

Review of some probability concepts: random variables

(A quick tour)

N. Torelli, G. Di Credico, V. Gioia

Fall 2023

University of Trieste

Random variables¹

Discrete distributions²

Continuous distributions³

C.d.f. and quantile functions⁴

¹Agresti, Kateri: sec 2.1 - 2.3

²Agresti, Kateri: 2.4

³Agresti, Kateri: 2.5

⁴Agresti, Kateri: 2.2.5-2.5.6-2.5.7

Random variables

Random variables

Statistics is about the extraction of information from data that contain an *unpredictable* component.

Random variables (r.v.) are the mathematical devices employed to build *models* of this variability.

A r.v. takes a different value at *random* each time is observed.

Distribution of a r.v.

The main tools used to describe the **distribution** of values taken by a r.v. are:

1. Probability (mass) functions (pmf)
2. (Probability) density functions (pdf)
3. Cumulative distribution functions (cdf)
4. Quantile functions

Discrete distributions

1. Probability functions

Discrete r.v. take values in a discrete set.

The **probability (mass) function** of a discrete r.v. X is the function $f(x)$ such that

$$f(x) = \Pr(X = x).$$

with $0 \leq f(x) \leq 1$ and $\sum_i f(x_i) = 1$.

The probability function defines the **distribution** of X .

Mean and variance of a discrete r.v.

For many purposes, the first two moments of a distribution provide a useful summary.

The **mean (expected value)** of a discrete r.v. X is

$$E(X) = \sum_i x_i f(x_i),$$

and the definition is extended to any function g of X

$$E\{g(X)\} = \sum_i g(x_i) f(x_i).$$

The special case $g(X) = (X - \mu)^2$, with $\mu = E(X)$, is the **variance** of X

$$\text{var}(X) = E\{(X - \mu)^2\} = E(X^2) - \mu^2.$$

The **standard deviation** is just given by $\sqrt{\text{var}(X)}$.

Notable discrete random variables

Discrete r.v. often used in applications:

- Binomial (and Bernoulli) distribution
- Poisson distribution
- Negative binomial distribution
- Geometric distribution
- Hypergeometric distribution

Let us give a closer look to some of them.

The binomial distribution

Consider n independent binary trials each with success probability p , $0 < p < 1$. The r.v. X that counts the number of successes has **binomial distribution** with probability function

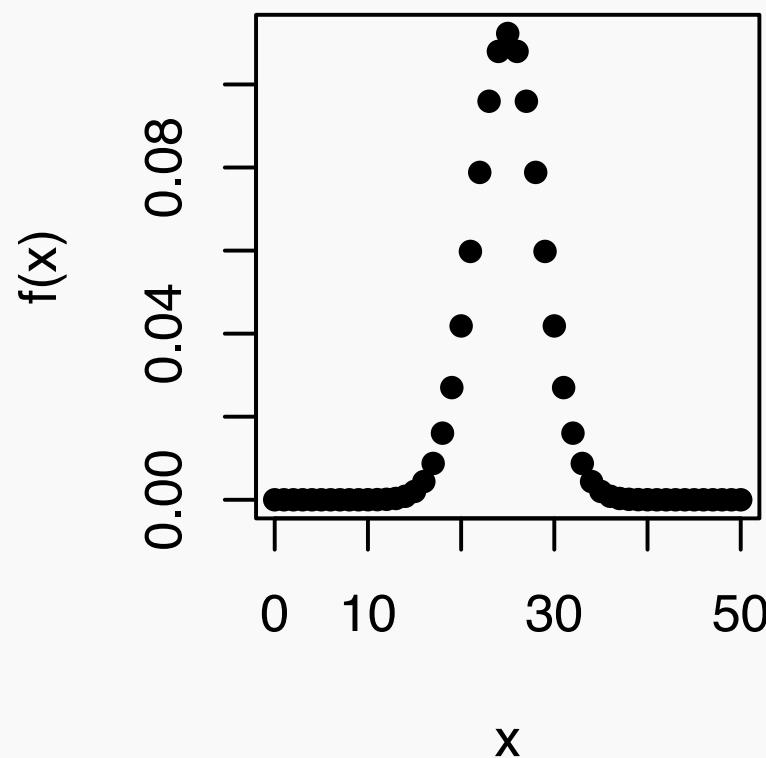
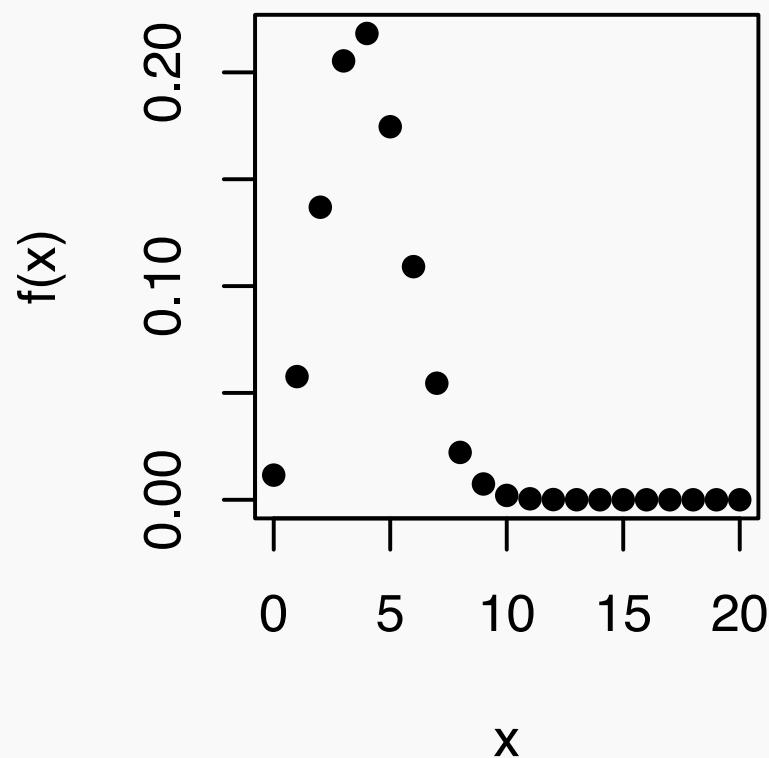
$$\Pr(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}, \quad x = 0, \dots, n.$$

The notation is $X \sim \mathcal{B}_i(n, p)$, and $E(X) = n p$, $\text{var}(X) = n p (1 - p)$.

The case when $n = 1$ is known as **Bernoulli distribution** and a single binary trial is called **Bernoulli trial**.

R lab: the binomial distribution

```
par(mfrow=c(1,2), pty="s", pch = 16)
plot(0:20, dbinom(0:20, 20, 0.2), xlab = "x", ylab = "f(x)")
plot(0:50, dbinom(0:50, 50, 0.5), xlab ="x", ylab = "f(x)")
```



The Poisson distribution

The special case the binomial distribution with $n \rightarrow \infty$ and $p \rightarrow 0$, while their product is held constant at $\lambda = n p$, yields the **Poisson distribution**.

Used for counts of events that occur randomly over time when: (1) counts of events in disjoint periods are independent, (2) it is essentially impossible to have two or more events simultaneously, (3) the rate of occurrence is constant.

The probability function is

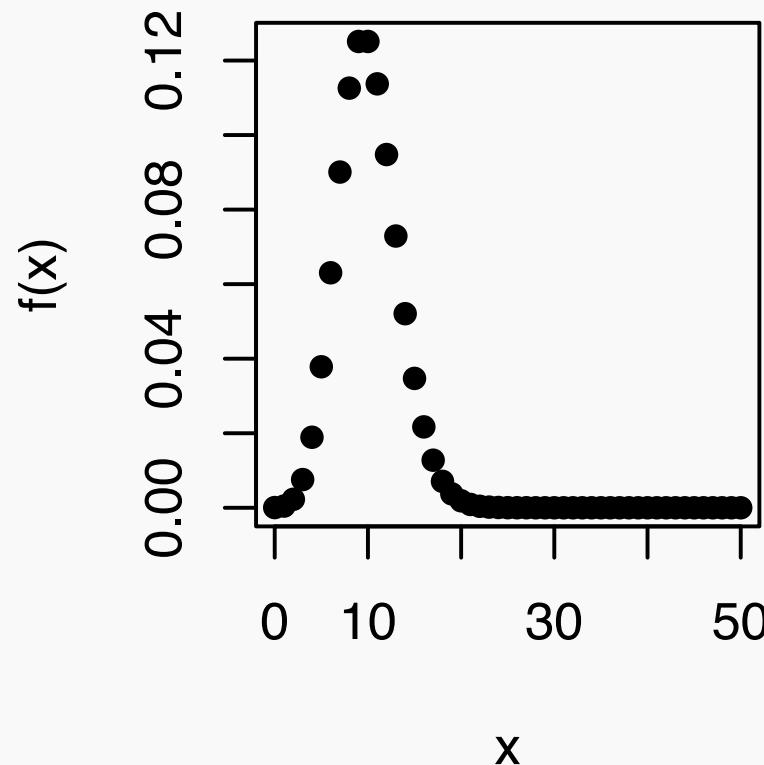
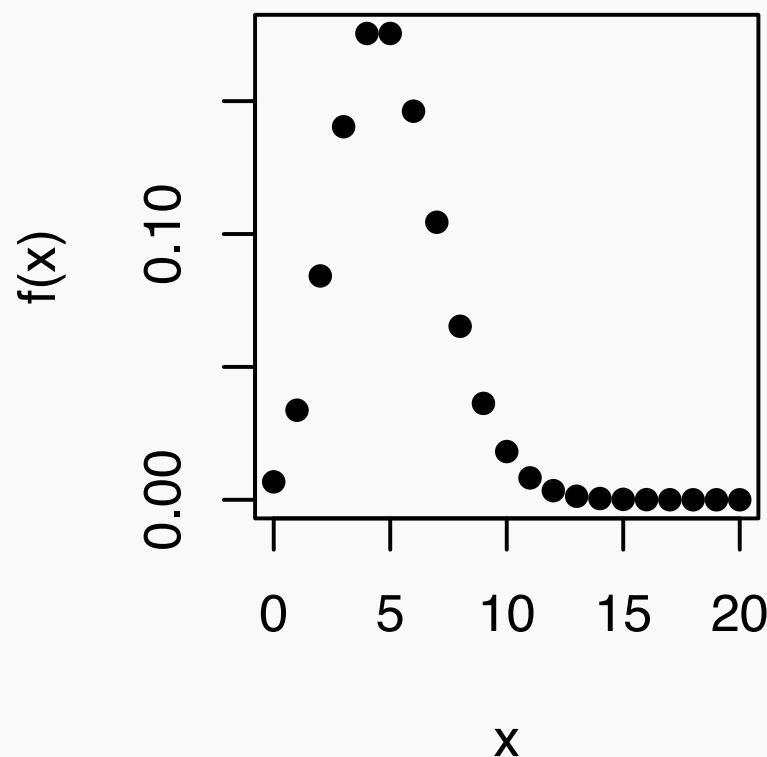
$$\Pr(X = x) = \frac{e^{-\lambda} \lambda^x}{x!}, \quad x = 0, 1, 2, \dots$$

with $\lambda > 0$.

The notation is $X \sim \mathcal{P}(\lambda)$, and $E(X) = \text{var}(X) = \lambda$.

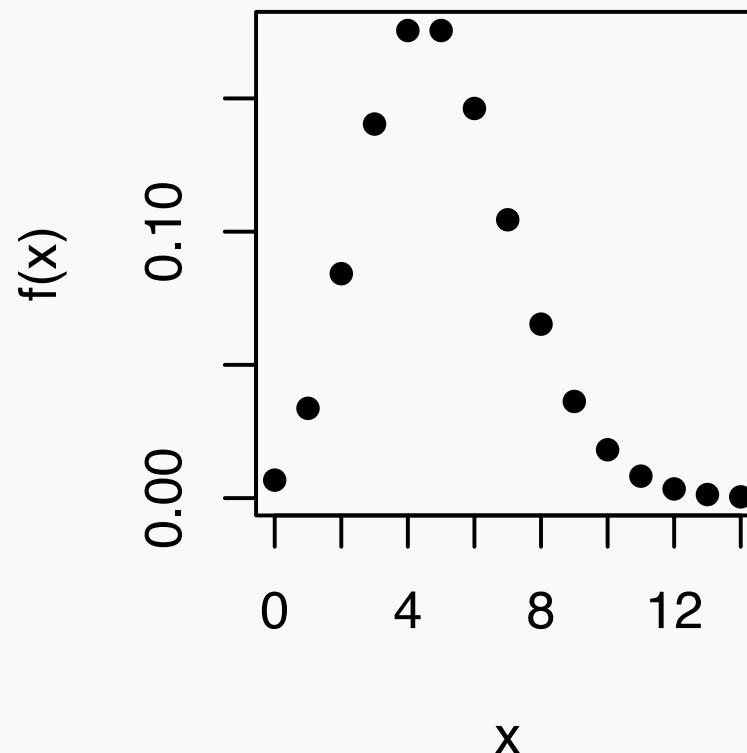
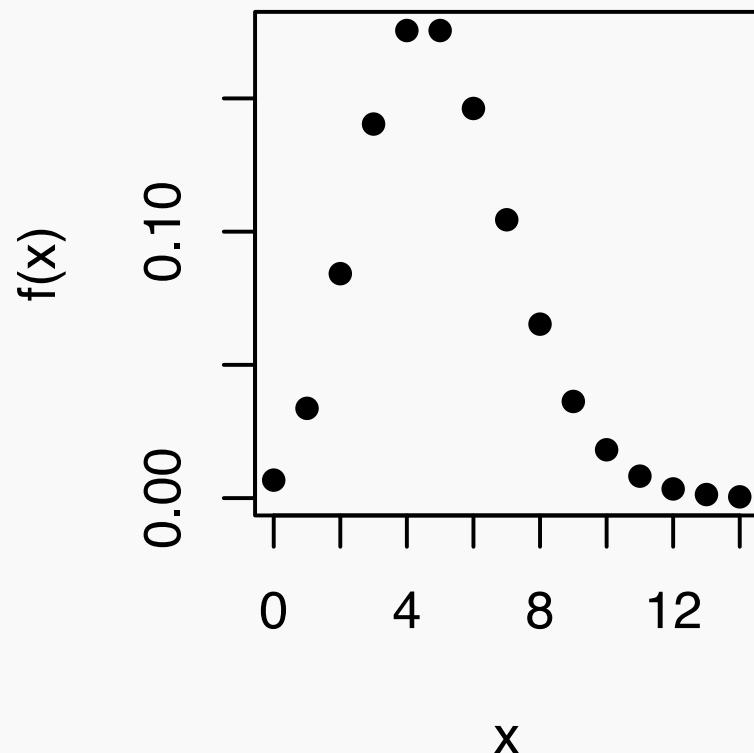
R lab: the Poisson distribution

```
par(mfrow=c(1,2), pty="s", pch = 16)
plot(0:20, dpois(0:20, 5), xlab = "x", ylab = "f(x)")
plot(0:50, dpois(0:50, 10), xlab = "x", ylab = "f(x)")
```



R lab: Poisson distribution and Binomial distribution

```
par(mfrow=c(1,2), pty="s", pch = 16)
plot(0:14, dpois(0:14, 5), xlab = "x", ylab = "f(x)")
plot(0:14, dbinom(0:14, 50000000, 0.0000001),
     xlab ="x", ylab = "f(x)")
```



Negative binomial distribution

Let us consider a sequence of independent Bernoulli trials with success probability p , let X be the count of trials necessary to observe the r -th success. Then X has a **Negative binomial** (or Pascal) distribution with parameters p and r .

The probability function is

$$\Pr(X = x) = \binom{x-1}{r-1} p^r (1-p)^{x-r} \quad x = r, r+1, r+2, \dots$$

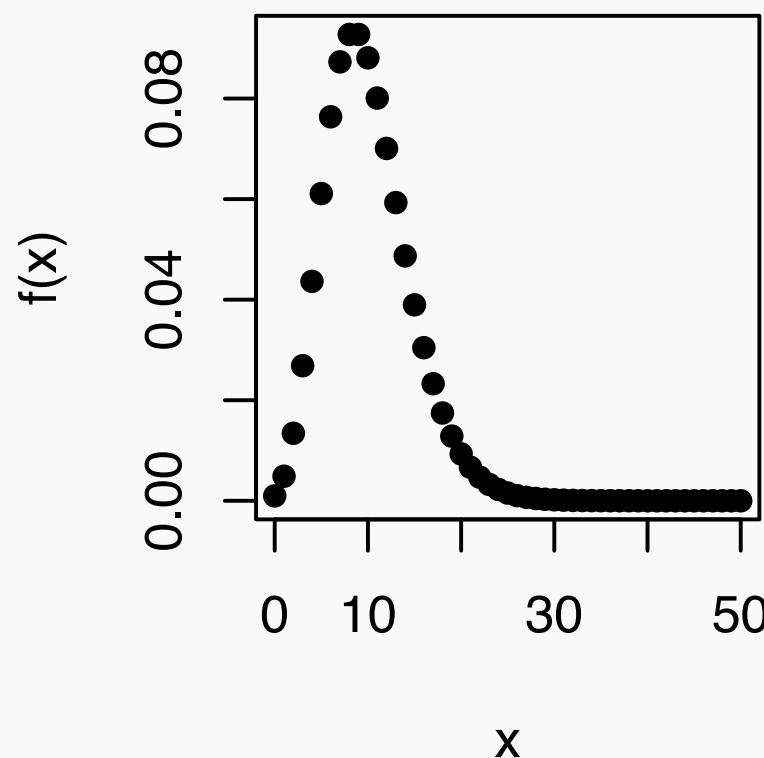
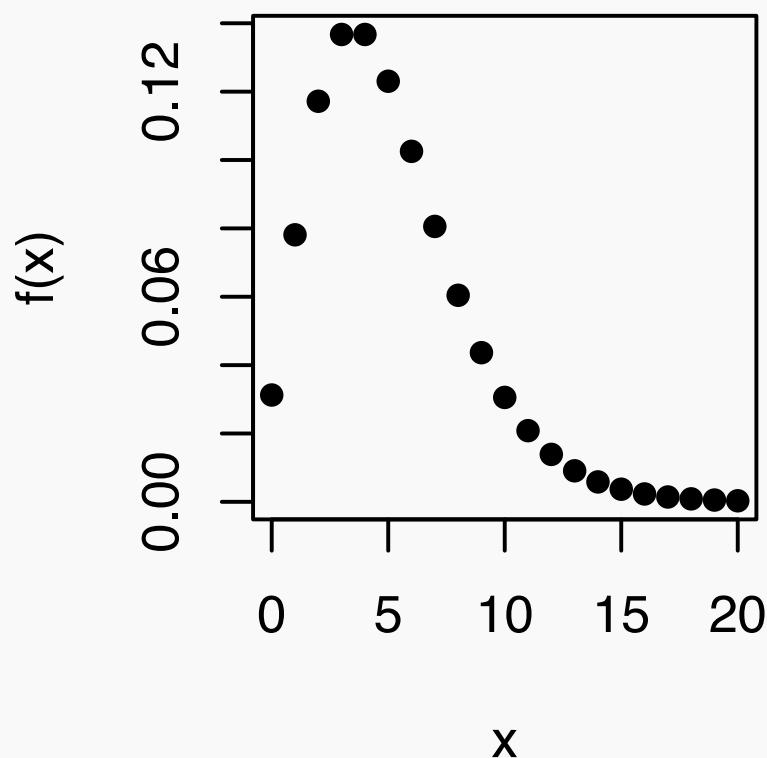
The notation is $X \sim \mathcal{NB}_i(p, r)$, and $E(X) = \frac{r}{p}$, $\text{var}(X) = \frac{r(1-p)}{p^2}$.

It can also be defined with support the Natural numbers by simply considering the variable $Y = X - r$

The case for $r = 1$ is known as the **Geometric** distribution.

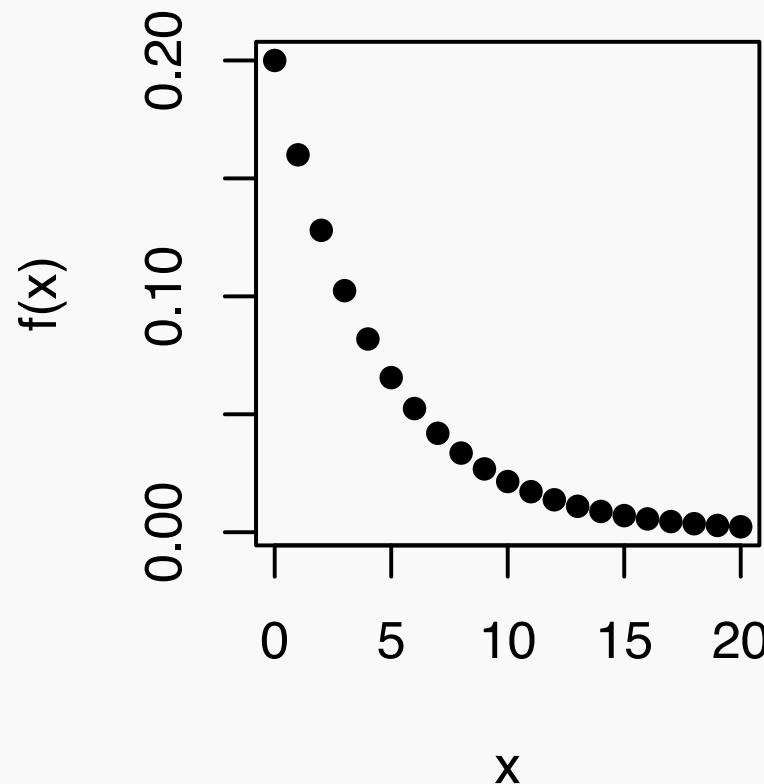
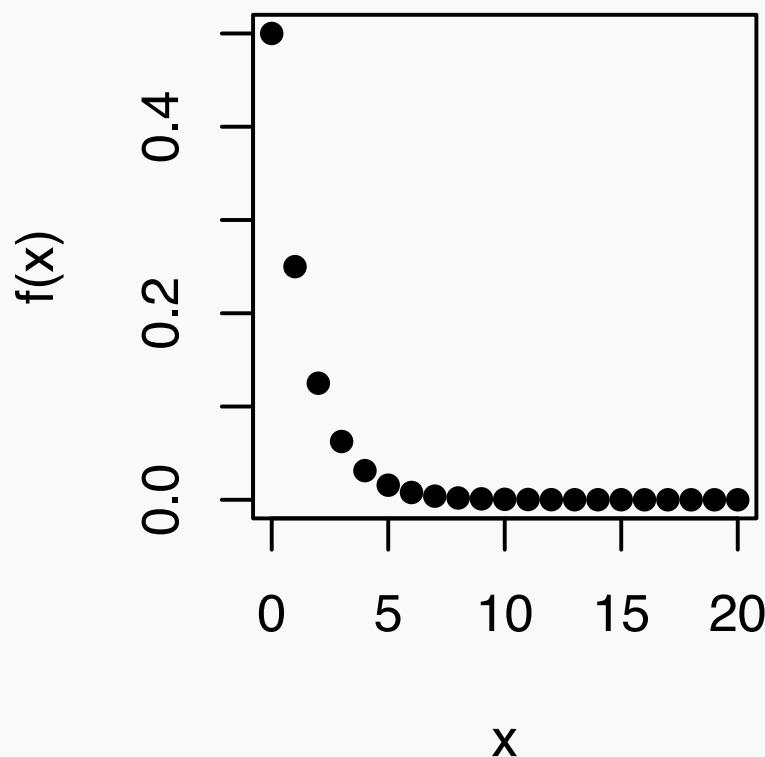
R lab: the Negative Binomial distribution

```
par(mfrow=c(1,2), pty="s", pch = 16)
plot(0:20, dnbinom(0:20, 5, 0.5), xlab = "x", ylab = "f(x)")
plot(0:50, dnbinom(0:50, 10, 0.5), xlab ="x", ylab = "f(x)")
```



R lab: the Geometric distribution

```
par(mfrow=c(1,2), pty="s", pch = 16)
plot(0:20, dnbinom(0:20, 1, 0.5), xlab = "x", ylab = "f(x)")
plot(0:20, dnbinom(0:20, 1, 0.2), xlab = "x", ylab = "f(x)")
```



Continuous distributions

2. Density functions

Continuous r.v. take values from intervals on the real line.

The **(probability) density function** (p.d.f.) of a continuous r.v. X is the function $f(x)$ such that, for any constants $a \leq b$

$$\Pr(a \leq X \leq b) = \int_a^b f(x)dx .$$

Note that $f(x) \geq 0$ and $\int_{-\infty}^{\infty} f(x)dx = 1$.

The probability density function defines the **distribution** of X .

Mean and variance of a continuous r.v.

The definitions given in the discrete case are readily extended.

The **mean (expected value)** of a continuous r.v. X is

$$E(X) = \int_{-\infty}^{\infty} x f(x) dx ,$$

and the definition is extended to any function g of X

$$E\{g(X)\} = \int_{-\infty}^{\infty} g(x) f(x) dx .$$

This includes the **variance** as a special case.

Two results, quite useful for continuous r.v., apply to a *linear transformation* $a + bX$, with a, b constants:

$$\begin{aligned} E(a + bX) &= a + b E(X) \\ \text{var}(a + bX) &= b^2 \text{var}(X) . \end{aligned}$$

Notable continuous random variables

Important continuous distributions include:

- Normal distribution
- Gamma, exponential and χ^2 distribution
- F distribution
- t and Cauchy distributions
- Beta distribution

The normal distribution has a major role in statistics. The χ^2 , t and F distributions are *relative* of the normal distribution.

The normal distribution

A r.v. X has a normal (or *Gaussian*) distribution if it has p.d.f.

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

The notation is $X \sim \mathcal{N}(\mu, \sigma^2)$, and $E(X) = \mu$ and $\text{var}(X) = \sigma^2$, $\sigma^2 > 0$, $\mu \in \mathbb{R}$.

An important property is that for any constants a, b

$$a + bX \sim \mathcal{N}(a + b\mu, b^2\sigma^2),$$

so that $Z = (X - \mu)/\sigma \sim \mathcal{N}(0, 1)$, the **standard normal distribution**.

