Statistical modelling: theory and practice

Linear regression

Gilles Guillot

gigu@dtu.dk

August 27, 2013

Example

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

....

Various questions

Various questions

- Does concentration increase with age? How much?

Various questions

- Does concentration increase with age? How much?
  Parameter estimation (aka inference)

Various questions

- Does concentration increase with age? How much?
  Parameter estimation (aka inference)
- How much confidence should we place in the answer?

Various questions

- Does concentration increase with age? How much?
  Parameter estimation (aka inference)
- How much confidence should we place in the answer?
  Testing significance

Various questions

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

Various questions

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

The simple regression model

The simple regression model

Notation: $x_1, ..., x_n$ age of the various individiduals, $y_1, ..., y_n$ DDT concentration

The simple regression model

Notation: $x_1, ..., x_n$ age of the various individiduals, $y_1, ..., y_n$ DDT concentration

The simple linear model

The simple regression model

Notation: $x_1, ..., x_n$ age of the various individiduals, $y_1, ..., y_n$ DDT concentration

The simple linear model

- $y_i = ax_i + b + \varepsilon_i$

The simple regression model

Notation: $x_1, ..., x_n$ age of the various individiduals, $y_1, ..., y_n$ DDT concentration

The simple linear model

- $y_i = ax_i + b + \varepsilon_i$
- the x_is are deterministic variables

The simple regression model

Notation: $x_1, ..., x_n$ age of the various individiduals, $y_1, ..., y_n$ DDT concentration

The simple linear model

- $y_i = ax_i + b + \varepsilon_i$
- the x_is are deterministic variables
  (a somehow arbitrary modelling choice)

The simple regression model

Notation: $x_1, ..., x_n$ age of the various individiduals, $y_1, ..., y_n$ DDT concentration

The simple linear model

- $y_i = ax_i + b + \varepsilon_i$
- the x_is are deterministic variables
  (a somehow arbitrary modelling choice)
- a and b are unknown deterministic coeffficients

The simple regression model

Notation: $x_1, ..., x_n$ age of the various individiduals, $y_1, ..., y_n$ DDT concentration

The simple linear model

- $y_i = ax_i + b + \varepsilon_i$
- the x_is are deterministic variables
  (a somehow arbitrary modelling choice)
- a and b are unknown deterministic coeffficients
- the ε_is are independent realisations of a $\mathcal{N}(0, \sigma^2)$ variable

The simple regression model

Notation: $x_1, ..., x_n$ age of the various individiduals, $y_1, ..., y_n$ DDT concentration

The simple linear model

- $y_i = ax_i + b + \varepsilon_i$
- the x_is are deterministic variables
  (a somehow arbitrary modelling choice)
- a and b are unknown deterministic coeffficients
- the ε_is are independent realisations of a $\mathcal{N}(0, \sigma^2)$ variable
  (made for convienience, consistency with data has to be checked, see later)

The simple regression model

Notation: $x_1, ..., x_n$ age of the various individiduals, $y_1, ..., y_n$ DDT concentration

The simple linear model

- $y_i = ax_i + b + \varepsilon_i$
- the x_is are deterministic variables
  (a somehow arbitrary modelling choice)
- a and b are unknown deterministic coeffficients
- the ε_is 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

Mininum sum of squared errors estimation

Mininum sum of squared errors 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

Mininum sum of squared errors 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

The MSSE principle

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

Mininum sum of squared errors 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

The MSSE principle

The unknown parameters (a, b) are estimated as $(\hat{a}, \hat{b})$ that jointly minimizes $\sum_i (y_i - ax_i - b)^2$.
In math style:

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

Remarks and questions on the MSSE principle

Remarks and questions on the MSSE principle

- The vector $(\hat{a}, \hat{b})$ is an *estimate* of (a, b)

Remarks and questions on the MSSE principle

- The vector $(\hat{a}, \hat{b})$ is an *estimate* of (a, b)
- The procedure Data $\longrightarrow (\hat{a}, \hat{b})$, i.e. the generic process associating an estimate to a dataset is an *estimator*

Remarks and questions on the MSSE principle

- The vector $(\hat{a}, \hat{b})$ is an *estimate* of (a, b)
- The procedure Data $\longrightarrow (\hat{a}, \hat{b})$, 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

Remarks and questions on the MSSE principle

- The vector $(\hat{a}, \hat{b})$ is an *estimate* of (a, b)
- The procedure Data $\longrightarrow (\hat{a}, \hat{b})$, 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 $= (x_i, y_i)_{i=1,\ldots,n}$

