Prelims Statistics and Data Analysis
Lectures 1–10

Christl Donnelly
christl.donnelly@stats.ox.ac.uk

Trinity Term 2024

Course info

Christl Donnelly, Department of Statistics,
christl.donnelly@stats.ox.ac.uk

16 lectures over 4 weeks

As usual in Prelims there will be one sheet per two lectures. So for this course this means two sheets per week – this should be about the right amount for one tutorial.

Introduction

Probability

We start with a probability model P, and we deduce properties of P.

Imagine flipping a coin 5 times. Assuming that flips are *fair* and *independent*, what is the probability of getting 2 heads?

Statistics

In statistics we have *data*, which we regard as having been generated from some *unknown* probability model P. We want to be able to say some useful things about P.

We flip a coin 1000 times and observe 530 heads. Is the coin fair? Is the probability of heads greater than $\frac{1}{2}$? or could it be equal to $\frac{1}{2}$?

Precision of estimation

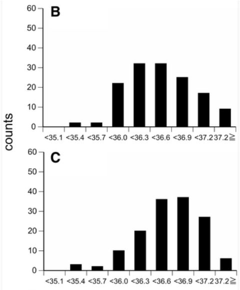
Question: What is the average body temperature in humans?

Collect data: let $x_1, \dots, x_n$ be the body temperatures of n randomly chosen individuals, i.e. our *model* is that individuals are chosen independently from the general population.

Then estimate: $\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$ could be our *estimate*, i.e. the *sample* average.

How precise is $\overline{x}$ as an estimate of the population mean?

Statistics Publications - Axillary temperature in young, healthy adults

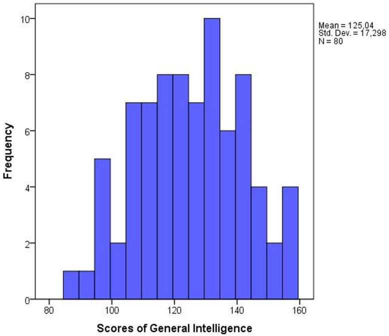


Histograms of axillary temperature B) measured for 10 seconds; C) measured for 10 minutes in °C (axillary = under the armpit)

Source: Marui, S., Misawa, A., Tanaka, Y. et al. J Physiol Anthropol (2017) 36: 18.

https://doi.org/10.1186/s40101-017-0133-y

Statistics Publications - IQ of University Students



Histogram showing the distribution of IQ among university students involved in the present study (40 men and 40 women).

Source: Kleisner K, Chvátalová V, Flegr J (2014) Perceived Intelligence Is Associated with Measured Intelligence in Men but Not Women. *PLoS ONE* 9(3): e81237.

https://doi.org/10.1371/journal.pone.0081237

Relationships between observations

Let y_i = price of house in month x_i, for $i = 1, \ldots, n$.

- Is it reasonable that $y_i = \alpha + \beta x_i$ + "error"?
- Is $\beta > 0$?
- How accurately can we estimate α and β, and how do we estimate them?

Relationship between time and house prices

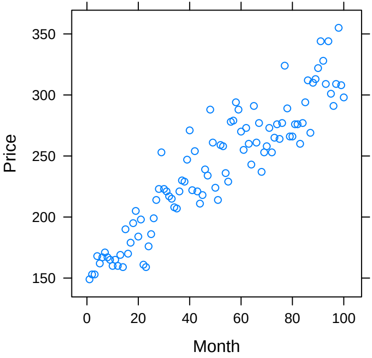
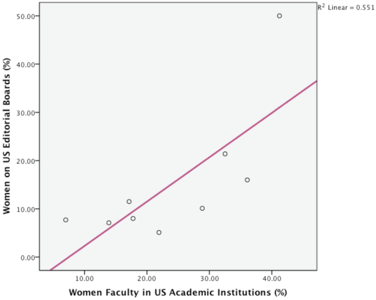


Figure: Scatterplot of some Oxford house price data.

Statistics Publications - Underrepresentation of women on boards



The association of women on editorial/advisory board and in the corresponding academic specialty.

Source: Ioannidou E, Rosania A (2015) Under-Representation of Women on Dental Journal Editorial Boards. *PLoS ONE* 10(1): e0116630. https://doi.org/10.1371/journal.pone.0116630

Notation and conventions

Lower case letters $x_1, \ldots, x_n$ denote *observations*: we regard these as the observed values of random variables $X_1, \ldots, X_n$.

Let $\mathbf{x} = (x_1, \ldots, x_n)$ and $\mathbf{X} = (X_1, \ldots, X_n)$.

It is convenient to think of x_i as

- the observed value of X_i
- or sometimes as a possible value that X_i can take.

Since x_i is a possible value for X_i we can calculate probabilities, e.g. if $X_i \sim \text{Poisson}(\lambda)$ then

$$P(X_i = x_i) = \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}, \quad x_i = 0, 1, \ldots.$$

1. Random Samples

Random Samples

A *random sample* of size n is a set of random variables $X_1, \dots, X_n$ which are independent and identically distributed (i.i.d.).

Let $X_1, \dots, X_n$ be a random sample from a Poisson(λ) distribution.

e.g. X_i = # traffic accidents on St Giles' in year i.

We'll write $f(\mathbf{x})$ to denote the joint probability mass function (p.m.f.) of $X_1, \dots, X_n$. Then

$$\begin{aligned}
f(\mathbf{x}) &= P(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n) && \text{definition of joint p.m.f.} \\
&= P(X_1 = x_1)P(X_2 = x_2) \dots P(X_n = x_n) && \text{by independence} \\
&= \frac{e^{-\lambda}\lambda^{x_1}}{x_1!} \cdot \frac{e^{-\lambda}\lambda^{x_2}}{x_2!} \cdots \frac{e^{-\lambda}\lambda^{x_n}}{x_n!} && \text{since } X_i \text{ Poisson} \\
&= \frac{e^{-n\lambda}\ \lambda^{(\sum_{i=1}^n x_i)}}{\prod_{i=1}^n x_i!}.
\end{aligned}$$

Random Samples

Let $X_1, \ldots, X_n$ be a random sample from an exponential distribution with probability density function (p.d.f.) given by

$$f(x) = \frac{1}{\mu} e^{-x/\mu}, \quad x \geq 0.$$

e.g. X_i = time until the first breakdown of machine i in a factory.

We'll write $f(\mathbf{x})$ to denote the joint p.d.f. of $X_1, \ldots, X_n$. Then

$$\begin{aligned} f(\mathbf{x}) &= f(x_1) \ldots f(x_n) \qquad \text{since the } X_i \text{ are independent} \\ &= \prod_{i=1}^{n} \frac{1}{\mu} e^{-x_i/\mu} \\ &= \frac{1}{\mu^n} \exp\left(-\frac{1}{\mu} \sum_{i=1}^{n} x_i \right). \end{aligned}$$

① We use f to denote a p.m.f. in the first example and to denote a p.d.f. in the second example. It is convenient to use the same letter (i.e. f) in both the discrete and continuous cases. (In introductory probability you may often see p for p.m.f. and f for p.d.f.)

We could write $f_{X_i}(x_i)$ for the p.m.f./p.d.f. of X_i, and $f_{X_1,\ldots,X_n}(\mathbf{x})$ for the joint p.m.f./p.d.f. of $X_, \ldots, X_n$. However it is convenient to keep things simpler by omitting subscripts on f.

② In the second example $E(X_i) = \mu$ and we say "X_i has an exponential distribution with mean μ" (i.e. expectation μ).

Sometimes, and often in probability, we work with "an exponential distribution with parameter λ" where $\lambda = 1/\mu$. To change the parameter from μ to λ all we do is replace μ by $1/\lambda$ to get

$$f(x) = \lambda e^{-\lambda x}, \quad x \geq 0.$$

Sometimes (often in statistics) we parametrise the distribution using μ, sometimes (often in probability) we parametrise it using λ.

In *probability* we assume that the parameters λ and μ in our two examples are known. In *statistics* we wish to estimate λ and μ from data.

- What is the best way to estimate them? And what does "best" mean?
- For a given method, how precise is the estimation?

2. Summary Statistics

Summary Statistics

The expected value of X, $\mathrm{E}(X)$, is also called its mean. This is often denoted μ.

The variance of X, $\mathrm{var}(X)$, is $\mathrm{var}(X) = E[(X-\mu)^2]$. This is often denoted σ^2. The standard deviation of X is σ.

Let $X_1, \ldots, X_n$ be a random sample. The *sample mean* $\overline{X}$ is defined by

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

The *sample variance* S^2 is defined by

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

The *sample standard deviation* is $S = \sqrt{\text{sample variance}}$.

Statistics Publications - Descriptive statistics - biochemistry

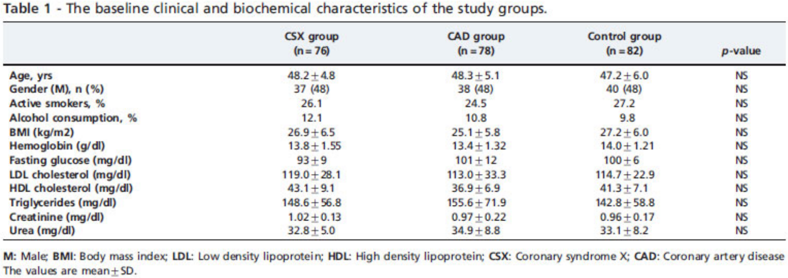


Table 1 - The baseline clinical and biochemical characteristics of the study groups.

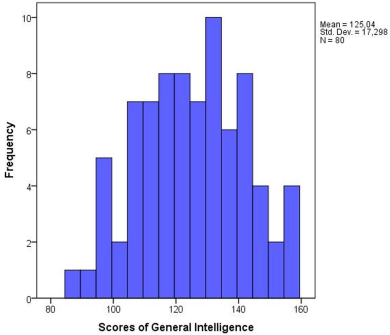
	CSX group (n = 76)	CAD group (n = 78)	Control group (n = 82)	*p*-value
Age, yrs	48.2±4.8	48.3±5.1	47.2±6.0	NS
Gender (M), n (%)	37 (48)	38 (48)	40 (48)	NS
Active smokers, %	26.1	24.5	27.2	NS
Alcohol consumption, %	12.1	10.8	9.8	NS
BMI (kg/m2)	26.9±6.5	25.1±5.8	27.2±6.0	NS
Hemoglobin (g/dl)	13.8±1.55	13.4±1.32	14.0±1.21	NS
Fasting glucose (mg/dl)	93±9	101±12	100±6	NS
LDL cholesterol (mg/dl)	119.0±28.1	113.0±33.3	114.7±22.9	NS
HDL cholesterol (mg/dl)	43.1±9.1	36.9±6.9	41.3±7.1	NS
Triglycerides (mg/dl)	148.6±56.8	155.6±71.9	142.8±58.8	NS
Creatinine (mg/dl)	1.02±0.13	0.97±0.22	0.96±0.17	NS
Urea (mg/dl)	32.8±5.0	34.9±8.8	33.1±8.2	NS

M: Male; **BMI:** Body mass index; **LDL:** Low density lipoprotein; **HDL:** High density lipoprotein; **CSX:** Coronary syndrome X; **CAD:** Coronary artery disease
The values are mean±SD.

Means ± standard deviations are reported for the continuous variables.

Source: Sait Demirkol et al. (2012) Evaluation of the mean platelet volume in patients with cardiac syndrome X. Clinics vol.67 no.9 http://dx.doi.org/10.6061/clinics/2012(09)06

Statistics Publications - IQ of University Students - Mean (125.04) and Standard deviation (17.298)



Histogram showing the distribution of IQ among university students

Source: Kleisner K, Chvátalová V, Flegr J (2014) Perceived Intelligence Is Associated with Measured Intelligence in Men but Not Women. *PLoS ONE* 9(3): e81237. https://doi.org/10.1371/journal.pone.0081237

Note

1. The denominator in the definition of S^2 is $n-1$, not n.
2. $\overline{X}$ and S^2 are random variables. Their distributions are called the *sampling distributions* of $\overline{X}$ and S^2.
3. Given observations $x_1, , \ldots, x_n$ we can compute the *observed values* $\overline{x}$ and s^2.
4. The sample mean $\overline{x}$ is a summary of the *location* of the sample.
5. The sample variance s^2 (or the sample standard deviation s) is a summary of the *spread* of the sample about $\overline{x}$.

The Normal Distribution

The random variable X has a normal distribution with mean μ and variance σ^2, written $X \sim N(\mu, \sigma^2)$, if the p.d.f. of X is

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(- \frac{(x-\mu)^2}{2\sigma^2} \right), \quad -\infty < x < \infty.$$

Properties: $E(X) = \mu$, $\mathrm{var}(X) = \sigma^2$.

Standard Normal Distribution

If $\mu = 0$ and $\sigma^2 = 1$, then $X \sim N(0,1)$ is said to have a *standard normal distribution*. Properties:

- If $Z \sim N(0,1)$ and $X = \mu + \sigma Z$, then $X \sim N(\mu, \sigma^2)$.
- If $X \sim N(\mu, \sigma^2)$ and $Z = (X - \mu)/\sigma$, then $Z \sim N(0,1)$.
- The cumulative distribution function (c.d.f.) of the *standard* normal distribution is

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-u^2/2}\, du.$$

This cannot be written in a closed form, but can be found by numerical integration to an arbitrary degree of accuracy.

Standard Normal

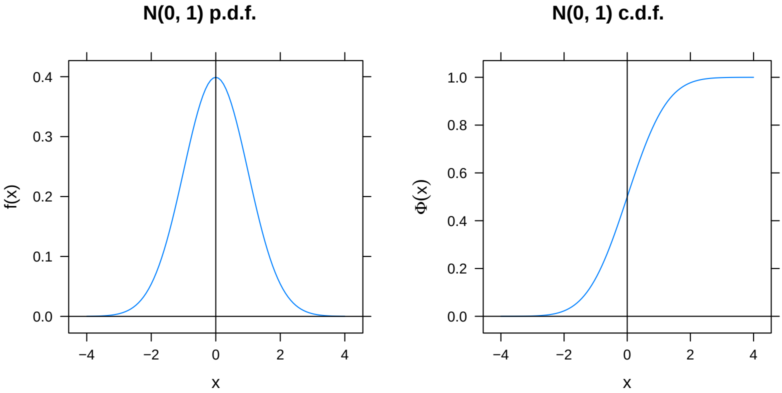
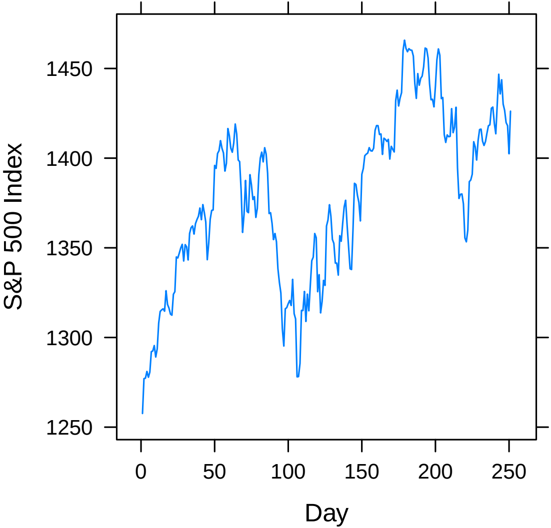
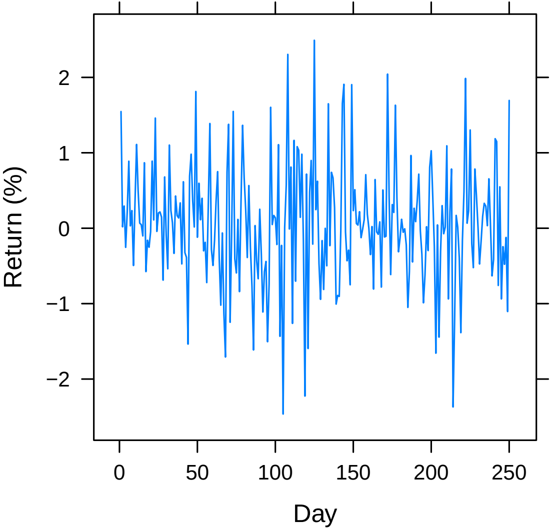


Figure: Standard normal p.d.f. and c.d.f.

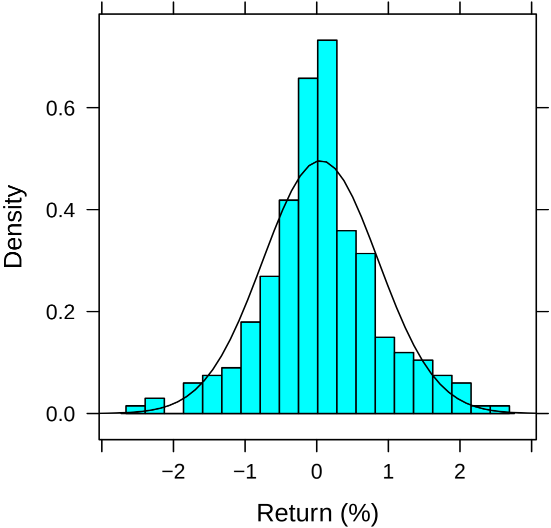
Standard and Poors 500 stock index (2012 data)



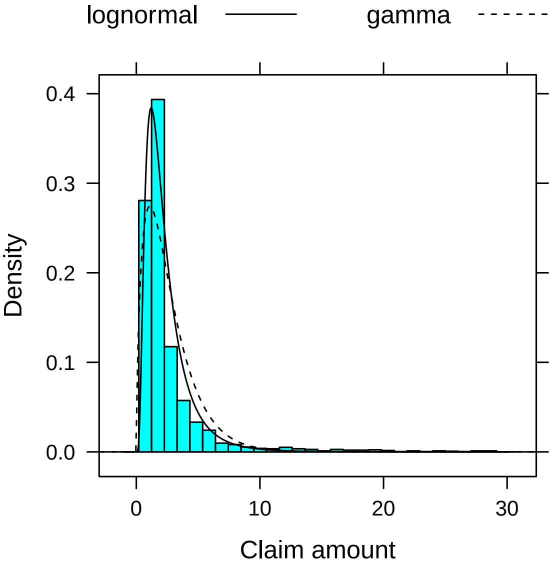
S&P 500 daily returns (%)



S&P 500 daily returns (%)

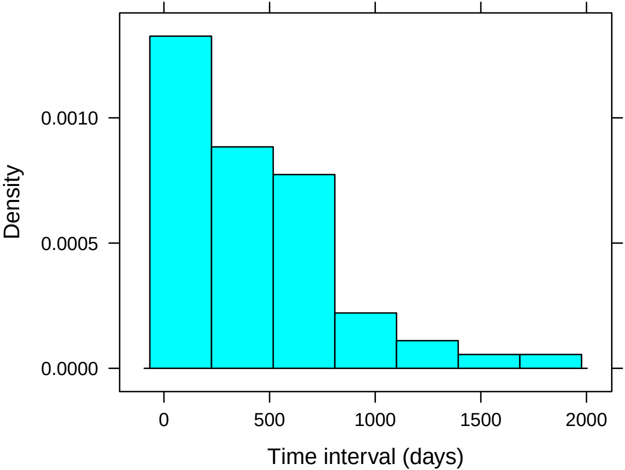


Insurance claim amounts



Time intervals between major earthquakes

Histogram of time intervals between 62 major earthquakes 1902–77: an exponential density looks plausible.