Finally, $Y = e^X$ has a **lognormal distribution**, useful for asymmetric variables with occasional right-tail outliers.

Joint prediction of parameters.

Intercept

If the intercept doesn't make sense w/ our variables then we must take a representative value or interpret it instead, like the mean.

$$E[Y_i | X = \text{representative value}] \sim \beta_0$$

Residuals:

	Min	1Q	Median	3Q	Max
-1.6880	-0.3285	-0.0060	0.3120	1.3120	

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	5.0060	0.0728	68.762	<2e-16 ***
Speciesversicolor	0.9300	0.1030	9.033	8.77e-16 ***
Speciesvirginica	1.5820	0.1030	15.366	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5148 on 147 degrees of freedom
Multiple R-squared: 0.6187, Adjusted R-squared: 0.6135
F-statistic: 119.3 on 2 and 147 DF, p-value: < 2.2e-16

> summary(lm(Sepal.Length ~ Species, data = iris)) → w/ intercept

> summary(lm(Sepal.Length ~ -1 + Species, data = iris)) → w/out intercept

Call:
lm(formula = Sepal.Length ~ -1 + Species, data = iris)

Residuals:

	Min	1Q	Median	3Q	Max
-1.6880	-0.3285	-0.0060	0.3120	1.3120	

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
Speciessetosa	5.0060	0.0728	68.76	<2e-16 ***
Speciesversicolor	5.9360	0.0728	81.54	<2e-16 ***
Speciesvirginica	6.5880	0.0728	90.49	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residuals:

	Min	1Q	Median	3Q	Max
-1.3891	-0.3043	-0.0472	0.2528	1.3358	

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
Speciessetosa	4.78044	0.08308	57.54	<2e-16 ***
Speciesversicolor	4.72019	0.26590	17.75	<2e-16 ***
Speciesvirginica	4.73036	0.39860	11.87	<2e-16 ***
Petal.Width	0.91690	0.19386	4.73	5.25e-06 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.481 on 146 degrees of freedom
Multiple R-squared: 0.9935, Adjusted R-squared: 0.9934
F-statistic: 5608 on 4 and 146 DF, p-value: < 2.2e-16

This applies when we have categorical variables
but what if all of variables are numerical?

$$Y_i = \beta_0 + \beta_1 I(F = \text{Versicolor}) + \beta_2 I(F = \text{Virginica}) + \varepsilon_i$$

$$E[Y_i | F = \text{Setosa}] = \beta_0$$

$$E[Y_i | F = \text{Versicolor}] = \beta_0 + \beta_1$$

$$E[Y_i | F = \text{Virginica}] = \beta_0 + \beta_1 + \beta_2$$

$$Y_i = \beta_0 I(F = S) + \beta_1 I(F = \text{Ver.}) + \beta_2 I(F = \text{Vir.})$$

$$E[Y_i | F = S] = \beta_0$$

⋮

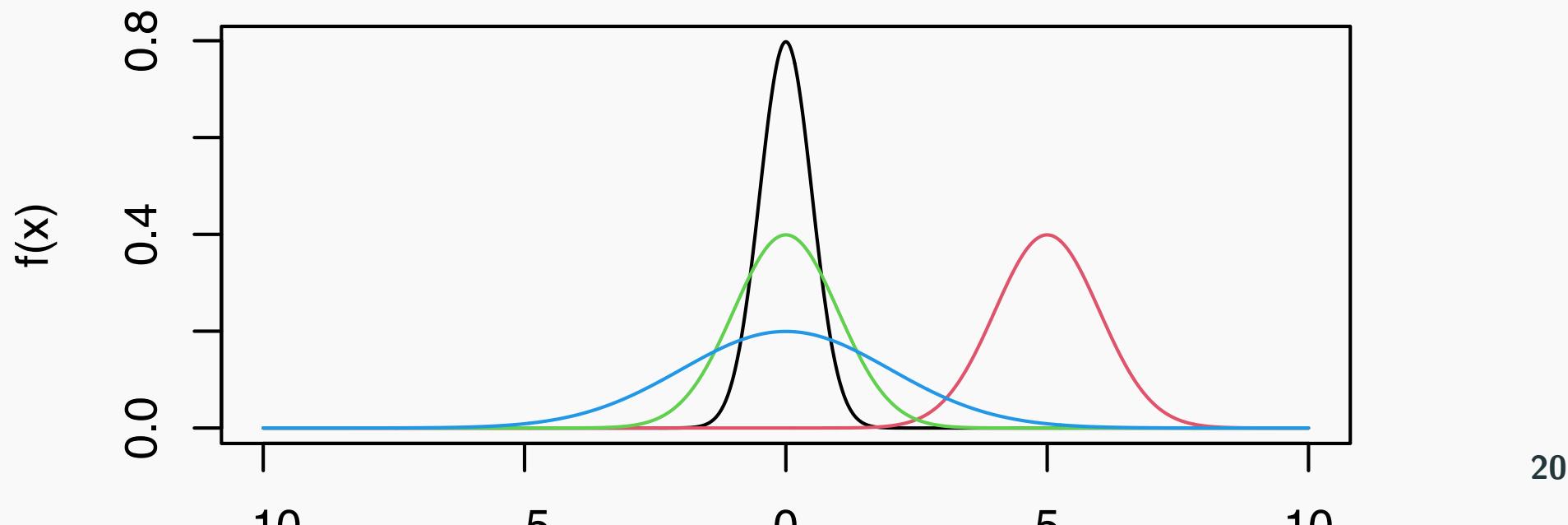
We still need to interpret the parameters

In this case we choose a value of Petal.Width such that we can interpret

$$E[Y_i | \text{Sepal.length} = \text{Petal.Width representative}] = \beta_0 + \beta_3$$

R lab: the normal distribution

```
xx <- seq(-10, 10, l=1000)
plot(xx, dnorm(xx, 0, 0.5), xlab ="x", ylab ="f(x)", type ="l")
lines(xx, dnorm(xx, 5, 1), col = 2)
lines(xx, dnorm(xx, 0, 1), col = 3)
lines(xx, dnorm(xx, 0, 2), col = 4)
```



The Gamma and the exponential distributions

A r.v. X has a Gamma distribution if it has the following pdf

$$f(x) = \frac{\lambda^\alpha x^{\alpha-1} e^{-\lambda x}}{\Gamma(\alpha)}, \quad x \geq 0$$

where $\lambda, \alpha > 0$ and $\Gamma(\alpha) = \int_0^\infty e^{-x} x^{\alpha-1} dx$.

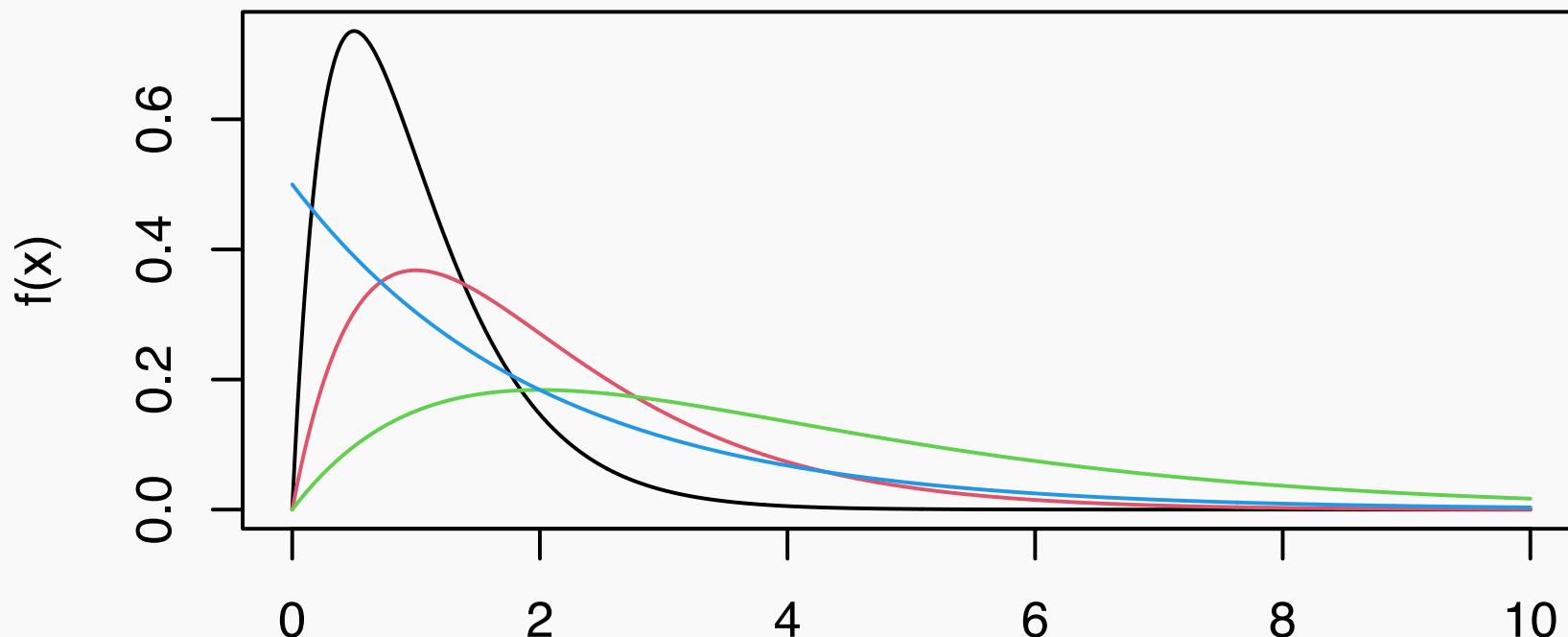
The notation is $X \sim Ga(\alpha, \lambda)$, $E(X) = \frac{\alpha}{\lambda}$ and $\text{var}(X) = \frac{\alpha}{\lambda^2}$.

When α is an integer it is also called **Erlang** distribution.

When $\alpha = 1$ it is called **exponential** distribution. The exponential distribution is related to the Poisson r.v. since X represents the waiting times between two arrivals in a Poisson process (The process which generates the Poisson rv)

Rlab: The Gamma and the exponential distributions

```
xx <- seq(0, 10, l=1000)
plot(xx, dgamma(xx, 2, 2), xlab ="x", ylab ="f(x)", type ="l")
lines(xx, dgamma(xx, 2, 1), col = 2)
lines(xx, dgamma(xx, 2, .5), col = 3)
lines(xx, dgamma(xx, 1, .5), col = 4) # exponential distribution
```



The Beta (and the uniform) distribution

A r.v. X has a Beta distribution if it has the following pdf

$$f(x) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1}, \quad 0 < x < 1$$

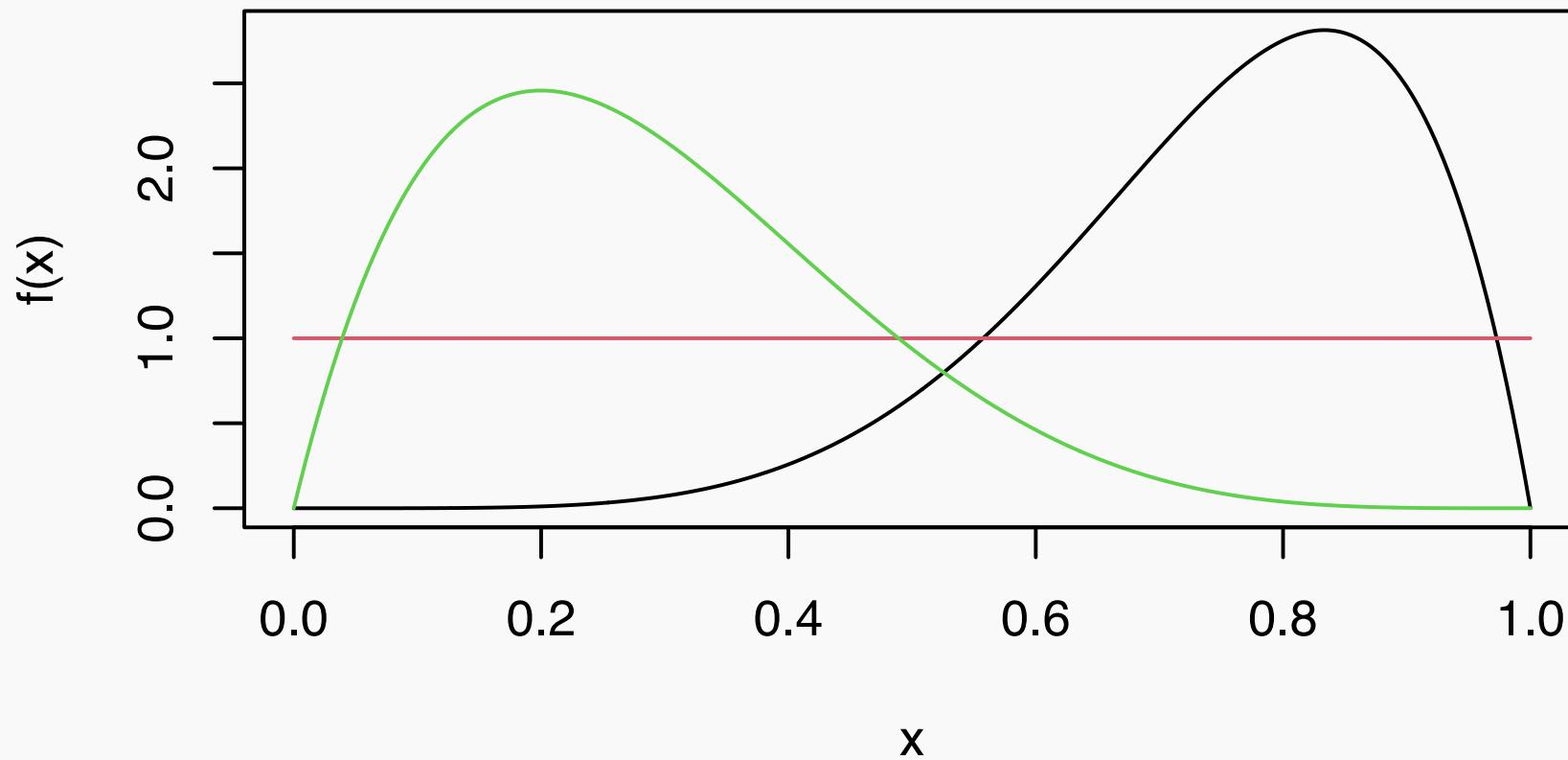
$$\alpha, \beta > 0$$

The notation is $X \sim Be(\alpha, \beta)$, $E(X) = \frac{\alpha}{\alpha+\beta}$ and $\text{var}(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$.

The **Uniform** distribution on $[0, 1]$ is a special case when $\alpha = 1$ and $\beta = 1$.

R lab: the Beta distribution

```
xx <- seq(0, 1, l=1000)
plot(xx, dbeta(xx, 6,2), xlab ="x", ylab ="f(x)", type ="l")
lines(xx, dbeta(xx, 1,1), col = 2)
lines(xx, dbeta(xx, 2, 5), col = 3)
```



The χ^2 distribution

Let Z_1, \dots, Z_k be a set of independent $\mathcal{N}(0, 1)$ r.v., then $X = \sum_{i=1}^k Z_i^2$ is a r.v. with a **χ^2 distribution with k degrees of freedom**.

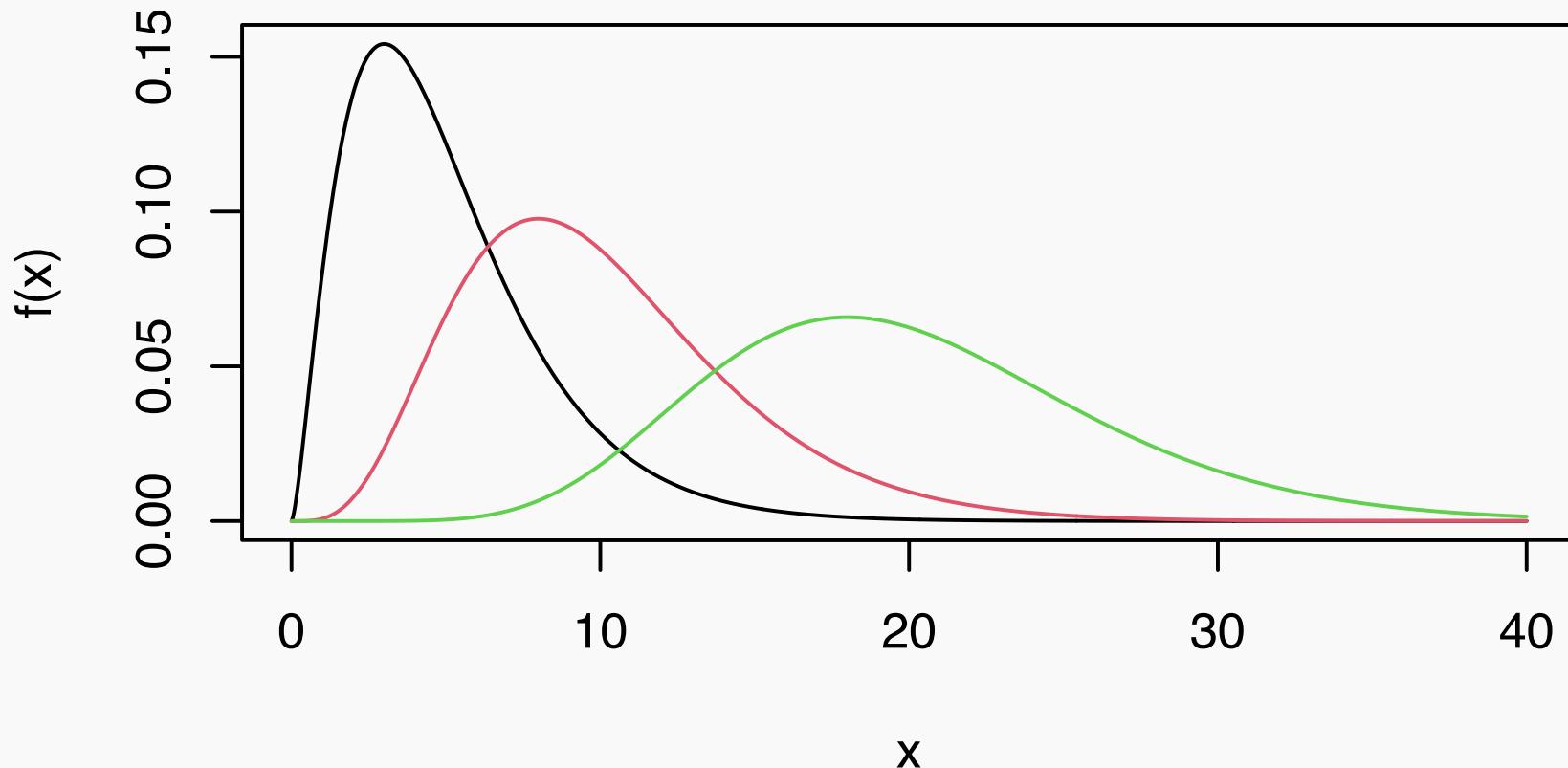
The notation is $X \sim \chi_k^2$, $E(X) = k$ and $\text{var}(X) = 2k$.

It is a special case of the Gamma distribution. In fact a χ^2 distribution with k degrees of freedom is a Gamma distribution with parameters $\alpha = k/2$ and $\lambda = 1/2$.

It plays an important role in the theory of hypothesis testing in statistics.

R lab: the χ^2 distribution

```
xx <- seq(0, 40, l=1000)
plot(xx, dchisq(xx, 5), xlab ="x", ylab ="f(x)", type ="l")
lines(xx, dchisq(xx, 10), col = 2)
lines(xx, dchisq(xx, 20), col = 3)
```



The F distribution

Let $X \sim \chi_n^2$ and $Y \sim \chi_m^2$, independent, then the r.v.

$$F = \frac{X/n}{Y/m}$$

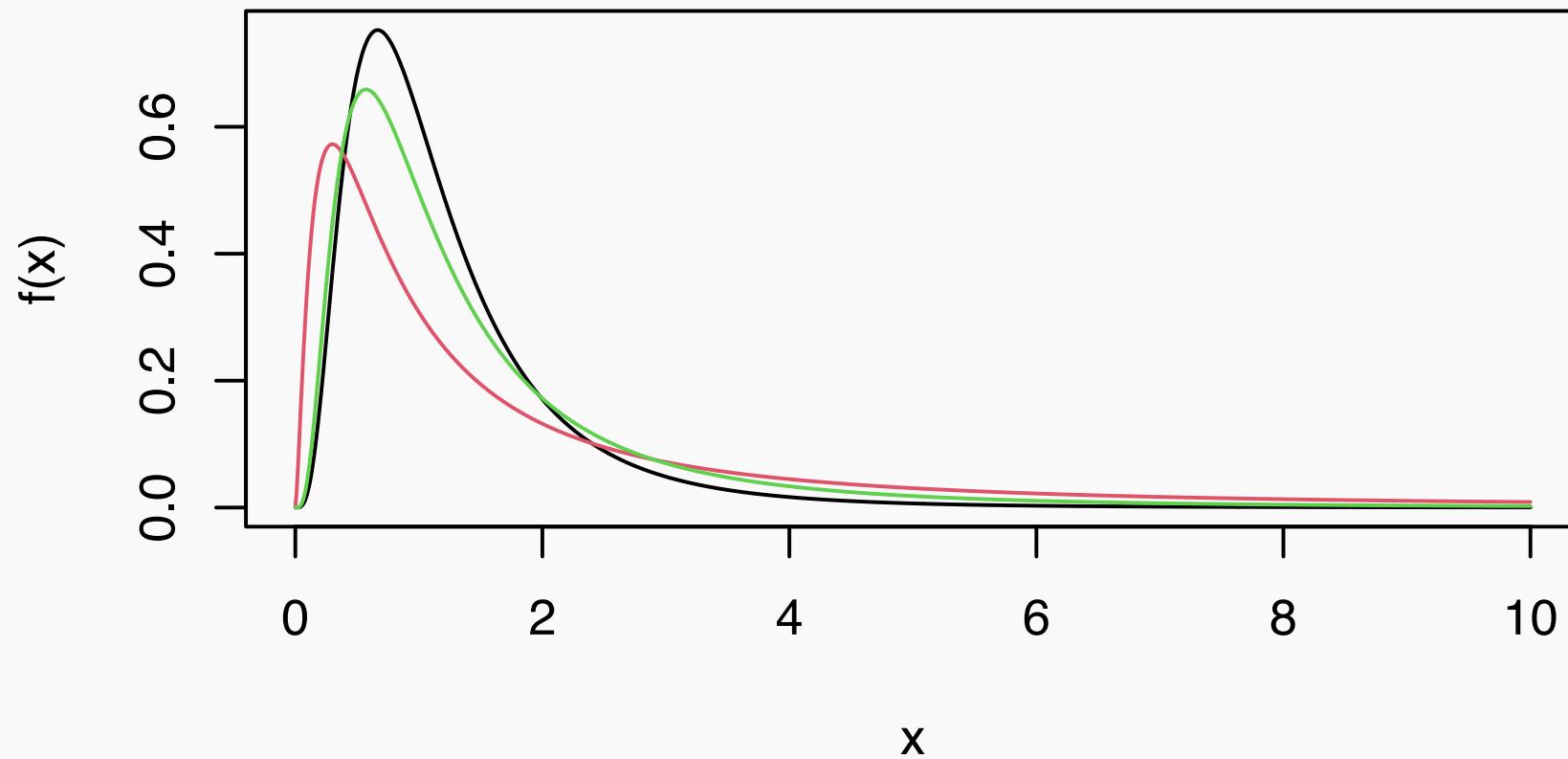
has an **F distribution with n and m degrees of freedom**.

The notation is $F \sim \mathcal{F}_{n,m}$, and $E(F) = m/(m - 2)$ provided that $m > 2$.

The distribution is almost never used as a model for observed data, but it has a central role in hypothesis testing involving linear models.

R lab: the F distribution

```
xx <- seq(0, 10, l=1000)
plot(xx, df(xx, 10, 10), xlab ="x", ylab ="f(x)", type ="l")
lines(xx, df(xx, 5, 2), col = 2)
lines(xx, df(xx, 10, 5), col = 3)
```



The t and Cauchy distributions

Let $Z \sim \mathcal{N}(0, 1)$ and $X \sim \chi_n^2$, independent, then the r.v.

$$T = \frac{Z}{\sqrt{\frac{X}{n}}}$$

has an **t distribution with n degrees of freedom**.

The notation is $T \sim t_n$, and $E(T) = 0$ provided that $n > 1$, whereas $\text{var}(T) = n/(n - 2)$ provided that $n > 2$.

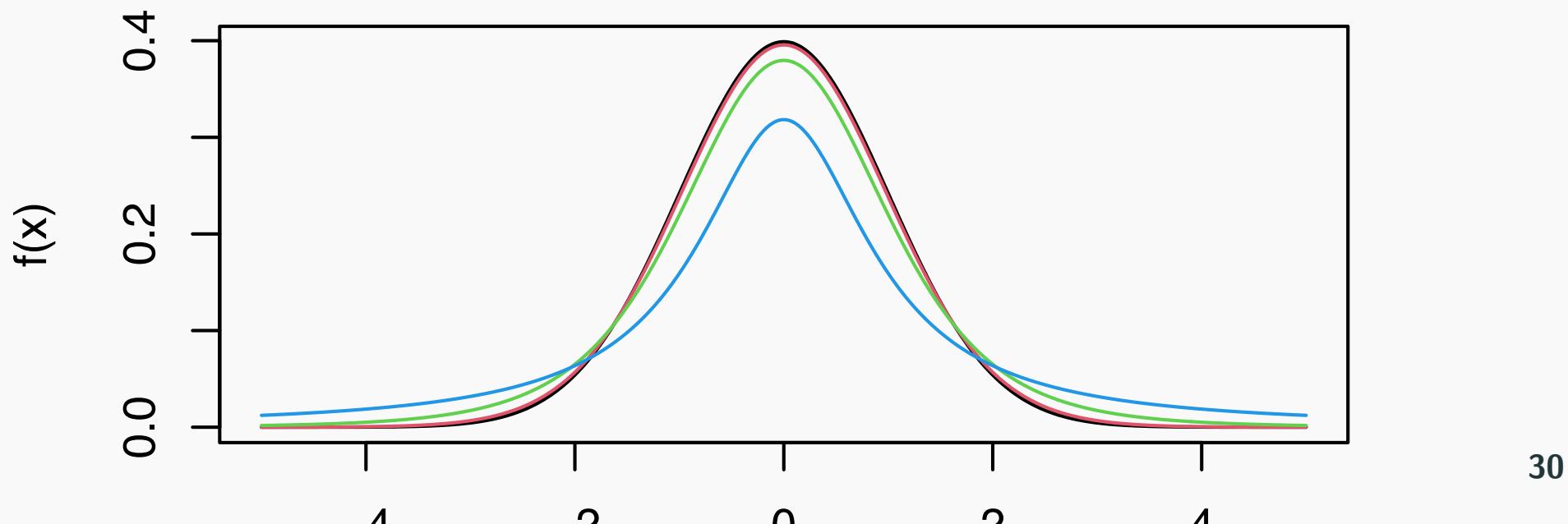
t_∞ is $\mathcal{N}(0, 1)$, while for n finite the distribution has heavier tails than the standard normal distribution.