Remarks and questions on the MSSE principle

- The vector $(\hat{a}, \hat{b})$ is an *estimate* of (a, b)
- The procedure Data $\longrightarrow (\hat{a}, \hat{b})$, 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 ε_i is a random variable

Remarks and questions on the MSSE principle

- The vector $(\hat{a}, \hat{b})$ is an *estimate* of (a, b)
- The procedure Data $\longrightarrow (\hat{a}, \hat{b})$, 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 ε_i is a random variable Data are random therefore $(\hat{a}, \hat{b})$ sould be seen as random

Remarks and questions on the MSSE principle

- The vector $(\hat{a}, \hat{b})$ is an *estimate* of (a, b)
- The procedure Data $\longrightarrow (\hat{a}, \hat{b})$, 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 ε_i is a random variable Data are random therefore $(\hat{a}, \hat{b})$ sould be seen as random

Miscellaneous questions

Remarks and questions on the MSSE principle

- The vector $(\hat{a}, \hat{b})$ is an *estimate* of (a, b)
- The procedure Data $\longrightarrow (\hat{a}, \hat{b})$, 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 ε_i is a random variable Data are random therefore $(\hat{a}, \hat{b})$ sould be seen as random

Miscellaneous questions

- Why the squared error?
  What if we attempt to minimize $\sum |y_i - ax_i - b|$?

Remarks and questions on the MSSE principle

- The vector $(\hat{a}, \hat{b})$ is an *estimate* of (a, b)
- The procedure Data $\longrightarrow (\hat{a}, \hat{b})$, 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 ε_i is a random variable Data are random therefore $(\hat{a}, \hat{b})$ sould be seen as random

Miscellaneous questions

- Why the squared error?
  What if we attempt to minimize $\sum |y_i - ax_i - b|$?
- What if we swap x and y?

Remarks and questions on the MSSE principle

- The vector $(\hat{a}, \hat{b})$ is an *estimate* of (a, b)
- The procedure Data $\longrightarrow (\hat{a}, \hat{b})$, 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 ε_i is a random variable Data are random therefore $(\hat{a}, \hat{b})$ sould be seen as random

Miscellaneous questions

- Why the squared error? 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?

Empirical covariance and correlation

Definition: Empirical covariance

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

Empirical covariance and correlation

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".

Empirical covariance and correlation

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

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: $Cov(ax, y) = aCov(x, y)$

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

Empirical covariance and correlation

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: $Cov(ax, y) = aCov(x, y)$

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

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

Empirical covariance and correlation

Definition: Empirical covariance

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

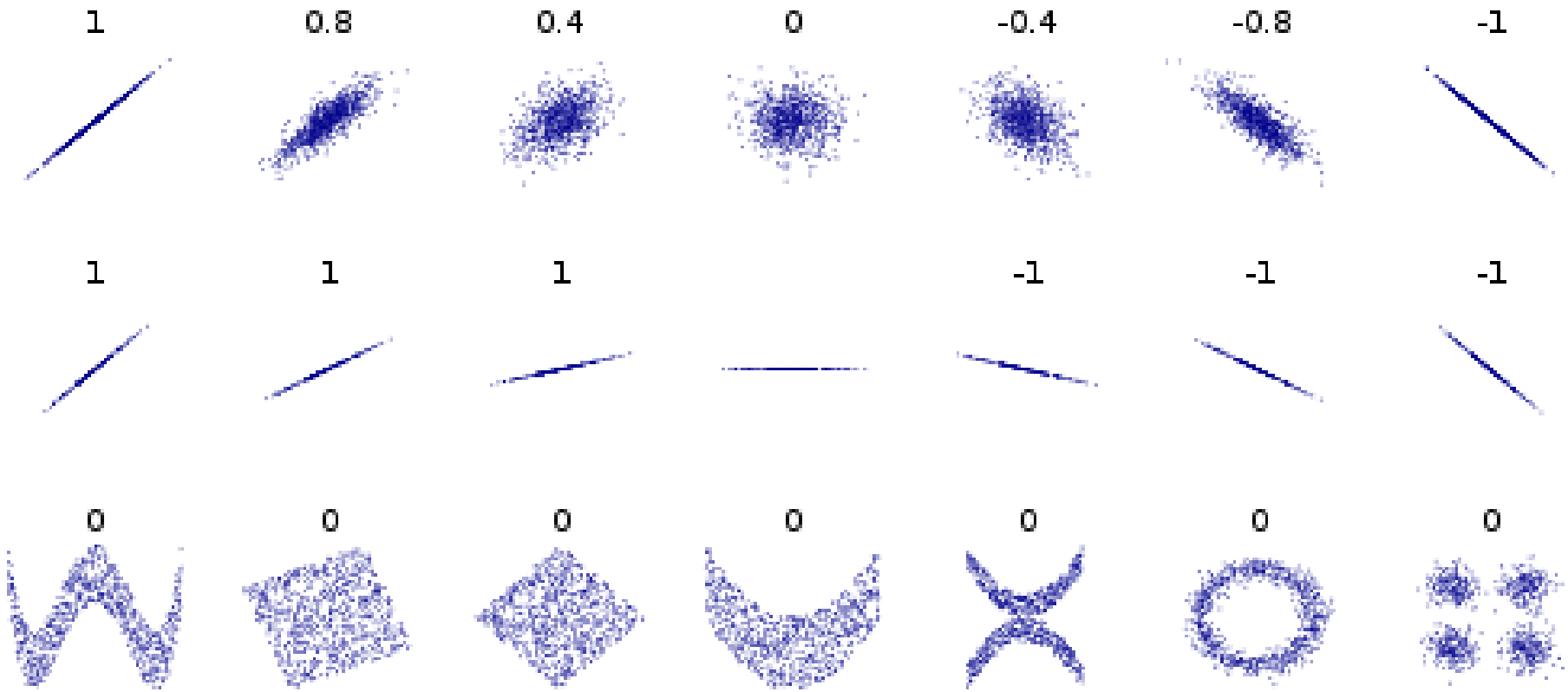
Tends to be large when x_i and y_i are large simultaneously. Hence quantifies how much x and y "co-vary together".
The covariance is scale-dependent: $Cov(ax, y) = aCov(x, y)$