3. Maximum Likelihood

Maximum Likelihood Estimation

Maximum likelihood estimation is a general method for estimating unknown parameters from data. This turns out to be the method of choice in many contexts, though this isn't obvious at this stage.

Suppose e.g. that $x_1, \ldots, x_n$ are time intervals between major earthquakes. Assume these are observations of $X_1, \ldots, X_n$ independently drawn from an exponential distribution with mean μ, so that each X_i has p.d.f.

$$f(x;\mu) = \frac{1}{\mu} e^{-x/\mu}, \quad x \geq 0.$$

We have written $f(x;\mu)$ to indicate that the p.d.f. f depends on μ.

- How do we estimate μ?
- Is $\overline{X}$ a good estimator for μ?
- Is there a general principle we can use?

Maximum Likelihood Estimation

In general we write $f(x;\theta)$ to indicate that the p.d.f./p.m.f. f, which is a function of x, depends on a parameter θ. Similarly, $f(\mathbf{x};\theta)$ denotes the joint p.d.f./p.m.f. of $\mathbf{X}$. (Sometimes $f(x;\theta)$ is written $f(x|\theta)$.)

The parameter θ may be a vector, e.g. $\theta = (\mu, \sigma^2)$ in the earlier $N(\mu, \sigma^2)$ example.

Maximum Likelihood Estimation

If we regard θ as unknown, then we need to estimate it using $x_1, \dots, x_n$.

Let $X_1, \dots, X_n$ have joint p.d.f./p.m.f. $f(\mathbf{x}; \theta)$. Given observed values $x_1, \dots, x_n$ of $X_1, \dots, X_n$, the *likelihood* of θ is the function

$$L(\theta) = f(\mathbf{x}; \theta). \tag{1}$$

The *log-likelihood* is $\ell(\theta) = \log L(\theta)$.

So $L(\theta)$ is the joint p.d.f./p.m.f. of the observed data *regarded as a function of* θ, for fixed $\mathbf{x}$.

In the definition of $\ell(\theta)$, log means log to base e, i.e. $\log = \log_e = \ln$.

Maximum Likelihood Estimation

Often we assume that $X_1, \ldots, X_n$ are a random sample from $f(x;\theta)$, so that

$$\begin{aligned} L(\theta) &= f(\mathbf{x};\theta) \\ &= \prod_{i=1}^{n} f(x_i;\theta) \qquad \text{since the } X_i \text{ are independent.} \end{aligned}$$

Sometimes we have independent X_i whose distributions differ, say X_i is from $f_i(x;\theta)$. Then the likelihood is

$$L(\theta) = \prod_{i=1}^{n} f_i(x_i;\theta).$$

The *maximum likelihood estimate* (MLE) $\widehat{\theta}(\mathbf{x})$ is the value of θ that maximises $L(\theta)$ for the given $\mathbf{x}$.

The idea of maximum likelihood is to estimate the parameter by the value of θ that gives the greatest likelihood to observations $x_1, \ldots, x_n$. That is, the θ for which the probability or probability density is maximised.

Since taking logs is monotone, $\widehat{\theta}(\mathbf{x})$ also maximises $\ell(\theta)$. Finding the MLE by maximising $\ell(\theta)$ is often more convenient.

[Example continued] In our exponential mean μ example, the parameter is μ and

$$L(\mu) = \prod_{i=1}^{n} \frac{1}{\mu} e^{-x_i/\mu} = \mu^{-n} \exp\left(-\frac{1}{\mu} \sum_{i=1}^{n} x_i \right)$$

$$\ell(\mu) = \log L(\mu) = -n \log \mu - \frac{1}{\mu} \sum_{i=1}^{n} x_i.$$

To find the maximum

$$\frac{d\ell}{d\mu} = -\frac{n}{\mu} + \frac{\sum_{i=1}^{n} x_i}{\mu^2}.$$

So

$$\frac{d\ell}{d\mu} = 0 \iff \frac{n}{\mu} = \frac{\sum_{i=1}^{n} x_i}{\mu^2} \iff \mu = \overline{x}.$$

This is a maximum since

$$\left.\frac{d^2\ell}{d\mu^2}\right|_{\mu=\overline{x}} = \frac{n}{\overline{x}^2} - \frac{2\sum_{i=1}^{n} x_i}{\overline{x}^3} = -\frac{n}{\overline{x}^2} < 0.$$

So the MLE is $\widehat{\mu}(\mathbf{x}) = \overline{x}$. Often we'll just write $\widehat{\mu} = \overline{x}$.

In this case the maximum likelihood *estimator* of μ is $\widehat{\mu}(\mathbf{X}) = \overline{X}$, which is a random variable. (More on the difference between estimates and estimators later.)

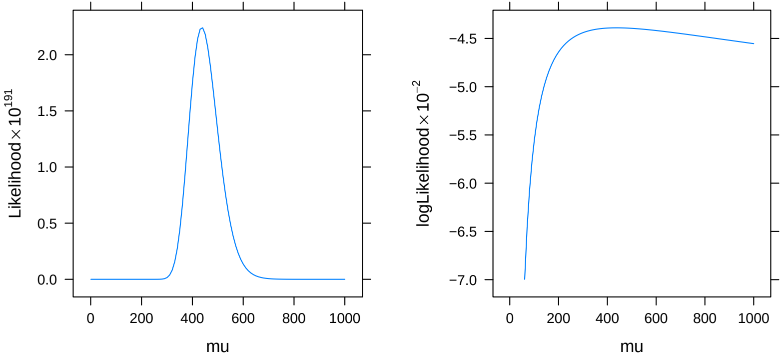


Figure: Likelihood $L(\mu)$ and log-likelihood $\ell(\mu)$ for exponential (earthquake) example.

Note that both the likelihood and the log-likelihood are not plotted on the natural scale.

If $\widehat{\theta}(\mathbf{x})$ is the maximum likelihood estimate of θ, then the *maximum likelihood estimator* (MLE) is defined by $\widehat{\theta}(\mathbf{X})$.

Note: both maximum likelihood *estimate* and maximum likelihood *estimator* are often abbreviated to MLE.

Opinion poll example

Suppose n individuals are drawn independently from a large population. Let

$$X_i = \begin{cases} 1 & \text{if individual } i \text{ is a Labour voter} \\ 0 & \text{otherwise.} \end{cases}$$

Let p be the proportion of Labour voters, so that

$$P(X_i = 1) = p, \quad P(X_i = 0) = 1 - p.$$

This is a Bernoulli distribution, for which the p.m.f. can be written

$$f(x;p) = P(X = x) = p^x (1-p)^{1-x}, \quad x = 0, 1.$$

The likelihood is

$$\begin{aligned} L(p) &= \prod_{i=1}^{n} p^{x_i} (1-p)^{1-x_i} \\ &= p^r (1-p)^{n-r} \end{aligned}$$

where $r = \sum_{i=1}^{n} x_i$.

Opinion poll example (continued)

So the log-likelihood is

$$\begin{aligned}\ell(p) &= \log L(p)\\ &= r\log p + (n-r)\log(1-p)\end{aligned}$$

For a maximum, differentiate and set to zero:

$$\frac{r}{p} - \frac{n-r}{1-p} = 0 \iff \frac{r}{p} = \frac{n-r}{1-p} \iff r - rp = np - rp$$

and so $p = r/n$. This is a maximum since $\ell''(p) < 0$.

So $\widehat{p} = r/n$, i.e. the MLE is $\widehat{p} = \sum_{i=1}^{n} X_i/n$ which is the proportion of Labour voters in the sample.

Genetics example

Suppose we test randomly chosen individuals at a particular locus on the genome. Each chromosome can be type A or a. Every individual has two chromosomes (one from each parent), so the genotype can be AA, Aa or aa. (Note that order is not relevant, there is no distinction between Aa and aA.)

Hardy-Weinberg law: under plausible assumptions,

$$P(AA) = p_1 = \theta^2, \quad P(Aa) = p_2 = 2\theta(1-\theta), \quad P(aa) = p_3 = (1-\theta)^2$$

for some θ with $0 \leq \theta \leq 1$.

Now suppose the random sample of n individuals contains:

$$x_1 \text{ of type } AA, \quad x_2 \text{ of type } Aa, \quad x_3 \text{ of type } aa$$

where $x_1 + x_2 + x_3 = n$ and these are observations of X_1, X_2, X_3.

Genetics example (continued)

Then

$$\begin{aligned} L(\theta) &= P(X_1 = x_1, X_2 = x_2, X_3 = x_3) \\ &= \frac{n!}{x_1!\, x_2!\, x_3!} p_1^{x_1} p_2^{x_2} p_3^{x_3} \\ &= \frac{n!}{x_1!\, x_2!\, x_3!} (\theta^2)^{x_1} [2\theta(1-\theta)]^{x_2} [(1-\theta)^2]^{x_3}. \end{aligned}$$

This is a *multinomial distribution*.

So

$$\begin{aligned} \ell(\theta) &= \text{constant} + 2x_1 \log\theta + x_2[\log 2 + \log\theta + \log(1-\theta)] + 2x_3 \log(1-\theta) \\ &= \text{constant} + (2x_1 + x_2)\log\theta + (x_2 + 2x_3)\log(1-\theta) \end{aligned}$$

(where the constants depend on x_1, x_2, x_3 but not θ).

Genetics example (continued)

Then $\ell'(\widehat{\theta}) = 0$ gives

$$\frac{2x_1 + x_2}{\widehat{\theta}} - \frac{x_2 + 2x_3}{1 - \widehat{\theta}} = 0$$

or

$$(2x_1 + x_2)(1 - \widehat{\theta}) = (x_2 + 2x_3)\widehat{\theta}.$$

So

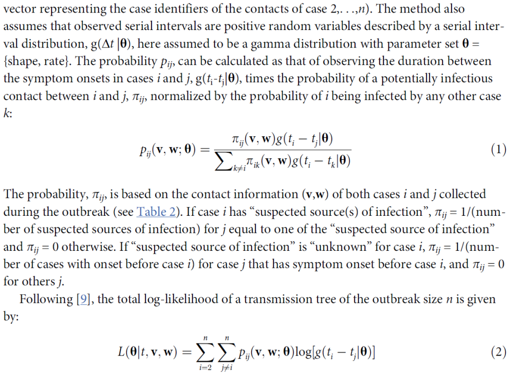
$$\widehat{\theta} = \frac{2x_1 + x_2}{2(x_1 + x_2 + x_3)} = \frac{2x_1 + x_2}{2n}.$$

Maximum likelihood approach

Steps:

- Write down the (log) likelihood
- Find the maximum (usually by differentiation, but not quite always)
- Rearrange to give the parameter estimate in terms of the data.

Statistics Publications - Likelihood of a hepatitis A transmission tree



vector representing the case identifiers of the contacts of case 2,. . .,n). The method also assumes that observed serial intervals are positive random variables described by a serial interval distribution, $g(\Delta t\,|\boldsymbol{\theta})$, here assumed to be a gamma distribution with parameter set $\boldsymbol{\theta}$ = {shape, rate}. The probability p_{ij}, can be calculated as that of observing the duration between the symptom onsets in cases i and j, $g(t_i\text{-}t_j|\boldsymbol{\theta})$, times the probability of a potentially infectious contact between i and j, π_{ij}, normalized by the probability of i being infected by any other case k:

$$p_{ij}(\mathbf{v},\mathbf{w};\boldsymbol{\theta}) = \frac{\pi_{ij}(\mathbf{v},\mathbf{w})g(t_i - t_j|\boldsymbol{\theta})}{\sum_{k\neq i}\pi_{ik}(\mathbf{v},\mathbf{w})g(t_i - t_k|\boldsymbol{\theta})} \tag{1}$$

The probability, π_{ij}, is based on the contact information ($\mathbf{v}$,$\mathbf{w}$) of both cases i and j collected during the outbreak (see Table 2). If case i has "suspected source(s) of infection", π_{ij} = 1/(number of suspected sources of infection) for j equal to one of the "suspected source of infection" and π_{ij} = 0 otherwise. If "suspected source of infection" is "unknown" for case i, π_{ij} = 1/(number of cases with onset before case i) for case j that has symptom onset before case i, and π_{ij} = 0 for others j.

Following [9], the total log-likelihood of a transmission tree of the outbreak size n is given by:

$$L(\boldsymbol{\theta}|t,\mathbf{v},\mathbf{w}) = \sum_{i=2}^{n}\sum_{j\neq i}^{n} p_{ij}(\mathbf{v},\mathbf{w};\boldsymbol{\theta})\log[g(t_i - t_j|\boldsymbol{\theta})] \tag{2}$$

v = possible infector; **w** = contacts; θ = parameters of serial interval dist.

Source: Zhang X-S, Iacono GL (2018) Estimating human-to-human transmissibility of hepatitis A virus in an outbreak at an elementary school in China, 2011. PLoS ONE 13(9): e0204201. https://doi.org/10.1371/journal.pone.0204201

Estimating multiple parameters

Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$ where both μ and σ^2 are unknown.

[Here $\overset{\text{iid}}{\sim}$ means "are independent and identically distributed as."]

The likelihood is

$$L(\mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2} \right)$$

$$= (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \right)$$

with log-likelihood

$$\ell(\mu, \sigma^2) = -\frac{n}{2} \log 2\pi - \frac{n}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2.$$

Estimating multiple parameters (continued)

We maximise ℓ jointly over μ and σ^2:

$$\frac{\partial \ell}{\partial \mu} = \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu)$$

$$\frac{\partial \ell}{\partial (\sigma^2)} = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2} \sum_{i=1}^{n} (x_i - \mu)^2$$

and solving $\frac{\partial \ell}{\partial \mu} = 0$ and $\frac{\partial \ell}{\partial (\sigma^2)} = 0$ simultaneously we obtain

$$\widehat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i = \overline{x}$$

$$\widehat{s}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \widehat{\mu})^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})^2.$$

Hence: the MLE of μ is the sample mean, but the MLE of σ^2 is $(n-1)s^2/n$. (More later.)

“Estimate” and “estimator”

Estimator:

- A rule for constructing an estimate.
- A function of the random variables **X** involved in the random sample.
- Itself a random variable.

Estimate:

- The numerical value of the estimator for the particular data set.
- The value of the function evaluated at the data $x_1, \ldots, x_n$.

Statistics Publications - Influenza transmission in Mexico in 2009

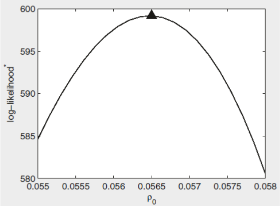


Figure 6 Log-likelihood profile for the standard (SD) model as a function of the marginal transmission rate (per contact) ρ_0. The triangle represents the maximum likelihood estimate from Table 2. The star on log-likelihood indicates that the constant term from the log-likelihood has been excluded.

Log likelihood profile of the influenza transmission rate

Source: Michael Springborn, Gerardo Chowell, Matthew MacLachlan, Eli P Fenichel (2015) Accounting for behavioral responses during a flu epidemic using home television viewing. BMC Infectious Diseases 15:21 https://doi.org/10.1186/s12879-014-0691-0

Statistics Publications - Influenza transmission in Mexico in 2009

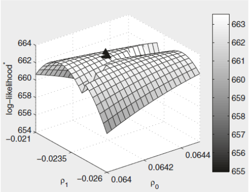


Figure 7 Log-likelihood profile for the fixed effect (FE) model as a function of the marginal transmission rate (per contact) ρ_0, and the shift in the marginal baseline transmission rate during the intervention window, ρ_1. The triangle represents the maximum likelihood estimates from Table 2. The star on log-likelihood indicates that the constant term from the log-likelihood has been excluded.

Sometimes the multidimensional log likelihood surface is complex

Source: Michael Springborn, Gerardo Chowell, Matthew MacLachlan, Eli P Fenichel (2015) Accounting for behavioral responses during a flu epidemic using home television viewing. BMC Infectious Diseases 15:21 https://doi.org/10.1186/s12879-014-0691-0

Statistics Publications - Influenza transmission in Mexico in 2009

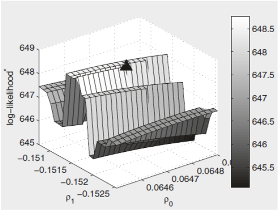


Figure 8 Log-likelihood profile for the dynamic behavioral (DB) model as a function of the marginal transmission rate (per contact) ρ_0, and the behavioral response, ρ_1. The triangle represents the maximum likelihood estimates from Table 2. The star on log-likelihood indicates that the constant term from the log-likelihood has been excluded.

Often these surfaces have to be searched numerically to identify the global maximum

Source: Michael Springborn, Gerardo Chowell, Matthew MacLachlan, Eli P Fenichel (2015) Accounting for behavioral responses during a flu epidemic using home television viewing. BMC Infectious Diseases 15:21 https://doi.org/10.1186/s12879-014-0691-0

Statistics Publications - The maximum log likelihood is easy to spot - Influenza transmission in Mexico in 2009

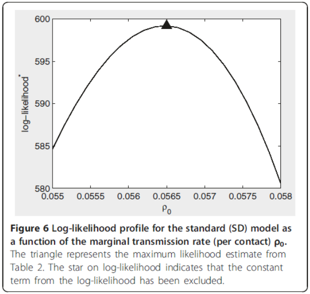
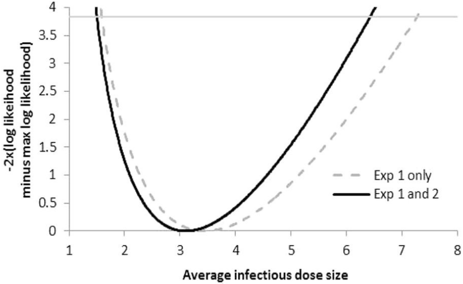


Figure 6 Log-likelihood profile for the standard (SD) model as a function of the marginal transmission rate (per contact) ρ_0. The triangle represents the maximum likelihood estimate from Table 2. The star on log-likelihood indicates that the constant term from the log-likelihood has been excluded.

Log likelihood profile of the influenza transmission rate

Source: Michael Springborn, Gerardo Chowell, Matthew MacLachlan, Eli P Fenichel (2015) Accounting for behavioral responses during a flu epidemic using home television viewing. BMC Infectious Diseases 15:21 https://doi.org/10.1186/s12879-014-0691-0

Statistics Publications - Sometimes a maximum looks like a minimum - Influenza transmission in ferrets



Sometimes the log likelihood profile is flipped and even given relative to its maximum

Source: Rebecca Frise et al. (2016) Contact transmission of influenza virus between ferrets imposes a looser bottleneck than respiratory droplet transmission allowing propagation of antiviral resistance. Scientific Reports volume 6, Article number: 29793

4. Parameter Estimation

Parameter Estimation - Earthquake example

Recall the earthquake example:

$$X_1, \dots, X_n \overset{\text{iid}}{\sim} \text{exponential distribution, mean } \mu.$$

Possible estimators of μ:

- $\overline{X}$
- $\frac{1}{3}X_1 + \frac{2}{3}X_2$
- $X_1 + X_2 - X_3$
- $\frac{2}{n(n+1)}(X_1 + 2X_2 + \cdots + nX_n)$.

How should we choose?

In general, suppose $X_1, \dots, X_n$ is a random sample from a distribution with p.d.f./p.m.f. $f(x;\theta)$. We want to estimate θ from observations $x_1, \dots, x_n$.