The case t_1 is the **Cauchy distribution**.

The distribution has a central role in statistical inference; at times it is used for modelling phenomena presenting *outliers*.

R lab: the t and Cauchy distributions

```
xx <- seq(-5, 5, l=1000)
plot(xx, dnorm(xx, 0, 1), xlab ="x", ylab ="f(x)", type ="l")
lines(xx, dt(xx, 30), col = 2)
lines(xx, dt(xx, 5), col = 3)
lines(xx, dt(xx, 1), col = 4)
```



C.d.f. and quantile functions

3. Cumulative distribution functions

The **cumulative distribution function** (c.d.f.) of a r.v. X is the function $F(x)$ such that

$$F(x) = \Pr(X \leq x),$$

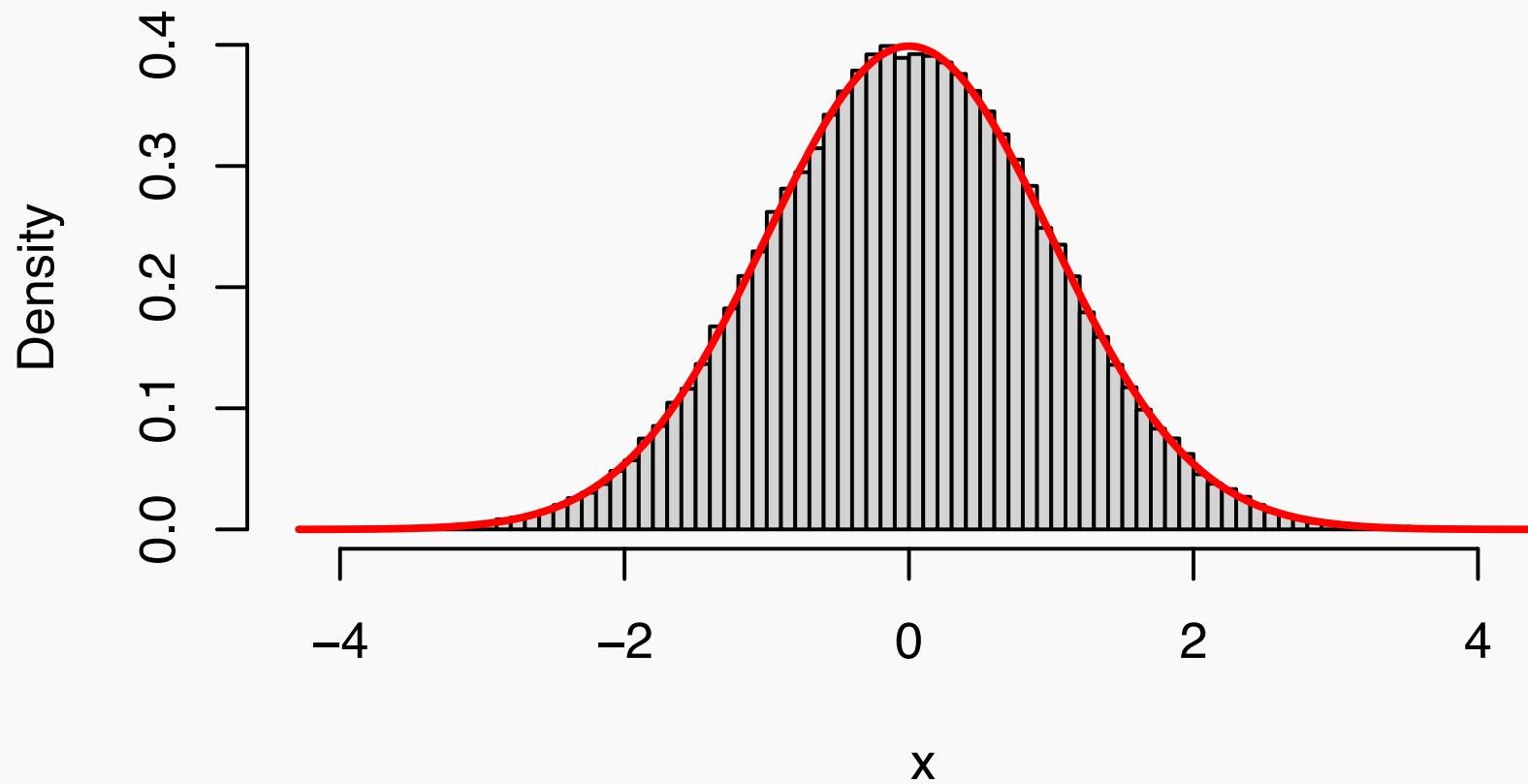
and it can be obtained from the probability function or the density function: the c.d.f. *identifies* the distribution.

From the definition of F it follows that $F(-\infty) = 0$, $F(\infty) = 1$, $F(x)$ is monotonic.

A useful property is that if F is a continuous function then $U = F(X)$ has a uniform distribution.

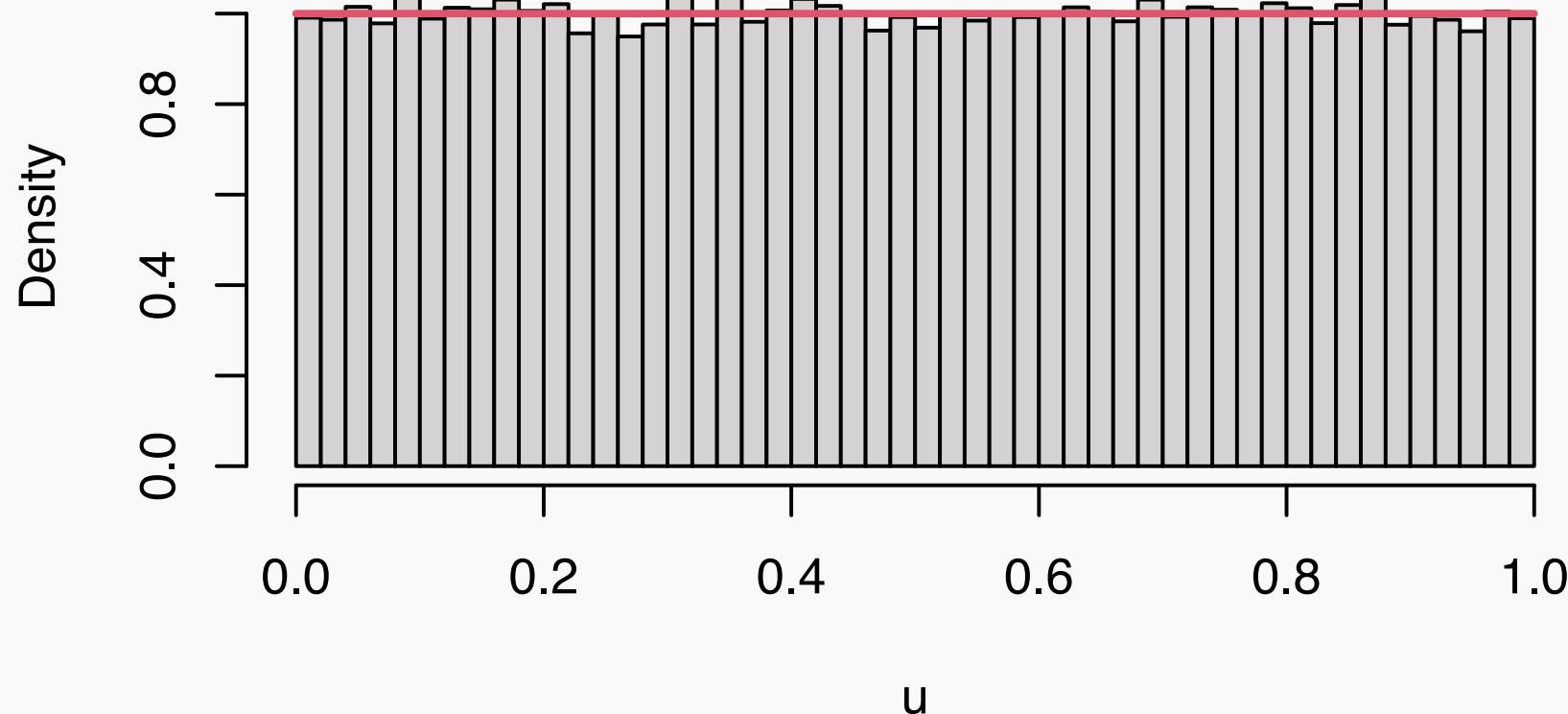
R lab: uniform transformation

```
x <- rnorm(10^5)    ### simulate values from N(0, 1)
xx <- seq(min(x), max(x), l = 1000)
hist.scott(x, main = "") ### from MASS package
lines(xx, dnorm(xx), col = "red", lwd = 2)
```



R lab: uniform transformation (cont'd.)

```
u <- pnorm(x)      ### that's the uniform transformation  
hist.scott(u, prob = TRUE, main="")  
segments(0, 1, 1, 1, col = 2, lwd = 2)
```



The quantile function

The inverse of the c.d.f. is defined as

$$F^-(p) = \min(x | F(x) \geq p), \quad 0 \leq p \leq 1.$$

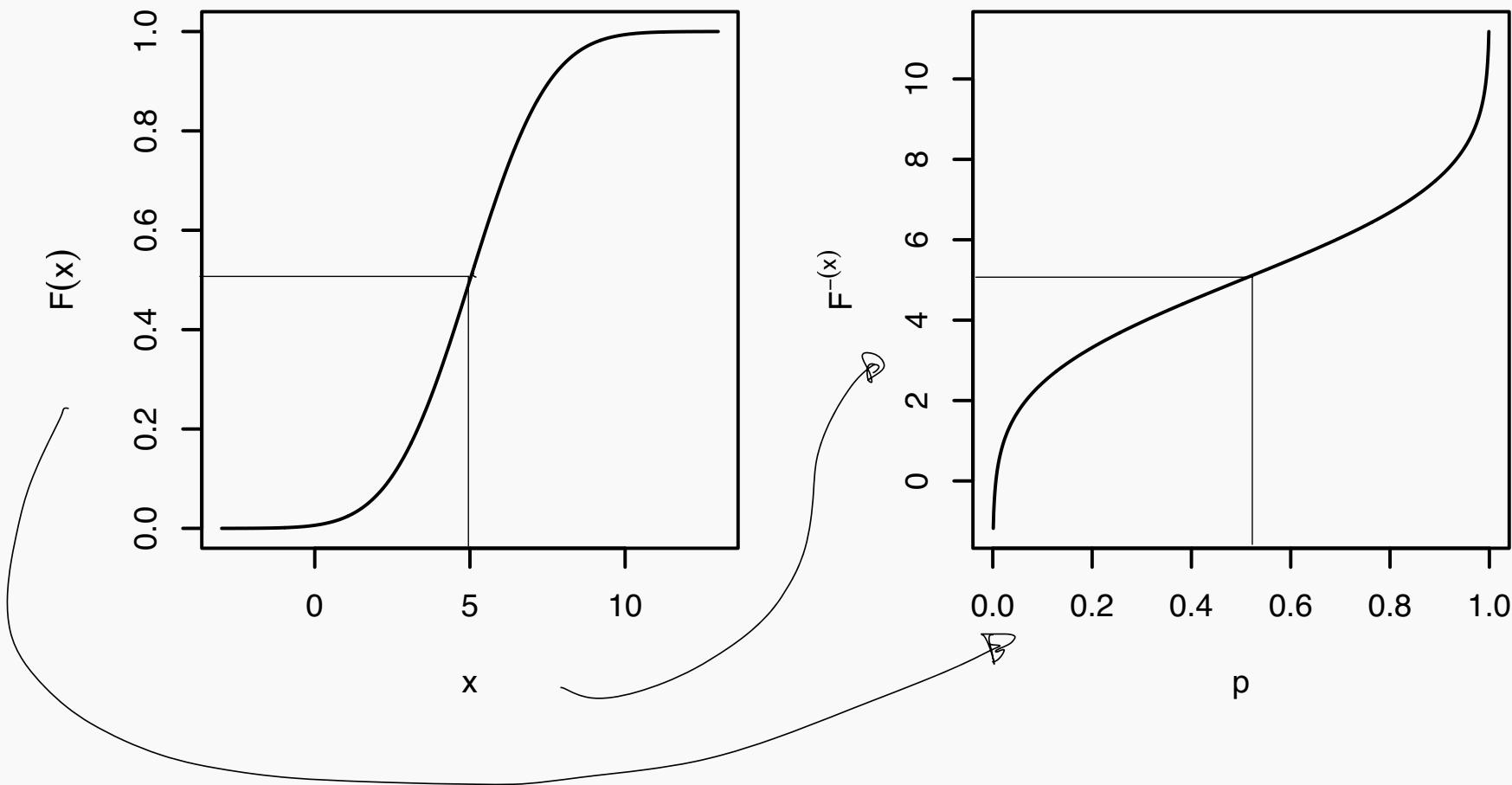
This is the usual inverse function of F when F is continuous.

Another useful property is that if $U \sim \mathcal{U}(0, 1)$, namely it has a *uniform distribution* in $[0, 1]$, then the r.v. $X = F^-(U)$ has c.d.f. F .

This provides a simple method to generate random numbers from a distribution with known quantile function: it is the **inversion sampling method**, that only requires the ability to simulate from a uniform distribution.

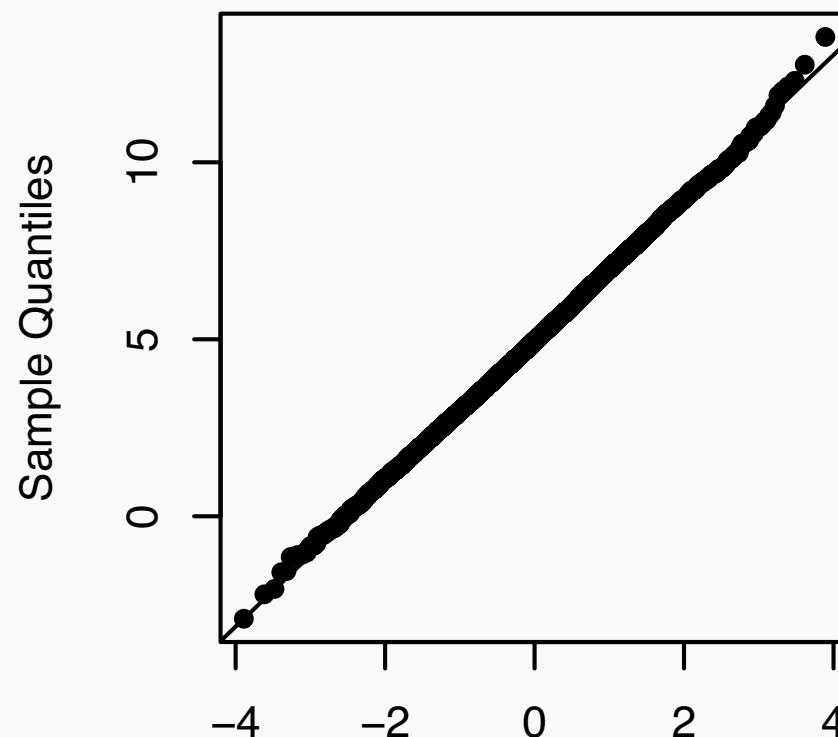
Example: normal cdf and quantile functions

Let us consider the case of $X \sim \mathcal{N}(5, 2^2)$, with c.d.f. and quantile functions given by `pnorm` and `qnorm`



R lab: inversion sampling

```
u <- runif(10^4); y <- qnorm(u, m = 5, s = 2)
par(pty = "s", cex = 0.8)
qqnorm(y, pch = 16, main = "")
qqline(y)
```



Side note: quantile-quantile plot

The previous slide demonstrated the usage of the quantile function to build a tool for **model goodness-of-fit**.

The *quantile-quantile plot* visualizes the plausibility of a theoretical distribution for a set of observations $y = (y_1, \dots, y_n)$.

This is done by comparing the quantile function of the assumed model with the sample quantiles, which are the points that lie on the inverse of the **empirical distribution function**

$$\hat{F}_n(t) = \frac{\text{number of elements of } y \leq t}{n}.$$

If the agreement between the data and the theoretical distribution is good, the points on the plot would approximately lie on a line.

START FROM P. 38 ↑

Review of some probability concepts: random vectors, large-sample results

(A quick tour)

N. Torelli, G. Di Credico, V. Gioia

Fall 2023

University of Trieste

Random vectors¹

The multivariate normal distribution²

Statistics³

Complements & large-sample results⁴

¹Agresti, Kateri: sec 2.6

²Agresti, Kateri: sec 2.7

³Agresti, Kateri: sec 3.1-3.2

⁴Agresti, Kateri: sec 3.3-3.4

Random vectors

Random vectors

In statistics multiple variables are usually observed, and vectors of random variables (**random vectors**) are required. The two-dimensional case is useful to illustrate the main concepts, and will be used here.

For continuous r.v., the **joint (probability) density function** extends the one-dimensional case: it is the $f(x, y)$ function such that, for any $A \subseteq \mathbb{R}^2$

$$\Pr\{(X, Y) \in A\} = \int \int_A f(x, y) dx dy .$$

Note that $f(x, y) \geq 0$ and $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dx dy = 1$.

The probability density function defines the **joint distribution** of the random vector (X, Y) .

Marginal distribution

The joint distribution embodies information about each components, so that the distribution of X , ignoring Y , can be obtained from $f(x, y)$.

The *marginal* density function of X is given by

$$f(x) = \int_{-\infty}^{\infty} f(x, y) dy ,$$

and similarly for the other variable.

(Note: here and elsewhere we always use the symbol f for any p.d.f., identifying the specific case by the argument of the function).

Conditional distribution

The *conditional density function* of Y given $X = x_0$ updates the distribution of Y by incorporating the information that $X = x_0$.

It is given by the important formula

$$f(y|X = x_0) = \frac{f(x_0, y)}{f(x_0)}, \quad \text{provide } f(x_0) > 0.$$

The simplified notation $f(y|x_0)$ is often employed.

The conditional p.d.f. is properly defined, since $f(y|X = x_0) \geq 0$ and $\int_{-\infty}^{\infty} f(y|x_0)dy = 1$.

A symmetric definition applies to X given $Y = y_0$.

Conditional distribution: useful properties

In the two dimensional case, it is readily possible to write

$$f(x, y) = f(x) f(y|x).$$

Extensions to higher dimensions require some care:

$$f(x, y, z) = f(x, y|z) f(z)$$

$$f(x, y|z) = f(x|z) f(y|x, z)$$

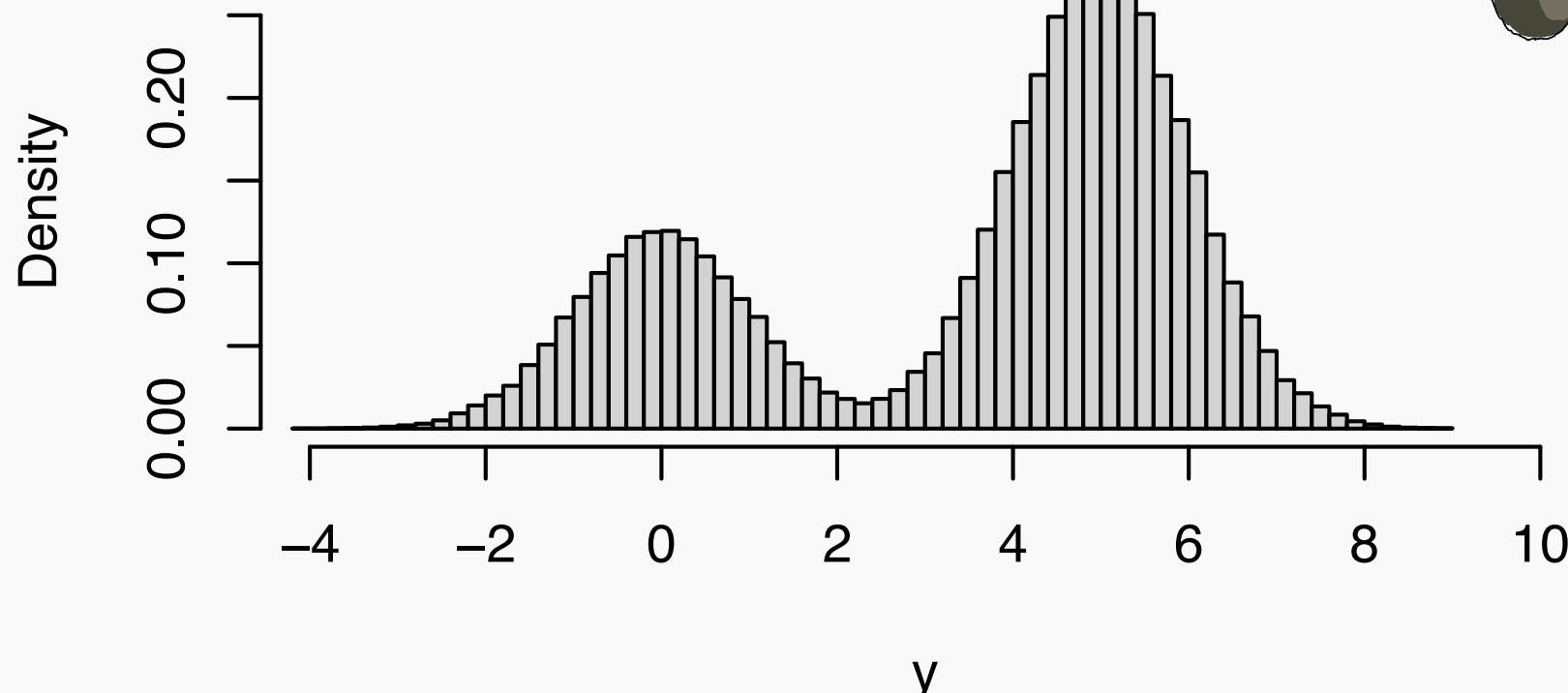
$$f(x, y, z) = f(x|y, z) f(y, z)$$

$$f(x, y, z) = f(x|y, z) f(y|z) f(z)$$

$$f(x_1, x_2, \dots, x_n) = f(x_1) f(x_2|x_1) f(x_3|x_2, x_1) \dots f(x_n|x_{n-1}, \dots, x_2, x_1)$$

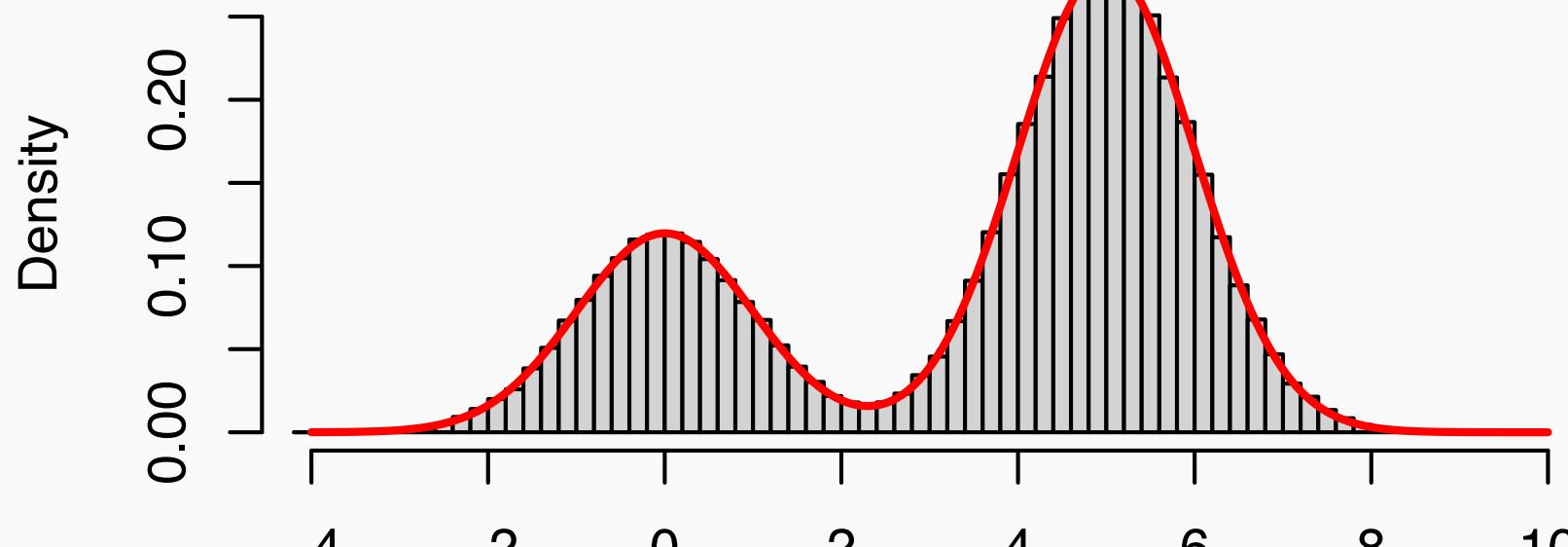
R lab: simulation from joint distributions (a mixture)

```
x <- rbinom(10^5, size = 1, prob = 0.7)
y <- rnorm(10^5, m = x * 5, s = 1) ### Y| X = x ~ N(x * 5, 1)
hist.scott(y, main = "", xlim = c(-4, 10))
```



R lab: simulation from joint distributions (cont'd.)

```
xx <- seq(-4, 10, l = 1000)
ff <- 0.3 * dnorm(xx, 0) + 0.7 * dnorm(xx, 5)
### This is a mixture of normal distributions
hist.scott(y, main = "", xlim = c(-4, 10))
lines(xx, ff, col = "red", lwd = 2)
```



Bayes theorem

From the factorization of the joint distribution it readily follows that

$$f(x, y) = f(x) f(y|x) = f(y) f(x|y)$$

from which we obtain the **Bayes theorem**

$$f(x|y) = \frac{f(x) f(y|x)}{f(y)} .$$

This is a cornerstone of statistics, leading to an entire school of statistical modelling.