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

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

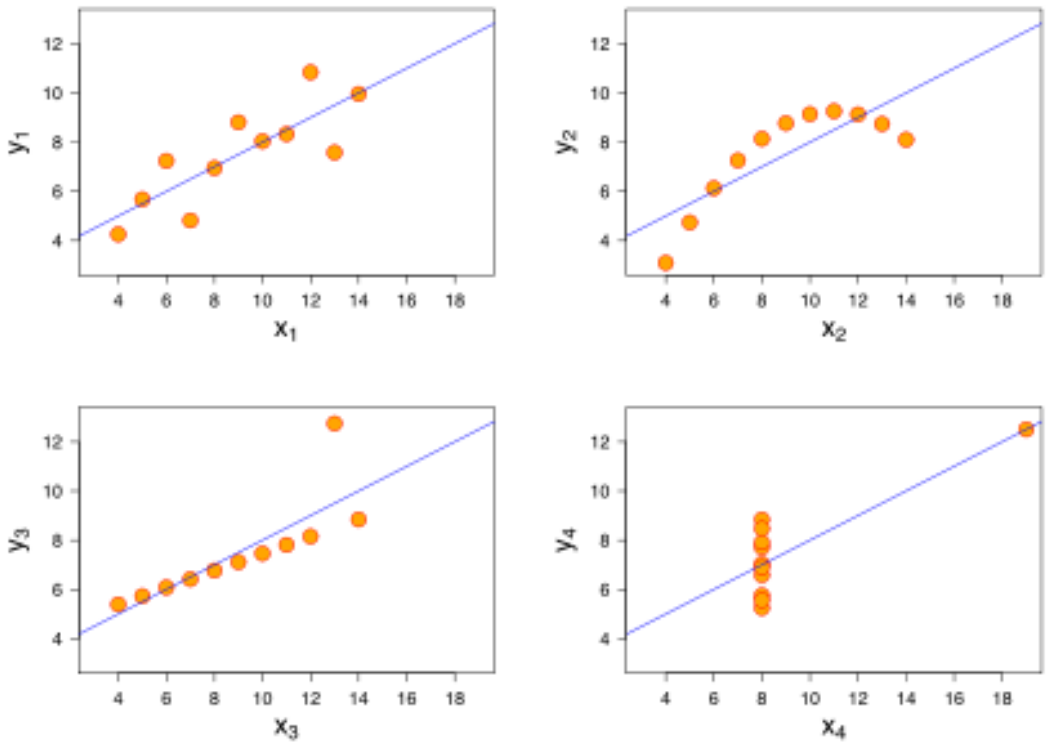
Cor. = Cov. rescaled by standard deviations