A *statistic* is any function $T(\mathbf{X})$ of $X_1, \dots, X_n$ that does not depend on θ.

An *estimator* of θ is any statistic $T(\mathbf{X})$ that we might use to estimate θ.

$T(\mathbf{x})$ is the *estimate* of θ obtained via T from observed values $\mathbf{x}$.

$T(\mathbf{X})$ is a random variable, e.g. $\overline{X}$.

$T(\mathbf{x})$ is a fixed number, based on data, e.g. $\overline{x}$.

We can choose between estimators by studying their properties. A good estimator should take values close to θ.

The estimator $T = T(\mathbf{X})$ is said to be *unbiased* for θ if, whatever the true value of θ, we have $E(T) = \theta$.

This means that “on average” T is correct.

Example: Earthquakes

Possible estimators $\overline{X}$, $\frac{1}{3}X_1 + \frac{2}{3}X_2$, etc.

Since $E(X_i) = \mu$, we have

$$E(\overline{X}) = \frac{1}{n}\sum_{i=1}^{n}\mu = \mu$$

$$E(\tfrac{1}{3}X_1 + \tfrac{2}{3}X_2) = \tfrac{1}{3}\mu + \tfrac{2}{3}\mu = \mu.$$

Similar calculations show that $X_1 + X_2 - X_3$ and $\frac{2}{n(n+1)}\sum_{j=1}^{n} jX_j$ are also unbiased.

Example: Normal variance

Suppose $X_1, \dots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$, with μ and σ^2 unknown, and let $T = \frac{1}{n}\sum(X_i - \overline{X})^2$. Then T is the MLE of σ^2. Is T unbiased?

Let $Z_i = (X_i - \mu)/\sigma$. So the Z_i are independent and $N(0,1)$, $E(Z_i) = 0$, $\operatorname{var}(Z_i) = E(Z_i^2) = 1$.

$E[(X_i - \overline{X})^2]$

$= E[\sigma^2(Z_i - \overline{Z})^2]$

$= \sigma^2 \operatorname{var}(Z_i - \overline{Z})$ since $E(Z_i - \overline{Z}) = 0$

$$= \sigma^2 \operatorname{var}\left(-\frac{1}{n}Z_1 - \frac{1}{n}Z_2 - \cdots + \frac{n-1}{n}Z_i + \cdots - \frac{1}{n}Z_n \right)$$

$$= \sigma^2\left(\frac{1}{n^2}\operatorname{var}(Z_1) + \frac{1}{n^2}\operatorname{var}(Z_2) + \cdots + \frac{(n-1)^2}{n^2}\operatorname{var}(Z_i) + \cdots + \frac{1}{n^2}\operatorname{var}(Z_n)\right)$$

since $\operatorname{var}(\sum a_i U_i) = \sum a_i^2 \operatorname{var}(U_i)$ for indep U_i

$$= \sigma^2\left((n-1) \times \frac{1}{n^2} + \frac{(n-1)^2}{n^2}\right) = \frac{(n-1)}{n}\sigma^2.$$

So

$$E(T) = \frac{1}{n}\sum_{i=1}^{n} E[(X_i - \overline{X})^2] = \frac{(n-1)}{n}\sigma^2 < \sigma^2.$$

Hence T is *not* unbiased, T will *underestimate* σ^2 on average.

But

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{n}{n-1}T$$

so the sample variance satisfies

$$E(S^2) = \frac{n}{n-1}E(T) = \sigma^2.$$

So S^2 is unbiased for σ^2.

Uniform distribution – some unusual features!

Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Uniform}[0, \theta]$, where $\theta > 0$, i.e.

$$f(x;\theta) = \begin{cases} \dfrac{1}{\theta} & \text{if } 0 \le x \le \theta \\ 0 & \text{otherwise.} \end{cases}$$

What is the MLE for θ? Is the MLE unbiased?

Calculate the likelihood:

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{n} f(x;\theta) \\ &= \begin{cases} \dfrac{1}{\theta^n} & \text{if } 0 \le x_i \le \theta \text{ for all } i \\ 0 & \text{otherwise} \end{cases} \\ &= \begin{cases} 0 & \text{if } 0 < \theta < \max x_i \\ \dfrac{1}{\theta^n} & \text{if } \theta \ge \max x_i. \end{cases} \end{aligned}$$

Note: $\theta \ge x_i$ for all $i \iff \theta \ge \max x_i$. (And $\max x_i$ means $\max_{1 \le i \le n} x_i$).

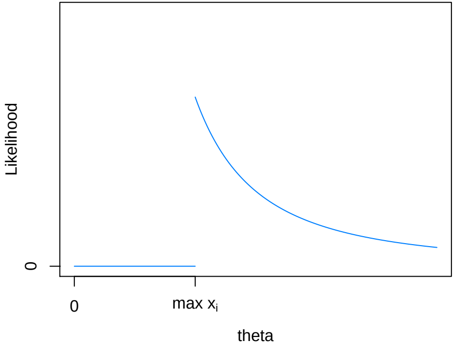


Figure: Likelihood $L(\theta)$ for Uniform$[0, \theta]$ example.

From the diagram:

- the max occurs at $\widehat{\theta} = \max x_i$
- this is *not* a point where $\ell'(\theta) = 0$
- taking logs doesn't help.

Consider the range of values of x for which $f(x;\theta) > 0$, i.e. $0 \le x \le \theta$. The thing that makes this example different to our previous ones is that this range *depends* on θ (and we must take this into account because the likelihood is a *function* of θ).

The MLE of θ is $\widehat{\theta} = \max X_i$. What is $E(\widehat{\theta})$?

Find the c.d.f. of $\widehat{\theta}$:

$$\begin{aligned}
F(y) &= P(\widehat{\theta} \le y) \\
&= P(\max X_i \le y) \\
&= P(X_1 \le y, X_2 \le y, \dots, X_n \le y) \\
&= P(X_1 \le y)P(X_2 \le y)\dots P(X_n \le y) \qquad \text{since } X_i \text{ independent} \\
&= \begin{cases} (y/\theta)^n & \text{if } 0 \le y \le \theta \\ 1 & \text{if } y > \theta. \end{cases}
\end{aligned}$$

So, differentiating the c.d.f., the p.d.f. is

$$f(y) = \frac{ny^{n-1}}{\theta^n}, \quad 0 \le y \le \theta.$$

So

$$\begin{aligned} E(\widehat{\theta}) &= \int_0^\theta y \cdot \frac{ny^{n-1}}{\theta^n}\, dy \\ &= \frac{n}{\theta^n} \int_0^\theta y^n\, dy \\ &= \frac{n\theta}{n+1}. \end{aligned}$$

So $\widehat{\theta}$ is *not* unbiased. But note that it *is* asymptotically unbiased: $E(\widehat{\theta}) \to \theta$ as $n \to \infty$.

In fact under mild assumptions MLEs are always asymptotically unbiased.

Further Properties of Estimators

The *mean squared error* (MSE) of an estimator T is defined by

$$\mathrm{MSE}(T) = E[(T - \theta)^2].$$

The *bias* $b(T)$ of T is defined by

$$b(T) = E(T) - \theta.$$

1. Both $\mathrm{MSE}(T)$ and $b(T)$ may depend on θ.
2. MSE is a measure of the "distance" between T and θ, so is a good overall measure of performance.
3. T is unbiased if $b(T) = 0$ for all θ.

Theorem 4.1 $\mathrm{MSE}(T) = \mathrm{var}(T) + [b(T)]^2$.

Proof

Let $\mu = E(T)$. Then

$$\begin{aligned}
\mathrm{MSE}(T) &= E[\{(T-\mu) + (\mu - \theta)\}^2] \\
&= E[(T-\mu)^2 + 2(\mu-\theta)(T-\mu) + (\mu-\theta)^2] \\
&= E[(T-\mu)^2] + 2(\mu-\theta)E[T-\mu] + (\mu-\theta)^2 \\
&= \mathrm{var}(T) + 2(\mu-\theta) \times 0 + (\mu-\theta)^2 \\
&= \mathrm{var}(T) + [b(T)]^2.
\end{aligned}$$

So an estimator with small MSE needs to have small variance *and* small bias. Unbiasedness alone is not particularly desirable – it is the combination of small variance and small bias which is important.

Reminder

Suppose $a_1, \ldots, a_n$ are constants. It is always the case that

$$E(a_1 X_1 + \cdots + a_n X_n) = a_1 E(X_1) + \cdots + a_n E(X_n).$$

If $X_1, \ldots, X_n$ are independent then

$$\text{var}(a_1 X_1 + \cdots + a_n X_n) = a_1^2 \text{var}(X_1) + \cdots + a_n^2 \text{var}(X_n).$$

In particular, if $X_1, \ldots, X_n$ is a random sample with $E(X_i) = \mu$ and $\text{var}(X_i) = \sigma^2$, then

$$E(\overline{X}) = \mu \quad \text{and} \quad \text{var}(\overline{X}) = \frac{\sigma^2}{n}.$$

Uniform distribution

Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Uniform}[0, \theta]$, i.e.

$$f(x;\theta) = \begin{cases} \dfrac{1}{\theta} & \text{if } 0 \le x \le \theta \\ 0 & \text{otherwise.} \end{cases}$$

We will consider two estimators of θ:

- $T = 2\overline{X}$, the natural estimator based on the sample mean (because the mean of the distribution is $\theta/2$)
- $\widehat{\theta} = \max X_i$, the MLE.

Now $E(T) = 2E(\overline{X}) = \theta$, so T is unbiased. Hence

$$\begin{aligned} \text{MSE}(T) &= \text{var}(T) \\ &= 4\,\text{var}(\overline{X}) \\ &= \frac{4\,\text{var}(X_1)}{n}. \end{aligned}$$

We have $E(X_1) = \theta/2$ and

$$E(X_1^2) = \int_0^\theta x^2 \cdot \frac{1}{\theta}\, dx = \frac{\theta^2}{3}$$

so

$$\operatorname{var}(X_1) = \frac{\theta^2}{3} - \left(\frac{\theta}{2}\right)^2 = \frac{\theta^2}{12}$$

hence

$$\text{MSE}(T) = \frac{4\operatorname{var}(X_1)}{n} = \frac{\theta^2}{3n}.$$

Previously we showed that $\widehat{\theta}$ has p.d.f.

$$f(y) = \frac{n y^{n-1}}{\theta^n}, \quad 0 \le y \le \theta$$

and $E(\widehat{\theta}) = n\theta/(n+1)$. So $b(\widehat{\theta}) = n\theta/(n+1) - \theta = -\theta/(n+1)$. Also,

$$E(\widehat{\theta^2}) = \int_0^\theta y^2 \cdot \frac{n y^{n-1}}{\theta^n}\, dy = \frac{n\theta^2}{n+2}$$

so

$$\text{var}(\widehat{\theta}) = \theta^2 \left(\frac{n}{n+2} - \frac{n^2}{(n+1)^2} \right) = \frac{n\theta^2}{(n+1)^2(n+2)}$$

hence

$$
\begin{aligned}
\mathrm{MSE}(\widehat{\theta}) &= \mathrm{var}(\widehat{\theta}) + [b(\widehat{\theta})]^2 \\
&= \frac{2\theta^2}{(n+1)(n+2)} \\
&< \frac{\theta^2}{3n} \qquad \text{for } n \geq 3 \\
&= \mathrm{MSE}(T).
\end{aligned}
$$

- $\mathrm{MSE}(\widehat{\theta}) \ll \mathrm{MSE}(T)$ for large n, so $\widehat{\theta}$ is much better – its MSE decreases like $1/n^2$ rather than $1/n$.

Remember: $T = 2\overline{X}$ and $\widehat{\theta} = \max X_i$

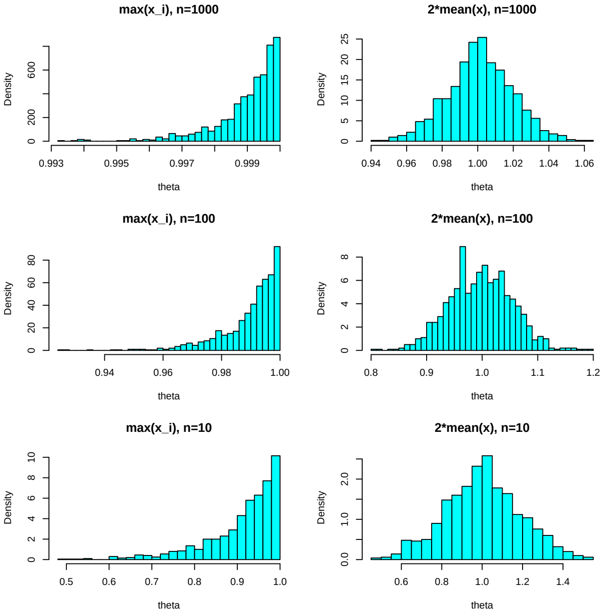
- Note that $(\frac{n+1}{n})\widehat{\theta}$ is unbiased and

$$\begin{aligned}\mathrm{MSE}\left(\frac{n+1}{n}\widehat{\theta}\right) &= \mathrm{var}\left(\frac{n+1}{n}\widehat{\theta}\right)\\ &= \frac{(n+1)^2}{n^2}\,\mathrm{var}(\widehat{\theta})\\ &= \frac{\theta^2}{n(n+2)}\\ &< \mathrm{MSE}(\widehat{\theta}) \qquad \text{for } n \geq 2.\end{aligned}$$

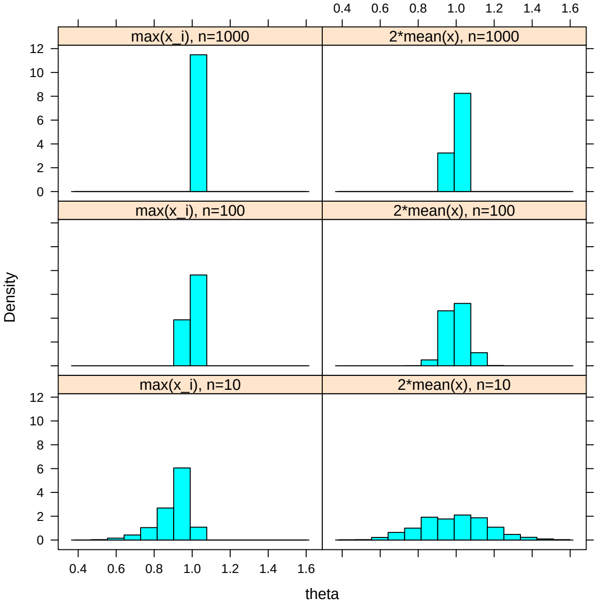
However, among all estimators of the form $\lambda\widehat{\theta}$, the MSE is minimized by $(\frac{n+2}{n+1})\widehat{\theta}$.

[To show this: note $\mathrm{var}(\lambda\widehat{\theta}) = \lambda^2\,\mathrm{var}(\widehat{\theta})$ and $b(\lambda\widehat{\theta}) = \frac{\lambda n\theta}{n+1} - \theta$. Now plug in formulae and minimise over λ.]

Estimating the parameter of Uniform$[0, \theta]$



Estimating the parameter of Uniform$[0, \theta]$



Estimation so far

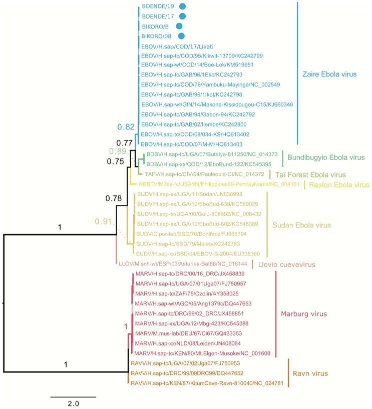
So far we've considered getting an estimate which is a single number – a *point* estimate – for the parameter of interest: e.g. $\overline{x}$, $\max x_i$, s^2, ...

Maximum likelihood is a good way (usually) of producing an estimate (but we did better when the range of the distribution depends on θ – fairly unusual).

MLEs are usually asymptotically unbiased, and have MSE decreasing like $1/n$ for large n.

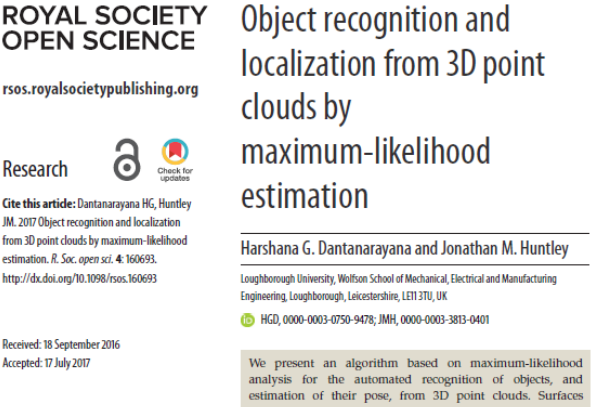
MLEs can be found in quite general situations.

Statistics Publications - Maximum likelihood in phylogenetics



Source: Placide Mbala-Kingebeni et al. (2019) Rapid Confirmation of the Zaire Ebola Virus in the Outbreak of the Equateur Province in the Democratic Republic of Congo: ... , Clinical Infectious Diseases 68(2): 330–333, https://doi.org/10.1093/cid/ciy527

Statistics Publications - Maximum likelihood in object recognition



ROYAL SOCIETY
OPEN SCIENCE

rsos.royalsocietypublishing.org

Research

Cite this article: Dantanarayana HG, Huntley JM. 2017 Object recognition and localization from 3D point clouds by maximum-likelihood estimation. *R. Soc. open sci.* **4**: 160693. http://dx.doi.org/10.1098/rsos.160693

Received: 18 September 2016
Accepted: 17 July 2017

Object recognition and localization from 3D point clouds by maximum-likelihood estimation

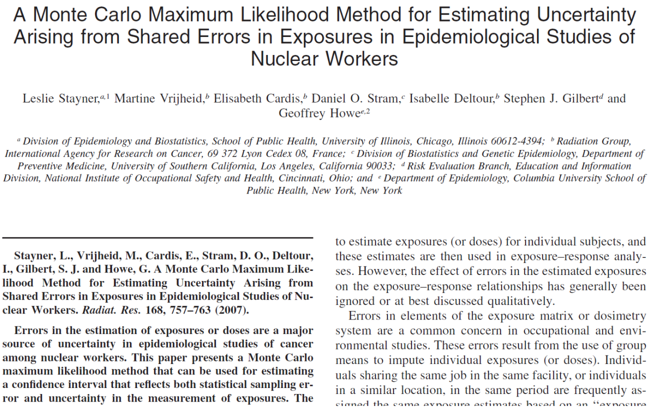
Harshana G. Dantanarayana and Jonathan M. Huntley

Loughborough University, Wolfson School of Mechanical, Electrical and Manufacturing Engineering, Loughborough, Leicestershire, LE11 3TU, UK

HGD, 0000-0003-0750-9478; JMH, 0000-0003-3813-0401

We present an algorithm based on maximum-likelihood analysis for the automated recognition of objects, and estimation of their pose, from 3D point clouds. Surfaces

Statistics Publications - Radiation research



A Monte Carlo Maximum Likelihood Method for Estimating Uncertainty Arising from Shared Errors in Exposures in Epidemiological Studies of Nuclear Workers

Leslie Stayner,[a,1] Martine Vrijheid,[b] Elisabeth Cardis,[b] Daniel O. Stram,[c] Isabelle Deltour,[b] Stephen J. Gilbert[d] and Geoffrey Howe[e,2]

[a] Division of Epidemiology and Biostatistics, School of Public Health, University of Illinois, Chicago, Illinois 60612-4394; [b] Radiation Group, International Agency for Research on Cancer, 69 372 Lyon Cedex 08, France; [c] Division of Biostatistics and Genetic Epidemiology, Department of Preventive Medicine, University of Southern California, Los Angeles, California 90033; [d] Risk Evaluation Branch, Education and Information Division, National Institute of Occupational Safety and Health, Cincinnati, Ohio; and [e] Department of Epidemiology, Columbia University School of Public Health, New York, New York

---

Stayner, L., Vrijheid, M., Cardis, E., Stram, D. O., Deltour, I., Gilbert, S. J. and Howe, G. A Monte Carlo Maximum Likelihood Method for Estimating Uncertainty Arising from Shared Errors in Exposures in Epidemiological Studies of Nuclear Workers. *Radiat. Res.* 168, 757–763 (2007).