Independence and conditional independence

When $f(y|x)$ does not depend on the value of x , the r.v. X and Y are *independent*, and

$$f(x, y) = f(y) f(x)$$

More in general, n r.v. are independent if and only if

$$f(x_1, x_2, \dots, x_n) = f(x_1) f(x_2) \dots f(x_n).$$

Conditional independence arises when two r.v. are independent given a third one:

$$f(y, x|z) = f(x|z) f(y|z)$$

An important part of statistical modelling exploits some sort of conditional independence.

Example of conditional independence: the Markov property

The general factorization defined above

$$f(x_1, x_2, \dots, x_n) = f(x_1) f(x_2|x_1) f(x_3|x_2, x_1) \dots f(x_n|x_{n-1}, \dots, x_2, x_1)$$

will simplify considerably when the *first order Markov property* holds:

$$f(x_i|x_1, \dots, x_{i-1}) = f(x_i|x_{i-1})$$

which means that X_i is independent of X_1, \dots, X_{i-2} given X_{i-1} . We get

$$f(x_1, x_2, \dots, x_n) = f(x_1) \prod_{i=2}^n f(x_i|x_{i-1}).$$

When the variables are observed over time, this means that the process has *short memory*, a property quite useful in the statistical modelling of **time series**.

Mean and variance of linear transformations

For two r.v. X and Y and two constants a, b we get

$$E(aX + bY) = aE(X) + bE(Y).$$

The result follows from the more general one

$$E\{g(X, Y)\} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f(x, y) dx dy.$$

For variances we need first to introduce the **covariance** between X and Y

$$\text{cov}(X, Y) = E\{(X - \mu_x)(Y - \mu_y)\} = E(XY) - \mu_x \mu_y,$$

where $\mu_x = E(X)$ and $\mu_y = E(Y)$. Then

$$\text{var}(aX + bY) = a^2 \text{var}(X) + b^2 \text{var}(Y) + 2ab \text{cov}(X, Y).$$

Note: for X, Y independent it follows that $\text{cov}(X, Y) = 0$. The reverse is not true, unless the joint distribution of X and Y is multivariate normal.

Mean vector

For a random vector $\mathbf{X} = (X_1, X_2, \dots, X_n)^\top$, the **mean vector** is just

$$E(\mathbf{X}) = \begin{pmatrix} E(X_1) \\ E(X_2) \\ \vdots \\ E(X_n) \end{pmatrix}.$$

The mean vector has the same properties of the scalar case, so that for example $E(\mathbf{X} + \mathbf{Y}) = E(\mathbf{X}) + E(\mathbf{Y})$, and for \mathbf{A} and \mathbf{b} a $n \times n$ matrix and a $n \times 1$ vector, respectively, it follows that

$$E(\mathbf{A}\mathbf{X} + \mathbf{b}) = \mathbf{A}E(\mathbf{X}) + \mathbf{b}.$$

Variance-covariance matrix

The variance-covariance matrix of the random vector \mathbf{X} collects all the variances (on the main) diagonal and all the pairwise covariances (off the main diagonal), being the $n \times n$ symmetric semi-definite matrix

$$\Sigma = E\{(\mathbf{X} - \mu_x)(\mathbf{X} - \mu_x)^\top\} = \begin{pmatrix} \text{var}(X_1) & \text{cov}(X_1, X_2) & \cdots & \text{cov}(X_1, X_n) \\ \text{cov}(X_1, X_2) & \text{var}(X_2) & \cdots & \text{cov}(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \text{cov}(X_1, X_n) & \text{cov}(X_2, X_n) & \cdots & \text{var}(X_n) \end{pmatrix}$$

Useful properties:

$$\Sigma_{A\mathbf{X}+\mathbf{b}} = A\Sigma A^\top$$

$$\Sigma_{\mathbf{X}^\top A\mathbf{X}} = \mu_x^\top A \mu_x + \text{tr}(A\Sigma)$$

Transformation of random variables and random vectors

Given a continuous r.v. X and a transformation $Y = g(X)$, with g an invertible function, it readily follows that

$$f_y(y) = f_x\{g^{-1}(y)\} \left| \frac{dx}{dy} \right|.$$

The result is extended to two continuous random vectors with the same dimension

$$f_{\mathbf{Y}}(\mathbf{Y}) = f_{\mathbf{X}}\{g^{-1}(\mathbf{Y})\} |\mathbf{J}|,$$

with $J_{ij} = \partial x_i / \partial y_j$.

For discrete r.v., the results are simpler, with no need of including the Jacobian of the transformation.

$$\text{Ex: } Y = e^X \quad , \quad X \sim N(\mu, \sigma^2)$$

$$\text{support: } S_Y = \left(\lim_{x \rightarrow -\infty} e^x, \lim_{x \rightarrow \infty} e^x \right) = (0, \infty)$$

$$g^{-1}(y) = \log y \quad , \quad \frac{dx}{dy} = \frac{d}{dy} \log y \approx \frac{1}{y}$$

$$f_Y(y) = \frac{1}{\sqrt{2\pi}\sigma} \exp \left\{ -\frac{(\log y - \mu)^2}{2\sigma^2} \right\} \frac{1}{y}$$

The multivariate normal distribution

The multivariate normal distribution

Start from a set of n i.i.d. $Z_i \sim \mathcal{N}(0, 1)$, so that $E(\mathbf{z}) = \mathbf{0}$ and covariance matrix \mathbf{I}_n . If \mathbf{B} is $m \times n$ matrix of coefficients and $\boldsymbol{\mu}$ a m -vector of coefficients, then the m -dimensional random vector \mathbf{X}

$$\mathbf{X} = \mathbf{B}\mathbf{z} + \boldsymbol{\mu}$$

has a **multivariate normal distribution** with covariance matrix $\boldsymbol{\Sigma} = \mathbf{B}\mathbf{B}^\top$.

The notation is

$$\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}).$$

Joint p.d.f.

Using basic results on transformation of random vectors, starting from the joint p.d.f of Z_1, Z_2, \dots, Z_n we obtain

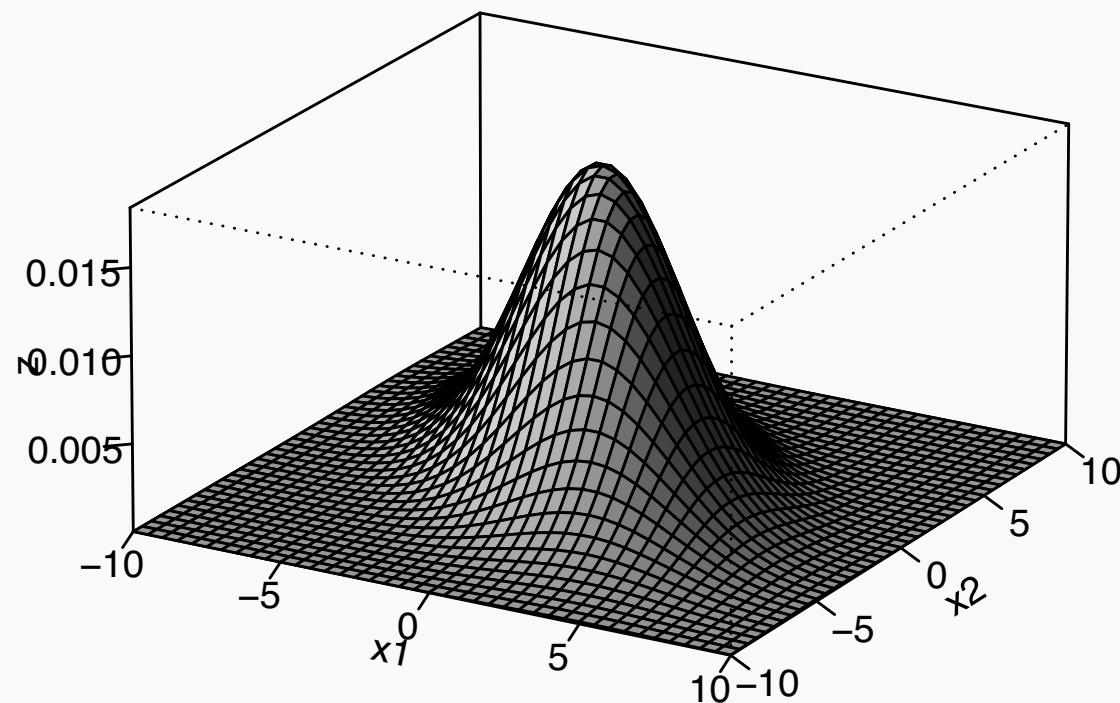
$$f_{\mathbf{X}}(\mathbf{X}) = \frac{1}{\sqrt{(2\pi)^m |\boldsymbol{\Sigma}|}} \exp \left\{ -\frac{1}{2} (\mathbf{X} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{X} - \boldsymbol{\mu}) \right\}, \quad \text{for } \mathbf{X} \in \mathbb{R}^m,$$

provide that $\boldsymbol{\Sigma}$ has full rank m . The result can be extended to *singular* $\boldsymbol{\Sigma}$ by recourse to the *pseudo-inverse* of $\boldsymbol{\Sigma}$: this is used, for example, in the analysis of *compositional data*.

A useful property which holds only for this distribution: *two r.v. with multivariate normal distribution and zero covariance are independent.*

Example: bivariate case

We take $\mu_1 = \mu_2 = 0$, $\sigma_1^2 = 10$, $\sigma_2^2 = 10$, $\sigma_{12} = 15$



Linear transformations

It is simple to verify that if $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and \mathbf{A} is a $k \times m$ matrix of constants then

$$\mathbf{A}\mathbf{X} \sim \mathcal{N}(\mathbf{A}\boldsymbol{\mu}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^\top).$$

A special case is obtained when $k = 1$, in that for a m -dimensional vector \mathbf{a}

$$\mathbf{a}^\top \mathbf{X} \sim \mathcal{N}(\mathbf{a}^\top \boldsymbol{\mu}, \mathbf{a}^\top \boldsymbol{\Sigma} \mathbf{a}).$$

Note that for suitable choices of \mathbf{a} (when all the elements 0s or 1s) it follows that **the marginal distribution of any subvector of \mathbf{X} is multivariate normal**.

Normality of the marginal distributions, instead, does not imply multivariate normality.

Conditional distributions

Consider two random vectors \mathbf{X} and \mathbf{Y} with multivariate normal joint distribution, and partition their joint covariance matrix as

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{xx} & \boldsymbol{\Sigma}_{xy} \\ \boldsymbol{\Sigma}_{yx} & \boldsymbol{\Sigma}_{yy} \end{pmatrix},$$

and similarly for the mean vector $\boldsymbol{\mu} = (\boldsymbol{\mu}_x, \boldsymbol{\mu}_y)^\top$.

Using results on *partitioned matrices*, it follows that the **conditional distributions are multivariate normal**.

For instance

$$\mathbf{Y}|\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}_y + \boldsymbol{\Sigma}_{yx} \boldsymbol{\Sigma}_{xx}^{-1} (\mathbf{X} - \boldsymbol{\mu}_x), \boldsymbol{\Sigma}_{yy} - \boldsymbol{\Sigma}_{yx} \boldsymbol{\Sigma}_{xx}^{-1} \boldsymbol{\Sigma}_{xy}).$$

Statistics

Random sample

The collection of r.v. X_1, X_2, \dots, X_n is said to be a **random sample** of size n if they are *independent and identically distributed*, that is

- X_1, X_2, \dots, X_n are independent r.v.
- They have the same distribution, namely the same c.d.f.

The concept is central in statistics, and it is the suitable mathematical model for the outcome of sampling units from a very large population. The definition is, however, more general.

(For more details: https://www.probabilitycourse.com/chapter8/8_1_1_random_sampling.php)

Statistics

A **statistic** is a r.v. defined as a function of a set of r.v.

Obvious examples are the sample mean and variance of data y_1, y_2, \dots, y_n

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i, \quad s^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2.$$

Consider a random vector \mathbf{Y} with p.d.f. $f_\theta(\mathbf{Y})$ depending on a vector θ (which is the *parameter*, as we will see).

If a statistic $t(\mathbf{Y})$ is such that $f_\theta(\mathbf{Y})$ can be written as

$$f_\theta(\mathbf{Y}) = h(\mathbf{Y}) g_\theta\{t(\mathbf{Y})\},$$

where h does not depend on θ , and g depends on \mathbf{Y} only through $t(\mathbf{Y})$, then t is a **sufficient statistic** for θ : all the *information* available on θ contained in \mathbf{Y} is supplied by $t(\mathbf{Y})$.

The concepts of information and sufficiency are central in statistical inference.

Example: sufficient statistic for the normal distribution

Given a vector of independent normal r.v. $Y_i \sim \mathcal{N}(\mu, \sigma^2)$, it follows that $\theta = (\mu, \sigma^2)$ and

$$\begin{aligned} f_{\theta}(\mathbf{Y}) &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} \exp \left\{ -\frac{1}{2\sigma^2}(y_i - \mu)^2 \right\} \\ &= \frac{1}{(\sqrt{2\pi})^n \sigma^n} \exp \left\{ -\frac{1}{2\sigma^2} \sum_i (y_i - \mu)^2 \right\}. \end{aligned}$$

By some simple algebra, it is possible to show that the two-dimensional statistic $t(\mathbf{Y}) = (\bar{y}, s^2)$ is sufficient for (μ, σ^2) .

Complements & large-sample results

Moment generating function

The **moment generating function** (m.g.f.) characterises the distribution of a r.v. X , and it is defined as

$$M_X(t) = E(e^{tX}), \quad \text{for } t \text{ real}.$$

The name derives from the fact the k^{th} derivative of the m.g.f. at $t = 0$ gives the k^{th} uncentered moment:

$$\frac{d^k M_X(t)}{d t^k} \Big|_{t=0} = E(X^k).$$

Two useful properties:

- If $M_X(t) = M_Y(t)$ for some small interval around $t = 0$, then X and Y have the same distribution.
- If X and Y are independent, $M_{X+Y}(t) = M_X(t) M_Y(t)$.

The central limit theorem

For i.i.d. r.v. X_1, X_2, \dots, X_n with mean μ and finite variance σ^2 , the **central limit theorem** states that for large n the distribution of the r.v. $\bar{X}_n = \sum_{i=1}^n X_i/n$ is approximately

$$\bar{X}_n \sim \mathcal{N}(\mu, \sigma^2/n).$$

More formally, the theorem says that for any $x \in \mathbb{R}$ the c.d.f. of $Z_n = (\bar{X}_n - \mu)/\sqrt{\sigma^2/n}$ satisfies

$$\lim_{n \rightarrow \infty} F_{Z_n}(x) = \int_{-\infty}^x \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz.$$

The proof is simple, and it uses the m.g.f.

The theorem generalizes to multivariate and non-identical settings.

It has a central importance in statistics, since it supports the normal approximation to the distribution of a r.v. that can be viewed as the sum of other r.v.

The law of large numbers

Consider i.i.d. (independent and identically distributed) r.v. X_1, \dots, X_n , with mean μ and $(E|X_i|) < \infty$.

The **strong law of large numbers** states that, for any positive ϵ

$$\Pr\left(\lim_{n \rightarrow \infty} |\bar{X}_n - \mu| < \epsilon\right) = 1,$$

namely \bar{X}_n converges almost surely to μ .

With the further assumption $\text{var}(X_i) = \sigma^2$, the **weak law of large numbers** follows

$$\lim_{n \rightarrow \infty} \Pr(|\bar{X}_n - \mu| \geq \epsilon) = 0.$$

Proof of the weak law of large numbers

First we recall the *Chebyshev's inequality*: given a r.v. X such that $E(X^2) < \infty$ and a constant $a > 0$, then

$$\Pr(|X| \geq a) \leq \frac{E(X^2)}{a^2}.$$

We apply the inequality to the case of interest, so that

$$\Pr(|\bar{X}_n - \mu| \geq \epsilon) \leq \frac{E\{(\bar{X}_n - \mu)^2\}}{\epsilon^2} = \frac{\text{var}(\bar{X}_n)}{\epsilon^2} = \frac{\sigma^2}{n \epsilon^2},$$

which tends to zero when $n \rightarrow \infty$.

The result may hold also for non-i.i.d. cases, provided $\text{var}(\bar{X}_n) \rightarrow 0$ for large n .

Jensen's inequality

This is another useful result, that states that for a r.v. X and a concave function g

$$g\{E(X)\} \geq E\{g(X)\}.$$

(Note: a concave function is such that

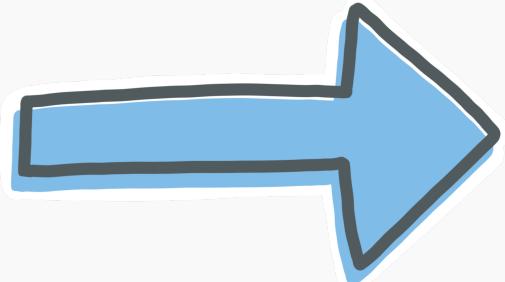
$$g\{\alpha x_1 + (1 - \alpha) x_2\} \geq \alpha g(x_1) + (1 - \alpha) g(x_2),$$

for any x_1, x_2 , and $0 \leq \alpha \leq 1$).

An example is $g(x) = -x^2$, so that

$$-E(X)^2 \geq -E(X^2) \quad \Rightarrow \quad E(X)^2 \leq E(X^2),$$

which is obviously true since $E(X^2) = \text{var}(X) + E(X)^2$.



Maybe brief Review

Statistical Models

N. Torelli, G. Di Credico, V. Gioia

Fall 2023

University of Trieste

The concept of statistical model

Simulation from a statistical model

The problems of statistical inference: an overview

The concept of statistical model

Intro: Aim of statistical inference

Statistics aims to **extract information from data**, and in particular on the process that generated the data.

Two intrinsic difficulties:

- It may be hard to infer what we wish to know from the data available;
- Most data contain some **random variability**: by replicating the data-gathering process several times we would obtain different data on each occasion.

We search for conclusions drawn from a single data set that are **generally valid**, and not the result of random peculiarities of that data set.

Role of statistical models

Statistics is able to draw conclusions from random data mainly through the use of **statistical models**.

A statistical model can be thought as a *mathematical cartoon* describing how our data might have been generated, if the unknown features of the data-generating process were actually known.

If the unknowns were known, a good model *can generate data* resembling the main features of observed data.

The purpose of **statistical inference** is to use the statistical model to go in the *reverse direction*: to infer the model unknowns that are consistent with the observed data.

Mathematical aspects

Notation:

- y random vector containing the observed data
- θ vector of parameters of unknown value

We assume that knowing the parameters would answer the question of interest about the process generating the data.

The model specifies how data akin to y may be simulated, implicitly defining the **distribution** of y and how it depends on θ .

Moreover, a statistical model may depend on some known parameters γ and some further data x , treated as known and denoted as *covariates* or *predictor variables*.

Visually

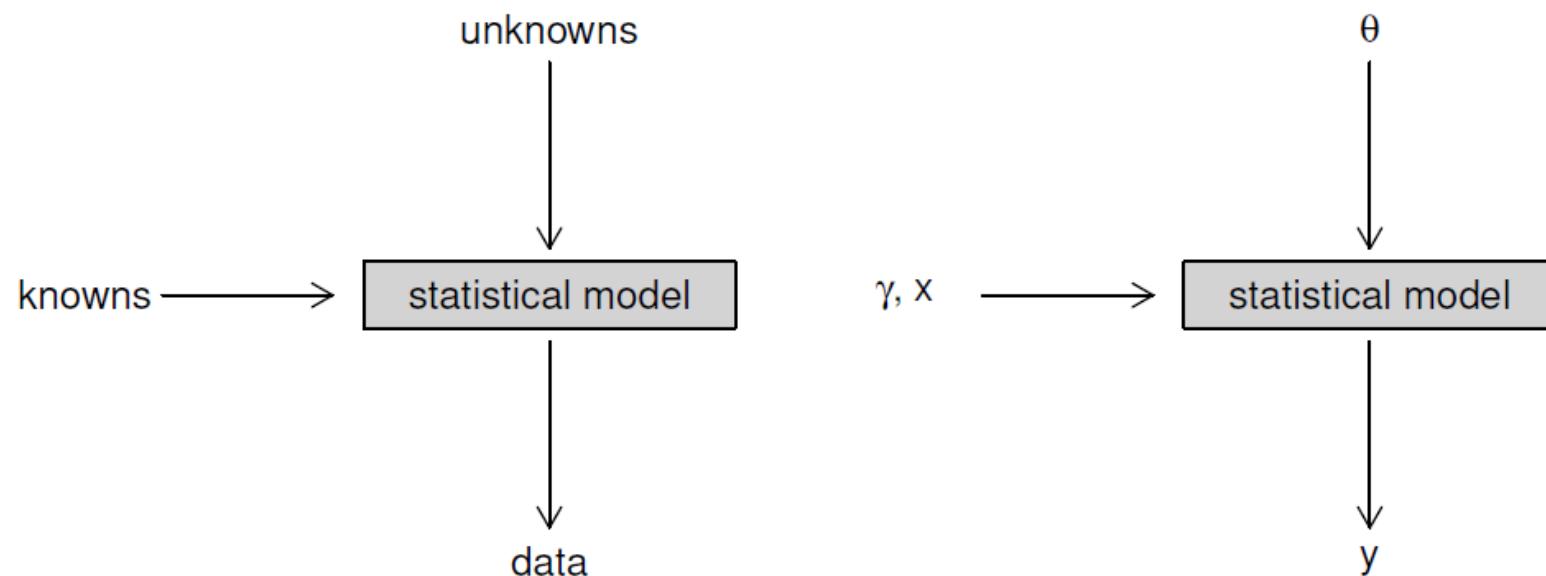
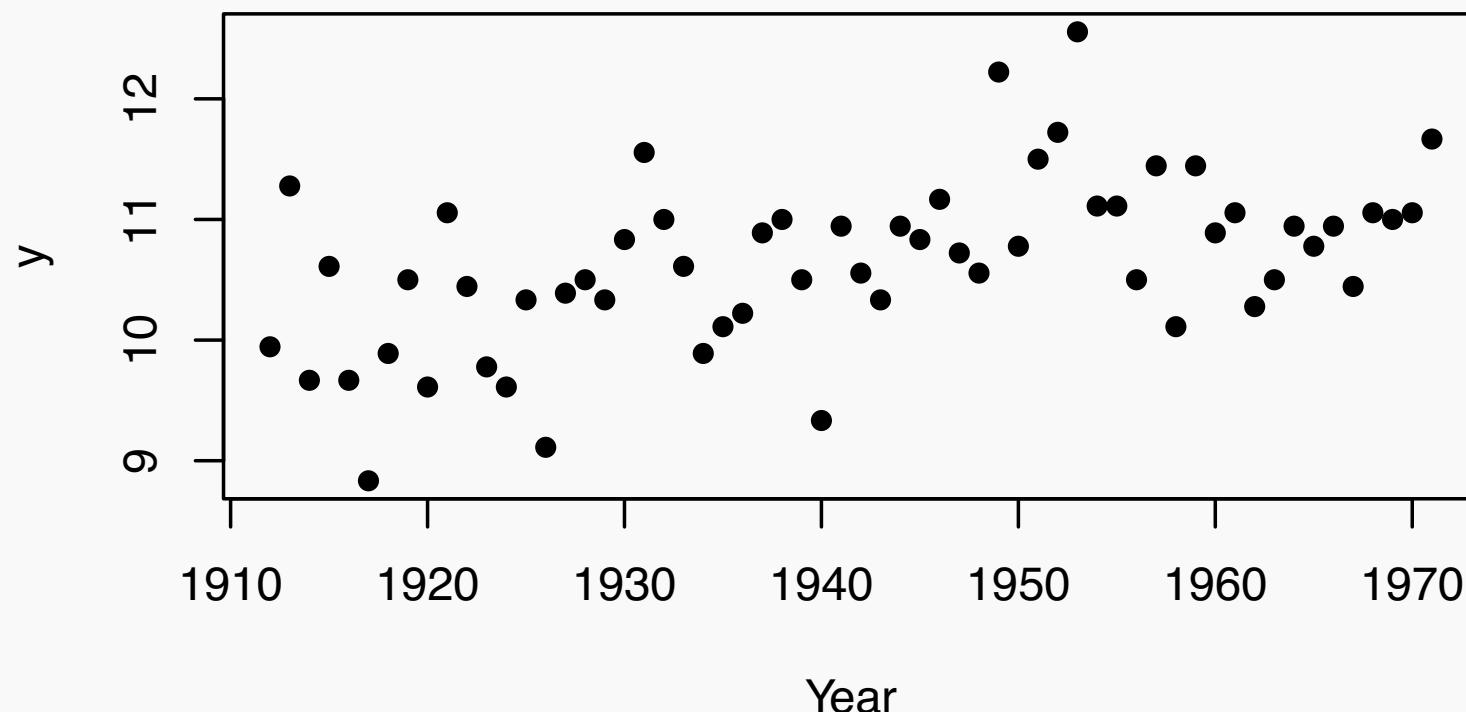


Figure 1: Taken from Wood's book, page 20

An example

Consider the following record of 60 mean annual temperatures in New Haven, expressed in $^{\circ}\text{C}$

```
y <- (nhtemp - 32) / 1.8  
plot(1912:1971, y, pch = 16, xlab = "Year", ylab = "y")
```



Example: Model 1

A first model simply assumes that the data are a random sample from a normal distribution namely they are the observation of i.i.d. r.v. from $\mathcal{N}(\mu, \sigma^2)$.