Correlation coefficient: interpretation and pitfalls



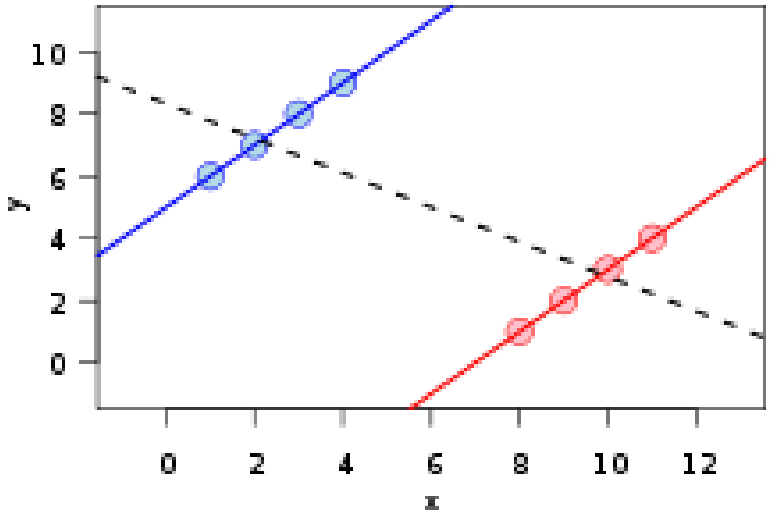
Correlation for various data patterns (reprinetd from wikipedia)

The correlation coefficient ρ 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 ρ.

Structure effect (aka Simpson's paradox)



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

Finding explicitly $(\hat{a}, \hat{b})$

Finding explicitly $(\hat{a}, \hat{b})$

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

Finding explicitly $(\hat{a}, \hat{b})$

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

Finding explicitly $(\hat{a}, \hat{b})$

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

- $SSE(a, b) = \sum_i (y_i - ax_i - b)^2$ is a second order polynomial in a and b

Finding explicitly $(\hat{a}, \hat{b})$

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

- $SSE(a,b) = \sum_i (y_i - ax_i - b)^2$ 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:

Finding explicitly $(\hat{a}, \hat{b})$

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

- $SSE(a,b) = \sum_i (y_i - ax_i - b)^2$ 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:

Finding explicitly $(\hat{a}, \hat{b})$

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

- $SSE(a,b) = \sum_i (y_i - ax_i - b)^2$ 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:

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

Finding explicitly $(\hat{a}, \hat{b})$

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

- $SSE(a,b) = \sum_i (y_i - ax_i - b)^2$ 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:

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

and

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

Estimating the variance σ^2 of residuals

Estimating the variance σ^2 of residuals

Unbiased estimate of σ^2:

Estimating the variance σ^2 of residuals

Unbiased estimate of σ^2:

We define

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

with $\hat{\varepsilon}_i = y_i - \hat{a}x + \hat{b}$

Estimating the variance σ^2 of residuals

Unbiased estimate of σ^2:

We define

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

with $\hat{\varepsilon}_i = y_i - \hat{a}x + \hat{b}$

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

Goodness of fit

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

Goodness of fit

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

Coefficient of determination:

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

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

- $0 \leq R^2 \leq 1$

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

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 $R^2 = \rho^2$

Parameter estimation in practice with R

Parameter estimation in practice with R

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

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)
```

Geometric interpretation of linear regression

Geometric interpretation of linear regression

- $x = (x_1, ..., x_n)$ and $y = (y_1, ..., y_n)$ are vectors in $\mathbb{R}^n$

Geometric interpretation of linear regression

- $x = (x_1, ..., x_n)$ and $y = (y_1, ..., y_n)$ are vectors in $\mathbb{R}^n$
- the values $ax_i + b$ can be seen as the entries of the vector $ax + b1$ in $\mathbb{R}^n$

Geometric interpretation of linear regression

- $x = (x_1, ..., x_n)$ and $y = (y_1, ..., y_n)$ are vectors in $\mathbb{R}^n$
- the values $ax_i + b$ can be seen as the entries of the vector $ax + b1$ in $\mathbb{R}^n$
- $\sum_i (y_i - ax_i - b)^2$ is the square of the norm of $y - ax - b1$

Geometric interpretation of linear regression

- $x = (x_1, ..., x_n)$ and $y = (y_1, ..., y_n)$ are vectors in $\mathbb{R}^n$
- the values $ax_i + b$ can be seen as the entries of the vector $ax + b1$ in $\mathbb{R}^n$
- $\sum_i (y_i - ax_i - b)^2$ is the square of the norm of $y - ax - b1$
- $\hat{a}x + \hat{b}1$ is the projection of y on $Span(1, x)$

Geometric interpretation of linear regression

- $x = (x_1, ..., x_n)$ and $y = (y_1, ..., y_n)$ are vectors in $\mathbb{R}^n$
- the values $ax_i + b$ can be seen as the entries of the vector $ax + b1$ in $\mathbb{R}^n$
- $\sum_i (y_i - ax_i - b)^2$ is the square of the norm of $y - ax - b1$
- $\hat{a}x + \hat{b}1$ is the projection of y on $Span(1, x)$
- $cor(x, y)$ is the cosine of the angle formed by x and y in $\mathbb{R}^n$

Testing $H_0 : a = a_0$

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

Testing $H_0 : a = a_0$

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.

Testing $H_0 : a = a_0$

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 $H_0 : a = a_0$

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?

Student test

Testing $H_0 : a = a_0$

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?

Student test

Under the assumptions given earlier here then

Testing $H_0 : a = a_0$

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?

Student test

Under the assumptions given earlier here then

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

Testing $H_0 : a = a_0$

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?

Student test

Under the assumptions given earlier here then

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

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

Checking model assumptions

Checking model assumptions

- Visual check of residuals

Checking model assumptions

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

Prediction

Prediction

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

Prediction

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

Prediction

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

Prediction

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

Prediction

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

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

The pike data analysis in R

Link to the R script used in this lecture.

Example I: The Pine Processionary Caterpillar

Harmfulness still debated but

- Weaken the trees
- Cause strong irritating reactions to humans and pets

What environmental factors favour their development?

What environmental factors favour their development?

Experimental design: 33 forest plots (of 500 sq meters each) in the South of France

What environmental factors favour their development?

Experimental design: 33 forest plots (of 500 sq meters each) in the South of France

Response variable: average number of nests per tree in the plot (last column of the file)

What environmental factors favour their development?

Experimental design: 33 forest plots (of 500 sq meters each) in the South of France

Response variable: average number of nests per tree in the plot (last column of the file)

Explanatory variables (col. 1-10):

What environmental factors favour their development?

Experimental design: 33 forest plots (of 500 sq meters each) in the South of France
Response variable: average number of nests per tree in the plot (last column of the file)
Explanatory variables (col. 1-10):

- Altitude (meters)

What environmental factors favour their development?

Experimental design: 33 forest plots (of 500 sq meters each) in the South of France
Response variable: average number of nests per tree in the plot (last column of the file)
Explanatory variables (col. 1-10):

- Altitude (meters)
- Slope (degrees)

What environmental factors favour their development?

Experimental design: 33 forest plots (of 500 sq meters each) in the South of France
Response variable: average number of nests per tree in the plot (last column of the file)