Errors in the estimation of exposures or doses are a major source of uncertainty in epidemiological studies of cancer among nuclear workers. This paper presents a Monte Carlo maximum likelihood method that can be used for estimating a confidence interval that reflects both statistical sampling error and uncertainty in the measurement of exposures. The

to estimate exposures (or doses) for individual subjects, and these estimates are then used in exposure–response analyses. However, the effect of errors in the estimated exposures on the exposure–response relationships has generally been ignored or at best discussed qualitatively.

Errors in elements of the exposure matrix or dosimetry system are a common concern in occupational and environmental studies. These errors result from the use of group means to impute individual exposures (or doses). Individuals sharing the same job in the same facility, or individuals in a similar location, in the same period are frequently assigned the same exposure estimates based on an "exposure

Source: Leslie Stayner et al. (2007) A Monte Carlo Maximum Likelihood Method for Estimating Uncertainty Arising from Shared Errors in Exposures in Epidemiological Studies of Nuclear Workers. Radiation Research, 168(6):757-763. https://doi.org/10.1667/RR0677.1

https://bioone.org/journals/Radiation-Research/volume-168/issue-6/RR0677.1/A-Monte-Carlo-Maximum-Likelihood-Method-for-Estimating-Uncertainty-Arising/10.1667/RR0677.1.full

5. Accuracy of Estimation

Accuracy of estimation: Confidence Intervals

A crucial aspect of statistics is not just to estimate a quantity of interest, but to assess how *accurate* or *precise* that estimate is. One approach is to find an interval, called a *confidence interval* (CI) within which we think the true parameter falls.

Theorem 5.1

Suppose $a_1, \ldots, a_n$ are constants and that $X_1, \ldots, X_n$ are independent with $X_i \sim N(\mu_i, \sigma_i^2)$. Let $Y = \sum_{i=1}^n a_i X_i$. Then

$$Y \sim N\Big(\sum_{i=1}^{n} a_i \mu_i, \sum_{i=1}^{n} a_i^2 \sigma_i^2\Big).$$

Proof: See Sheet 1.

We know from Prelims Probability how to calculate the mean and variance of Y. The additional information here is that Y has a *normal* distribution, i.e. "a linear combination of normals is itself normal."

Example

Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma_0^2)$ where μ is unknown and σ_0^2 is known. What can we say about μ?

By Theorem 5.1,

$$\sum_{i=1}^{n} X_i \sim N(n\mu, \, n\sigma_0^2)$$

$$\overline{X} \sim N\left(\mu, \, \frac{\sigma_0^2}{n}\right).$$

So, standardising $\overline{X}$,

$$\overline{X} - \mu \sim N\left(0, \, \frac{\sigma_0^2}{n}\right)$$

$$\frac{\overline{X} - \mu}{\sigma_0/\sqrt{n}} \sim N(0,1).$$

Now if $Z \sim N(0,1)$, then $P(-1.96 < Z < 1.96) = 0.95$.

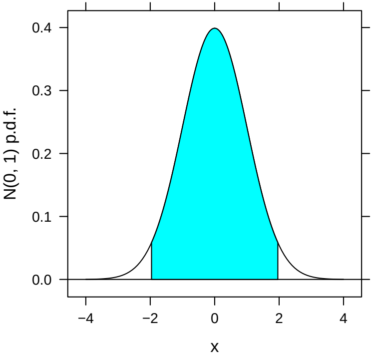


Figure: Standard normal p.d.f.: the shaded area, i.e. the area under the curve from $x = -1.96$ to $x = 1.96$, is 0.95.

So

$$P\left(-1.96 < \frac{\overline{X}-\mu}{\sigma_0/\sqrt{n}} < 1.96\right) = 0.95$$

$$P\left(-1.96\frac{\sigma_0}{\sqrt{n}} < \overline{X}-\mu < 1.96\frac{\sigma_0}{\sqrt{n}}\right) = 0.95$$

$$P\left(\overline{X}-1.96\frac{\sigma_0}{\sqrt{n}} < \mu < \overline{X}+1.96\frac{\sigma_0}{\sqrt{n}}\right) = 0.95$$

$$P\left(\text{the interval } \left(\overline{X} \pm 1.96\frac{\sigma_0}{\sqrt{n}}\right) \text{ contains } \mu\right) = 0.95.$$

Note that we write $\left(\overline{X} \pm 1.96\frac{\sigma_0}{\sqrt{n}}\right)$ to mean the interval

$$\left(\overline{X}-1.96\frac{\sigma_0}{\sqrt{n}},\ \overline{X}+1.96\frac{\sigma_0}{\sqrt{n}}\right).$$

This interval is a *random* interval since its endpoints involve $\overline{X}$ (and $\overline{X}$ is a random variable). It is an example of a *confidence interval*.

Confidence Intervals: general definition

Definition

If $a(\mathbf{X})$ and $b(\mathbf{X})$ are two statistics, and $0 < \alpha < 1$, the interval $\big(a(\mathbf{X}), b(\mathbf{X})\big)$ is called a *confidence interval* for θ with confidence level $1 - \alpha$ if, for all θ,

$$P\big(a(\mathbf{X}) < \theta < b(\mathbf{X})\big) = 1 - \alpha.$$

The interval $\big(a(\mathbf{X}), b(\mathbf{X})\big)$ is also called a $100(1 - \alpha)\%$ CI, e.g. a "95% confidence interval" if $\alpha = 0.05$.

Usually we are interested in small values of α: the most commonly used values are 0.05 and 0.01 (i.e. confidence levels of 95% and 99%) but there is nothing special about any confidence level.

The interval $\big(a(\mathbf{x}), b(\mathbf{x})\big)$ is called an *interval estimate* and the random interval $\big(a(\mathbf{X}), b(\mathbf{X})\big)$ is called an *interval estimator*.

Note: $a(\mathbf{X})$ and $b(\mathbf{X})$ do *not* depend on θ.

We would like to construct $a(\mathbf{X})$ and $b(\mathbf{X})$ so that:

- the width of the interval $\big(a(\mathbf{X}), b(\mathbf{X})\big)$ is small
- the probability $P\big(a(\mathbf{X}) < \theta < b(\mathbf{X})\big)$ is large.

Percentage points of normal distribution

For any α with $0 < \alpha < 1$, let z_α be the constant such that $\Phi(z_\alpha) = 1 - \alpha$, where Φ is the $N(0,1)$ c.d.f. (i.e. if $Z \sim N(0,1)$ then $P(Z > z_\alpha) = \alpha$).

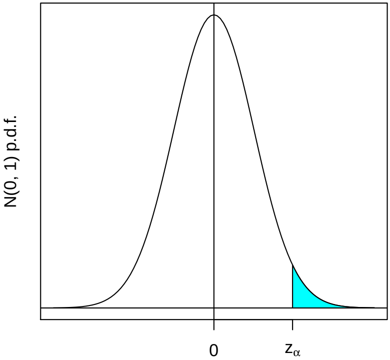


Figure: Standard normal p.d.f.: the shaded area, i.e. the area under the curve to the right of z_α, is α.

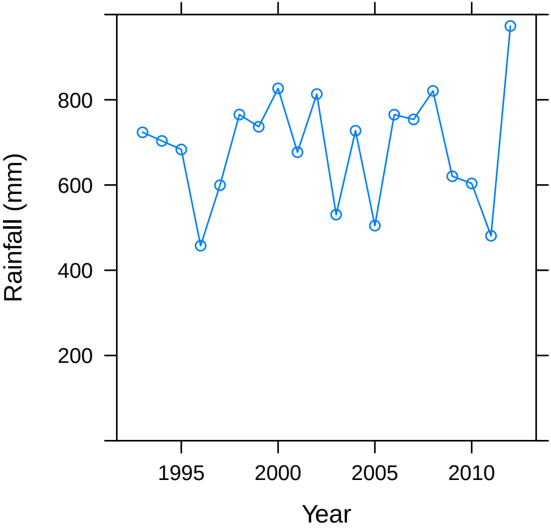
We call z_α the "$1-\alpha$ quantile of $N(0,1)$."

α	0.1	0.05	0.025	0.005
z_α	1.28	1.64	1.96	2.58

By the same argument as before, if $X_1, \dots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma_0^2)$ with σ_0^2 known, then a level $1-\alpha$ confidence interval for μ is

$$\left(\overline{X} \pm \frac{z_{\alpha/2}\sigma_0}{\sqrt{n}}\right).$$

Oxford rainfall, annual, 20 years



Oxford rainfall, annual, 20 years

We have rainfall amounts (in mm) $x_1, \ldots, x_n$ where $n = 20$ and $\overline{x} = 688.4$. We assume $\sigma_0 = 130$.

- The endpoints of a 95% CI are

$$\overline{x} \pm 1.96 \frac{\sigma_0}{\sqrt{n}} = 688.4 \pm 57.0.$$

  So a 95% CI for μ is $(631.4, 745.4)$.

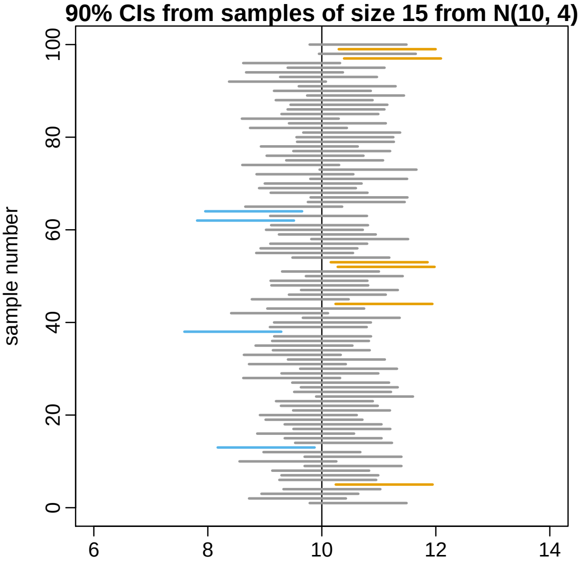
- A 99% CI is $(\overline{x} \pm 2.58 \frac{\sigma_0}{\sqrt{n}}) = (613.4, 763.4)$.

Confidence interval example

When $X_1, \dots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma_0^2)$, a 90% CI for μ is

$$\big(a(\mathbf{X}), b(\mathbf{X})\big) = \left(\overline{X} - 1.64\frac{\sigma_0}{\sqrt{n}},\ \overline{X} + 1.64\frac{\sigma_0}{\sqrt{n}}\right).$$

So a 90% interval estimate based on data $\mathbf{x}$ is $\big(a(\mathbf{x}), b(\mathbf{x})\big)$.


90% CIs from samples of size 15 from N(10, 4)
sample number
0
20
40
60
80
100
6
8
10
12
14

The symmetric confidence interval for μ

$$\overline{x} \pm 1.96 \frac{\sigma_0}{\sqrt{n}}$$

is called a *central* confidence interval for μ.

Suppose now that c and d are any constants such that

$$P(-c < Z < d) = 1 - \alpha$$

for $Z \sim N(0, 1)$.

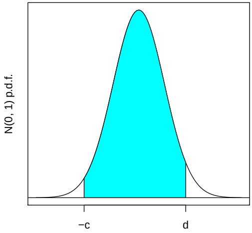


Figure: Standard normal p.d.f.: the shaded area under the curve from $-c$ to d, is $1 - \alpha$.

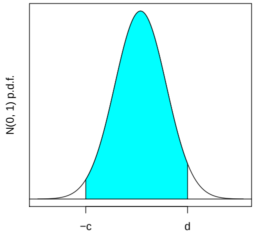


Figure: Standard normal p.d.f.: the shaded area under the curve from $-c$ to d, is $1-\alpha$.

Then

$$P\left(\overline{X} - \frac{d\sigma_0}{\sqrt{n}} < \mu < \overline{X} + \frac{c\sigma_0}{\sqrt{n}}\right) = 1 - \alpha.$$

The choice $c = d = z_{\alpha/2}$ gives the *shortest* such interval.

One-sided confidence limits

Continuing our normal example we have

$$P\left(\frac{\overline{X}-\mu}{\sigma_0/\sqrt{n}} > -z_\alpha\right) = 1-\alpha$$

so

$$P\left(\mu < \overline{X} + \frac{z_\alpha \sigma_0}{\sqrt{n}}\right) = 1-\alpha$$

and so $(-\infty, \overline{X} + \frac{z_\alpha\sigma_0}{\sqrt{n}})$ is a “one-sided” confidence interval. We call $\overline{X} + \frac{z_\alpha\sigma_0}{\sqrt{n}}$ an *upper* $1-\alpha$ *confidence limit* for μ.

Similarly

$$P\left(\mu > \overline{X} - \frac{z_\alpha \sigma_0}{\sqrt{n}}\right) = 1-\alpha$$

and $\overline{X} - \frac{z_\alpha\sigma_0}{\sqrt{n}}$ is a *lower* $1-\alpha$ *confidence limit* for μ.

Interpretation of a Confidence Interval

- The parameter θ is fixed but unknown.
- If we imagine repeating our experiment then we'd get new data, $\mathbf{x}' = (x'_1, \dots, x'_n)$ say, and hence we'd get a new confidence interval $\big(a(\mathbf{x}'), b(\mathbf{x}')\big)$. If we did this repeatedly we would "catch" the true parameter value about 95% of the time, for a 95% confidence interval: i.e. about 95% of our intervals would contain θ.
- The confidence level is a *coverage probability*, the probability that the *random* confidence interval $\big(a(\mathbf{X}), b(\mathbf{X})\big)$ covers the true θ. (It's a random interval because the endpoints $a(\mathbf{X}), b(\mathbf{X})$ are random variables.)

But note that the interval $\big(a(\mathbf{x}), b(\mathbf{x})\big)$ is *not* a random interval, e.g. $\big(a(\mathbf{x}), b(\mathbf{x})\big) = (631.4, 745.4)$ in the rainfall example. So it is wrong to say that $\big(a(\mathbf{x}), b(\mathbf{x})\big)$ contains θ with probability $1 - \alpha$: this interval, e.g. $(631.4, 745.4)$, *either* definitely does *or* definitely does not contain θ, but we can't say which of these two possibilities is true as θ is unknown.

The Central Limit Theorem (CLT)

We know that if $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$ then

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1).$$

Theorem 5.2 Let $X_1, \ldots, X_n$ be i.i.d. from *any* distribution with mean μ and variance $\sigma^2 \in (0, \infty)$. Then, for all x,

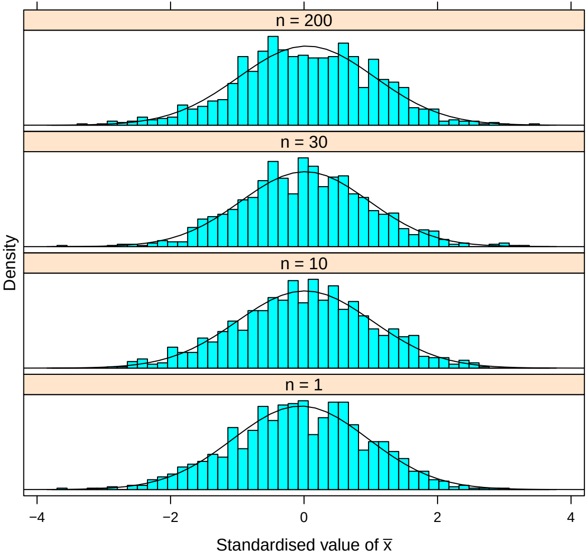
$$P\left(\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \leq x\right) \to \Phi(x) \quad \text{as } n \to \infty.$$

Here, as usual, Φ is the $N(0,1)$ c.d.f. So for large n, *whatever the distribution* of the X_i,

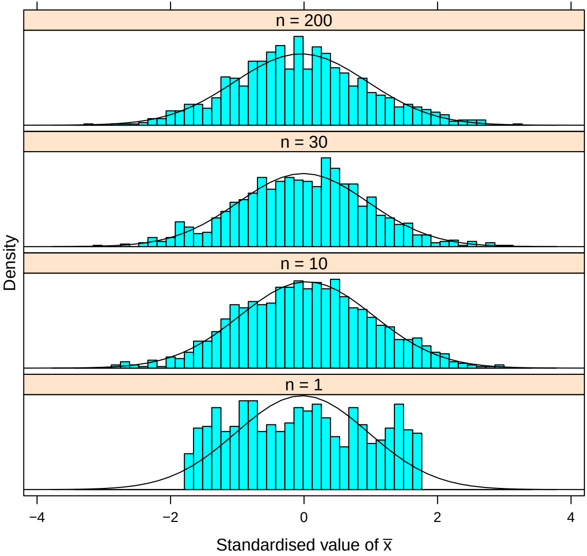
$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \approx N(0,1)$$

where $\approx$ means "has approximately the same distribution as." Usually $n > 30$ is ok for a reasonable approximation.

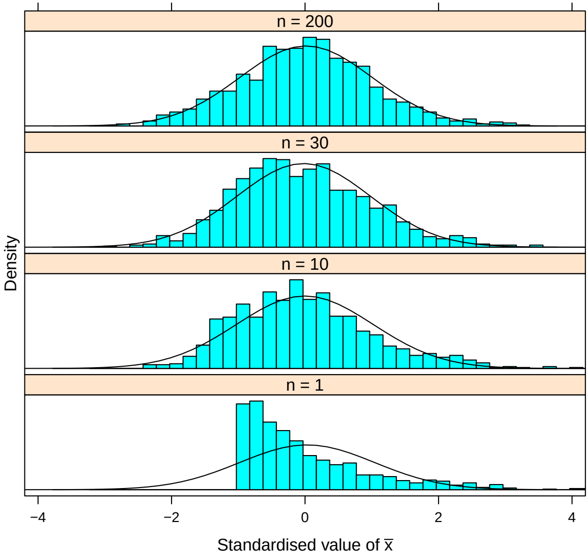
CLT with $X_i \sim N(7, 9)$



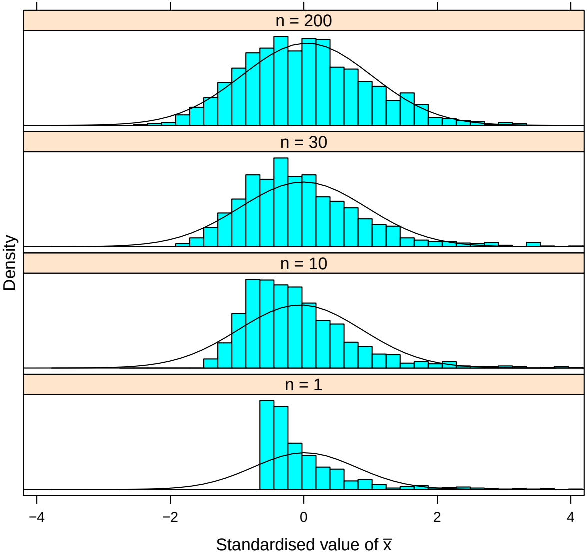
CLT with $X_i \sim U(0,1)$



CLT with $X_i \sim$ Exponential(1)



CLT with $X_i \sim$ Pareto$(1, 3)$



With $X_1, X_2, \ldots$ i.i.d. from *any* distribution with $E(X_i) = \mu$, $\text{var}(X_i) = \sigma^2$:

- the weak law of large numbers (Prelims Probability) tells us that the distribution of $\overline{X}$ concentrates around μ as n becomes large,

  i.e. for $\epsilon > 0$, we have $P(|\overline{X} - \mu| > \epsilon) \to 0$ as $n \to \infty$
- the CLT adds to this
  - the fluctuations of $\overline{X}$ around μ are of order $1/\sqrt{n}$
  - the asymptotic distribution of these fluctuations is normal.