It follows that the distribution for the entire data vector is the product of the single contributions

$$f(\mathbf{y}) = \prod_{i=1}^n \frac{1}{\sigma} \phi \left\{ (y_i - \mu) / \sigma \right\},$$

$$\begin{aligned}\phi(z) &= \frac{1}{\sqrt{2\pi}} \exp(-z^2) \\ f(x | \mu, \sigma^2) &= \frac{1}{\sigma} \phi\left(\frac{x-\mu}{\sigma}\right)\end{aligned}$$

where ϕ is the $\mathcal{N}(0, 1)$ p.d.f.

Example: Model 2

A second model retains the random sample assumption, but replaces the normal distribution with a heavier-tailed t_5 distribution, assuming

$$\frac{Y_i - \mu}{\sigma} \sim t_5 .$$

The distribution of the data becomes

$$f(\mathbf{y}) = \prod_{i=1}^n \frac{1}{\sigma} f_{t_5} \left\{ (y_i - \mu) / \sigma \right\} ,$$

where f_{t_5} is the t_5 p.d.f.

Example: Model 3

The third model relaxes the assumption of identical distribution, assuming a linear trend over time: after setting $t_i = \text{year}_i - 1911$, $i = 1, \dots, 60$; we then take

$$y_i = \beta_0 + \beta_1 t_i + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2).$$

The independence between observations still holds, so that

$$f(\mathbf{y}) = \prod_{i=1}^n \frac{1}{\sigma} \phi \left\{ (y_i - \beta_0 - \beta_1 t_i) / \sigma \right\}.$$

Example: Model 4

The last model maintains the trend assumption, but also includes *autocorrelation* for the error term, meaning that we assume

$$y_i = \beta_0 + \beta_1 t_i + \varepsilon_i, \quad \varepsilon_i = \rho \varepsilon_{i-1} + v_i,$$

with $v_i \sim \mathcal{N}(0, \sigma^2)$, and the autocorrelation $\rho \in (-1, 1)$.

The model also requires to specify the distribution for the first observation, here taken as $Y_1 \sim \mathcal{N}\{\beta_0, \sigma^2/(1 - \rho^2)\}$, so that all the variables in the sample have the same variance.

Example: Model 4 (cont'd.)

The model is an instance of a **linear regression model with autocorrelated errors**. The r.v. of the sample are not longer independent, yet the distribution of \mathbf{Y} can be found with some algebra.

It is possible to verify that \mathbf{Y} is multivariate normal, with mean vector given by the linear trend

$$E(Y_i) = \mu_i = \beta_0 + \beta_1 t_i ,$$

and covariance matrix

$$\boldsymbol{\Sigma} = \frac{\sigma^2}{(1 - \rho^2)} \begin{pmatrix} 1 & \rho & \cdots & \rho^{n-1} \\ \rho & 1 & \cdots & \rho^{n-2} \\ \vdots & \vdots & \vdots & \vdots \\ \rho^{n-1} & \rho^{n-2} & \cdots & 1 \end{pmatrix},$$

so that $f(\mathbf{y}) = \phi_n(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$, being ϕ_n the multivariate normal p.d.f.

Example: model parameters

It is useful to write down the vector parameters θ for each of the four model specifications proposed:

- Model 1: $\theta = (\mu, \sigma^2)$
- Model 2: $\theta = (\mu, \sigma^2)$
- Model 3: $\theta = (\beta_0, \beta_1, \sigma^2)$
- Model 4: $\theta = (\beta_0, \beta_1, \rho, \sigma^2)$

Note that the meaning of each parameter depends on the chosen model:
 $\sigma^2 = \text{var}(Y_i)$ in Model 1, but $\sigma^2 = 0.6 \text{ var}(Y_i)$ in Model 2.

Simulation from a statistical model

Simulation from a statistical model

A decent model would allow to simulate data sets reproducing some of the features of the observed data, with better models providing more realistic results.

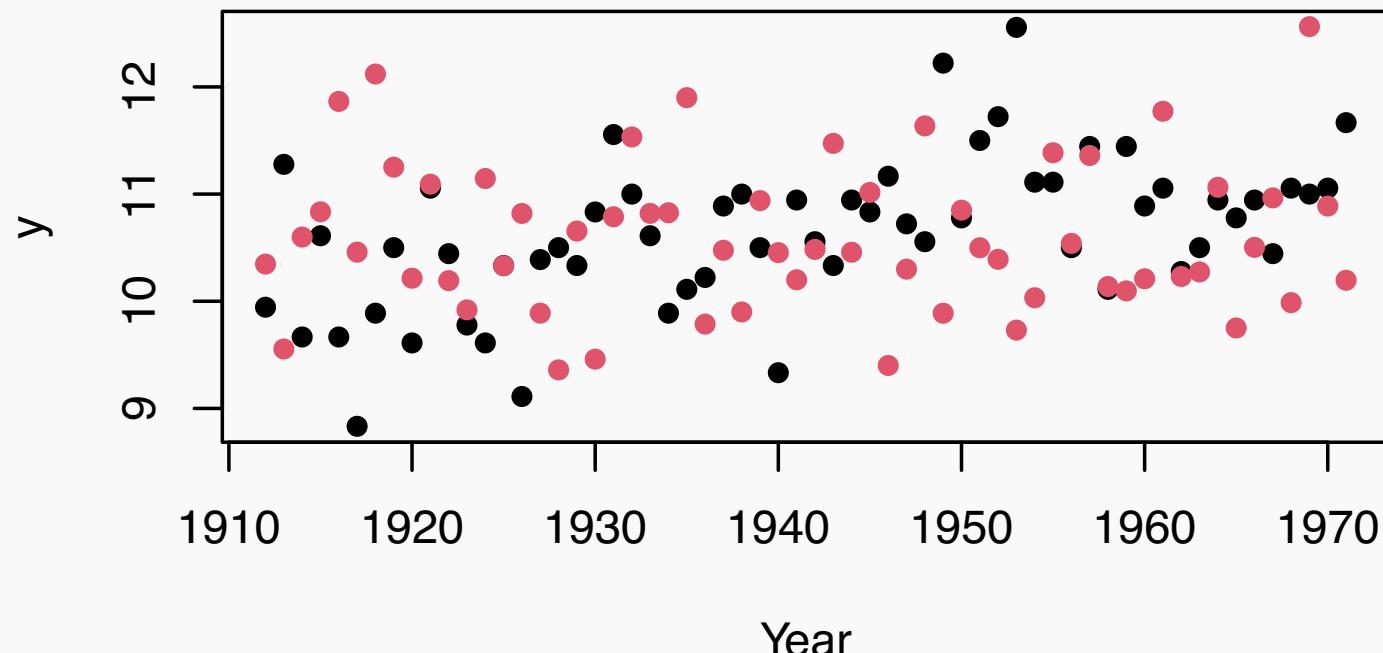
Simulation is an essential part of modern statistical inference. Its role is not only for the assessment of a candidate statistical model, but also to obtain **predictions** based on a chosen model.

Simulation requires that a value for the model parameters θ is chosen beforehand. This task is accomplished by **parameter estimation**, which would be illustrated later on.

Example: Model 1

For Model 1, the parameters μ and σ^2 are readily estimated by \bar{y} and s^2 . Then, a further dataset can be simulated using such values

```
set.seed(2018); ysim <- rnorm(length(y), m = mean(y), s = sd(y))
plot(1912:1971, y, pch = 16, xlab = "Year", ylab = "y")
points(1912:1971, ysim, col = 2, pch = 16)
```



Example: what should we look for?

In order to evaluate whether the simulated dataset is similar to the observed one, we should focus on some important features.

For example, climate changes over time may suggest that the temperature of a given year may be positively correlated with the temperature of the subsequent year, an example of *positive autocorrelation*.

We can quantify this point by computing the **sample autocorrelation**

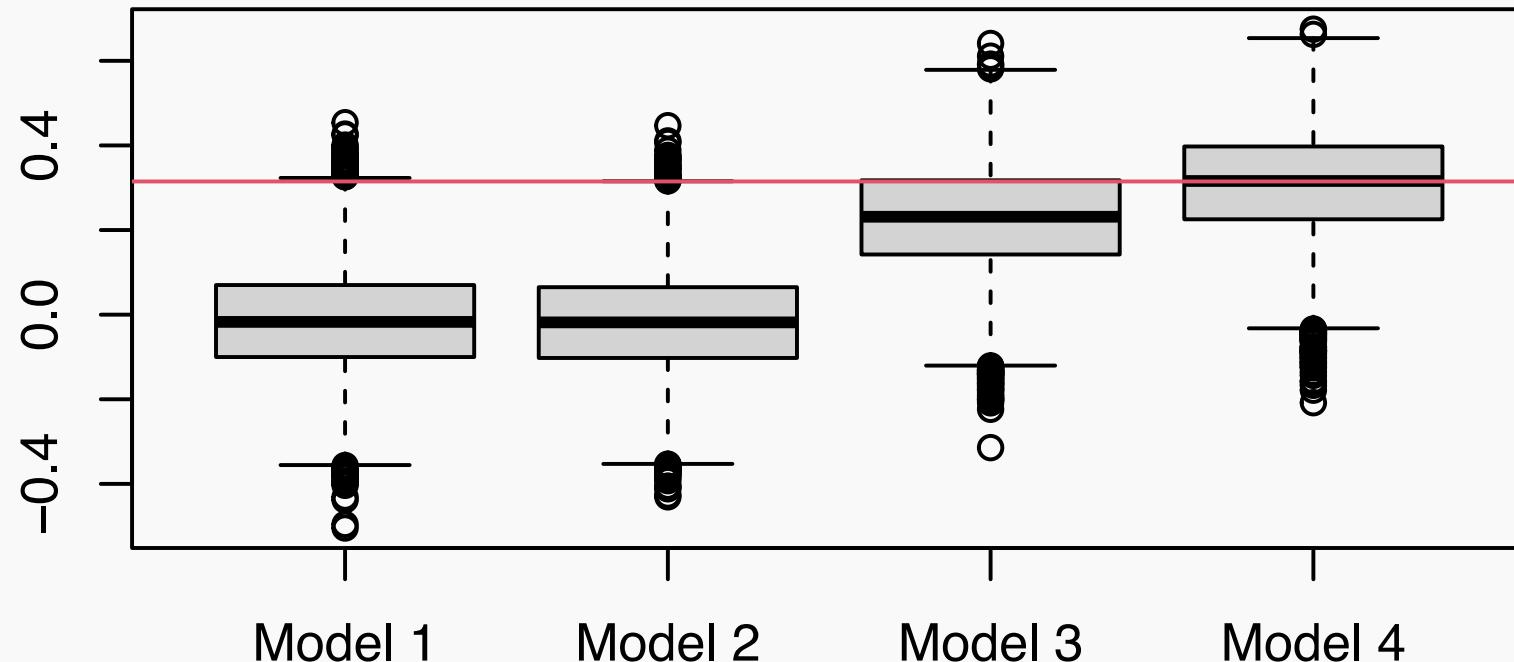
$$r_1 = \frac{\sum_{i=1}^{n-1} (y_i - \bar{y})(y_{i+1} - \bar{y})}{\sum_{i=1}^n (y_i - \bar{y})^2},$$

which is computed by the R function `acf`.

For the original data set $r_1 = 0.31$, whereas for the simulated data from Model 1 $r_1 = -0.12$. This is just a single data set, though.

Example: Simulated sample autocorrelation

We simulate 10,000 samples from each of the four models, and each time we compute the r_1 coefficient. The sample distributions obtained are displayed in the plot below. Model 4 is better at reproducing autocorrelation, as expected.



The problems of statistical inference: an overview

Inferential questions

Given a statistical model for data \mathbf{y} , with model parameters θ , there are some basic questions to ask (pasted from the CS book):

1. What values of θ are most consistent with \mathbf{y} ? [*Point estimation*]
2. What range of values of θ are consistent with \mathbf{y} ? [*Interval estimation*]
3. Is some prespecified restriction on θ consistent with \mathbf{y} ? [*Hypothesis testing*]
4. Is the model consistent with the data for any values of θ at all?
[*Model checking*]

Question 4 can be enlarged to include *which of several alternative models is most consistent with \mathbf{y}* ? This is point of *model selection*, which partially overlaps with model checking.

The central issue is the acknowledgment of the **intrinsic uncertainty** inherent in trying to learn about θ .

A further question

For settings where some control over the data-gathering process is possible, a further question arises:

5. How might the data-gathering process be organized to produce data that enables answers to the preceding questions to be as accurate and precise as possible?

This is the *core of experimental and survey design methods*.

It represents an often neglected question, of central importance in many traditional fields where statistics is routinely applied (medical sciences, industrial research, biosciences . . .). It is also very relevant for business and web analytics, like in *A/B testing*.

Approaches to statistical inference

There are two classes of methods providing an answer to questions 1-4, namely the **frequentist** and **Bayesian** approach.

They differ mainly for the role of model parameters θ , which are treated as fixed constants in the former approach and as r.v. in the latter one.

The difference may appear remarkable, and there has been controversy over the years about the merits of each approach.

Yet, from a *practical perspective* the two approaches have much in common, and tend to give similar answers when properly applied, especially when compared to approaches that are not based on a statistical model.

Roadmap

In the rest of this course, a brief overview of classical frequentist methods for point estimation, interval estimation, hypothesis testing and design will be provided. The important idea of the bootstrap will be also illustrated.

Afterwards, the most important frequentist class of methods, given by **likelihood-based methods**, will be covered. This is rather comprehensive methodology, that provides also some tools for model selection.

Model checking will be illustrated with reference to some specific class of statistical models, such as **linear and generalized linear regression models**, whose theory will be covered in the course as well. We will skim over some important extensions, such as **nonparametric regression and mixed models**.

Some (limited) space will be devoted also to the main ideas of the Bayesian approach.

A first look at model diagnostics

Model diagnostics, a basic tool for model checking, it also has a role for simple models, like those of our illustrative example.

A basic tool is given by quantile-quantile plots, already briefly introduced, which can be used to verify whether the data \mathbf{y} are consistent with an assumed model.

This is straightforward for i.i.d. models, like Model 1 and 2, where the fact that the assumed distribution for y_i depends on μ and σ is rather inconsequential.

A first look at model diagnostics (cont'd.)

For more complex settings, such as Model 3 and 4, the general idea is as follows.

Assume that according to the fitted model the expected value and covariance matrix of \mathbf{y} are $\hat{\boldsymbol{\mu}}_\theta$ and $\hat{\boldsymbol{\Sigma}}_\theta$.

Then the **standardized residuals** are

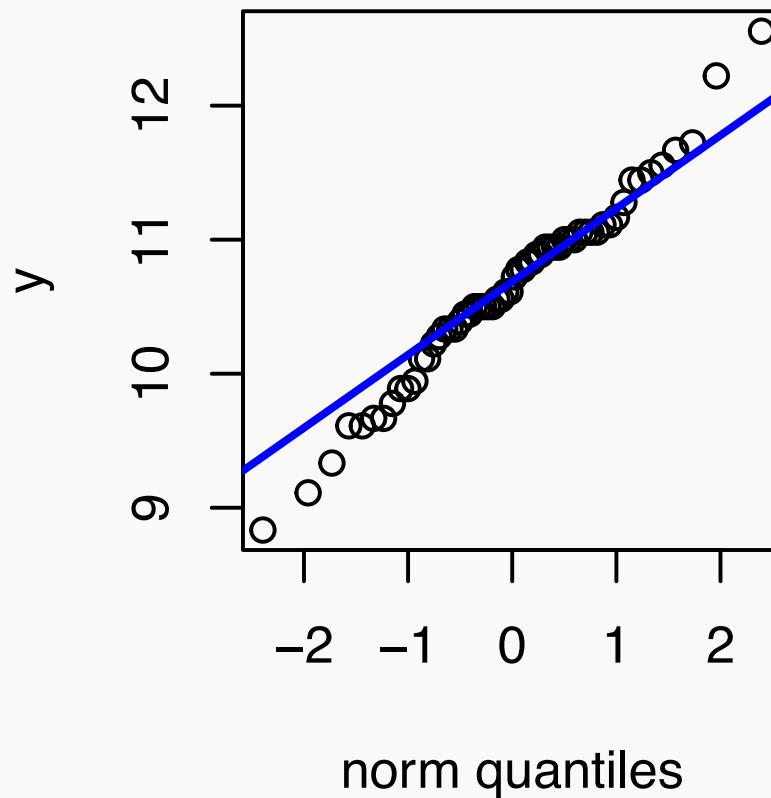
$$\hat{\boldsymbol{\varepsilon}} = \hat{\boldsymbol{\Sigma}}_\theta^{-1/2} (\mathbf{y} - \hat{\boldsymbol{\mu}}_\theta),$$

where $\hat{\boldsymbol{\Sigma}}_\theta^{-1/2}$ is any matrix *square root* of $\hat{\boldsymbol{\Sigma}}_\theta^{-1}$, such as its Choleski factor.

If the model is correct, $\hat{\boldsymbol{\varepsilon}}$ should appear approximately independent, with zero mean and unit variance, and roughly normal if the model assumes normality.

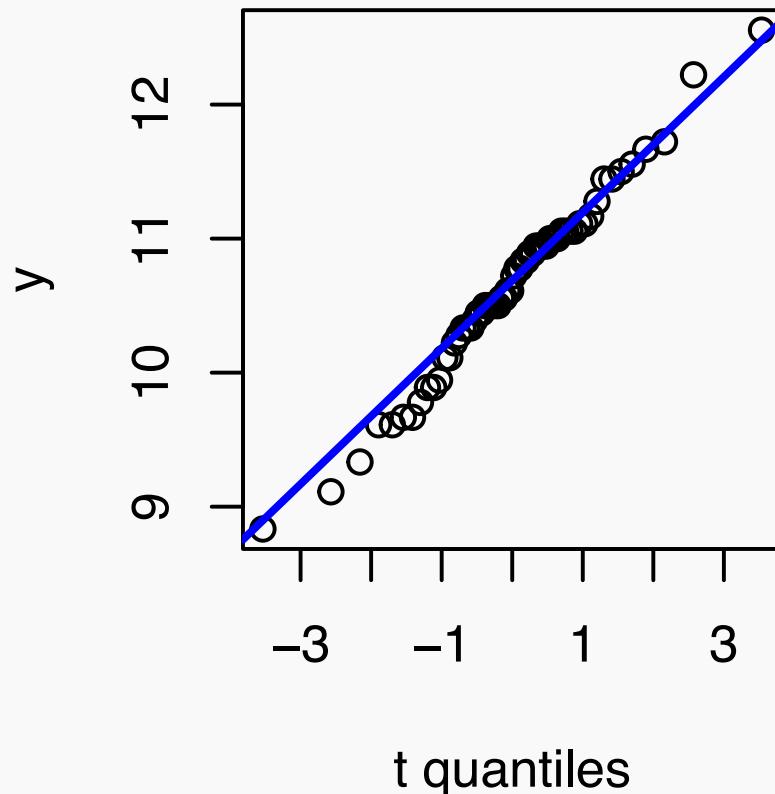
Example: model checking for Model 1

```
par(pty="s")
library(car)
qqPlot(y, dist="norm", envelope=FALSE, grid=FALSE, id=FALSE)
```



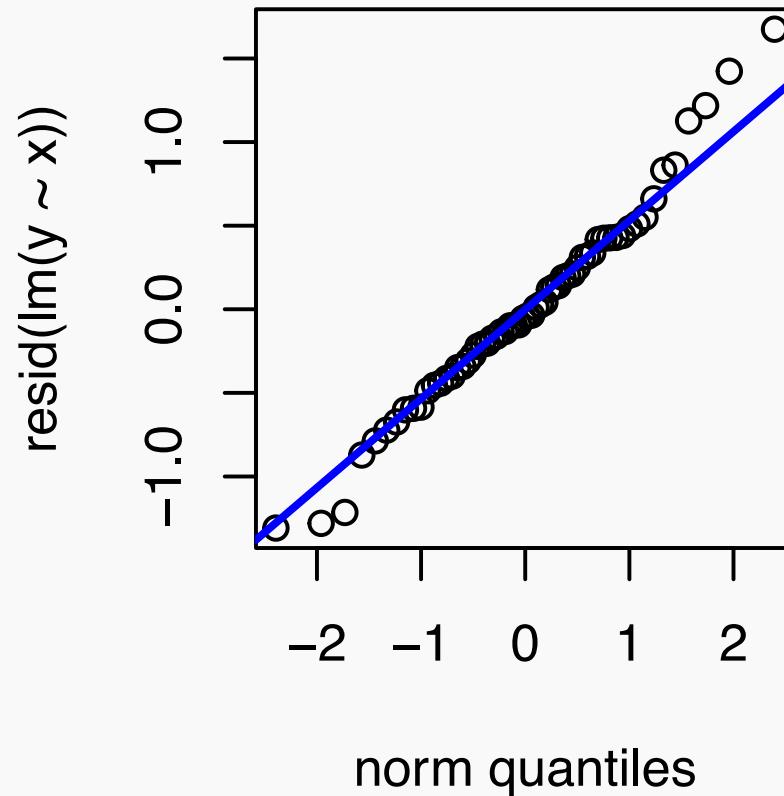
Example: model checking for Model 2

```
par(pty="s")
qqPlot(y, dist="t", df=5, envelope=FALSE, grid=FALSE, id=FALSE)
```



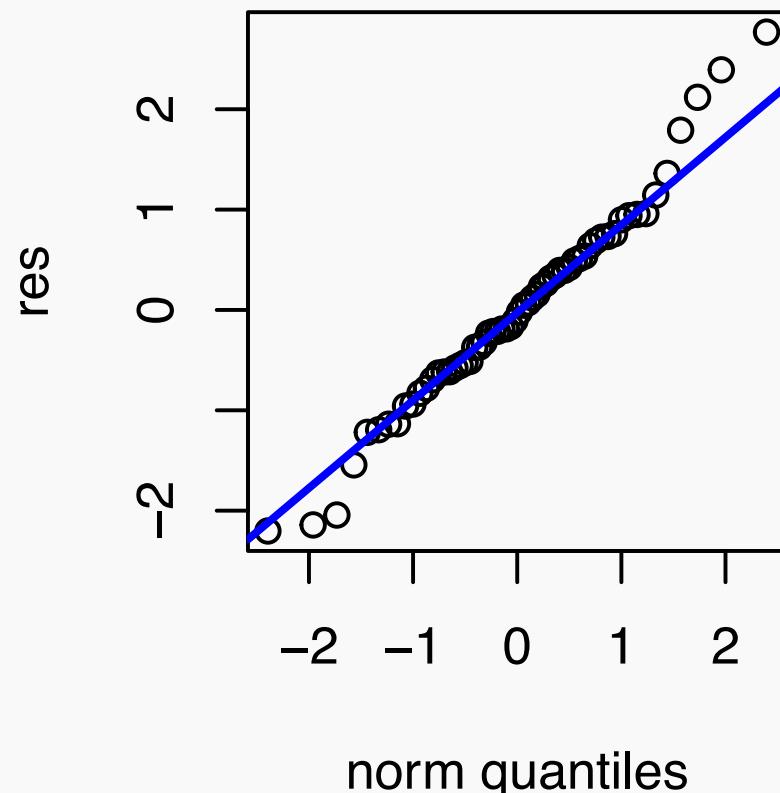
Example: model checking for Model 3

```
par(pty="s")
x <- 1912:1971-1911
qqPlot(resid(lm(y~x)), envelope=FALSE, grid=FALSE, id=FALSE)
```



Example: model checking for Model 4

```
par(pty="s")
qqPlot(res, dist="norm", envelope=FALSE, grid=FALSE, id=FALSE)
```

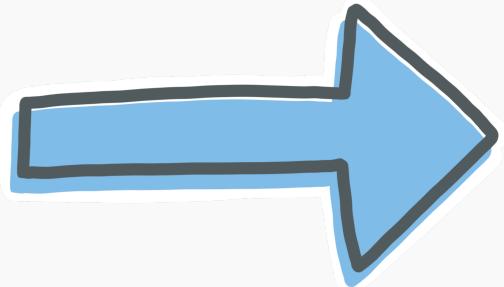


Example: winding up

The example shows that no model gives a perfect fit for this data set, a fact that we ought to accept in broad generality.

Model 3 and Model 4 both provide an acceptable fit, with the latter slightly better in reproducing some of the autocorrelation observed in data.

More sophisticated models may give better results, but **simple models** conform to the **Occam's Razor principle**, that for statistical modelling argues in favor of *simple models for simple problems*, moving to more complex models when simple models are inappropriate.



Parameter Estimation

N. Torelli, G. Di Credico, V. Gioia

Fall 2023

University of Trieste

Point estimation¹

Interval estimation²

¹Agresti, Kateri: sec 3-4 - 4.1

²Agresti, Kateri: sec 4.3 - 4.4

Point estimation

The aim of point estimation

Given a model for the data \mathbf{y} , with parameter θ , **point estimation** is concerned with finding a reasonable parameter estimate from the data.

There are several methods for doing this, and the problem can be simply stated as *finding the parameter value most consistent with the data*, a definition that leads to the method of **maximum likelihood estimation**.

We will delve into the details of maximum likelihood estimation in due time, but here we focus on some general aspects of point estimation.

Example: sample mean and sample variance

A very simple model assumes that the data are a random sample from a normal distribution namely they are the observations of i.i.d. r.v. from $\mathcal{N}(\mu, \sigma^2)$.

Straightforward estimates of μ and σ^2 are given by **the sample mean**

$$\hat{\mu} = \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$$

and by the **sample variance**

$$\hat{\sigma}^2 = s^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2.$$

Such estimates are actually sensible anytime we are interested in estimating the mean and variance of an i.i.d. sample.

Estimation properties

To figure out what could be a good estimate, we need to consider *repeated estimation under repeated replication of the data-generating process.*

This makes fully sense whenever the available data are a random sample obtained from a large population, like in industrial or social surveys, so that it would perfectly possible to iterate the sampling and obtain further data with the same structure of y .

However, we apply the same logic even when repetition is just the result of an idealization, like in the case of the temperatures recorded in New Haven of the previous lecture.

The point is: what do we expect to find if we repeat the same analysis to many data sets generated from the same model?

Unbiasedness

If we replicate the random data and we repeat the estimation process, the result will be a different value of $\hat{\theta}$ for each replicate.

The values are observations of a random vector, the **estimator** of θ , which is usually also denoted by $\hat{\theta}$ (the context will make clear whether we are referring to the estimator or to the estimate for a given sample).