Explanatory variables (col. 1-10):

- Altitude (meters)
- Slope (degrees)
- Height of the tree at the centre of the plot

What environmental factors favour their development?

Experimental design: 33 forest plots (of 500 sq meters each) in the South of France

Response variable: average number of nests per tree in the plot (last column of the file)

Explanatory variables (col. 1-10):

- Altitude (meters)
- Slope (degrees)
- Height of the tree at the centre of the plot
- Diameter of the tree at the centre of the plot

What environmental factors favour their development?

Experimental design: 33 forest plots (of 500 sq meters each) in the South of France
Response variable: average number of nests per tree in the plot (last column of the file)
Explanatory variables (col. 1-10):

- Altitude (meters)
- Slope (degrees)
- Height of the tree at the centre of the plot
- Diameter of the tree at the centre of the plot
- Index of tree density in the plot

What environmental factors favour their development?

Experimental design: 33 forest plots (of 500 sq meters each) in the South of France
Response variable: average number of nests per tree in the plot (last column of the file)
Explanatory variables (col. 1-10):

- Altitude (meters)
- Slope (degrees)
- Height of the tree at the centre of the plot
- Diameter of the tree at the centre of the plot
- Index of tree density in the plot
- Orientation (from 1 if southbound to 2)

What environmental factors favour their development?

Experimental design: 33 forest plots (of 500 sq meters each) in the South of France
Response variable: average number of nests per tree in the plot (last column of the file)
Explanatory variables (col. 1-10):

- Altitude (meters)
- Slope (degrees)
- Height of the tree at the centre of the plot
- Diameter of the tree at the centre of the plot
- Index of tree density in the plot
- Orientation (from 1 if southbound to 2)
- Height of the dominant tree (meters)

What environmental factors favour their development?

Experimental design: 33 forest plots (of 500 sq meters each) in the South of France
Response variable: average number of nests per tree in the plot (last column of the file)
Explanatory variables (col. 1-10):

- Altitude (meters)
- Slope (degrees)
- Height of the tree at the centre of the plot
- Diameter of the tree at the centre of the plot
- Index of tree density in the plot
- Orientation (from 1 if southbound to 2)
- Height of the dominant tree (meters)
- Number of vegetation strata

What environmental factors favour their development?

Experimental design: 33 forest plots (of 500 sq meters each) in the South of France
Response variable: average number of nests per tree in the plot (last column of the file)
Explanatory variables (col. 1-10):

- Altitude (meters)
- Slope (degrees)
- Height of the tree at the centre of the plot
- Diameter of the tree at the centre of the plot
- Index of tree density in the plot
- Orientation (from 1 if southbound to 2)
- Height of the dominant tree (meters)
- Number of vegetation strata
- Mixing index of vegetation cover: from 1 (not mixed) to 2 (mixed)

Questions brought up by the caterpillar data

Questions brought up by the caterpillar data

- How do local environmental conditions affect caterpillar spread?

Questions brought up by the caterpillar data

- How do local environmental conditions affect caterpillar spread?
- Where should caterpillars be expected?

Data available as text file caterpillar.dat on course web page.

Example II: relation between salary and experience

A available as text file profsalary.txt on Sheather's book web site.

Example II: relation between salary and experience

A available as text file profsalary.txt on Sheather's book web site.

- EDA and simple model fit with R...

Example II: relation between salary and experience

A available as text file profsalary.txt on Sheather's book web site.

- EDA and simple model fit with R...
- A simple regression does not capture the whole pattern

Example II: relation between salary and experience

A available as text file profsalary.txt on Sheather's book web site.

- EDA and simple model fit with R...
- A simple regression does not capture the whole pattern
- A model along the line of

$$y = a_0 + a_1 x + a_2 x + \varepsilon$$

would be more accurate

The multiple regression model I

Setting:

The multiple regression model II

Model assumptions:

- The explanatory variables x_{ij}-s are observed and non random
- The response variables y_i-s are observed and random
- y_i relates linearly to the x_{ij}-s:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} + \varepsilon_i$$

- The coefficients $\beta_0, ...\beta_p$ are deterministic (and unknown)
- The error terms are independent and indentically distributed (iid) as a $\mathcal{N}(0, \sigma_e^2)$
- The error variance σ_e^2 is deterministic (and unknown)

The multiple regression model III

Multiple regression equation in matrix form:

Denoting

- $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p)^t$
- $\boldsymbol{X} = (\mathbf{1}, X_1, ..., X_p)$
- $E = (\varepsilon_1, ..., \varepsilon_n)^t$

We have:

$$\begin{array}{ccc} Y & = & \boldsymbol{X}\boldsymbol{\beta} + E \\ (n \times 1) & & (n \times p + 1)(p + 1 \times 1)(n \times 1) \end{array}$$

In the above, $\boldsymbol{X}$ is known as the *design* matrix.

Stems from the situations where the scientist choses the x_{ij} values in view of maximising the amount of information obtained for a given cost (design of experiment or DoE).

Least square estimation I

Least square estimation II

Geometric interpretation:

- $\beta_1 \mathbf{1} + \beta_1 X_1 + ... + \beta_p X_p$ is an element of $Span(\mathbf{1}, X_1, ..., X_p)$
- $\sum_i [y_i - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip})]^2 = ||Y - \boldsymbol{X}\boldsymbol{\beta}||^2$
- $\boldsymbol{X}\hat{\boldsymbol{\beta}}$ is the orthogonal projection of Y on $Span(\mathbf{1}, X_1, ..., X_p)$

Deriving $\hat{\boldsymbol{\beta}}$ explictly: the normal equations

It can be shown that $\hat{\boldsymbol{\beta}}$ satisfies:

$$(\boldsymbol{X}^t \boldsymbol{X})\hat{\boldsymbol{\beta}} = \boldsymbol{X}^t y$$

which is known as the *normal equations*.

Least square estimation III

The LSE estimate of $\boldsymbol{\beta}$

If $(\boldsymbol{X}^t\boldsymbol{X})$ if of full rank, normal equations give:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t y$$

$\hat{\boldsymbol{\beta}}$ defined above is an *unbiased* estimate of β.

Note: the LSE estimate exists if the columns of X are linearly independent.

Least square estimation IV

LSE estimate of the residual variance σ^2:

$$S^2 = \frac{RSS}{n-p-1} = \frac{1}{n-p-1}\sum_i (\hat{y}_i - y_i)^2$$

is an unbiased estimate of σ^2.

Model fitting in R

R function `lm`:

Model fitting in R

R function `lm`:

- `lm(y ~ 1 + x1 + x2 + x3)`

Model fitting in R

R function `lm`:

- `lm(y ~ 1 + x1 + x2 + x3)`
- `lm(y ~ x1 + x2 + x3) # includes a constant term by default`

Model fitting in R

R function `lm`:

- `lm(y ~ 1 + x1 + x2 + x3)`
- `lm(y ~ x1 + x2 + x3) # includes a constant term by default`
- `lm(y ~ 0 +x1 + x2 + x3) # no constant term`

Model fitting in R

R function `lm`:

- `lm(y ~ 1 + x1 + x2 + x3)`
- `lm(y ~ x1 + x2 + x3) # includes a constant term by default`
- `lm(y ~ 0 +x1 + x2 + x3) # no constant term`
- `lm(y ~ .,data=mydata)`
  `# include all variables in the rhs from data.frame mydata`

Model fitting in R

R function `lm`:

- `lm(y ~ 1 + x1 + x2 + x3)`
- `lm(y ~ x1 + x2 + x3) # includes a constant term by default`
- `lm(y ~ 0 +x1 + x2 + x3) # no constant term`
- `lm(y ~ .,data=mydata)`
  `# include all variables in the rhs from data.frame mydata`
- `lm(y ~ .  - x2 ,data=mydata) # drop variable x2`

By default, the `lm` function does much more than just parameter estimation.

Testing global significance I

We are interested in testing whether the response y is pure noise and does not relate linearly to any of the putative explanatory variables. This is a global test of significance and answers to the question "does the model explain anything at all?"

Testing global significance II

We define

- $SST = \sum_i (y_i - \bar{y})^2$
- $SSM = \sum_i (\hat{y}_i - \bar{y})^2$
- $SSE = \sum_i (\hat{y}_i - y_i)^2$

If one of the parameters β_j is $\neq 0$, the model should "do good" with $SSM \approx SST$ and $SSE \approx 0$.

Test statistic in global significance testing

Under H_0, $F = \dfrac{SSM/p}{SSE/(n-p-1)} \sim F(p, n-p-1)$. A large value of F suggests that the model "explains something" and leads to reject H_0.

Testing global significance III

In R, the p-value for the global F test can be found at the first line of the analysis of variance table (`anova(lm(formula))`) or on the last line of `summary(lm(formula))`.

Testing individual coefficients I

If H_0 is rejected in the global test above, we want to identify wich coefficient β_j is non zero.
Mor generally, we can be interested in testing $H_0 : \beta_j = \beta_j^0$ for some specific β_j^0 values.

Testing individual coefficients II

The previous test of $\beta_j = 0$ is performed under the assumption that other parameters are non zero. It measures the improvement brought by X_j in a model containing the other variables.

The result of the test of $\beta_j = 0$ depends on which other variables are included in the model. Cf. cubic term for salaries with and without quadratic term.

Model assessment and variable selection I

Model assessment and variable selection II

Adjusted coefficient of determination R^2_{adj}:

A coefficient that includes a penalty term in p

$$R^2_{adj} = 1 - \frac{SSE/(n-p-1)}{SST/(n-1)}$$

R^2_{adj} makes it possible (in principle) to compare models of different dimensions.

Other criteria for model selection include Akaike Information Criterion (AIC) and Bayeisan Information Criterion (BIC). To be defined after chapters on likelihood and Bayesian theory.

Suggested reading

- Chapter 6: Regression and correlation, *Introductory statistics with R* , P. Dalgaard, Series Statistics and Computing, Springer, 2008.
- Sheather, chapter 5. Multiple linear regression.

Simple regression

- 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.

Caterpillar case: suggested steps I

- Load the data and give the variables some names, e.g.