Example Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim}$ exponential mean μ, e.g. X_i = survival time of patient i. So

$$f(x;\mu) = \frac{1}{\mu}e^{-x/\mu}, \quad x \geq 0$$

and $E(X_i) = \mu$, $\operatorname{var}(X_i) = \mu^2$.

For large n, by CLT,

$$\frac{\overline{X} - \mu}{\mu/\sqrt{n}} \approx N(0,1).$$

So

$$P\left(-z_{\alpha/2} < \frac{\overline{X} - \mu}{\mu/\sqrt{n}} < z_{\alpha/2}\right) \approx 1 - \alpha$$

$$P\left(\mu\left(1 - \frac{z_{\alpha/2}}{\sqrt{n}}\right) < \overline{X} < \mu\left(1 + \frac{z_{\alpha/2}}{\sqrt{n}}\right)\right) \approx 1 - \alpha$$

$$P\left(\frac{\overline{X}}{1 + \frac{z_{\alpha/2}}{\sqrt{n}}} < \mu < \frac{\overline{X}}{1 - \frac{z_{\alpha/2}}{\sqrt{n}}}\right) \approx 1 - \alpha.$$

Hence an approximate $1-\alpha$ CI for μ is

$$\left(\frac{\overline{X}}{1+\frac{z_{\alpha/2}}{\sqrt{n}}},\ \frac{\overline{X}}{1-\frac{z_{\alpha/2}}{\sqrt{n}}}\right).$$

Numerically, if we have $n = 100$ patients and $\alpha = 0.05$ (so $z_{\alpha/2} = 1.96$), then

$$(0.84\overline{x}, 1.24\overline{x})$$

is an approximate 95% CI for μ.

Example: Opinion poll

In a opinion poll, suppose 321 of 1003 voters said they would vote for the Party X. What's the underlying level of support for Party X?

Let $X_1, \ldots, X_n$ be a random sample from the Bernoulli(p) distribution, i.e.

$$P(X_i = 1) = p, \quad P(X_i = 0) = 1 - p.$$

The MLE of p is $\overline{X}$. Also $E(X_i) = p$ and $\mathrm{var}(X_i) = p(1-p) = \sigma^2(p)$ say.

For large n, by CLT,

$$\frac{\overline{X} - p}{\sigma(p)/\sqrt{n}} \approx N(0,1).$$

So

$$1 - \alpha \approx P\left(-z_{\alpha/2} < \frac{\overline{X} - p}{\sigma(p)/\sqrt{n}} < z_{\alpha/2} \right)$$

$$= P\left(\overline{X} - z_{\alpha/2}\frac{\sigma(p)}{\sqrt{n}} < p < \overline{X} + z_{\alpha/2}\frac{\sigma(p)}{\sqrt{n}} \right).$$

The interval $\left(\overline{X} \pm \frac{z_{\alpha/2}}{\sqrt{n}}\sigma(p)\right)$ has approximate probability $1 - \alpha$ of containing the true p, but it is *not* a confidence interval since its endpoints depend on p via $\sigma(p)$.

To get an approximate confidence interval:

- either, solve the inequality to get $P\big(a(\mathbf{X}) < p < b(\mathbf{X})\big) \approx 1 - \alpha$ where $a(\mathbf{X}), b(\mathbf{X})$ *don't* depend on p
- or, estimate $\sigma(p)$ by $\sigma(\widehat{p}) = \sqrt{\widehat{p}(1-\widehat{p})}$ where $\widehat{p} = \overline{x}$ the MLE. This gives endpoints

  $$\widehat{p} \pm z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}.$$

  For $n = 1003$ and $\widehat{p} = \overline{x} = 321/1003$, this gives a 95% CI for p of $(0.29, 0.35)$ using the two approximations: (i) CLT and (ii) $\sigma(p)$ approximated by $\sigma(\widehat{p})$.

Opinion polls often mention "±3% error."

Note that for any p,

$$\sigma^2(p) = p(1-p) \leq \frac{1}{4}$$

since $p(1-p)$ is maximised at $p = \frac{1}{2}$. Then we have

$$\begin{aligned} P(\widehat{p} - 0.03 < p < \widehat{p} + 0.03) &= P\left(\frac{-0.03}{\sigma(p)/\sqrt{n}} < \frac{\widehat{p} - p}{\sigma(p)/\sqrt{n}} < \frac{0.03}{\sigma(p)/\sqrt{n}} \right) \\ &\approx \Phi\left(\frac{0.03}{\sigma(p)/\sqrt{n}} \right) - \Phi\left(\frac{-0.03}{\sigma(p)/\sqrt{n}} \right) \\ &\geq \Phi(0.03\sqrt{4n}) - \Phi(-0.03\sqrt{4n}). \end{aligned}$$

For this probability to be at least 0.95 we need $0.03\sqrt{4n} \geq 1.96$, or $n \geq 1068$. Opinion polls typically use $n \approx 1100$.

Standard errors

Definition Let $\widehat{\theta}$ be an estimator of θ based on $\mathbf{X}$. The *standard error* $\mathrm{SE}(\widehat{\theta})$ of $\widehat{\theta}$ is defined by

$$\mathrm{SE}(\widehat{\theta}) = \sqrt{\mathrm{var}(\widehat{\theta})}.$$

Example

- Let $X_1, \dots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$. Then $\widehat{\mu} = \overline{X}$ and $\mathrm{var}(\widehat{\mu}) = \sigma^2/n$. So $\mathrm{SE}(\widehat{\mu}) = \sigma/\sqrt{n}$.
- Let $X_1, \dots, X_n \overset{\text{iid}}{\sim} \mathrm{Bernoulli}(p)$. Then $\widehat{p} = \overline{X}$ and $\mathrm{var}(\widehat{p}) = p(1-p)/n$. So $\mathrm{SE}(\widehat{p}) = \sqrt{p(1-p)/n}$.

Sometimes $\mathrm{SE}(\widehat{\theta})$ depends on θ itself, meaning that $\mathrm{SE}(\widehat{\theta})$ is unknown. In such cases we have to plug-in parameter estimates to get the *estimated* standard error. e.g. plug-in to get estimated standard errors $\mathrm{SE}(\overline{X}) = \widehat{\sigma}/\sqrt{n}$ and $\mathrm{SE}(\widehat{p}) = \sqrt{\widehat{p}(1-\widehat{p})/n}$.

The values plugged-in ($\widehat{\sigma}$ and $\widehat{p}$ above) could be maximum likelihood, or other, estimates.

(We could write $\widehat{\mathrm{SE}}(\widehat{p})$, i.e. with a hat on the SE, to denote that $\widehat{p}$ has been plugged-in, but this is ugly so we won't, we'll just write $\mathrm{SE}(\widehat{p})$.)

If $\widehat{\theta}$ is unbiased, then $\text{MSE}(\widehat{\theta}) = \text{var}(\widehat{\theta}) = [\text{SE}(\widehat{\theta})]^2$. So the standard error (or estimated standard error) gives some quantification of the accuracy of estimation.

If in addition $\widehat{\theta}$ is approximately $N(\theta, \text{SE}(\widehat{\theta})^2)$ then, by the arguments used above, an approximate $1-\alpha$ CI for θ is given by $\big(\widehat{\theta} \pm z_{\alpha/2}\,\text{SE}(\widehat{\theta})\big)$ where again we might need to plug-in to obtain the estimated standard error. Since, roughly, $z_{0.025} = 2$ and $z_{0.001} = 3$,

$$\begin{array}{ll} \text{(estimate} \pm 2 \text{ estimated std errors)} & \text{is an approximate 95\% CI} \\ \text{(estimate} \pm 3 \text{ estimated std errors)} & \text{is an approximate 99.8\% CI.} \end{array}$$

The CLT justifies the normal approximation for $\widehat{\theta} = \overline{X}$, but $\widehat{\theta} \approx N(\theta, \text{SE}(\widehat{\theta})^2)$ is also appropriate for more general MLEs by other theory (see Part A).

Statistics Publications - ADHD in psychiatric patients

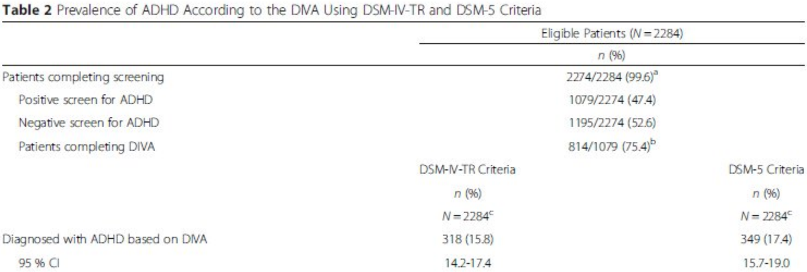


Table 2 Prevalence of ADHD According to the DIVA Using DSM-IV-TR and DSM-5 Criteria

	Eligible Patients (*N* = 2284)	
	n (%)	
Patients completing screening	2274/2284 (99.6)[a]	
Positive screen for ADHD	1079/2274 (47.4)	
Negative screen for ADHD	1195/2274 (52.6)	
Patients completing DIVA	814/1079 (75.4)[b]	
	DSM-IV-TR Criteria	DSM-5 Criteria
	n (%)	*n* (%)
	N = 2284[c]	*N* = 2284[c]
Diagnosed with ADHD based on DIVA	318 (15.8)	349 (17.4)
95 % CI	14.2-17.4	15.7-19.0

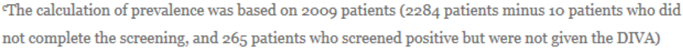
Attention-deficit/hyperactivity disorder diagnoses

Walter Deberdt et al. (2015) Prevalence of ADHD in nonpsychotic adult psychiatric care (ADPSYC): A multinational cross-sectional study in Europe. BMC Psychiatry15:242

https://doi.org/10.1186/s12888-015-0624-5

Statistics Publications - ADHD in psychiatric patients

We can reconstruct the confidence intervals reported for the percentage diagnosed with ADHD. There was an additional note at the bottom of the table:



[c]The calculation of prevalence was based on 2009 patients (2284 patients minus 10 patients who did not complete the screening, and 265 patients who screened positive but were not given the DIVA)

Let us consider those diagnosed using the DSM-IV-TR criteria. There were 2009 patients analysed and it was reported that 15.8% were diagnosed with ADHD.

We know that for $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Bernoulli}(p)$.

$\widehat{p} = \overline{X} = 318/2009 = 0.1582$

$\text{SE}(\widehat{p}) = \sqrt{0.1582(1 - 0.1582)/2009} = 0.0081.$

Using $\widehat{p} \pm 1.96\,\text{SE}(\widehat{p})$ gives the interval (0.1423,0.1742) matching what was reported in the paper.

Example Suppose $X_1, \dots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$ with μ and σ^2 unknown.

The MLEs are $\widehat{\mu} = \overline{X}$, $\widehat{\sigma}^2 = \frac{1}{n} \sum (X_i - \overline{X})^2$, and $\text{SE}(\widehat{\mu}) = \sigma/\sqrt{n}$ is unknown because σ is unknown. So to use $\widehat{\mu} \pm z_{\alpha/2}\, \text{SE}(\widehat{\mu})$ as the basis for a confidence interval we need to estimate σ. One possibility is to use $\widehat{\sigma}$ and so get the interval

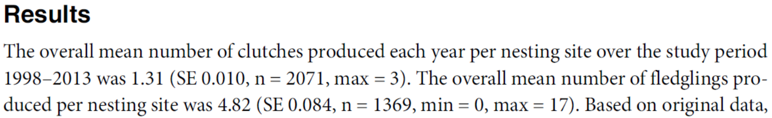
$$\left(\widehat{\mu} \pm z_{\alpha/2} \frac{\widehat{\sigma}}{\sqrt{n}}\right).$$

However we can improve on this:

(i) use s (the sample standard deviation) instead of $\widehat{\sigma}$ (better as $E(S^2) = \sigma^2$ whereas $E(\widehat{\sigma}^2) < \sigma^2$)

(ii) use a critical value from a t-distribution instead of $z_{\alpha/2}$ – see Part A Statistics (better as $(\overline{X} - \mu)/(S/\sqrt{n})$ has a t-distribution, exactly, whereas using the CLT is an approximation).

Statistics Publications - Barn Owl Productivity

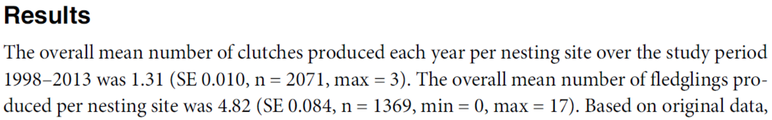


Results

The overall mean number of clutches produced each year per nesting site over the study period 1998–2013 was 1.31 (SE 0.010, n = 2071, max = 3). The overall mean number of fledglings produced per nesting site was 4.82 (SE 0.084, n = 1369, min = 0, max = 17). Based on original data,

Pavluvčík P, Poprach K, Machar I, Losík J, Gouveia A, Tkadlec E (2015) Barn Owl Productivity Response to Variability of Vole Populations. PLoS ONE 10(12): e0145851. https://doi.org/10.1371/journal.pone.0145851

Statistics Publications - Barn Owl Productivity



Results

The overall mean number of clutches produced each year per nesting site over the study period 1998–2013 was 1.31 (SE 0.010, n = 2071, max = 3). The overall mean number of fledglings produced per nesting site was 4.82 (SE 0.084, n = 1369, min = 0, max = 17). Based on original data,

So for clutches we know that $\widehat{\mu} = 1.31$, $n = 2071$ and $SE(\widehat{\mu}) = 0.010$.

Thus, the 95% confidence interval is $(1.31 \pm 1.96 \times 0.010) = (1.29, 1.33)$.

And $\widehat{\sigma} = 0.010 \times \sqrt{2071} = 0.455$.

Pavluvčík et al. 2015 https://doi.org/10.1371/journal.pone.0145851

6. Linear Regression

Linear Regression

Suppose we measure two variables in the same population:

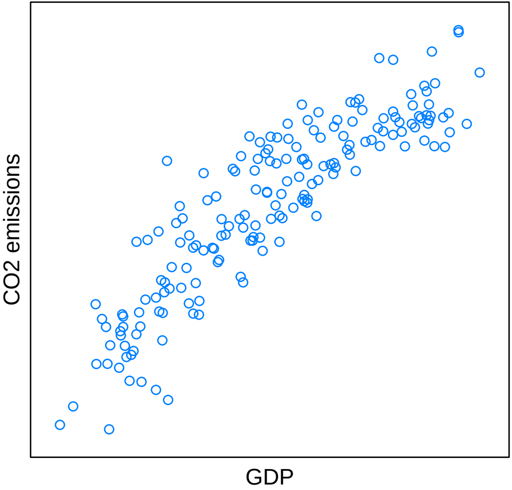
x the *explanatory* variable

y the *response* variable.

Other possible names for x are the *predictor* or *feature* or *input variable* or *independent variable*. Other possible names for y are the *output variable* or *dependent variable*.

CO_2 emissions versus GDP

Let x measure the GDP per head, and y the CO_2 emissions per head, for 178 countries.



Questions of interest:

For fixed x, what is the average value of y?

How does that average value change with x?

A simple model for the dependence of y on x is

$$y = \alpha + \beta x + \text{"error"}.$$

Note: a linear relationship like this does not necessarily imply that x *causes* y.

More precise model

We regard the values of x as being fixed and known, and we regard the values of y as being the observed values of random variables.

We suppose that

$$Y_i = \alpha + \beta x_i + \epsilon_i, \quad i = 1, \ldots, n \tag{2}$$

where

$x_1, \ldots, x_n$ are known constants

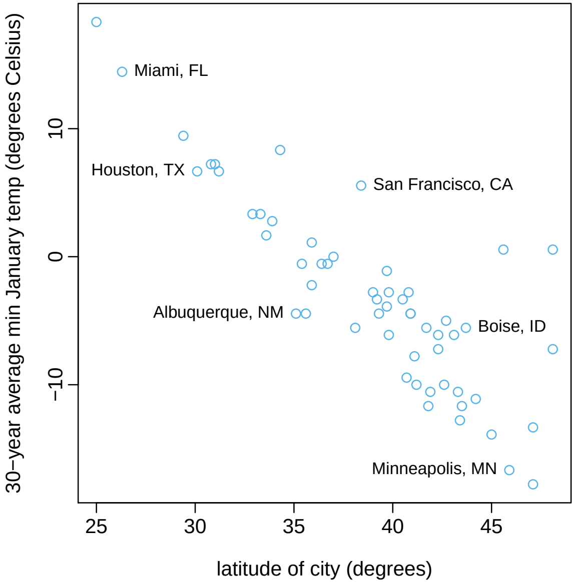
$\epsilon_1, \ldots, \epsilon_n$ are i.i.d. $N(0, \sigma^2)$ "random errors"

α, β are unknown parameters.

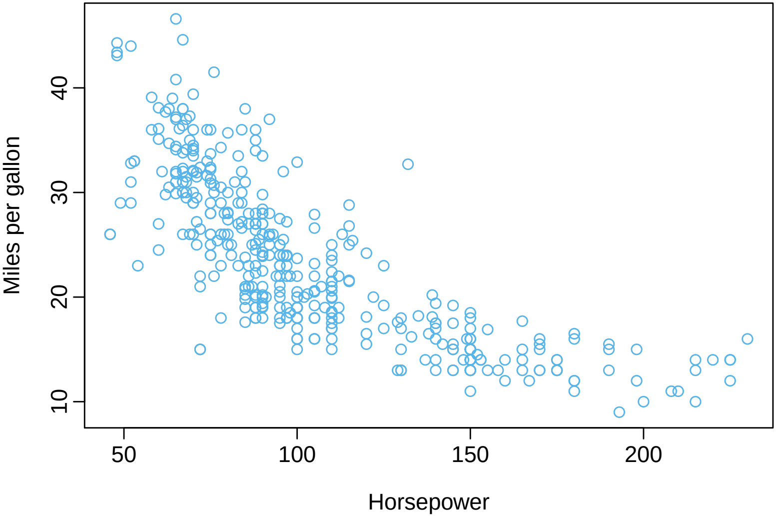
The “random errors” $\epsilon_1, \ldots, \epsilon_n$ represent random scatter of the points (x_i, y_i) about the line $y = \alpha + \beta x$, we do not expect these points to lie on a perfect straight line.