Since, the estimator is a r.v., it makes fully sense to compute its mean.

For an **unbiased** estimator

$$E(\hat{\theta}) = \theta .$$

Unbiasedness is a desirable property, and we would also like the estimator to have **low variance**.

Mean Squared Error

There is *tradeoff* between unbiasedness and low variance, so we usually seek to get both (to some extent): ideally we would target a small **Mean Squared Error (MSE)**

$$\text{MSE}(\hat{\theta}) = E\{(\hat{\theta} - \theta)^2\}.$$

With some algebra, we obtain

$$\text{MSE}(\hat{\theta}) = \{E(\hat{\theta}) - \theta\}^2 + \text{var}(\hat{\theta}) = \text{Squared bias} + \text{Variance}.$$

Example: normal random sample

For a normal random sample, it is straightforward to verify that

$$E(\bar{Y}) = \mu, \quad \text{var}(\bar{Y}) = \frac{\sigma^2}{n} = \text{MSE}(\bar{Y}).$$

For the sample variance, we use the property that

$$\frac{(n-1)}{\sigma^2} S^2 \sim \chi_{n-1}^2,$$

to obtain

$$E(S^2) = \sigma^2, \quad \text{var}(S^2) = \frac{2\sigma^4}{(n-1)} = \text{MSE}(S^2).$$

The unbiasedness of the sample mean and variance is a general property, holding also for non-normal samples.

Consistency

A (scalar) estimator is said to be **(weakly) consistent** if, for any $\epsilon > 0$

$$\Pr(|\hat{\theta} - \theta| > \epsilon) \rightarrow 0, \quad \text{as } n \rightarrow \infty.$$

A sufficient condition for this is that $MSE(\hat{\theta}) \rightarrow 0$ for large samples, which requires that both bias and variance become negligible.

The law of large samples implies that the sample mean is a consistent estimator for the true mean in random samples.

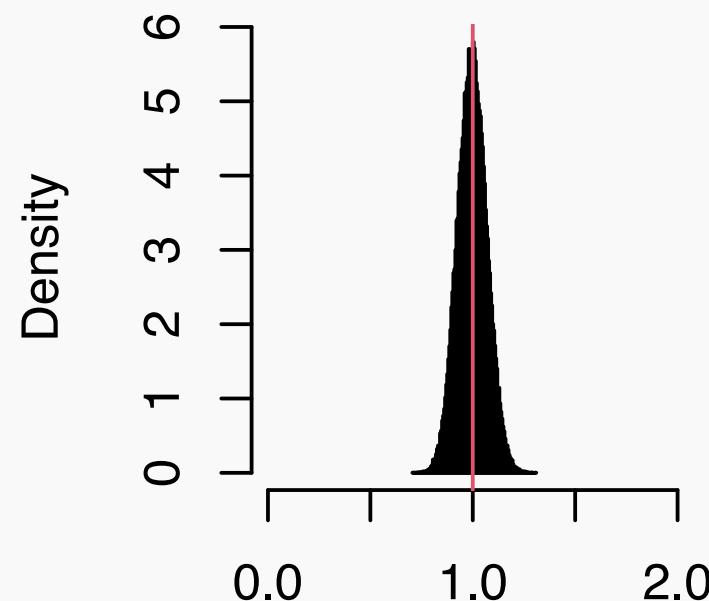
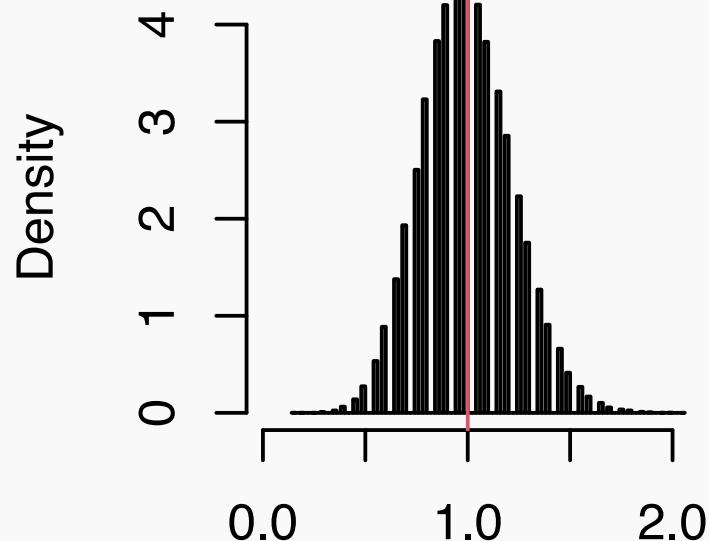
R lab: consistency of the sample mean

```
M <- 100000; n1 <- 20; n2 <- 200; y1 <- y2 <- rep(NA, M)
for(i in 1:M) {y1[i] <- mean(rpois(n1, 1))
                y2[i] <- mean(rpois(n2, 1))}

par(mfrow=c(1,2))

hist.scott(y1, xlim=c(0,2), main="", xlab="")
abline(v=1,col=2)

hist.scott(y2, xlim=c(0,2), main="", xlab="")
abline(v=1,col=2)
```



Efficiency

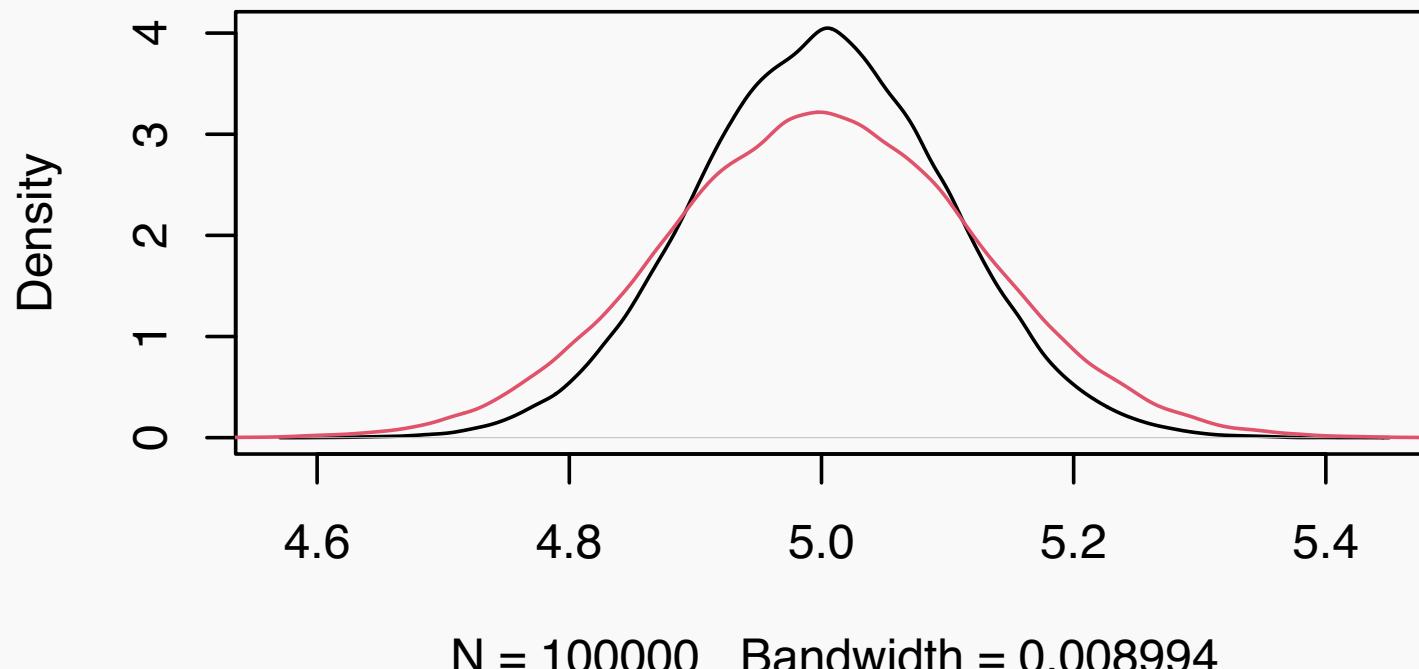
An **efficient estimator** is an estimator that estimates the parameter of interest in some *optimal* manner.

Among estimators with negligible bias, efficiency is associated to small variance. Since this is the case of consistent estimators, they are usually compared in terms of their variance.

R lab: efficiency of the sample mean

For a normal random sample, both the sample mean and sample median are consistent estimators of μ . The mean is more efficient.

```
M <- 100000; n <- 100; mat.y <- matrix(NA, nrow = M, ncol = 2)
for(i in 1:M) {y <- rnorm(n, 5)
  mat.y[i,] <- c(mean(y), median(y))}
plot(density(mat.y[,1]), type="l", main="")
lines(density(mat.y[,2]), col=2)
```



Standard Error

An important quantity defined for a (scalar) estimator is given by its **standard error**, defined as

$$\text{SE}(\hat{\theta}) = \sqrt{\text{var}(\hat{\theta})}.$$

Once a sample is observed, and a numerical estimate of θ obtained, then the estimated standard error is obtained by replacing θ by $\hat{\theta}$.

An example is the **standard error of the mean** $\text{SE}(\bar{Y}) = \sigma/\sqrt{n}$, which is estimated by s/\sqrt{n} .

In applications, the estimated standard error is routinely reported along with the estimate, since it quantifies the **estimation precision**.

The delta method

Suppose that we are interested in a parameter which is a function of a scalar parameter θ , namely

$$\psi = g(\theta), \quad \text{for a continuous and differentiable function } g.$$

If $\widehat{\theta}$ is a consistent estimator of θ , then the **continuous mapping theorem** ensures that $g(\widehat{\theta})$ is consistent for ψ .

Its standard error is provided by the **delta method**, stating that

$$\text{SE}(\widehat{\psi}) \doteq \text{SE}(\widehat{\theta}) |g'(\theta)|,$$

with the approximation becoming more accurate for larger samples.

The result can be extended to settings with multiple parameters.

Robust estimation

A **robust** estimator has good performances across a wide range of statistical models for the data.

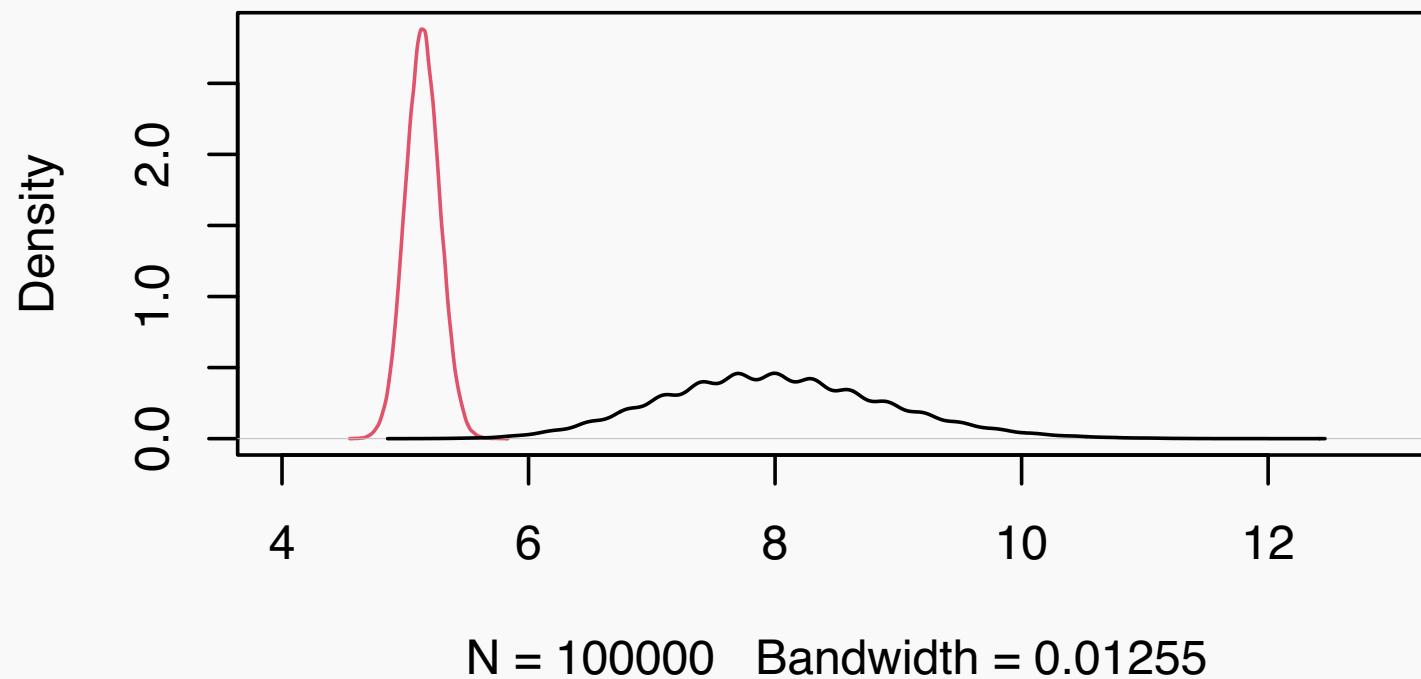
The **sample median** is a robust estimation of location, not affected by possible outlying data, quite the opposite of the sample mean.

Robust estimation trades some efficiency with resistance to outliers, and they are often a sensible choice for semi-automatic data analyses.

R lab: robustness of the sample median

```
M <- 100000; n <- 100; mat.y <- matrix(NA, nrow = M, ncol = 2)
for(i in 1:M) { x <- rbinom(n, size = 1, prob = 0.9)
                  y <- x * rnorm(n, 5) + (1 - x) * rnorm(n, 35)
                  mat.y[i,] <- c(mean(y), median(y))}

plot(density(mat.y[,2]), type="l", main="", xlim=c(4, 13),
      col = 2)
lines(density(mat.y[,1]), col=1)
```



Interval estimation

The aim of interval estimation

Confidence intervals provide more satisfactory estimation results than point estimates alone, giving an entire set of values to estimate the model parameter.

They are built by considering a single parameter at a time.

Extensions to multidimensional *confidence regions* exist, but they are seldom used in practice.

Pivots

Confidence intervals make suitable usage of **pivots**, which are **functions of the data and the parameter whose distribution is known**.

A notable example is the following one for a random sample from a $\mathcal{N}(\mu, \sigma^2)$ distribution, when the parameter of interest is the mean μ , and σ^2 is not known (so that $\theta = (\mu, \sigma^2)$):

$$T(\mu) = \frac{\bar{Y} - \mu}{\sqrt{\frac{S^2}{n}}} \sim t_{n-1}, \quad \forall \mu \in \mathbb{R}, \sigma^2 > 0$$

Obtaining a confidence interval

In the normal random sample example, from the previous pivot property it follows that (for $0 < \alpha < 1$)

$$\Pr(t_{n-1;\alpha/2} \leq T(\mu) \leq t_{n-1;1-\alpha/2}) = 1 - \alpha,$$

where $t_{n-1;\alpha}$ is the α quantile of a t_{n-1} distribution; due to symmetry of the latter, $t_{n-1;\alpha/2} = -t_{n-1;1-\alpha/2}$.

With some simple algebra, the previous property is equivalent to

$$\Pr\left(\bar{Y} - t_{n-1;1-\alpha/2} \sqrt{\frac{S^2}{n}} \leq \mu \leq \bar{Y} + t_{n-1;1-\alpha/2} \sqrt{\frac{S^2}{n}}\right) = 1 - \alpha.$$

Definition of confidence interval

Hence the *random interval* with endpoints

$$\bar{Y} - t_{n-1;1-\alpha/2} \sqrt{\frac{S^2}{n}}, \quad \bar{Y} + t_{n-1;1-\alpha/2} \sqrt{\frac{S^2}{n}}$$

contains μ with probability $(1 - \alpha)$.

This interval is called a $(1 - \alpha) \times 100\%$ **confidence interval**.

Common choices are $(1 - \alpha) = 0.95$ or $(1 - \alpha) = 0.99$.

Interpretation

Given a particular set of data y_1, \dots, y_n we calculate the confidence interval by replacing \bar{Y} and S^2 with their observed values \bar{y} and s^2

$$\bar{y} - t_{n-1;1-\alpha/2} \sqrt{\frac{s^2}{n}}, \quad \bar{y} + t_{n-1;1-\alpha/2} \sqrt{\frac{s^2}{n}}$$

This interval *either does or does not contain the true value of μ .*

The probability interpretation previously introduced refers to an *hypothetical sequence of sets of data* generated from the statistical model.

R lab: confidence interval

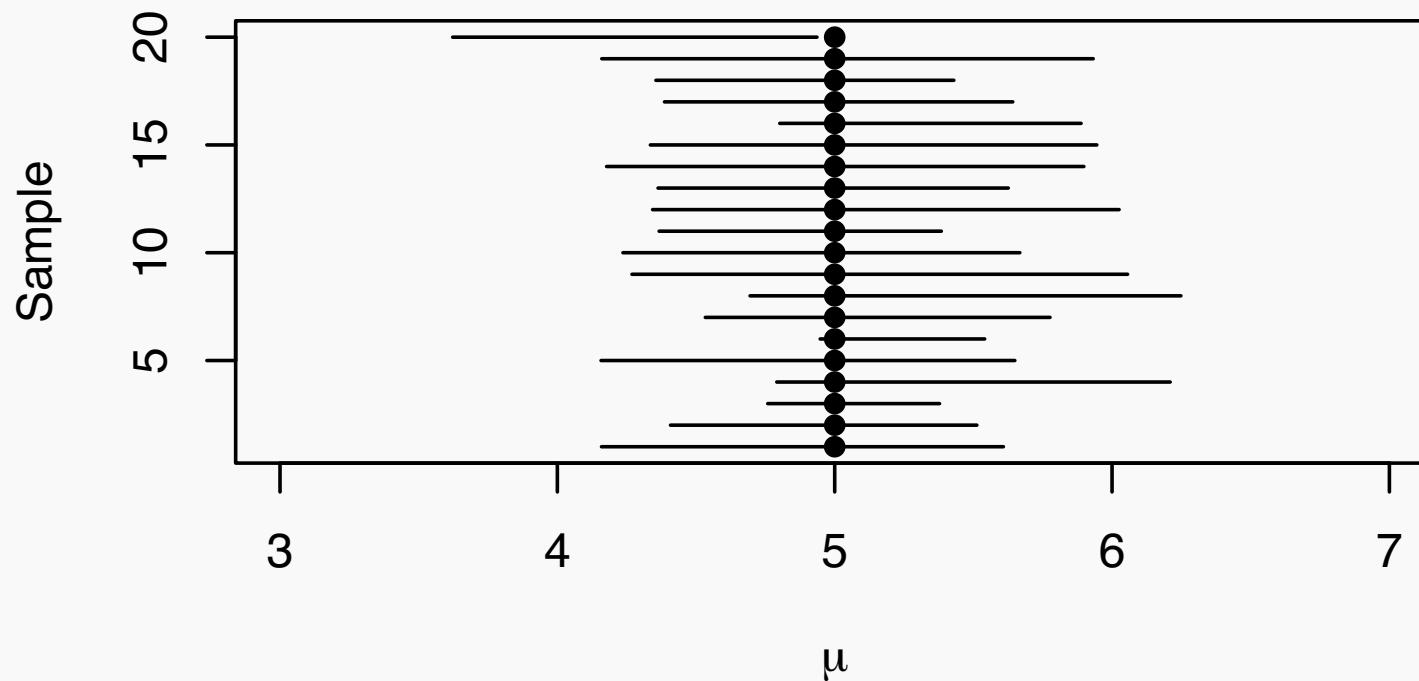
```
M <- 100000; n <- 10; mat.ci <- matrix(NA, nrow = M, ncol = 2)
for(i in 1:M) { y <- rnorm(n, 5)
                  se_t <- sqrt(var(y) / n) * qt(0.975, n-1)
                  mat.ci[i,] <- mean(y) + se_t * c(-1, 1)}
mean(mat.ci[,1] < 5 & mat.ci[,2] > 5)

## [1] 0.94909
```

R lab: visualizing confidence intervals

We can visualize the first 20 simulated confidence intervals, expecting that (on average) 19 out of 20 will include the true μ

```
plot(rep(5, 20), 1:20, pch = 16, ylab="Sample",
     xlab=expression(mu))
for(i in 1:20) segments(mat.ci[i,1], i, mat.ci[i,2], i)
```



One-sided confidence intervals

If we lift the equi-tailed condition, we can define infinitely many intervals such that

$$\Pr \left(\bar{Y} - t_{n-1;1-\alpha_1} \sqrt{\frac{S^2}{n}} \leq \mu \leq \bar{Y} + t_{n-1;1-\alpha_2} \sqrt{\frac{S^2}{n}} \right) = 1 - \alpha,$$

where $\alpha_1 + \alpha_2 = \alpha$.

Other than the standard choice $\alpha_1 = \alpha_2 = \alpha/2$, other notable choices are $\alpha_1 = 0$ (which makes the lower limit equal to $-\infty$) or $\alpha_2 = 0$ (which makes the upper limit equal to ∞).

They are called **one-sided confidence intervals**, and are sometimes employed in applications.

Approximate confidence intervals & coverage probability

Exact pivots are scarce, but approximate ones are easy to find.

A common one is the **Wald pivot** for a generic parameter of interest ψ , based on a consistent estimator which is approximately normally distributed for large samples

$$Z(\psi) = \frac{\hat{\psi} - \psi}{\text{SE}(\hat{\psi})} \stackrel{\sim}{\sim} \mathcal{N}(0, 1), \quad \forall \psi \in \Psi$$

The corresponding confidence interval is

$$\hat{\psi} - z_{1-\alpha/2} \text{SE}(\hat{\psi}), \quad \hat{\psi} + z_{1-\alpha/2} \text{SE}(\hat{\psi})$$

The Central Limit Theorem provides such a solution for random samples, when ψ corresponds to the mean of each variable.

R lab: approximate confidence intervals

```
M <- 100000; n <- 10; mat.ci <- matrix(NA, nrow = M, ncol = 2)
for(i in 1:M) { y <- rnorm(n, 5)
                  se_z <- sqrt(var(y) / n) * qnorm(0.975)
                  mat.ci[i,] <- mean(y) + se_z * c(-1, 1)}
mean(mat.ci[,1] < 5 & mat.ci[,2] > 5)

## [1] 0.91904

M <- 100000; n <- 100; mat.ci <- matrix(NA, nrow = M, ncol = 2)
for(i in 1:M) { y <- rnorm(n, 5)
                  se_z <- sqrt(var(y) / n) * qnorm(0.975)
                  mat.ci[i,] <- mean(y) + se_z * c(-1, 1)}
mean(mat.ci[,1] < 5 & mat.ci[,2] > 5)

## [1] 0.94676
```

Confidence interval for a proportion

The method for approximate intervals can be readily used for confidence intervals on a proportion π , the success probability of a random sample of n binary variables,

$$Y_i \sim \mathcal{B}(1, \pi), \quad i = 1, \dots, n.$$

Here the pivot is

$$Z(\pi) = \frac{\bar{Y} - \pi}{\sqrt{\frac{\bar{Y}(1 - \bar{Y})}{n}}} \stackrel{d}{\sim} \mathcal{N}(0, 1), \quad \forall \pi \in (0, 1),$$

since $\hat{\pi} = \bar{Y}$ and $\text{SE}(\hat{\pi}) = \sqrt{\frac{\pi(1 - \pi)}{n}}$, which is estimated by plugging-in $\hat{\pi}$ in place of π .

R lab: confidence interval for a proportion

```
M <- 100000; n <- 50; mat.ci <- matrix(NA, nrow = M, ncol = 2)
for(i in 1:M) { y <- rbinom(n, size = 1, prob = 0.25)
  p.hat <- mean(y)
  se_z <- sqrt(p.hat * (1 - p.hat) / n)
  se_qz <- se_z * qnorm(0.975)
  mat.ci[i,] <- mean(y) + se_qz * c(-1, 1)}
mean(mat.ci[,1] < 0.25 & mat.ci[,2] > 0.25)

## [1] 0.94063
```

Confidence interval for a difference of means

An important application concerns the computation of the confidence interval for the difference between two means $\delta = \mu_X - \mu_Y$.

For two independent (and large) random samples, the approximate normal pivot is

$$Z(\delta) = \frac{\hat{\delta} - \delta}{\text{SE}(\hat{\delta})},$$

with $\hat{\delta} = \bar{X} - \bar{Y}$ and $\text{SE}(\hat{\delta}) = \sqrt{\text{SE}(\bar{X})^2 + \text{SE}(\bar{Y})^2}$.

Again, for normal samples, exact solutions exist, both for the case of equal variances and for the case of unequal variances.

Likelihood theory: Maximum likelihood estimation

(An overview)

N. Torelli, G. Di Credico, V. Gioia

Fall 2023

University of Trieste

The likelihood function¹

Maximum likelihood estimation: theory

Some numerical aspects

¹Agresti, Kateri: sec 4.2

The likelihood function

The likelihood function

Introduced by Sir Ronald Fisher, the **likelihood function** for a certain statistical model $f_\theta(\mathbf{y})$ for the data \mathbf{y} is given by the following function of the parameter θ

$$\begin{aligned} L &: \Theta \rightarrow \mathbb{R}^+ \\ \theta &\rightarrow c(\mathbf{y}) f_\theta(\mathbf{y}), \end{aligned}$$

where $c(\mathbf{y}) > 0$ is an arbitrary constant of proportionality.

We may write $L(\theta; \mathbf{y})$ to stress the fact that the data enter the function, though its argument is given by θ .

Interpreting the likelihood function

The likelihood function assigns support (*credibility*) to possible values of θ , meaning that if $L(\theta_1) > L(\theta_2)$ then θ_1 is more supported by the observed data than θ_2 .

So the *likelihood ratio* $L(\theta_1)/L(\theta_2)$ allows for the comparison between θ_1 and θ_2 ; note that the constant $c(\mathbf{y})$ cancels out.

A mathematical justification for the above interpretation is given by the **Wald inequality**: if θ_t is the **true parameter value**, then

$$E_{\theta_t} \{\log L(\theta_t; \mathbf{Y})\} > E_{\theta_t} \{\log L(\theta; \mathbf{Y})\} \quad \theta \neq \theta_t .$$

The above fact can be proven by straightforward application of the Jensen's inequality.