```
cat.dat <- read.table("...")

colnames(cat.dat) <- c("Altitude",
                       "Slope",
                       "Nb.pines.in.area",
                       "Height.tree.center",
                       "Diameter.tree.center",
                       "Density.index",
                       "Orientation",
                       "Height.dominant.tree",
                       "Nb.vegetation.strata",
                       "Mixing.index",
                       "Nb.nests")
```

- Explore graphically the dataset (the R functions `plot`, `pairs` and `hist` are your best friends).
  You may want to print the correlation coefficients on the pairs plot with:

Caterpillar case: suggested steps II

```
panel.cor <- function(x, y, digits=2, prefix="", cex.cor, ...)
     {
          usr <- par("usr"); on.exit(par(usr))
          par(usr = c(0, 1, 0, 1))
          r <- abs(cor(x, y))
          txt <- format(c(r, 0.123456789), digits=digits)[1]
          txt <- paste(prefix, txt, sep="")
          if(missing(cex.cor)) cex.cor <- 0.8/strwidth(txt)
          text(0.5, 0.5, txt, cex = cex.cor * r)
     }

pairs(cat.dat, lower.panel=panel.smooth, upper.panel=panel.cor)
```

- The raw correlation coefficient can be misleading as some of the explanatory variables are correlated with each other. One can "get rid" of the effect of one variable using the partial correlation coefficient. It is defined as the correlation coefficent bewteen the residuals of a regression on the variable one wants to get rid of. It is obtained as:

Caterpillar case: suggested steps III

```
pcor <- function(v1, v2, v3)
        {
        c12 <- cor(v1, v2)
        c23 <- cor(v2, v3)
        c13 <- cor(v1, v3)
        partial <- (c12-(c13*c23))/(sqrt(1-(c13^2)) * sqrt(1-(c23^2)))
        return(partial)
        }
```

Use this to evaluate the correlation between the response variable and the explanatory variables given `Nb.vegetation.strata`. Compare to the plain correlation. Which variable would you expect to be useless in a multiple regression once `Nb.vegetation.strata` is included in the model?

- Fit a linear model explaining the number of nests with all the other variables as predictors.
  Give the exact math equation that describes the model fitted by the R command above. What are the parameters involved? What are the main assumptions involved?

Caterpillar case: suggested steps IV

- Can you reject with confidence the null hypothesis in a global test?
- Plot the residuals againtst the predicted values. Does it seem reasonable to assume that the variance of the noise is the same for all observations?
  Hints: The residuals are computed by the `lm` function and stored as a component of the returned list named residuals. The predicted values can be obtained by `predict(lm.all)`
- Can the model quality be improved by a regression of the logarithm of the number of nests (instead of the number of nests)?

From now on, we work with the logarithm of the number of nests as response variable.

- The last line printed by the R command `summary(lm.all)` contains the p-value for the null hypothesis of absence of correlation with any explanatoty variable. Can you accept safely this null hypothesis?
- Plot the predicted values against the observed response values and compute their coefficient of correlation. Does this seem to be a useful model?

Caterpillar case: suggested steps V

- Print the esimated coefficients and associated p-values. What variables stand out as good or poor predictors? Compare this to a linear model with Density.index as single predictor. How do you explain the difference?

- From the above, there are obsviously some variables that are useless in the model. Compute leave-one-out prediction error with all variables then with all variables except Mixing.index.
Hints:

```
## leave one out prediction with full model
mse.1 <- 0
loo.pred1 <- numeric(nrow(cat.dat))
for(i in 1:nrow(cat.dat))
  {
    ## fitting a model with all variables
    ## to the dataset without obs. i
    lm.res.loo <- lm(formula=log(Nb.nests) ~. , data=cat.dat[-i,])
    ## predicting Log.nb.nests[i] with this model
    ## (just a linear combination with estimated coef.):
    loo.pred1[i] <- sum(c(1,as.numeric(cat.dat[i,1:10]))*lm.res.loo$coefficients)
    ## squre error in prediction
    mse.1 <- mse.1 + (log(cat.dat$Nb.nests[i]) - loo.pred1[i])^2
  }

mse.1 <- mse.1/(nrow(cat.dat)-1)
rmse.1 <- sqrt(mse.1)
```

Caterpillar case: suggested steps VI

```
## leave one out prediction with all variables exept Mixing.index
mse.2 <- 0
loo.pred2 <- numeric(nrow(cat.dat))
for(i in 1:nrow(cat.dat))
  {
    ## fitting a model with all variables except Mixing.index
    ## to the dataset without obs. i
    lm.res.loo <- lm(formula=(log(Nb.nests) ~.  - Mixing.index), data=cat.dat[-i,])
    ## predicting Log.nb.nests[i] with this model
    ## (just a linear combination with estimated coef.):
    loo.pred2[i] <- sum(c(1,as.numeric(cat.dat[i,1:9]))*lm.res.loo$coefficients)
    ## squre error in prediction
    mse.2 <- mse.2 + (log(cat.dat$Nb.nests[i]) - loo.pred2[i])^2
  }
mse.2 <- mse.2/(nrow(cat.dat)-1)
rmse.2 <- sqrt(mse.2)

rmse.1
rmse.2

##
plot(log(cat.dat$Nb.nests),loo.pred1,xlab="Log nb  of nest",
     ylab="LOO prediction"); abline(0,1, col=3)
points(log(cat.dat$Nb.nests),loo.pred2,col=2,pch=2)
```

On the basis of the above, would you include `Mixing.index` in the model?