Sometimes we will refer to the above model as being $Y = \alpha + \beta x + \epsilon$.

US city temperature data



Auto data



$Y = \alpha + \beta x + \epsilon$ for the CO_2 example

❶ The values $y_1, \ldots, y_n$ (e.g. the CO_2 emissions per head in the various countries) are the observed values of random variables $Y_1, \ldots, Y_n$.

❷ The values $x_1, \ldots, x_n$ (e.g. the GDP per head in the various countries) do *not* correspond to random variables. They are *fixed* and *known* constants.

Questions:

- How do we estimate α and β?
- Does the mean of Y actually depend on the value of x? i.e. is $\beta \neq 0$?

We now find the MLEs of α and β, and we regard σ^2 as being known. The MLEs of α and β are the same if σ^2 is unknown. If σ^2 is unknown, then we simply maximise over σ^2 as well to obtain its MLE – this is no harder than what we do here (try it!). However, working out all of the properties of this MLE *is* harder and beyond what we can do in this course.

We have $Y_i \sim N(\alpha + \beta x_i, \sigma^2)$. So Y_i has p.d.f.

$$f_{Y_i}(y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y_i - \alpha - \beta x_i)^2 \right), \quad -\infty < y_i < \infty.$$

So the likelihood $L(\alpha, \beta)$ is

$$\begin{aligned} L(\alpha, \beta) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y_i - \alpha - \beta x_i)^2 \right) \\ &= (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2 \right) \end{aligned}$$

with log-likelihood

$$\ell(\alpha, \beta) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2.$$

$$\ell(\alpha, \beta) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2.$$

$$\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2 = \sum_{i=1}^{n}\left(y_i - (\alpha + \beta x_i)\right)^2.$$

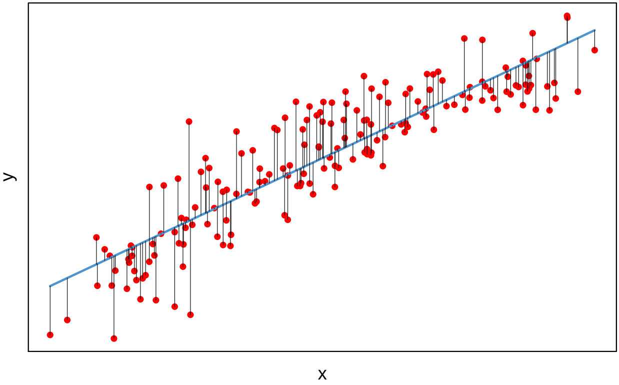
So *maximising* $\ell(\alpha, \beta)$ over α and β is equivalent to *minimising* the sum of squares

$$S(\alpha, \beta) = \sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2.$$

For this reason the MLEs of α and β are also called *least squares estimators*.

What does simple linear regression do?

Minimise the sum of squared vertical distances from the points to the line $y = \alpha + \beta x$.



Theorem 6.1

The MLEs (or, equivalently, the least squares estimates) of α and β are given by

$$\widehat{\alpha} = \frac{(\sum x_i^2)(\sum y_i) - (\sum x_i)(\sum x_i y_i)}{n \sum x_i^2 - (\sum x_i)^2}$$

$$\widehat{\beta} = \frac{n \sum x_i y_i - (\sum x_i)(\sum y_i)}{n \sum x_i^2 - (\sum x_i)^2}.$$

These sums are from $i = 1$ to n.

Proof of Theorem 6.1

To find $\widehat{\alpha}$ and $\widehat{\beta}$ we calculate

$$\frac{\partial S}{\partial \alpha} = -2 \sum (y_i - \alpha - \beta x_i)$$

$$\frac{\partial S}{\partial \beta} = -2 \sum x_i (y_i - \alpha - \beta x_i).$$

Putting these partial derivatives equal to zero, the minimisers $\widehat{\alpha}$ and $\widehat{\beta}$ satisfy

$$n\widehat{\alpha} + \widehat{\beta} \sum x_i = \sum y_i$$

$$\widehat{\alpha} \sum x_i + \widehat{\beta} \sum x_i^2 = \sum x_i y_i.$$

Solving this pair of simultaneous equations for $\widehat{\alpha}$ and $\widehat{\beta}$ gives the required MLEs.

Sometimes we consider the model

$$Y_i = a + b(x_i - \overline{x}) + \epsilon_i, \quad i = 1, \ldots, n$$

and find the MLEs of a and b by minimising $\sum(y_i - a - b(x_i - \overline{x}))^2$. This model is just an alternative parametrisation of our original model: the first model is

$$Y = \alpha + \beta x + \epsilon$$

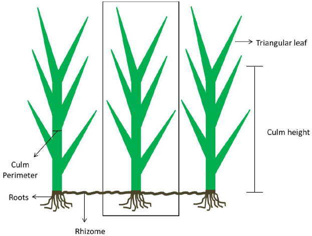
and the second is

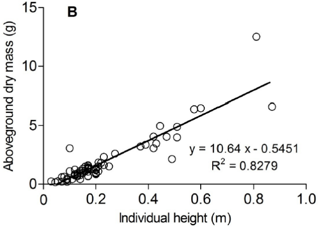
$$\begin{aligned} Y &= a + b(x - \overline{x}) + \epsilon \\ &= (a - b\overline{x}) + bx + \epsilon. \end{aligned}$$

Here Y, x denote general values of Y, x (and $\overline{x} = \frac{1}{n} \sum x_i$ is the mean of the n data values of x). Comparing the two model equations, $b = \beta$ and $a - b\overline{x} = \alpha$.

The interpretation of the parameters is that $\beta = b$ is the increase in $E(Y)$ when x increases by 1. The parameter α is the value of $E(Y)$ when x is 0; whereas a is the value of $E(Y)$ when x is $\overline{x}$.

Statistics Publications - Estimating dry mass in plants

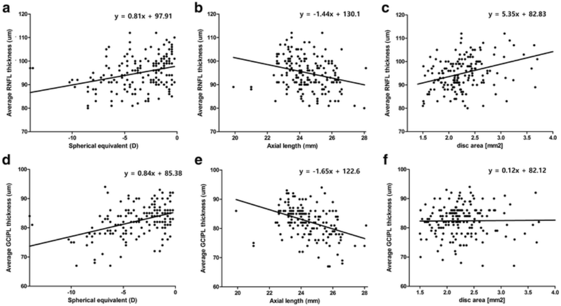




Estimation of the relationship between individual plant height and aboveground dry mass

Laís Samira Correia Nunes, Antonio Fernando Monteiro Camargo (2017) A simple non-destructive method for estimating aboveground biomass of emergent aquatic macrophytes. Acta Limnologica Brasiliensia vol.29 http://dx.doi.org/10.1590/s2179-975x6416

Statistics Publications - Eye measurements



Relationships between the peripapillary retinal nerve fiber layer (RNFL, a,b,c) and the thickness of the ganglion cell-inner plexiform layer (GCIPL, d,e,f) and other eye characteristics (important for evaluation of glaucoma)

Sam Seo et al. (2017) Ganglion cell-inner plexiform layer and retinal nerve fiber layer thickness according to myopia and optic disc area: a quantitative and three-dimensional analysis. BMC Ophthalmology 17:22 https://doi.org/10.1186/s12886-017-0419-1

Alternative expressions for $\widehat{\alpha}$ and $\widehat{\beta}$ are

$$\widehat{\beta} = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sum(x_i - \overline{x})^2} \tag{3}$$

$$= \frac{\sum(x_i - \overline{x})y_i}{\sum(x_i - \overline{x})^2} \tag{4}$$

$$\widehat{\alpha} = \overline{y} - \widehat{\beta}\,\overline{x}.$$

The above alternative for $\widehat{\alpha}$ follows directly from $\frac{\partial S}{\partial \alpha} = 0$.

To obtain the alternatives for $\widehat{\beta}$: Theorem 6.1 gives

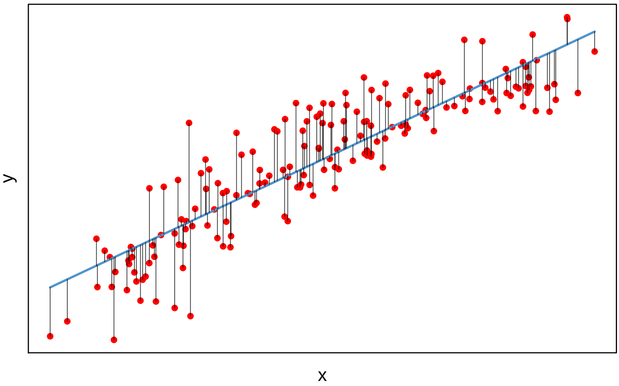
$$\widehat{\beta} = \frac{\sum x_i y_i - \frac{1}{n}(\sum x_i)(\sum y_i)}{\sum x_i^2 - \frac{1}{n}(\sum x_i)^2}$$

$$= \frac{\sum x_i y_i - n\overline{x}\,\overline{y}}{\sum x_i^2 - n\overline{x}^2}. \tag{5}$$

Now check that the numerators and denominators in (3) and (5) are the same. Then observe that the numerators of (3) and (4) differ by $\sum(x_i - \overline{x})\overline{y}$, which is 0.

The *fitted* regression line is the line $y = \widehat{\alpha} + \widehat{\beta}x$.

The point $(\overline{x}, \overline{y})$ always lies on this line.



Bias of regression parameter estimates

Let $w_i = x_i - \overline{x}$ and note $\sum w_i = 0$.

From (4) the maximum likelihood *estimator* of β is

$$\widehat{\beta} = \frac{1}{\sum w_i^2}\Big(\sum w_i Y_i\Big)$$

so

$$\begin{aligned}
E(\widehat{\beta}) &= \frac{1}{\sum w_i^2} E\Big(\sum w_i Y_i\Big) \\
&= \frac{1}{\sum w_i^2} \sum w_i E(Y_i).
\end{aligned}$$

Note $E(Y_i) = \alpha + \beta x_i = \alpha + \beta\overline{x} + \beta w_i$ (using $x_i = w_i + \overline{x}$), and so

$$\begin{aligned}
E(\widehat{\beta}) &= \frac{1}{\sum w_i^2} \sum w_i(\alpha + \beta\overline{x} + \beta w_i) \\
&= \frac{1}{\sum w_i^2} \Big[(\alpha + \beta\overline{x}) \sum w_i + \beta \sum w_i^2\Big] = \beta
\end{aligned}$$

since $\sum w_i = 0$ so $\widehat{\beta}$ is unbiased.

Bias of regression parameter estimates

Also $\widehat{\alpha} = \overline{Y} - \widehat{\beta}\,\overline{x}$ and so

$$\begin{aligned} E(\widehat{\alpha}) &= E(\overline{Y}) - \overline{x}E(\widehat{\beta}) \\ &= \frac{1}{n}\sum E(Y_i) - \beta\overline{x} \qquad \text{since } E(\widehat{\beta}) = \beta \\ &= \frac{1}{n}\sum(\alpha + \beta x_i) - \beta\overline{x} \\ &= \frac{1}{n}\cdot n(\alpha + \beta\overline{x}) - \beta\overline{x} \\ &= \alpha. \end{aligned}$$

So $\widehat{\alpha}$ and $\widehat{\beta}$ are unbiased.

Note the unbiasedness of $\widehat{\alpha}, \widehat{\beta}$ does **not** depend on the assumptions that the ϵ_i are independent, normal and have the same variance, only on the assumptions that the errors are additive and $E(\epsilon_i) = 0$.

Variance of regression parameter estimates

We are usually only interested in the variance of $\widehat{\beta}$:

$$\begin{aligned}
\operatorname{var}(\widehat{\beta}) &= \operatorname{var}\left(\frac{\sum w_i Y_i}{\sum w_i^2}\right) \\
&= \frac{1}{(\sum w_i^2)^2} \operatorname{var}\left(\sum w_i Y_i\right) \\
&= \frac{1}{(\sum w_i^2)^2} \sum w_i^2 \operatorname{var}(Y_i) \\
&= \frac{1}{(\sum w_i^2)^2} \sum w_i^2 \sigma^2 \\
&= \frac{\sigma^2}{\sum w_i^2}.
\end{aligned}$$

Since $\widehat{\beta}$ is a linear combination of the independent normal random variables Y_i, the estimator $\widehat{\beta}$ is itself normal: $\widehat{\beta} \sim N(\beta, \sigma_\beta^2)$ where $\sigma_\beta^2 = \sigma^2 / \sum w_i^2$.

So the standard error of $\widehat{\beta}$ is σ_β and if σ^2 is known, then a 95% CI for β is

$$(\widehat{\beta} \pm 1.96\sigma_\beta).$$

However, this is only a valid CI when σ^2 is known and, in practice, σ^2 is rarely known.

For σ^2 unknown we need to plug-in an estimate of σ^2, i.e. use $\widehat{\sigma}_\beta^2 = \widehat{\sigma}^2 / \sum w_i^2$ where $\widehat{\sigma}^2$ is some estimate of σ^2. For example we could use the MLE which is $\widehat{\sigma}^2 = \frac{1}{n} \sum (y_i - \widehat{\alpha} - \widehat{\beta} x_i)^2$. Using the $\widehat{\theta} \pm 2\,\mathrm{SE}(\widehat{\theta})$ approximation for a 95% confidence interval, we have that $(\widehat{\beta} \pm 2\widehat{\sigma}_\beta)$ is an approximate 95% confidence interval for β.

A better approach here, but beyond the scope of this course, is to estimate σ^2 using

$$s^2 = \frac{1}{n-2} \sum (y_i - \widehat{\alpha} - \widehat{\beta} x_i)^2$$

and to base the confidence interval on a t-distribution rather than a normal distribution. This estimator S^2 is unbiased for σ^2 (see Sheet 5), but details about its distribution and the t-distribution are beyond this course – see Parts A/B.

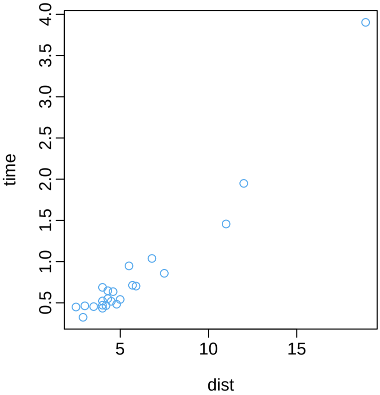
7. Multiple Linear Regression

Multiple linear regression – hill races data

Below are record times (in hours) for 23 hill races together with race distance (in miles) and amount of climb (in feet).

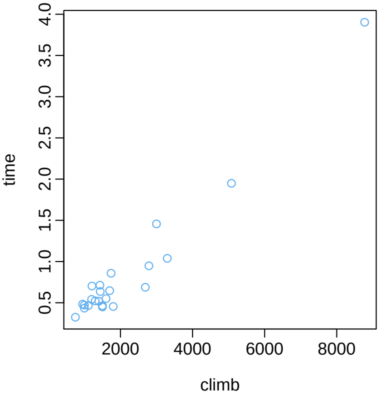
```
                        dist climb   time
Binevenagh               7.5  1740 0.8583
Slieve Gullion           4.2  1110 0.4667
Glenariff Mountain       5.9  1210 0.7031
Donard & Commedagh       6.8  3300 1.0386
McVeigh Classic          5.0  1200 0.5411
Tollymore Mountain       4.8   950 0.4833
Slieve Martin            4.3  1600 0.5506
Moughanmore              3.0  1500 0.4636
Hen & Cock               2.5  1500 0.4497
Annalong Horseshoe      12.0  5080 1.9492
Monument Race            4.0  1000 0.4717
Loughshannagh Horseshoe  4.3  1700 0.6469
Rocky                    4.0  1300 0.5231
Meelbeg Meelmore         3.5  1800 0.4544
Donard Forest            4.5  1400 0.5186
Slieve Donard            5.5  2790 0.9483
Flagstaff to Carling    11.0  3000 1.4569
Slieve Bearnagh          4.0  2690 0.6878
Seven Sevens            18.9  8775 3.9028
Lurig Challenge          4.0  1000 0.4347
Scrabo Hill Race         2.9   750 0.3247
Slieve Gallion           4.6  1440 0.6361
BARF Turkey Trot         5.7  1430 0.7131
```

Scatterplots



- Time increases with distance and also with amount of climb.
- Most data points are crowded into the bottom LH corner of each plot – we could transform `time`, `dist`, `climb` before fitting a model (and we will).

Scatterplots



- Time increases with distance and also with amount of climb.
- Most data points are crowded into the bottom LH corner of each plot – we could transform `time`, `dist`, `climb` before fitting a model (and we will).

Multiple linear regression

For the Hill races example, we could consider: Let Y = time, x_1 = distance, x_2 = climb.

We can consider a model in which Y depends on both x_1 and x_2, of the form

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon.$$

This model has one response variable Y (as usual), but now we have *two* explanatory variables x_1 and x_2, and *three* regression parameters $\beta_0, \beta_1, \beta_2$.

Let the ith race have time y_i, distance x_{i1} and climb x_{i2}. Then in more detail our model is

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i, \quad i = 1, \dots, n$$

where

$$\begin{aligned} &x_{i1}, x_{i2}, \text{ for } i = 1, \dots, n, \text{ are known constants} \\ &\epsilon_1, \dots, \epsilon_n \overset{\text{iid}}{\sim} N(0, \sigma^2) \\ &\beta_0, \beta_1, \beta_2 \text{ are unknown parameters} \end{aligned}$$

and (as usual) y_i denotes the observed value of the random variable Y_i.

As for simple linear regression we obtain the MLEs/least squares estimates of $\beta_0, \beta_1, \beta_2$ by minimising

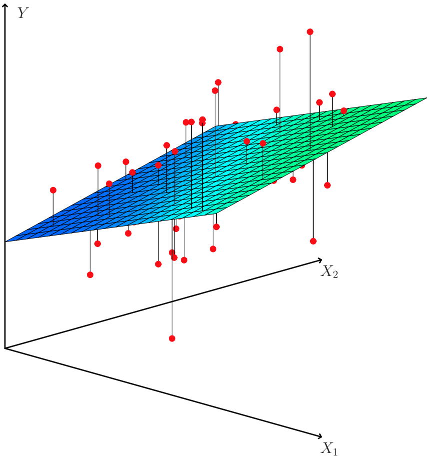
$$S(\beta) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2})^2$$

with respect to $\beta_0, \beta_1, \beta_2$, i.e. by solving $\frac{\partial S}{\partial \beta_k} = 0$ for $k = 0, 1, 2$.

As before, the only property of the ϵ_i needed to define the least squares estimates is $E(\epsilon_i) = 0$.

What does multiple linear regression do?

For two explanatory variables: minimise the sum of squared vertical distances from the points to the plane $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$.



[Figure from James, Witten, Hastie and Tibshirani (2013).]