The log likelihood function

In the previous slide the **log likelihood function** has been introduced, which is simply the logarithm of $L(\theta)$, namely

$$\ell(\theta) = \log L(\theta).$$

The log likelihood function carries the same information of the likelihood function, but it is much more manageable. Indeed, for a random sample

$$L(\theta) = \prod_{i=1}^n f_\theta(y_i)$$

but

$$\ell(\theta) = \sum_{i=1}^n \log f_\theta(y_i).$$

Notice that $\ell(\theta)$ is defined up to an additive constant, depending only on the data \mathbf{y} .

Example 1: the Poisson model

For a random sample y_1, \dots, y_n , with $Y_i \sim \mathcal{P}(\lambda)$ i.i.d., we readily get

$$L(\lambda) = \frac{\lambda^{\sum_{i=1}^n y_i} \exp\{-n\lambda\}}{\prod_{i=1}^n y_i!},$$

so that

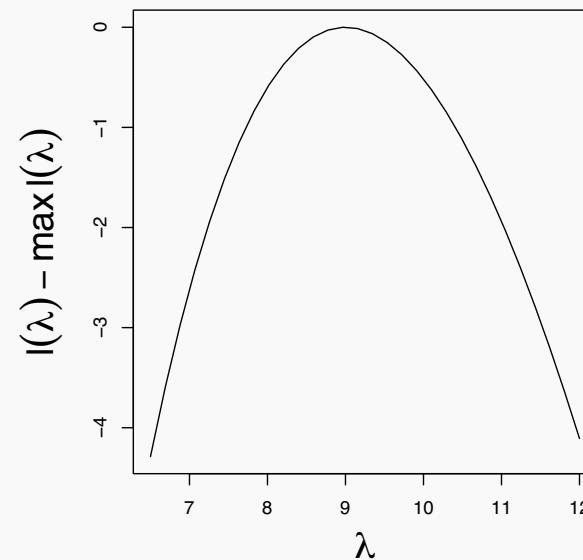
$$\ell(\lambda) = \log(\lambda) \sum_{i=1}^n y_i - n\lambda,$$

neglecting the term which does not depend on λ .

R lab: the Poisson log likelihood

Assume that for a sample $n = 10$ we observe $\sum_i y_i = 90$.

```
lik_pois <- function(lam, n, sumy) log(lam) * sumy - n * lam
xx <- seq(6.5, 12, l = 30)
ll <- sapply(xx, lik_pois, sumy = 90, n = 10)
par(pty = "s")
plot(xx, ll - max(ll), type = "l", xlab = expression(lambda),
     ylab = expression(l(lambda)-max(l(lambda))), cex.lab = 2)
```



Example 2: the normal model

For a random sample y_1, \dots, y_n , with $Y_i \sim \mathcal{N}(\mu, \sigma^2)$ i.i.d.

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

and then with some simple algebra

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

Sufficient statistics

The definition of sufficient statistic, given in the probability part, can be re-interpreted for the log likelihood function: $t(\mathbf{y})$ is **sufficient** for θ if $L(\theta)$ can be written as

$$L(\theta) = h(\mathbf{y}) g_\theta\{t(\mathbf{y})\}.$$

The **minimal sufficient statistic** allows for the maximal reduction of dimensionality, in the sense that a minimal sufficient statistic is a function of every other sufficient statistic.

For the Poisson model, the $\sum_i y_i$ (or, equivalently, the sample mean \bar{y}) is sufficient for λ , whereas for the normal model the sufficient statistic is given by the pair $(\sum_i y_i, \sum_i y_i^2)$ (or, equivalently, by the pair (\bar{y}, s^2)).

These two statistical models are an instance of an **exponential family**, an important model class that includes also other important elements, such as the binomial distribution. They play an important role in the theory of *generalized linear models*.

Maximum likelihood estimation

Given the interpretation of the (log) likelihood, the maximum of $\ell(\theta)$ is the value of the parameter which is most supported by the data.

A natural step is to take it as the point estimate, the **maximum likelihood estimate** (MLE) of θ

$$\hat{\theta} = \operatorname{argmax}_{\theta \in \Theta} \ell(\theta)$$

Notice that since $\ell(\theta)$ is also a function of \mathbf{y} , the MLE is a statistic.

The MLE in the two examples

For the Poisson model, simple calculus gives

$$\hat{\lambda} = \frac{1}{n} \sum_{i=1}^y y_i = \bar{y}.$$

For the normal model, we need to maximize a function of two variables, and we get

$$\begin{cases} \hat{\mu} = \bar{y} \\ \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2. \end{cases}$$

MLE: comments

Maximum likelihood estimation has a **central role** in modern statistics (and machine learning). There are several reasons for this:

1. The MLE algorithm is **automatic**: given a parametric statistical model for the data, the MLE follows from the chosen model.
2. The MLE of a function of a parameter $\psi = g(\theta)$ is defined by the simple plug-in rule $\hat{\psi} = g(\hat{\theta})$, which is very convenient for applications.
3. The MLE has **excellent properties**, which we illustrate in what follows.

Maximum likelihood estimation: theory

Likelihood quantities

The first two derivatives of $\ell(\theta)$ play an important role.

The vector of first derivatives is called the **score function**

$$U(\theta) = U(\theta; \mathbf{y}) = \frac{\partial \ell(\theta)}{\partial \theta}$$

The matrix of second derivatives, with negative sign, is called the **observed information matrix**:

$$J(\theta) = J(\theta; \mathbf{y}) = -\frac{\partial^2 \ell(\theta)}{\partial \theta \partial \theta^\top}$$

Some properties

The derivatives of the log likelihood function satisfy some important properties, provided that some **regularity conditions** hold (we shall return on them later on).

The proofs are simple, and they are reported in the CS book.

1. *Zero expected score*

$$E_{\theta} \{ U(\theta; \mathbf{Y}) \} = \mathbf{0}$$

2. *2nd Bartlett identity*

$$\text{cov}_{\theta} \{ U(\theta; \mathbf{Y}) \} = E_{\theta} \{ J(\theta; \mathbf{Y}) \} = \mathcal{I}(\theta)$$

The expected value $\mathcal{I}(\theta)$ of the observed information matrix is called the *Fisher information matrix* (or just the *expected information matrix*).

The Cramér-Rao lower bound

The third property is important, and we first state it for a one-parameter model (scalar θ).

3. *The Cramér-Rao lower bound:* the variance of *any unbiased estimator* $\tilde{\theta}$ cannot be smaller than the reciprocal of the expected information:

$$\text{var}_\theta\{\tilde{\theta}(\mathbf{Y})\} \geq \frac{1}{\mathcal{I}(\theta)}.$$

Actually, by differentiation of the unbiasedness condition with respect to θ it follows that $\text{cov}_\theta\{\tilde{\theta}, U(\theta; \mathbf{Y})\} = 1$, which readily implies the Cramér-lower bound.

The extension to multiparameter models is given by the condition that the matrix $\text{cov}(\tilde{\boldsymbol{\theta}}) = \mathcal{I}(\boldsymbol{\theta})^{-1}$ is positive semi-definite.

Consistency of MLE

We are ready to state the first crucial property of the MLE:

Maximum likelihood estimators are usually consistent, that is if the sample size tends to infinity $\hat{\theta}$ tends to θ_t , the **true** parameter value.

A justification for the result is given by the fact that in regular situations $\ell(\theta)/n \rightarrow E_{\theta}\{\ell(\theta)\}/n$ as $n \rightarrow \infty$, so that eventually the maximum of $\ell(\theta)$ and $E\{\ell(\theta)\}$ must coincide at θ_t by the Wald inequality.

The formal proof (typically) employs the law of large numbers.

Consistency can fail if the number of parameters increases with the sample size.

Large-sample distribution of MLE

We establish it by a Taylor expansion for the score function:

$$U(\hat{\theta}) \doteq U(\theta_t) - (\hat{\theta} - \theta_t) J(\theta_t),$$

with equality when $n \rightarrow \infty$ since $\hat{\theta} - \theta_t \rightarrow \mathbf{0}$.

From the definition of $\hat{\theta}$, we get $U(\hat{\theta}) = \mathbf{0}$. Under mild assumptions

$$\frac{J(\theta_t)}{n} \rightarrow \frac{\mathcal{I}(\theta_t)}{n},$$

whereas $U(\theta_t)$ is a random vector with mean vector $\mathbf{0}$ and covariance matrix $\mathcal{I}(\theta_t)$.

In the large sample limit

$$\hat{\theta} - \theta_t \stackrel{\text{d}}{\sim} \mathcal{I}(\theta_t)^{-1} U(\theta_t; \mathbf{y}),$$

implying that $E(\hat{\theta} - \theta_t) = \mathbf{0}$ and $\text{cov}(\hat{\theta} - \theta_t) = \mathcal{I}(\theta_t)^{-1}$.

Large-sample normality of MLE

In the case when the sample is formed by independent observations, it follows that the log likelihood is the sum of independent contributions: under mild conditions the central limit theorem applies, and in the large sample limit

$$\hat{\boldsymbol{\theta}} \stackrel{d}{\sim} \mathcal{N}\{\boldsymbol{\theta}_t, \mathcal{I}(\boldsymbol{\theta}_t)^{-1}\}.$$

Notice that whenever this holds, it would be possible (and recommendable, in some sense) to use $J(\boldsymbol{\theta}_t)$ in place of $\mathcal{I}(\boldsymbol{\theta}_t)$.

Again, since $\boldsymbol{\theta}_t$ is unknown, we replace it by $\hat{\boldsymbol{\theta}}$, obtaining the following estimated standard error for the k -th component of $\boldsymbol{\theta}$

$$\text{SE}(\hat{\boldsymbol{\theta}}_k) = \sqrt{\left[J(\hat{\boldsymbol{\theta}})^{-1}\right]_{kk}}$$

Note: for *regular models* (see next slide), the observed information is positive definite at $\hat{\boldsymbol{\theta}}$, so that the SE above is well defined.

Regularity conditions

We end the summary of the theory by mentioning the **regularity conditions**, which are some assumptions on the statistical model, required for the previous results to be valid.

The CS book lists the following ones:

1. The pdf of \mathbf{y} defined by different values of θ are distinct, namely the model is *identifiable*.
2. The true parameter value θ_t is interior to Θ .
3. Within some neighbourhood of θ_t , the first three derivatives of $\ell(\theta)$ exist and are bounded, while the expected information satisfies the 2nd Bartlett identity, is positive definite and finite.

These are mild conditions, which are generally valid in most cases.

Winding up

The previous results have illustrated that

1. The MLE is a **consistent estimator**.
2. The MLE is **asymptotic efficient**, since its asymptotic variance attains the Cramér-Rao lower bound.
3. The large sample distribution (aka the approximate distribution) of the MLE is **multivariate normal**, with standard error that can be estimated by the observed information evaluated at the parameter estimate.

Example 1: Poisson model

Here $\hat{\lambda} = \bar{y}$, and consistency follows from the law of large numbers, in agreement with likelihood theory.

Furthermore, the CLT states that for large n

$$\hat{\lambda} \stackrel{d}{\sim} \mathcal{N}(\lambda, \lambda/n).$$

This result can be obtained also from likelihood theory. Indeed, we get

$$J(\lambda) = \frac{\sum_i y_i}{\lambda^2}$$

so that $\mathcal{I}(\lambda) = n/\lambda$ and $\mathcal{I}(\lambda)^{-1} = \lambda/n$.

Example 2: normal example

Here we get

$$J(\mu, \sigma^2) = \begin{pmatrix} \frac{n}{\sigma^2} & \frac{n}{\sigma^4} (\bar{y} - \mu) \\ \frac{n}{\sigma^4} (\bar{y} - \mu) & -\frac{n}{2\sigma^4} + \frac{1}{\sigma^6} \sum_{i=1}^n (y_i - \mu)^2 \end{pmatrix}$$

and therefore

$$\mathcal{I}(\mu, \sigma^2) = \begin{pmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{pmatrix}$$

The implication is that $\hat{\mu}$ and $\hat{\sigma}^2$ are (asymptotically) uncorrelated, and the two estimated standard errors are

$$\text{SE}(\hat{\mu}) = \frac{\hat{\sigma}}{\sqrt{n}}, \quad \text{SE}(\hat{\sigma}^2) = \frac{\sqrt{2}\hat{\sigma}^2}{\sqrt{n}}.$$

Some numerical aspects

Numerical optimisation

The algorithmic nature of the MLE estimation method translates the statistical model into an optimisation problem: once a (sensible) statistical model has been specified for the data, we obtain parameter estimates with excellent properties by maximizing the log likelihood.

In some simple settings, like in the examples above, it is possible to find the analytical expression for the MLE, but in general we must resort to **numerical optimisation** of the log likelihood.

There are indeed several methods available for the task. Some knowledge of the most important issues related to it turns out particularly useful even for the application of off-the-shelf routines in R (or other environments).

Newton's method

Newton's method for optimisation is commonly used for minimization, in this case of the objective function $f(\theta) = -\ell(\theta)$.

The theory is well described in the CS book, here we mention the most important aspects. The idea is to locally approximate $f(\theta)$ as a quadratic function, which is repeatedly minimised.

The resulting method consists in an **iterative algorithm**, which is started with $k = 0$ and a *guesstimate* $\theta^{[0]}$, and iterates the following steps:

1. Evaluate $\ell(\theta^{[k]})$, $U(\theta^{[k]})$ and $J(\theta^{[k]})$.
2. If $U(\theta^{[k]}) \doteq \mathbf{0}$ and $J(\theta^{[k]})$ is positive definite then stop.
3. If $\mathbf{H} = J(\theta^{[k]})$ is not positive definite, perturb it so that it is.
4. Solve $\mathbf{H}\delta = U(\theta^{[k]})$ for the search direction δ .
5. If $\ell(\theta^{[k]} + \delta)$ is not $> \ell(\theta^{[k]})$, repeatedly halve δ until it is (*this is the step-length control*).
6. Set $\theta^{[k+1]} = \theta^{[k]} + \delta$, increment k by one and return to step 1.

Fisher scoring and Quasi-Newton.

Whenever available, it is always a good idea to replace the observed information with the expected information $\mathcal{I}(\boldsymbol{\theta}^{[k]})$ in the Newton's method.

The resulting algorithm has a long successful tradition in statistics, it is called **Fisher scoring** and, indeed, it has better convergence properties.

Another variant avoids the computation of either $J(\boldsymbol{\theta}^{[k]})$ or $\mathcal{I}(\boldsymbol{\theta}^{[k]})$, by building an approximation to the second derivative of $\ell(\boldsymbol{\theta})$ as the optimization proceeds. This is the approach of the **Quasi-Newton** methods, such as the widely used BFGS algorithm.

Quasi-Newton methods are implemented in several R functions and packages; see the CRAN Task View for *Optimisation* (<https://cran.r-project.org/web/views/Optimization.html>).

An example: logistic regression

We follow the MASS book for a simple example on a dose-response model.

Namely, we assume that y_i is the number of dead budworms (out of 20) for a dose of insecticide x_i^* . In particular, the statistical model is

$$Y_i \sim \mathcal{B}_i(20, \pi_i) \quad i = 1, \dots, 12, \text{ independent}$$

with

$$\pi_i(\alpha, \beta) = \frac{e^{\alpha + \beta x_i}}{1 + e^{\alpha + \beta x_i}}$$

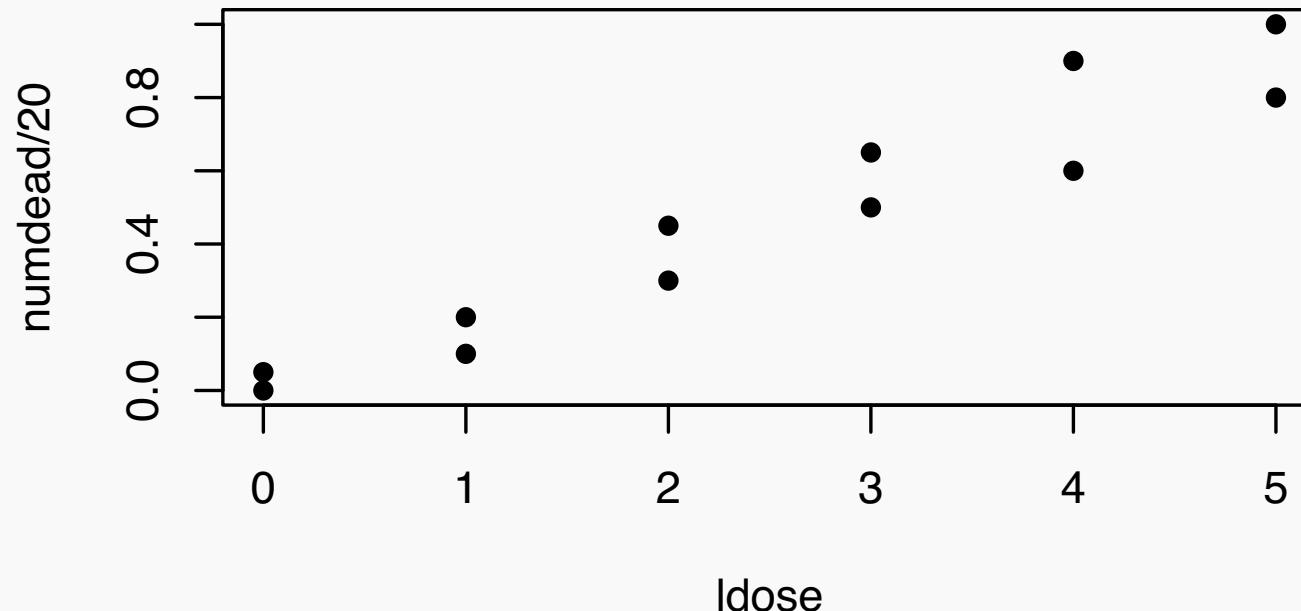
with $x_i = \log(x_i^*)$.

This is a simple instance of a *logistic regression model*.

R lab: budworm data

There are two observations at each dose (M/F budworms), but here for the sake of simplicity we ignore the different sex.

```
ldose <- rep(0:5, 2)  
numdead <- c(1, 4, 9, 13, 18, 20, 0, 2, 6, 10, 12, 16)  
plot(ldose, numdead / 20, pch=16)
```



Logistic regression: likelihood quantities

With some simple algebra we get:

$$\ell(\alpha, \beta) = \sum_i \left\{ y_i (\alpha + \beta x_i) - 20 \log(1 + e^{\alpha + \beta x_i}) \right\}$$

$$U(\alpha, \beta) = \begin{pmatrix} \sum_i \{y_i - 20 \pi_i(\alpha, \beta)\} \\ \sum_i \{y_i - 20 \pi_i(\alpha, \beta)\} x_i \end{pmatrix}$$

$$\mathcal{I}(\alpha, \beta) = \begin{pmatrix} \sum_i 20 \pi_i(\alpha, \beta) \{1 - \pi_i(\alpha, \beta)\} & \sum_i 20 \pi_i(\alpha, \beta) \{1 - \pi_i(\alpha, \beta)\} x_i \\ \sum_i 20 \pi_i(\alpha, \beta) \{1 - \pi_i(\alpha, \beta)\} x_i & \sum_i 20 \pi_i(\alpha, \beta) \{1 - \pi_i(\alpha, \beta)\} x_i^2 \end{pmatrix}$$

Notice that for this model $J(\alpha, \beta) = \mathcal{I}(\alpha, \beta)$.

R lab: likelihood and score functions

```
loglik <- function(theta, data){  
    eta <- theta[1] + theta[2] * data$x  
    out <- sum(data$y * eta - 20 * log(1+exp(eta)))  
    return(out)  
}  
  
score <- function(theta, data){  
    prob <- plogis(theta[1] + theta[2] * data$x)  
    out <- c(sum(data$y - prob * 20),  
            sum((data$y - prob * 20) * data$x))  
    return(out)  
}
```

R lab: information function

```
info <- function(theta, data){  
    prob <- plogis(theta[1] + theta[2] * data$x)  
    info11 <- sum(20 * prob * (1-prob))  
    info12 <- sum(20 * prob * (1-prob) * data$x)  
    info22 <- sum(20 * prob * (1-prob) * data$x^2)  
    out <- matrix(c(info11, info12, info12, info22), 2, 2)  
    return(out)  
}
```

R lab: starting point

Let's start from $\alpha = \beta = 0$: we obtain

```
theta0 <- c(0, 0); budw <- data.frame(y = numdead, x = ldose)

loglik(theta0, budw)

## [1] -166.3553

score(theta0, budw)

## [1] -9 105

info(theta0, budw)

##      [,1] [,2]
## [1,]    60   150
## [2,]   150   550
```

R lab: first step

```
H <- info(theta0, budw)
u0 <- score(theta0, budw)
delta <- solve(H, u0)
theta1 <- theta0 + delta

theta1
## [1] -1.9714286  0.7285714

loglik(theta1, budw)
## [1] -114.7219
```

which is clearly an improvement.

R lab: first 10 steps

```
## k = 1 theta= -1.971429 0.7285714 loglik= -114.7219
## k = 2 theta= -2.621436 0.9572079 loglik= -111.8192
## k = 3 theta= -2.760585 1.004947 loglik= -111.734
## k = 4 theta= -2.766079 1.006804 loglik= -111.7339
## k = 5 theta= -2.766087 1.006807 loglik= -111.7339
## k = 6 theta= -2.766087 1.006807 loglik= -111.7339
## k = 7 theta= -2.766087 1.006807 loglik= -111.7339
## k = 8 theta= -2.766087 1.006807 loglik= -111.7339
## k = 9 theta= -2.766087 1.006807 loglik= -111.7339
## k = 10 theta= -2.766087 1.006807 loglik= -111.7339
```

The algorithm converges quickly, and actually after 10 iterations

```
cat(score(theta10, budw), det(info(theta10, budw)),
    sqrt(diag(solve(info(theta10, budw)))))
```

```
## 1.776357e-15 5.329071e-15 2361.462 0.3701342 0.1235889
```

R lab: glm analysis

```
budworm.lg0 <- glm(cbind(y, 20-y) ~ x, binomial, budw)
summary(budworm.lg0, cor = FALSE)

##
## Call:
## glm(formula = cbind(y, 20 - y) ~ x, family = binomial, data =
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.7661    0.3701 -7.473 7.82e-14 ***
## x            1.0068    0.1236  8.147 3.74e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 124.876  on 11  degrees of freedom
## Residual deviance: 16.984  on 10  degrees of freedom
```

Likelihood theory: inferential results

(An overview)

N. Torelli, G. Di Credico, V. Gioia

Fall 2023

University of Trieste

Confidence intervals¹

Likelihood-based tests

Model selection

¹Agresti, Kateri: sec 5.7

Confidence intervals

Wald-type intervals

Since the theory of MLE provides a general formula for standard errors, Wald-type confidence intervals for a parameter of interest ψ are generally available (here given for $1 - \alpha = 0.95$):

$$\hat{\psi} \pm 1.96 \text{SE}(\hat{\psi})$$

The asymptotic normality of the MLE justifies the usage of normal quantiles.

Actually, the availability of a general formula for $\text{SE}(\hat{\psi})$ when $\hat{\psi}$ is the MLE supports the widespread usage of this kind of confidence intervals.

Performance of Wald-type confidence intervals

The biggest issue with Wald-type confidence intervals is that **their accuracy depends on the chosen parametrization.**

Eventually, the MLE is approximately normally distributed, but for finite sample the parametrization matters.

(That's why methods which are invariant, such as percentile bootstrap confidence intervals, are preferable).

R lab: Wald-type CI for a variance

Let us assess the coverage probability for Wald-type intervals for σ^2 of a normal random sample.

```
M <- 100000; n <- 20; mat.ci <- matrix(NA, nrow = M, ncol = 2)
for(i in 1:M) { y <- rnorm(n, s=sqrt(2)); s2 <- var(y) * (n-1)/n
  se_s2 <- sqrt(2/n) * s2 * qnorm(0.975)
  mat.ci[i,] <- s2 + se_s2 * c(-1, 1)}
mean(mat.ci[,1] < 2 & mat.ci[,2] > 2)       $\tilde{\sigma}^2 = \frac{1}{J(S)} \approx \frac{1}{\frac{n}{2G^4}}$ 
## [1] 0.86881

M <- 100000; n <- 100; mat.ci <- matrix(NA, nrow = M, ncol = 2)
for(i in 1:M) { y <- rnorm(n, s=sqrt(2)); s2 <- var(y) * (n-1)/n
  se_s2 <- sqrt(2/n) * s2 * qnorm(0.975)
  mat.ci[i,] <- s2 + se_s2 * c(-1, 1)}
mean(mat.ci[,1] < 2 & mat.ci[,2] > 2)
## [1] 0.93249
```

R lab: Wald-type CI for σ^2

Things get better, given these parameter values, if we choose $\psi = \sigma$ and then re-transform the intervals:

```
M <- 100000; n <- 20; mat.ci <- matrix(NA, nrow = M, ncol = 2)
for(i in 1:M) { y <- rnorm(n, s=sqrt(2)); s2 <- var(y) * (n-1)/n
  se_s <- sqrt(s2 / (n * 2)) * qnorm(0.975)
  mat.ci[i,] <- (sqrt(s2) + se_s * c(-1, 1))^2}
mean(mat.ci[,1] < 2 & mat.ci[,2] > 2)
## [1] 0.89632
```

$\Psi = \sigma \rightarrow \text{SEE}[\hat{\Psi}] = \frac{1}{2\hat{\sigma}} \frac{\hat{\sigma}_{\text{MLE}}^2}{\sqrt{n}} = \frac{\hat{\sigma}}{2\sqrt{n}}$

$\Psi = \sqrt{\sigma^2}$

$g(x) = \sqrt{x} = x^{1/2} \rightarrow g'(x) = \frac{1}{2} x^{-1/2}$

```
M <- 100000; n <- 100; mat.ci <- matrix(NA, nrow = M, ncol = 2)
for(i in 1:M) { y <- rnorm(n, s=sqrt(2)); s2 <- var(y) * (n-1)/n
  se_s <- sqrt(s2 / (n * 2)) * qnorm(0.975)
  mat.ci[i,] <- (sqrt(s2) + se_s * c(-1, 1))^2}
mean(mat.ci[,1] < 2 & mat.ci[,2] > 2)
## [1] 0.9395
```

Alternative methods

There are also other approaches for confidence intervals based on the likelihood function.

