, ..., x_n)$ and $\boldsymbol{y} = (y_1, ..., y_n)$ are vectors in $\mathbb{R}^n$
- the values $ax_i + b$ can be seen as the entries of the vector $a\boldsymbol{x} + b\mathbf{1}$ in $\mathbb{R}^n$
- $\sum_i (y_i - ax_i - b)^2$ is the square of the norm of $\boldsymbol{y} - a\boldsymbol{x} - b\mathbf{1}$
- $\hat{a}\boldsymbol{x} + \hat{b}\mathbf{1}$ is the projection of $\boldsymbol{y}$ on $Span(\mathbf{1}, \boldsymbol{x})$
- $Cor(\boldsymbol{x}, \boldsymbol{y})$ is the cosine of the angle formed by $\boldsymbol{x}$ and $\boldsymbol{y}$ in $\mathbb{R}^n$

# Goodness of fit

The quality of the model can be assessed by the coefficent of determination:

**Coefficient of determination:**

$$R^2 \hat{=} 1 - \frac{SS_{err}}{SS_{tot}}$$

with $SS_{tot} = \sum_i (y_i - \bar{y})^2$ and $SS_{err} = \sum_i (\hat{y}_i - y_i)^2$

- $0 \leq R^2 \leq 1$
- "Good" model $\Leftrightarrow$ low $SS_{err}$ $\Leftrightarrow$ high $R^2$
- If the regression model includes an intercept, then $R^2 = \rho^2$

# Testing $H_0 : a = a_0$ I

We are often interested in assessing if $a$ is significantly different from a particular value $a_0$.

Often $H_0 : a = 0$ which corresponds to the absence of dependence between $x$ and $y$.

The question is: should the difference between $\hat{a}$ and $a_0$ be considered large enough to reject $H_0$?

# Testing $H_0 : a = a_0$ II

## The Student test

Under the assumptions given here then

$$T = \frac{\hat{a} - a_0}{\widehat{sd}(\hat{a})} \sim St_{n-2}$$

$H_0$ is rejected at level $\alpha$ if $|t|$ is larger than the quantile with probability $1 - \alpha/2$ of a $St_{n-2}$ distribution

# Checking model assumptions

The $\varepsilon_i$'s are assumed to be i.i.d. If $\widehat{(a, b)}$ estimates $(a, b)$ correctly, the $\hat{\varepsilon}_i$'s should be close to i.i.d.

A plot of $(\hat{\varepsilon}_i)_{i=1,\ldots,n}$ against $(x_i)_{i=1,\ldots,n}$ should not display any pattern.

- Visual check of residuals
- Visual check of standardised residuals `plot(lm(...))` in R

# Prediction

What value $y_{new}$ should be expected for an extra individual with observed explanatory variable $x_{new}$?

**Definition: prediction**

$$y_{new} = \hat{a}x_{new} + \hat{b}$$

Straightforward in R, see also use of the generic function `predict`.

# Parameter estimation in practice with R

Assuming data objects are named `x` and `y` in your R session

```
# fit a linear model and store the (long) output list
res.lm = lm(formula = y ~ x)
# extract estimated coef
res.coef = coefficients(res.lm)
```

# R code to fit a linear regression on the pike data

```
pike=read.table("http://www2.imm.dtu.dk/courses/02418/lecture3_simpl
    header=TRUE)

## fitting regression line
lm.res = lm(formula=DDT~1+Age,data=pike)
```

# Pike data analysis output in R

```
## display  the R object lm.res
summary(lm.res)
Call:
lm(formula = DDT ~ 1 + Age, data = pike)

Residuals:
     Min       1Q   Median       3Q      Max
-0.24133 -0.10500  0.01133  0.08300  0.30733

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.23533    0.11269  -2.088    0.057 .
Age          0.17133    0.02656   6.450 2.16e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1455 on 13 degrees of freedom
Multiple R-squared:  0.7619, Adjusted R-squared:  0.7436
F-statistic: 41.61 on 1 and 13 DF,  p-value: 2.165e-05
```

# R code

Link to the R script used in this lecture (and more)

.

# Exercise

- Bingham and Fry, exercise 1.3 p. 29.
  Data file here
  Hints:
  - download with `download.file(url="", destfile="./running.txt")`
  - read in R with `read.table("./running.txt",header=TRUE)`
  - model fit on log data can be obtained by `lm(log(y) ~ log(x))`
  - use R function `confint` for confidence interval
- Sheather, exercise 1 p. 38
  Data are available on the book web site as file playbill.csv
  Hints:
  For testing $\beta_0 = 10000$, use function `test.coef` available from here.

# References

Suggested reading Chapter Regression and correlation, *Introductory statistics with R* , P. Dalgaard, Series Statistics and Computing, Springer, 2008.