Hill races data

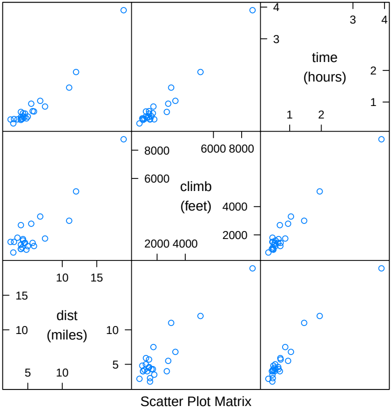
Maindonald and Braun (2010) suggest taking logarithms and considering $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$ where

$$Y = \log(\texttt{time}), \;\; x_1 = \log(\texttt{dist}), \;\; x_2 = \log(\texttt{climb}).$$

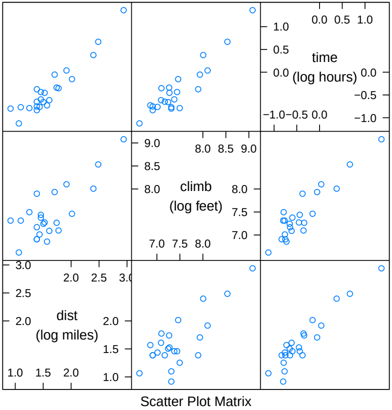
A couple of reasons why we might transform like this:

- The max value of `time` is more than 10 times the min value, with times bunched up towards zero, and with a long tail, similarly for `dist`, `climb`. A log-transform will lead to a more symmetric spread.
- The longest race has a `time` more than double the next largest, and the `dist` and `climb` of this race stand out similarly. Using untransformed variables this race will have much more say in determining the fitted model than any other, taking logs will reduce this. (Can we be more precise than "will have much more say"? Yes, see later.)

Untransformed scales



Logarithmic scales



The fitted model is

$$\begin{aligned} y &= -4.96 + 0.68 x_1 + 0.47 x_2 \\ \log(\texttt{time}) &= -4.96 + 0.68 \log(\texttt{dist}) + 0.47 \log(\texttt{climb}) \\ \texttt{time} &= 0.0070 \times \texttt{dist}^{0.68} \times \texttt{climb}^{0.47} \end{aligned}$$

Some interpretation: for a given value of climb, if the distance doubles then time increases by a factor of $2^{0.68} = 1.60$. Is this reasonable? If the distance doubles, don't we expect the time taken to be multiplied by more than 2?

The key thing is that `climb` is being held constant – so although doubling the length seems to make a race more than twice as hard, the gradient is halved (if climb is held constant), making the climbing aspect of the race easier. The estimated effect overall is the factor of 1.60.

- We can fit a model depending on distance only

$$\log(\texttt{time}) = -2.21 + 1.12 \log(\texttt{dist})$$

- or depending on climb only

$$\log(\texttt{time}) = -7.10 + 0.90 \log(\texttt{climb})$$

- or depending on both distance and climb

$$\log(\texttt{time}) = -4.96 + 0.68 \log(\texttt{dist}) + 0.47 \log(\texttt{climb}).$$

Observe that the estimated regression coefficients are different in each model, e.g. the coefficient of log(`dist`) is different depending on whether log(`climb`) is part of the model – because if log(`climb`) is included then it can help explain some of the variation in log(`time`), whereas if log(`climb`) is absent then log(`dist`) has to account for the variation in log(`time`) on its own.

A general multiple regression model has p explanatory variables $(x_1, \ldots, x_p)$,

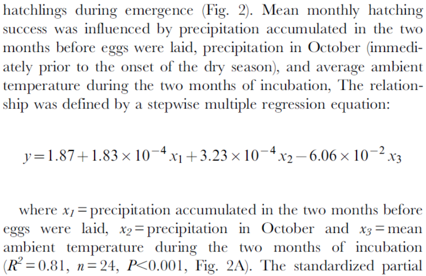
$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon$$

and the MLEs/least squares estimates are obtained by minimising

$$\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_p x_{ip})^2$$

with respect to $\beta_0, \beta_1, \ldots, \beta_p$. In this course we will focus on $p = 1$ or 2.

Statistics Publications - Multiple regression for monthly hatching success of leatherback turtles



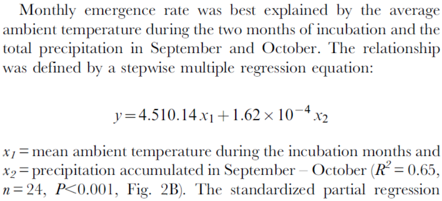
hatchlings during emergence (Fig. 2). Mean monthly hatching success was influenced by precipitation accumulated in the two months before eggs were laid, precipitation in October (immediately prior to the onset of the dry season), and average ambient temperature during the two months of incubation, The relationship was defined by a stepwise multiple regression equation:

$$y = 1.87 + 1.83 \times 10^{-4} x_1 + 3.23 \times 10^{-4} x_2 - 6.06 \times 10^{-2} x_3$$

where x_1 = precipitation accumulated in the two months before eggs were laid, x_2 = precipitation in October and x_3 = mean ambient temperature during the two months of incubation ($R^2 = 0.81$, $n = 24$, $P < 0.001$, Fig. 2A). The standardized partial

Santidrián Tomillo P, Saba VS, Blanco GS, Stock CA, Paladino FV, Spotila JR (2012) Climate Driven Egg and Hatchling Mortality Threatens Survival of Eastern Pacific Leatherback Turtles. PLoS ONE 7(5): e37602. https://doi.org/10.1371/journal.pone.0037602

Statistics Publications - Multiple regression for monthly emergence rate of leatherback turtles



Monthly emergence rate was best explained by the average ambient temperature during the two months of incubation and the total precipitation in September and October. The relationship was defined by a stepwise multiple regression equation:

$$y = 4.510.14\, x_1 + 1.62 \times 10^{-4}\, x_2$$

x_1 = mean ambient temperature during the incubation months and x_2 = precipitation accumulated in September – October ($R^2 = 0.65$, $n = 24$, $P < 0.001$, Fig. 2B). The standardized partial regression

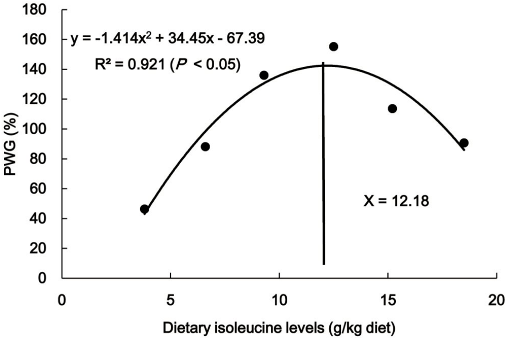
Note the absence of an intercept term. What does this imply?

Santidrián Tomillo P, Saba VS, Blanco GS, Stock CA, Paladino FV, Spotila JR (2012) Climate Driven Egg and Hatchling Mortality Threatens Survival of Eastern Pacific Leatherback Turtles. PLoS ONE 7(5): e37602. https://doi.org/10.1371/journal.pone.0037602

Quadratic regression

The relationship between Y and x may be approximately quadratic in which case we can consider the model $Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$. This is the case $p = 2$ with $x_1 = x$ and $x_2 = x^2$.

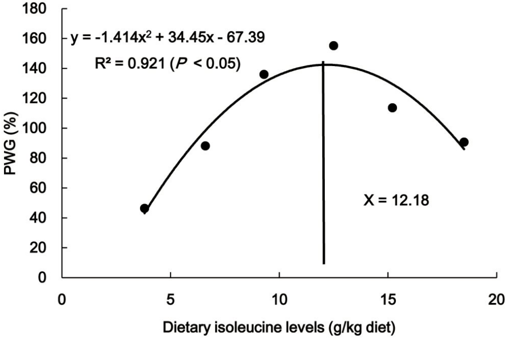
Statistics Publications - Weight gain and diet in fish



Quadratic regression analysis of Percentage Weight Gain (PWG) for grass carp fed graded levels of isoleucine.

Source: Gan L, Jiang W-D, Wu P, Liu Y, Jiang J, Li S-H, et al. (2014) Flesh Quality Loss in Response to Dietary Isoleucine Deficiency and Excess in Fish: ... .*PLoS ONE* 9(12): e115129. https://doi.org/10.1371/journal.pone.0115129

Statistics Publications - Weight gain and diet in fish

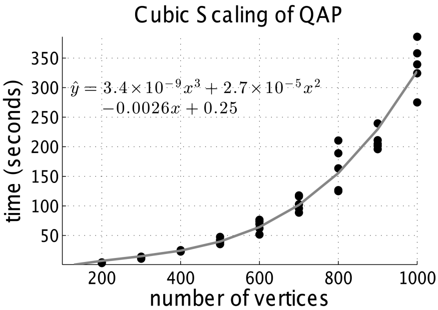


Source: Gan et al. 2014 https://doi.org/10.1371/journal.pone.0115129

What level of isoleucine will maximise the Percentage Weight Gain?

If $y = -1.414x^2 + 34.45x - 67.39$, then the maximum is found where $2 \times -1.414x + 34.45 = 0$ so $x = \frac{34.45}{2\times 1.414} = 12.18$.

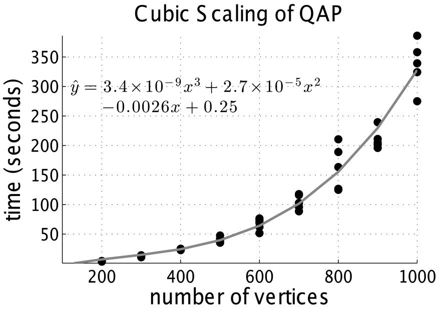
Statistics Publications - Cubic regression



Cubic regression analysis of the run time for a fast approximate quadratic assignment algorithm as a function of the number of vertices.

Source: Vogelstein JT, Conroy JM, Lyzinski V, Podrazik LJ, Kratzer SG, Harley ET, et al. (2015) Fast Approximate Quadratic Programming for Graph Matching. PLoS ONE 10(4): e0121002. https://doi.org/10.1371/journal.pone.0121002

Statistics Publications - Cubic regression



Cubic regression analysis of the run time for a fast approximate quadratic assignment algorithm as a function of the number of vertices.

Source: Vogelstein et al. 2015 https://doi.org/10.1371/journal.pone.0121002

What is the prediction of the run time if there are 50 vertices?

Does the variation in the observation appear to be constant?

For convenience write the two explanatory variables as x and z. So suppose

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \epsilon_i, \quad i = 1, \ldots, n$$

where $\epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$ and assume $\sum x_i = \sum z_i = 0$.

Then minimising $S(\beta)$ gives

$$\begin{aligned}
\widehat{\beta}_0 &= \frac{\sum y_i}{n} \\
\widehat{\beta}_1 &= \frac{1}{\Delta}\left(\sum z_i^2 \sum x_i y_i - \sum x_i z_i \sum z_i y_i\right) \\
\widehat{\beta}_2 &= \frac{1}{\Delta}\left(\sum x_i^2 \sum z_i y_i - \sum x_i z_i \sum x_i y_i\right)
\end{aligned}$$

where

$$\Delta = \sum x_i^2 \sum z_i^2 - \left(\sum x_i z_i\right)^2.$$

The method (solving $\frac{\partial S}{\partial \beta_k} = 0$ for $k = 0, 1, 2$, i.e. 3 equations in 3 unknowns) is straightforward, the algebra less so – the are more elegant ways to do some of this (using matrices, in 3rd year).

Interpretation of regression coefficients

Consider the model with $p = 2$, so two regressors x_1 and x_2:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon.$$

We interpret β_1 as the average effect on Y of a one unit increase in x_1, *holding the other regressor x_2 fixed*.

Similarly we interpret β_2 as the average effect on Y of a one unit increase in x_2, *holding the other regressor x_1 fixed*.

In these interpretations "average" means we are talking about the change in $E(Y)$ when changing x_1 (or x_2).

One important thing to note here is that x_1 and x_2 often change together. e.g. in the hill races example, a race whose distance is one mile greater will usually to have an increased value of climb as well. This makes interpretation more difficult.

8. Assessing the fit of a model

Assessing the fit of a model

Having fitted a model, we should consider how well it fits the data. A model is normally an approximation to reality: is the approximation sufficiently good that the model is useful? This question applies to mathematical models in general. In this course we will approach the question by considering the fit of a simple linear regression (generalisations are possible).

For the model $Y = \alpha + \beta x + \epsilon$ let $\widehat{\alpha}, \widehat{\beta}$ be the usual estimates of α, β based on the observation pairs $(x_1, y_1), \dots, (x_n, y_n)$.

From now on we consider this model, with the usual assumptions about ϵ, unless otherwise stated.

Definition The ith *fitted value* $\widehat{y}_i$ of Y is defined by $\widehat{y}_i = \widehat{\alpha} + \widehat{\beta} x_i$, for $i = 1, \dots, n$.

The ith *residual* e_i is defined by $e_i = y_i - \widehat{y}_i$, for $i = 1, \dots, n$.

The *residual sum of squares* RSS is defined by $\mathrm{RSS} = \sum e_i^2$.

The *residual standard error* RSE is defined by $\mathrm{RSE} = \sqrt{\frac{1}{n-2}\mathrm{RSS}}$.

The RSE is an estimate of the standard deviation σ. If the fitted values are close to the observed values, i.e. $\widehat{y}_i \approx y_i$ for all i (so that the e_i are small), then the RSE will be small. Alternatively if one or more of the e_i is large then the RSE will be higher.

We have $E(e_i) = 0$. In taking this expectation, we treat y_i as the random variable Y_i, and we treat $\widehat{y}_i$ as the random variable $\widehat{\alpha} + \widehat{\beta} x_i$ (in particular, $\widehat{\alpha}$ and $\widehat{\beta}$ are estimators, not estimates). Hence

$$\begin{aligned} E(e_i) &= E(Y_i - \widehat{\alpha} - \widehat{\beta} x_i) \\ &= E(Y_i) - E(\widehat{\alpha}) - E(\widehat{\beta}) x_i \\ &= E(\alpha + \beta x_i + \epsilon_i) - \alpha - \beta x_i \quad \text{since } \widehat{\alpha}, \widehat{\beta} \text{ are unbiased} \\ &= \alpha + \beta x_i + E(\epsilon_i) - \alpha - \beta x_i \\ &= 0 \quad \text{since } E(\epsilon_i) = 0. \end{aligned}$$

Potential problem: non-linearity

The model $Y = \alpha + \beta x + \epsilon$ assumes a straight-line relationship between Y (the response) and x (the predictor). If the true relationship is far from linear then any conclusions (e.g. predictions) we draw from the fit will be suspect.

A *residual plot* is a useful graphical tool for identifying non-linearity: for simple linear regression we can plot the residuals e_i against the fitted values $\widehat{y}_i$. Ideally the plot will show no pattern. The existence of a pattern may indicate a problem with some aspect of the linear model.

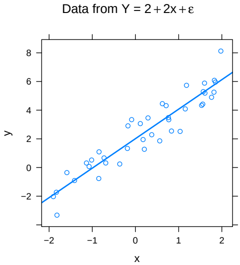
Note that for simple linear regression (i.e. the case $p = 1$) plotting e_i against x_i gives an equivalent plot, just with a different horizontal scale, since there is an exact linear relation between the x_i and $\widehat{y}_i$ (i.e. $\widehat{y}_i = \widehat{\alpha} + \widehat{\beta} x_i$).

[Plotting e_i against $\widehat{y}_i$ generalises better to multiple regression.]

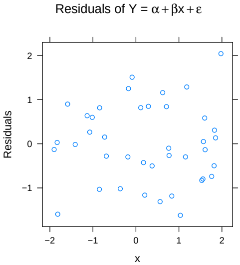
Residual plots

True model is $Y = 2 + 2x + \epsilon$. Fitted model is $Y = \alpha + \beta x + \epsilon$.

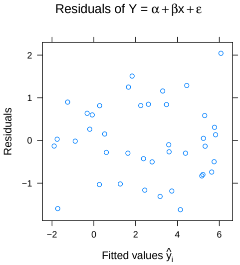
The right form of model has been fitted, the residual plot should show no pattern, just random scatter – it does.



Data from Y = 2 + 2x + ε

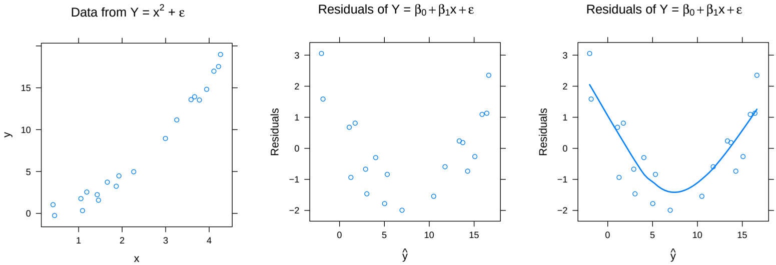


Residuals of Y = α + βx + ε



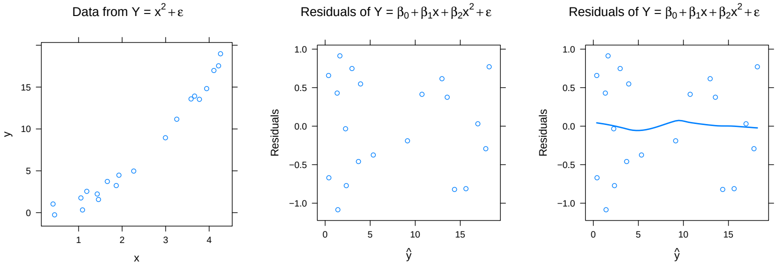
Residuals of Y = α + βx + ε

Plotting residuals against x_i, or against $\widehat{y}_i$, is basically the same (since $\widehat{y}_i = \widehat{\alpha} + \widehat{\beta} x_i$).



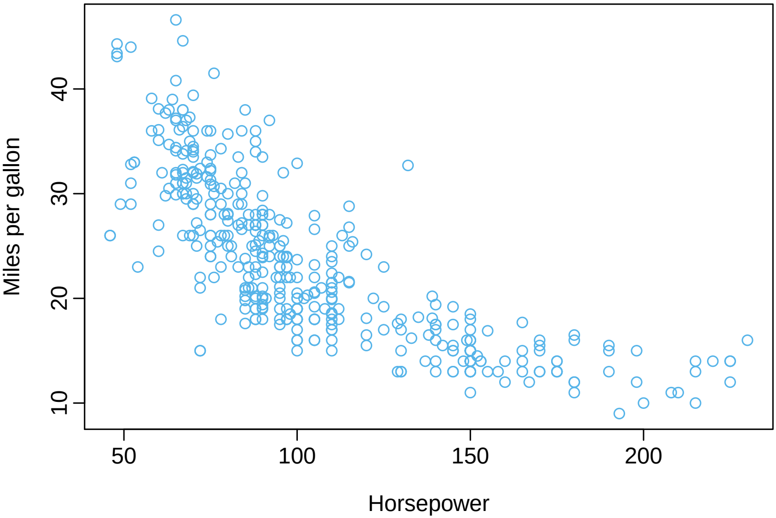
A straight-line model $Y = \beta_0 + \beta_1 x + \epsilon$ has been fitted when actually the relationship is quadratic $Y = x^2 + \epsilon$.

The residuals should indicate a problem – they do – there is a pattern, they are not randomly scattered. The curvature indicated in the right-hand plot is what we should notice in the middle plot. [How to fit the curve is beyond the scope of this course.]

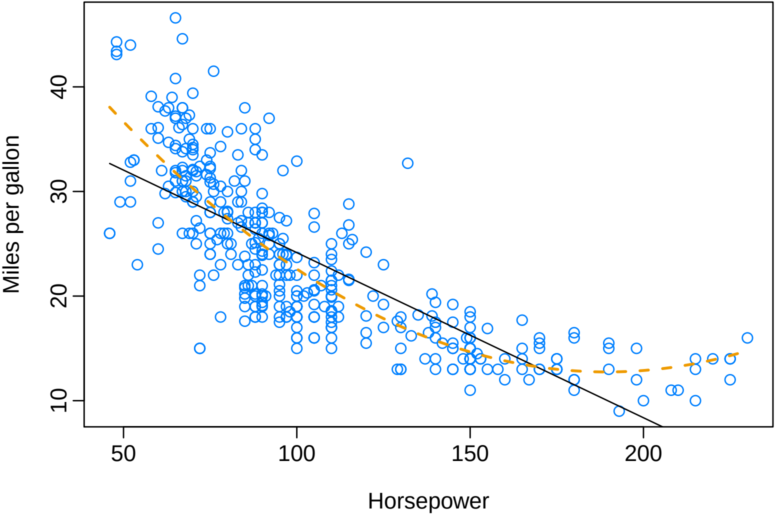


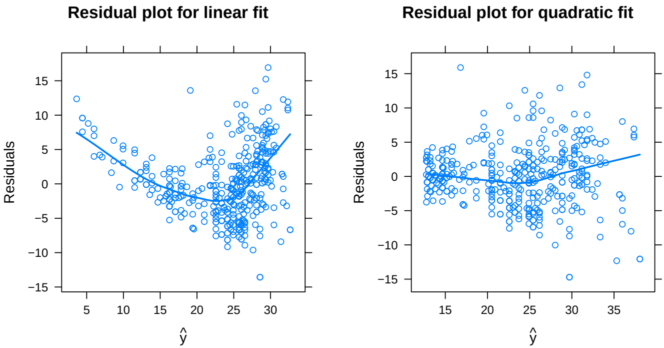
True model is $Y = x^2 + \epsilon$. Fitted model is $Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$, i.e. the correct form of model has been fitted. The residuals don't show any pattern.

Auto data



Auto data, linear fit and quadratic fit





Left: the pattern (curvature) in the residuals from the linear fit $Y = \beta_0 + \beta_1 x + \epsilon$ indicates non-linearity.

Right: little pattern remains in the residuals from the quadratic fit $Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$.

Potential problem: non-constant variance of errors

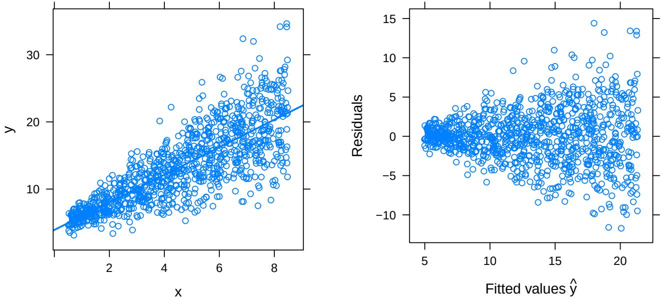
We have assumed that the errors have a constant variance, i.e. $\text{var}(Y_i) = \text{var}(\epsilon_i) = \sigma^2$. That is, the *same* variance *for all* i. Unfortunately, this is often not true. e.g. the variance of the error may increase as Y increases.

Non-constant variance is also called *heteroscedasticity*. We can identify this from the presence of a funnel-type shape in the residual plot.

Non-constant variance of errors

Plots with response variable Y, so the model is $Y = \alpha + \beta x + \epsilon$.

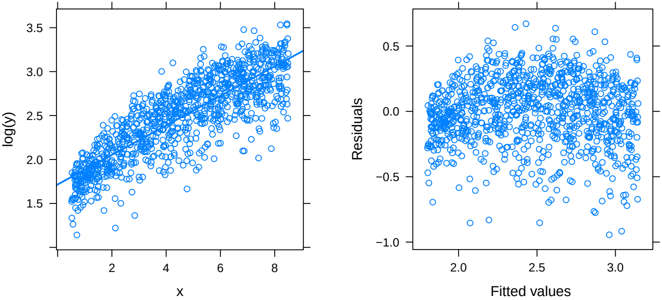
Both plots suggest non-constant variance of the error, the variability appears larger when Y (or x) is larger.



How might we deal with non-constant variance of the errors?

One possibility is to transform the response Y using a transformation such as $\log Y$ or $\sqrt{Y}$ (which shrinks larger responses more), leading to a reduction in heteroscedasticity.

Plots with $\log Y$ as the response variable, so the model is $\log Y = \alpha + \beta x + \epsilon$.



Sometimes we might have a good idea about of the variance of Y_i: we might think $\mathrm{var}(Y_i) = \mathrm{var}(\epsilon_i) = \frac{\sigma^2}{w_i}$ where σ^2 is unknown but where the w_i are known. e.g. if Y_i is actually the mean of n_i observations, where each of these n_i observations are made at $x = x_i$, then $\mathrm{var}(Y_i) = \frac{\sigma^2}{n_i}$. So $w_i = n_i$ in this case.

It is straightforward to show (exercise) that the MLEs of α, β are obtained by minimising

$$\sum_{i=1}^{n} w_i(y_i - \alpha - \beta x_i)^2. \tag{6}$$

The form of (6) is intuitively correct: if w_i is small then $\mathrm{var}(Y_i)$ is large, so there is a lot of uncertainty about observation i, so this observation shouldn't affect the fit too much – this is achieved in (6) by observation i being weighted by the small value of w_i. Hence this approach is called *weighted least squares*.

Potential problem: outliers

An *outlier* is a point for which y_i is far from the value $\widehat{y}_i$ predicted by the model. Outliers can arise for a variety of reasons, e.g. incorrect recording of an observation during data collection.

Residual plots can be used to identify outliers. Recall that $E(e_i) = 0$. But in practice it can be difficult to decide how large (i.e. how far from the expected value of zero) a residual needs to be before we consider it a possible outlier. To address this we can plot *studentized residuals* instead of residuals, where

$$i\text{th studentized residual} = \frac{e_i}{\mathrm{SE}(e_i)}.$$

Theorem 8.1 $\mathrm{var}(e_i) = \sigma^2(1 - h_i)$ where

$$h_i = \frac{1}{n} + \frac{(x_i - \overline{x})^2}{\sum_{j=1}^{n}(x_j - \overline{x})^2}.$$

Definition The ith *studentized residual* r_i is defined by

$$r_i = \frac{e_i}{s\sqrt{1 - h_i}}$$

where s = RSE is the residual standard error.

[Here we call the r_i "studentized" residuals. Some authors call the r_i "standardized" residuals and save the word "studentized" to mean something similar but different.]

So the r_i are all on a comparable scale, each having a standard deviation of about 1. We will say that observations with $|r_i| > 3$ are possible outliers.

If we believe an outlier is due to an error in data collection, then one solution is to simply remove the observation from the data and re-fit the model. However, an outlier may instead indicate a problem with the model, e.g. a nonlinear relationship between Y and x, so care must be taken.

Similarly, this kind of problem could arise if we have a missing regressor, i.e. we could be using $Y = \beta_0 + \beta_1 x_1 + \epsilon$ when we should really be using $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$.

Proof of Theorem 8.1

Idea of the proof: write $e_j = \sum_j a_j Y_j$ and use

$$\operatorname{var}\left(\sum_j a_j Y_j\right) = \sum_j a_j^2 \operatorname{var}(Y_j) = \sigma^2 \sum_j a_j^2 \tag{7}$$

since the Y_j are independent with $\operatorname{var}(Y_j) = \sigma^2$ for all j. Here and below, all sums are from 1 to n.

First recall

$$\widehat{\beta} = \frac{\sum_j (x_j - \overline{x}) Y_j}{S_{xx}} \quad \text{where } S_{xx} = \sum_k (x_k - \overline{x})^2$$

and

$$\begin{aligned}\widehat{\alpha} &= \overline{Y} - \widehat{\beta}\,\overline{x} \\ &= \frac{1}{n}\left(\sum_j Y_j\right) - \overline{x}\left(\frac{\sum_j (x_j - \overline{x}) Y_j}{S_{xx}}\right) \\ &= \sum_j \left(\frac{1}{n} - \frac{\overline{x}(x_j - \overline{x})}{S_{xx}}\right) Y_j .\end{aligned}$$

Proof of Theorem 8.1 ... continued

So

$$\begin{aligned}
\widehat{y}_i &= \widehat{\alpha} + \widehat{\beta} x_i \\
&= \sum_j \left(\frac{1}{n} - \frac{\overline{x}(x_j - \overline{x})}{S_{xx}} \right) Y_j + x_i \left(\frac{\sum_j (x_j - \overline{x}) Y_j}{S_{xx}} \right) \\
&= \sum_j \left(\frac{1}{n} + \frac{(x_i - \overline{x})(x_j - \overline{x})}{S_{xx}} \right) Y_j. \qquad (8)
\end{aligned}$$

We can write

$$Y_i = \sum_j \delta_{ij} Y_j \qquad (9)$$

where

$$\delta_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise.} \end{cases}$$

Proof of Theorem 8.1 ... continued

Note $\delta_{ij}^2 = \delta_{ij}$ and $\sum_j \delta_{ij}^2 = \sum_j \delta_{ij} = 1$. So

$$e_i = Y_i - \widehat{\alpha} - \widehat{\beta} x_i$$

and using (8) and (9)

$$= \sum_j \left(\delta_{ij} - \frac{1}{n} - \frac{(x_i - \overline{x})(x_j - \overline{x})}{S_{xx}} \right) Y_j.$$

Proof of Theorem 8.1 ... continued

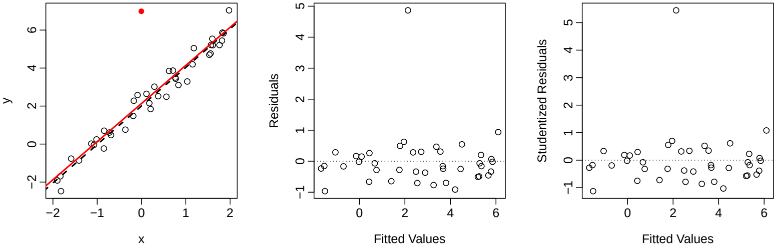
As the Y_j are independent, as at (7),

$$\begin{aligned}
\operatorname{var}(e_i) &= \sum_j \left(\delta_{ij} - \frac{1}{n} - \frac{(x_i - \overline{x})(x_j - \overline{x})}{S_{xx}} \right)^2 \operatorname{var}(Y_j) \\
&= \sigma^2 \sum_j \left(\delta_{ij}^2 + \frac{1}{n^2} + \frac{(x_i - \overline{x})^2 (x_j - \overline{x})^2}{S_{xx}^2} - 2\frac{1}{n}\delta_{ij} \right. \\
&\qquad \left. - 2\frac{(x_i - \overline{x})(x_j - \overline{x})}{S_{xx}}\delta_{ij} + 2\frac{1}{n}\frac{(x_i - \overline{x})(x_j - \overline{x})}{S_{xx}} \right) \\
&= \sigma^2 \left(1 + \frac{1}{n} + \frac{(x_i - \overline{x})^2}{S_{xx}} - \frac{2}{n} - 2\frac{(x_i - \overline{x})^2}{S_{xx}} + \frac{2}{n}\frac{(x_i - \overline{x})}{S_{xx}} \sum_j (x_j - \overline{x}) \right) \\
&= \sigma^2 \left(1 - \frac{1}{n} - \frac{(x_i - \overline{x})^2}{S_{xx}} \right) \\
&= \sigma^2 (1 - h_i). \quad \square
\end{aligned}$$

Plots with with studentized residuals

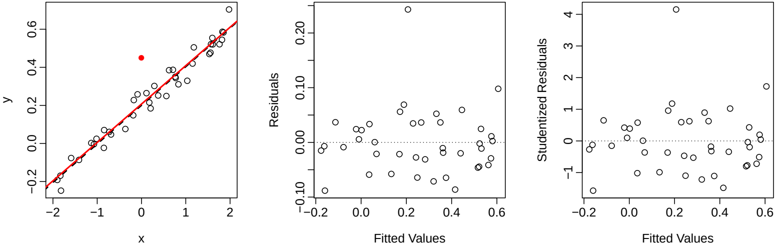
Left: dotted line = regression line based on the black points, red line = regression line with red point included.

Middle = residuals e_i. Right = studentized residuals r_i.



Residuals e_i and studentized residuals r_i are almost the same. The regression does not fit the red point well, but whether the red point is included or not has little effect on the fitted line.

Scale of y has changed, $|e_i|$ and $|r_i|$ are much smaller. Also, the red point isn't so extreme.



Studentized residual of red point is greater than 4, it is a possible outlier, again it has a tiny effect on the fitted line.

Potential problem: high leverage points

Outliers are observations for which the response y_i is unusual given the value of x_i. On the other hand, observations with *high leverage* have an unusual value of x_i.

Definition The *leverage* of the ith observation is h_i, where

$$h_i = \frac{1}{n} + \frac{(x_i - \overline{x})^2}{\sum_{j=1}^{n}(x_j - \overline{x})^2}.$$

Clearly h_i depends only on the values of the x_i (it doesn't depend on the y_i values at all). We see that h_i increases with the distance of x_i from $\overline{x}$.

High leverage points tend to have a sizeable impact on the regression line. Since $\text{var}(e_i) = \sigma^2(1 - h_i)$, a large leverage h_i will make $\text{var}(e_i)$ small. And then since $E(e_i) = 0$, this means that the regression line will be "pulled" close to the point (x_i, y_i).

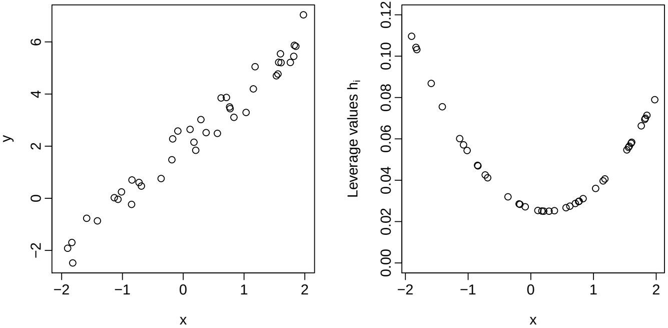
Note that $\sum h_i = 2$, so the average leverage is $\overline{h} = \frac{2}{n}$. One rule-of-thumb is to regard points with a leverage more than double this to be high leverage points, i.e. points with $h_i > 4/n$.

Why does this matter? We should be concerned if the regression line is heavily affected by just a couple of points, because any problems with these points might invalidate the entire fit. Hence it is important to identify high leverage observations.

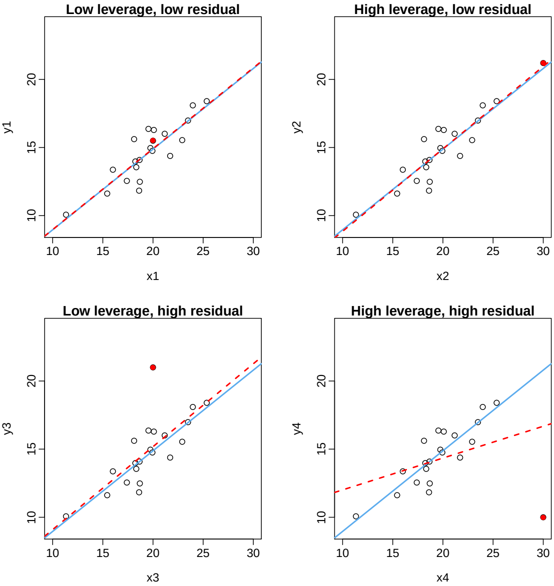
Leverage

Left: plot of (x_i, y_i), $i = 1, \ldots, n$.

Right: plot of leverage values h_i against x_i.

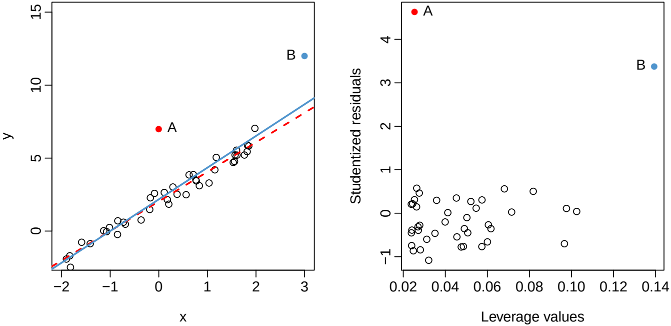


Blue line = excluding red point, dotted red line = including red point.



Left: red line = fitted line from black points + point A, blue line = fitted line from black points + point A + point B.

Right: studentized residuals r_i against leverage values h_i.

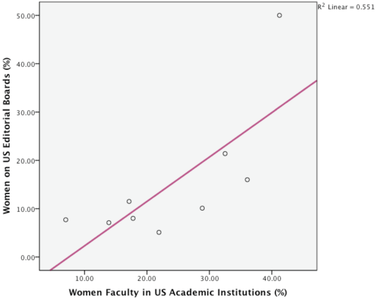


Removing point B – which has high leverage and a high residual – has a much more substantial impact on the regression line than removing an outlier with low leverage and/or low residual.

In the plot of studentized residuals against leverage:

- point B has high leverage and a high residual, it is having a substantial effect on the regression line yet it still has a high residual
- in contrast point A has a high residual, the model isn't fitting well at A, but t A isn't affecting the fit much at all – the reason that A has little effect on the fit is that it has low leverage.

Statistics Publications - Underrepresentation of women on boards

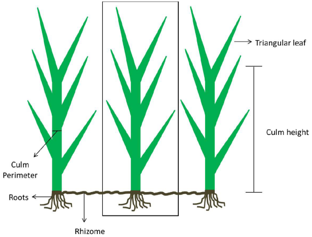


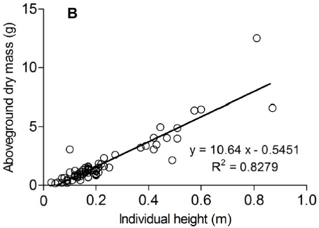
The association of women on editorial/advisory board and in the corresponding academic specialty.

Do you have any concerns here about fitting a simple linear regression? Spot the high leverage point!

Source: Ioannidou E, Rosania A (2015) Under-Representation of Women on Dental Journal Editorial Boards. *PLoS ONE* 10(1): e0116630. https://doi.org/10.1371/journal.pone.0116630

Statistics Publications - Estimating dry mass in plants



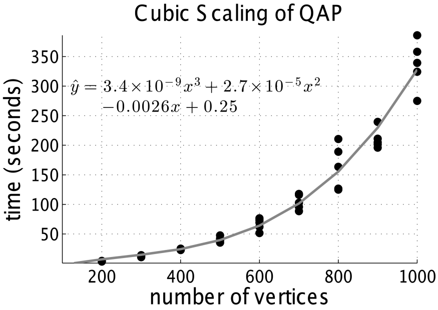


Estimation of the relationship between individual plant height and aboveground dry mass

Do you have any concerns here about fitting a simple linear regression? Spot the high leverage point!

Laís Samira Correia Nunes, Antonio Fernando Monteiro Camargo (2017) A simple non-destructive method for estimating aboveground biomass of emergent aquatic macrophytes. Acta Limnologica Brasiliensia vol.29 http://dx.doi.org/10.1590/s2179-975x6416

Statistics Publications - Cubic regression



Cubic regression analysis of the run time for a fast approximate quadratic assignment algorithm as a function of the number of vertices.

Source: Vogelstein et al. 2015 https://doi.org/10.1371/journal.pone.0121002

Does the variation in the observation appear to be constant?