They are based on **likelihood-based test statistics**, taking advantage of the relation existing between tests and confidence intervals, which is a general fact.

Likelihood-based tests

The likelihood ratio test*

We saw that the likelihood ratio makes possible to choose between different parameter values. Therefore, it is not strange that the likelihood ratio can be used as test statistic, being in some sense the optimal choice, as supported by the **Neyman-Pearson lemma**.

Formally, the lemma is valid for choosing between two *simple hypotheses* $H_0 : \theta = \theta_0$ vs $H_1 : \theta = \theta_1$, for any pair of parameter values θ_0 and θ_1 .

The **likelihood ratio test statistic** is given by

$$\lambda(\mathbf{y}) = \frac{L(\theta_1)}{L(\theta_0)} = \frac{f_{\theta_1}(\mathbf{y})}{f_{\theta_0}(\mathbf{y})}$$

with rejection region

$$\mathcal{R}_\alpha = \{\mathbf{y} : \lambda(\mathbf{y}) \geq k_\alpha\},$$

being the test's *power* $\beta(\theta_0) = \Pr_{\theta_0}\{\lambda(\mathbf{Y}) \geq k_\alpha\} = \alpha$.

The Neyman-Pearson lemma*

The lemma says that given another test statistic $\lambda^*(\mathbf{y})$, with rejection region \mathcal{R}_α^* and significance level $\leq \alpha$, namely

$$\beta^*(\boldsymbol{\theta}_0) = \Pr_{\boldsymbol{\theta}_0}(\mathbf{Y} \in \mathcal{R}_\alpha^*) \leq \alpha,$$

then the likelihood ratio test is the most powerful of the two tests at $\boldsymbol{\theta}_1$, $\beta(\boldsymbol{\theta}_1) \geq \beta^*(\boldsymbol{\theta}_1)$.

Sketch of the proof:

- Define the indicator function $\phi(\mathbf{y}) = 1$ if $\mathbf{y} \in \mathcal{R}_\alpha$ and 0 otherwise; similarly define $\phi^*(\mathbf{y})$ for the other statistic.
- We get $\{\phi(\mathbf{y}) - \phi^*(\mathbf{y})\} \{f_{\boldsymbol{\theta}_1}(\mathbf{y}) - k_\alpha f_{\boldsymbol{\theta}_0}(\mathbf{y})\} \geq 0$
- Therefore

$$\begin{aligned} 0 &\leq \int_{\mathcal{Y}} \{\phi(\mathbf{y}) - \phi^*(\mathbf{y})\} \{f_{\boldsymbol{\theta}_1}(\mathbf{y}) - k_\alpha f_{\boldsymbol{\theta}_0}(\mathbf{y})\} d\mathbf{y} \\ &= \beta(\boldsymbol{\theta}_1) - \beta^*(\boldsymbol{\theta}_1) - k_\alpha \{\beta(\boldsymbol{\theta}_0) - \beta^*(\boldsymbol{\theta}_0)\} \leq \beta(\boldsymbol{\theta}_1) - \beta^*(\boldsymbol{\theta}_1) \end{aligned}$$

which means $\Pr_{\boldsymbol{\theta}_1}(\mathbf{Y} \in \mathcal{R}_\alpha) \geq \Pr_{\boldsymbol{\theta}_1}(\mathbf{Y} \in \mathcal{R}_\alpha^*)$.

Three likelihood-based tests*

We first focus on a simple one-parameter model, and on the problem of testing $H_0 : \theta = \theta_0$ vs $H_1 : \theta \neq \theta_0$.

The following three tests are available:

- The **likelihood ratio test (LRT)**

$$W(\theta_0) = 2 \{ \ell(\hat{\theta}) - \ell(\theta_0) \}$$

- The **Wald test**

$$W_e(\theta_0) = (\hat{\theta} - \theta_0)^2 J(\hat{\theta}) = \frac{(\hat{\theta} - \theta_0)^2}{\text{SE}(\hat{\theta})^2}$$

- The **score test**

$$W_u(\theta_0) = \frac{U(\theta_0)^2}{I(\theta_0)}$$

In all the three cases, we reject H_0 for large values of the statistic, so that the p -value is (for instance) $p = \Pr_{\theta_0}(W \geq w_{obs})$.

Visually*

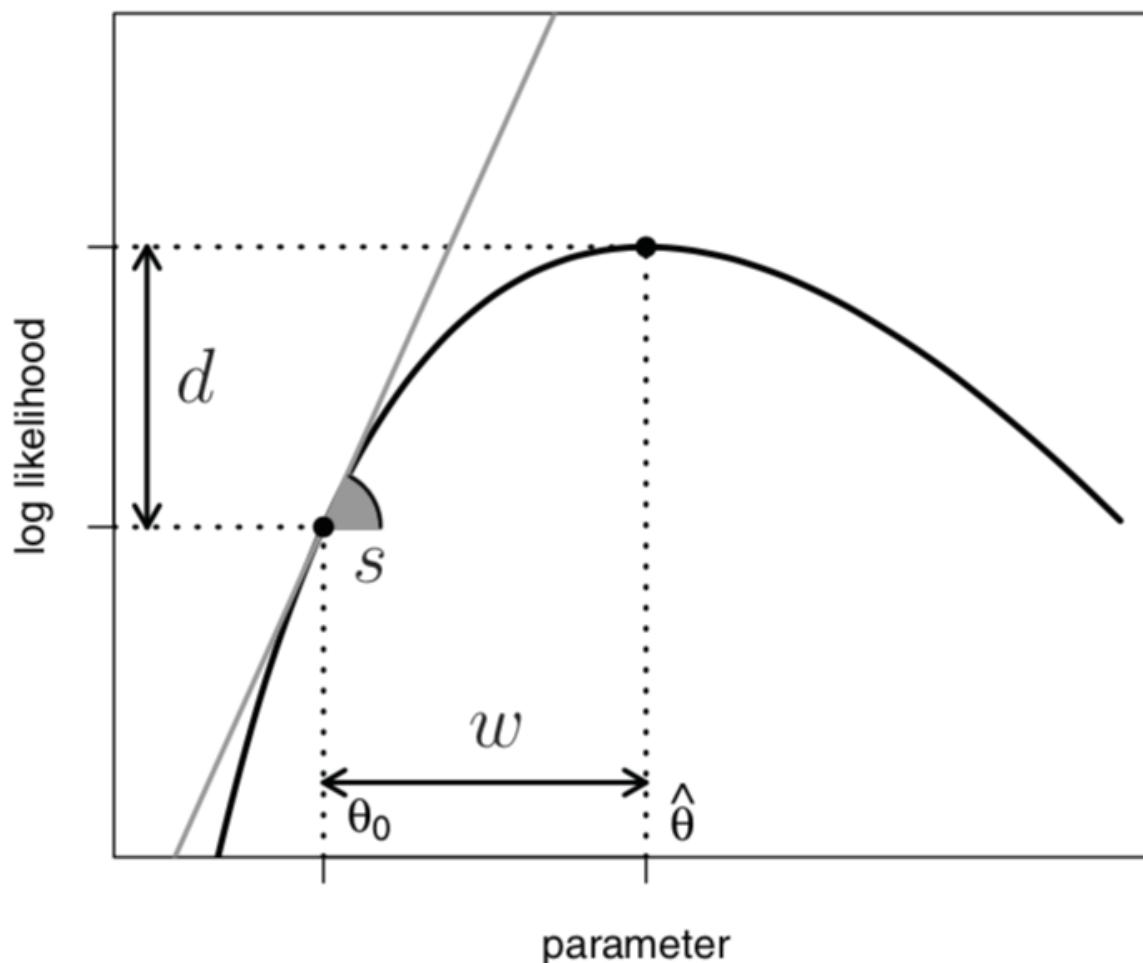


Figure 1. Comparing the three test statistics according to the traditional plot: Likelihood ratio is reported on the y scale, Wald on the x scale, and the score on the first derivative scale. The different scales do not favor understanding of the underlying connections.

Three likelihood-based tests: comments*

- Whenever available, the exact distribution of these tests can be employed.
- Which one is preferable? The likelihood ratio test is clearly an obvious choice, but for large samples the three statistics are equivalent: this fact can be proved by a Taylor expansion of $U(\hat{\theta})$ around θ_0 .
- From the asymptotic distribution of the MLE, it readily follows that the null distribution of W_e is approximately

$$W_e(\theta_0) \stackrel{d}{\sim} \chi_1^2$$

and since the two other tests are equivalent in large samples, the same result holds also for them.

- For one-sided alternatives such as $H_1 : \theta > \theta_0$, the signed squared-root versions of the test should be used, namely (for the LRT)

$$R(\theta_0) = \text{sgn}(\hat{\theta} - \theta_0) \sqrt{W(\theta_0)}$$

and, under H_0 , $R(\theta_0) \stackrel{d}{\sim} \mathcal{N}(0, 1)$.

Confidence intervals based on W

The confidence interval based on W is particularly appealing: using the relation between confidence intervals and tests, it can be written as

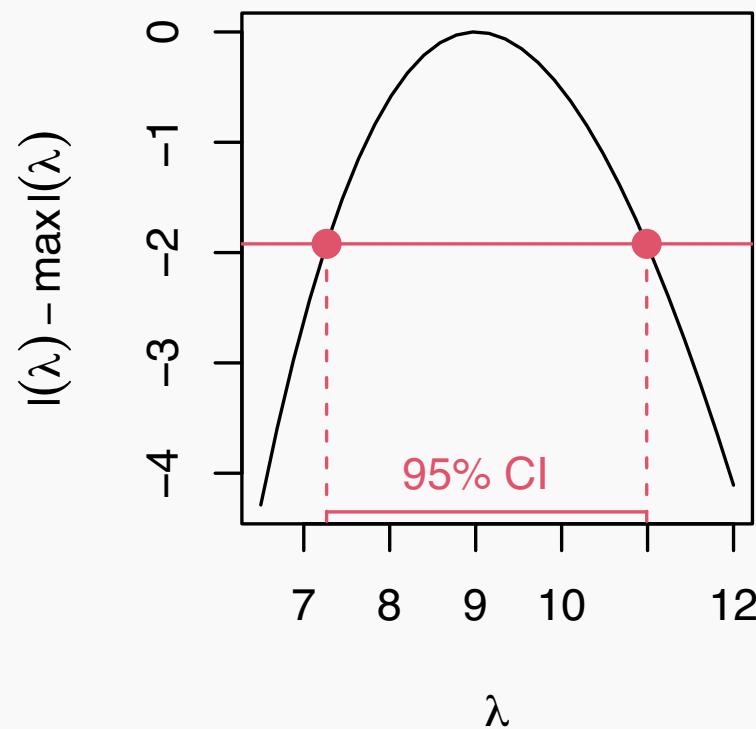
$$\{\theta : W(\theta) \leq \chi^2_{1;1-\alpha}\} = \left\{ \theta : \ell(\theta) \geq \ell(\hat{\theta}) - \frac{\chi^2_{1;1-\alpha}}{2} \right\}$$

so that the interval (which could actually be a union of intervals for multi-modal log likelihoods) includes all the parameter values with large log likelihood, i.e. the set of values most supported by the data.

The result does not depend on the parameterization (the chosen scale), differently from the Wald test.

R lab: visualizing the confidence interval based on the LRT

Back to the Poisson example (with $n = 10$ and $\sum_i y_i = 90$):



Parameter of interest and nuisance parameters*

The three tests introduced readily generalize to hypotheses on the entire p -dimensional parameter θ . For instance, the LRT would become

$$W(\theta_0) = 2 \{ \ell(\hat{\theta}) - \ell(\theta_0) \}$$

with asymptotic null distribution given by χ_p^2 .

At any rate, the typical (and most interesting) situation is where we wish to test an hypothesis on a q -dimensional subset of θ , with $q < p$.

Following the CS book, we write $\theta^\top = (\psi^\top, \gamma^\top)$, with the null and alternative hypotheses given by $H_0 : \psi = \psi_0$ vs $H_1 : \psi \neq \psi_0$.

Here ψ is denoted as the **parameter of interest** and γ is the **nuisance parameter**.

The profile likelihood*

Likelihood theory handles nuisance parameters by introducing the **profile likelihood**.

Denoted by $\hat{\gamma}_\psi$ the MLE of γ for fixed value of ψ , namely

$$\hat{\gamma}_\psi = \underset{\gamma \in \Gamma}{\operatorname{argmax}} \ell(\psi, \gamma)$$

then we define the profile likelihood for ψ as

$$L_P(\psi) = L(\psi, \hat{\gamma}_\psi).$$

Note that the maximum of $L_P(\psi)$ is given by the MLE of ψ .

Inference based on the profile likelihood*

A crucial point is the large-sample properties of the profile likelihood are **those of a bona-fide likelihood function** for the parameter of interest only.

In particular, the profile likelihood LRT

$$W_P(\psi) = 2 \{ \ell_P(\hat{\psi}) - \ell_P(\psi_0) \}$$

the asymptotic null distribution is given by χ_q^2 .

Note, however, that if the dimension of γ is large, the large-sample results may be poor. In such cases, the parametric bootstrap is a more accurate route to obtain the p -value.

The t -test as a likelihood-based method*

Many noteworthy tests can be derived from the LRT based on the profile likelihood.

A very important instance is the t -test on μ for a normal random sample, $Y_i \sim \mathcal{N}(\mu, \sigma^2)$. With some simple algebra

$$\ell_P(\hat{\mu}) - \ell_P(\mu_0) = \frac{n}{2} \log(\hat{\sigma}^2) - \frac{n}{2} \log(\hat{\sigma}_{\mu_0}^2),$$

and since $\hat{\sigma}_{\mu}^2 = \hat{\sigma}^2 + (\hat{\mu} - \mu)^2$, it follows

$$r_P(\mu_0) = \text{sgn}(\hat{\mu} - \mu_0) \sqrt{n \log \left\{ 1 + \frac{(\hat{\mu} - \mu_0)^2}{\hat{\sigma}^2} \right\}}.$$

Further simple algebra shows that $R_P(\mu_0)$ is a monotonic increasing function of the T test statistic $T(\mu_0) = (\bar{y} - \mu_0)/\sqrt{s^2/n}$, so that, for instance, $\Pr_{H_0}\{R(\mu_0) \geq r_{obs}\} = \Pr_{H_0}\{T(\mu_0) \geq t_{obs}\}$.

Other notable instances*

Several other tests can be derived as special cases of the LRT, such as the F test for one-way anova models, or exact tests employed in linear regression models.

Other famous tests are instead special cases of the score test. The most notable instance is the chi-squared test of independence for two-way contingency tables, and related tests. The underlying statistical model is the **multinomial distribution** for the observed frequencies.

The generalised likelihood ratio statistic*

In broad generality, for testing $H_0 : \theta \in \Theta_0$ vs $H_1 : \theta \in \Theta_1$ the LRT is most natural resolution

$$W(H_0) = 2 \{ \ell(\hat{\theta}) - \ell(\hat{\theta}_{H_0}) \}.$$

In broad generality, parametric bootstrap is the most convenient approach to approximate the null distribution and compute the p -value.

An approximate (large-sample) null distribution exists when H_0 can be expressed as

$$H_0 : \mathbf{R}(\theta) = 0$$

where \mathbf{R} is a vector-valued function of θ that imposes r restrictions on the parameter vector. In such case, under the null

$$W(H_0) \stackrel{d}{\sim} \chi_r^2.$$

Model selection

Choosing the best model

Several statistical tests are applied for choosing between two alternative specifications of a statistical model. For this sort of problem, more suitable techniques are available, which can also be extended to settings where the two models are not *nested* (i.e. one model is a special instance of the other).

The **Akaike's Information Criterion (AIC)** is perhaps the most commonly used method for choosing the *best* model.

A useful starting point is the **Kullback-Leibler divergence** between the true model f_t and the model under consideration

$$K(f_{\hat{\theta}}, f_t) = \int_{\mathcal{Y}} \{\log f_t(\mathbf{y}) - \log f_{\hat{\theta}}(\mathbf{y})\} f_t(\mathbf{y}) d\mathbf{y}$$

Selecting models to minimize (an estimate of) the expected value of K is equivalent to selecting the model that has the lowest value of

$$\text{AIC} = -2 \ell(\hat{\boldsymbol{\theta}}) + 2 p$$

with $p = \dim(\boldsymbol{\theta})$.

Derivation of the AIC

If we denote by θ_K the parameter value minimizing $K(f_\theta, f_t)$, then it is possible to show (see the CS book)

$$E_{f_t}\{K(f_{\hat{\theta}}, f_t)\} \simeq K(f_{\theta_K}, f_t) + p/2.$$

The next step is the approximation

$$K(f_{\theta_K}, f_t) \simeq E\{-\ell(\hat{\theta})\} + p/2 + \int_{\mathcal{Y}} \log\{f_t(\mathbf{y})\} f_t(\mathbf{y}) d\mathbf{y},$$

so that

$$\widehat{K(f_{\hat{\theta}}, f_t)} = -\ell(\hat{\theta}) + p + \int_{\mathcal{Y}} \log\{f_t(\mathbf{y})\} f_t(\mathbf{y}) d\mathbf{y}.$$

The AIC is just twice the last expression, neglecting the last term which depends only on the true model.

Model selection based on AIC: comments

- The point is that we cannot select the model based on the log likelihood only, since it always selects the more complex model. AIC overcomes this problem by a penalty for adding parameters.
- The **AIC is not consistent**: as $n \rightarrow \infty$, the probability of selecting the correct model does not tend to 1. Indeed, at least for nested models, twice the drop in the maximized log likelihood between an overly complex model and the true model follows (approximately) a χ_r^2 distribution. Since neither χ_r^2 nor $p/2$ depends on n , the probability of selecting the overly complex model by AIC is nonzero and independent of n (for n large).
- The practical implications of the previous point are less serious than it may seem: if all the models under consideration are wrong, then we will tend to select increasingly complex specifications as the sample size increases and the predictive disadvantages of complexity diminish.

R lab: annual mean temperatures in New Haven

We return on the example employed to introduce Statistical Models, and compute the AIC for the four proposed models.

```
y <- (nhtemp - 32) / 1.8; x <- 1912:1971-1
AIC.vals <- rep(NA, 4)
mle1 <- fitdistr(y, "normal")
AIC.vals[1] <- -2 * mle1$loglik + 2 * 2
mle2 <- fitdistr(y, "t", df = 5)
AIC.vals[2] <- -2 * mle2$loglik + 2 * 2
mle3 <- lm(y ~ x)
AIC.vals[3] <- AIC(mle3)
mle4 <- arima(y, xreg=x, order=c(1, 0, 0))
AIC.vals[4] <- AIC(mle4)
AIC.vals

## [1] 130.9961 130.3981 114.9645 116.2789
```

AIC as an alternative to Cross Validation

As stated in the CS book, an alternative approach starts from observing that the KL divergence only depends on the model via

$-\int_{\mathcal{Y}} \log f_{\hat{\theta}}(\mathbf{y}) f_t(\mathbf{y}) d\mathbf{y}$, where the expectation is taken over data not used to estimate $\hat{\theta}$.

An obvious direct estimator of this is the **cross-validation score**

$$CV = - \sum_i \log f_{\hat{\theta}^{[-i]}}(y_i)$$

where $\hat{\theta}^{[-i]}$ is the MLE based on the data with y_i omitted. We might multiply it by 2 to obtain something on the same scale of the AIC.

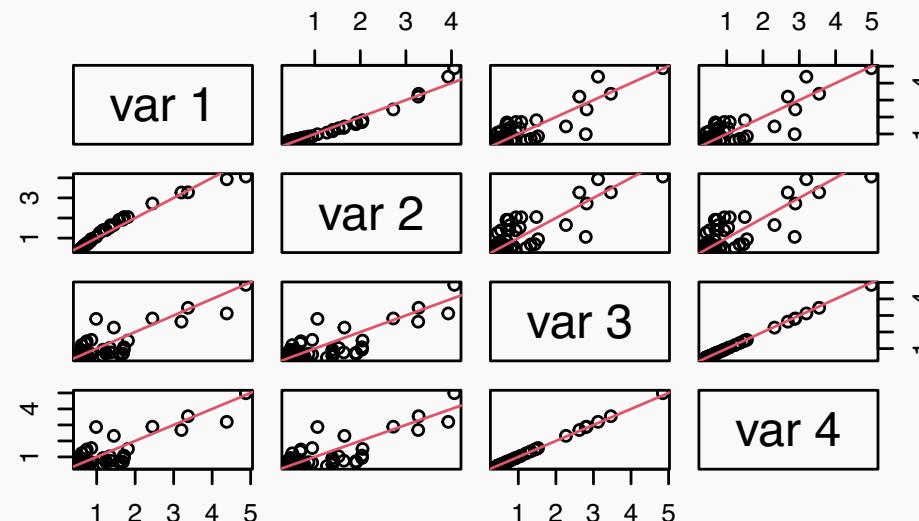
This estimates directly the **predictive accuracy** of the model, and it is a central quantity of *statistical learning methods*. Variants exist where more than one data point at a time are omitting from fitting, with 5 – 10 groups (*folds*) being a common choice. Clearly, the AIC is a much faster alternative.

R lab: CV scores for the example

```
n <- length(y); mat.CV1 <- matrix(0, nrow=n, ncol=4)
for(i in 1:n){
  mle1 <- fitdistr(y[-i], "normal")
  mat.CV1[i,1] <- -log(dnorm(y[i], mle1$est[1], mle1$est[2]))
  mle2 <- fitdistr(y[-i], "t", df = 5)
  mat.CV1[i,2] <- -log(dt((y[i] - mle2$est[1]) / mle2$est[2],
                            df = 5)) + log(mle2$est[2])
  mle3 <- lm(y[-i] ~ x[-i])
  mui <- mle3$coef[1] + mle3$coef[2] * x[i]
  si <- summary(mle3)$sigma
  mat.CV1[i,3] <- -log(dnorm(y[i], mui, si))
  mle4 <- arima(y[-i], xreg = x[-i], order = c(1, 0, 0))
  mui <- mle4$coef[2] + mle4$coef[3] * x[i]
  si <- sqrt(mle4$sigma2 / (1 - mle4$coef[1]^2))
  mat.CV1[i,4] <- -log(dnorm(y[i], mui, si))
}
```

R lab: CV scores for the example

```
my_line <- function(x,y){points(x,y); abline(a=0, b=1, col=2)}  
pairs(mat.CV1, panel = my_line)
```



```
colSums(mat.CV1) * 2
```

```
## [1] 131.9519 130.5076 115.9256 116.1472
```

Hypothesis Testing

N. Torelli, G. Di Credico, V. Gioia

Fall 2023

University of Trieste

Fundamentals of hypothesis testing¹

Some commonly used tests²

Relation between tests and confidence intervals³

Nonparametric tests⁴

Likelihood-based tests

¹Agresti, Kateri: sec 5.1-5.5

²Agresti, Kateri: sec 5.2-5.3-5.4

³Agresti, Kateri: sec 5.6

⁴Agresti, Kateri: sec 5.8

Fundamentals of hypothesis testing

The idea of hypothesis testing

The basic aim of hypothesis testing within a *parametric statistical model* $f_\theta(\mathbf{y})$ is to establish whether the data could be reasonably be generated from $f_{\theta_0}(\mathbf{y})$, where θ_0 is a specific value of the parameter.

This is simply denoted by the succinct notation

Example:

$$H_0 : \theta = \theta_0 , \quad \begin{cases} H_0 : & \pi = 0.4 \\ H_1 : & \pi > 0.4 \end{cases}$$

with H_0 being termed **null hypothesis**.

Complementary to the choice of H_0 , it is required to select a complementary **alternative hypothesis** H_1 , specifying the values of the parameter which become reasonable when H_0 does not hold.

Example: testing the mean of a normal sample

Assume the very simple model for independent observations y_1, y_2, \dots, y_n given by $Y_i \sim \mathcal{N}(\mu, 1)$. Then we may want to test

$$H_0 : \mu = 0$$

against

$$H_1 : \mu > 0$$

which amounts to testing the null hypothesis of data generated from a standard normal distribution, against the possibility that the true mean takes instead a positive value.

This choice of H_1 makes fully sense when we can rule out negative values of μ (**one-sided alternative**). If this is not the case, a better choice would be given by $H_1 : \mu \neq 0$ (**two-sided alternative**).

General formulation

In broad generality, hypothesis on a parameter θ can be cast in the form

$$H_0 : \theta \in \Theta_0$$

against

$$H_1 : \theta \in \Theta_1$$

where Θ_0 and Θ_1 form a bi-partition of the set containing all the possible values for the parameter θ , that is named the **parameter space Θ** .

The tools for addressing problems of such level of generality will be covered in the part of the course devoted to *likelihood methods*.

In what follows, instead, we will illustrate the main ideas by means of simple, yet important, instances.

Steps of hypothesis testing

The theory of hypothesis testing is rather articulated, so that it may help to go through the main parts of the theory in a systematic fashion.

Some noteworthy concepts are

- Test statistic
- Null and alternative distributions
- p -value
- Significance level, rejection and acceptance regions
- Errors and power

Test statistic

A **test statistic** is a **statistic** (namely, a function of the r.v. representing the available sample) which is used to carry out the test.

Large values (in absolute value) of the test statistic cast doubt on H_0 and on the theory underlying it. *Inconsistent data!!! Probably the assumed process didn't generate the observed data.*

Its choice depends on the problem under study. For the simple normal example mentioned above, a natural choice is to take as test statistic the (standardized) sample mean

$$Z = \frac{\bar{Y}}{\sqrt{\frac{1}{n}}} = \sqrt{n} \bar{Y}$$
$$Z = \frac{\bar{Y} - \mu}{\sqrt{\frac{\sigma^2}{n}}} = \frac{\bar{Y} - \mu}{\text{SE}[\bar{Y}]}$$

Null and alternative distributions

The distribution of a test statistic will generally depend on the "true value" of the parameter under testing.

In the example, if H_0 is true (*under H_0*), then

$$Z \sim \mathcal{N}(0, 1),$$

The distribution depends on the result of the hypothesis testing.

and this is called the **null distribution** of Z .

Instead, if H_1 holds (*under H_1*), it follows that

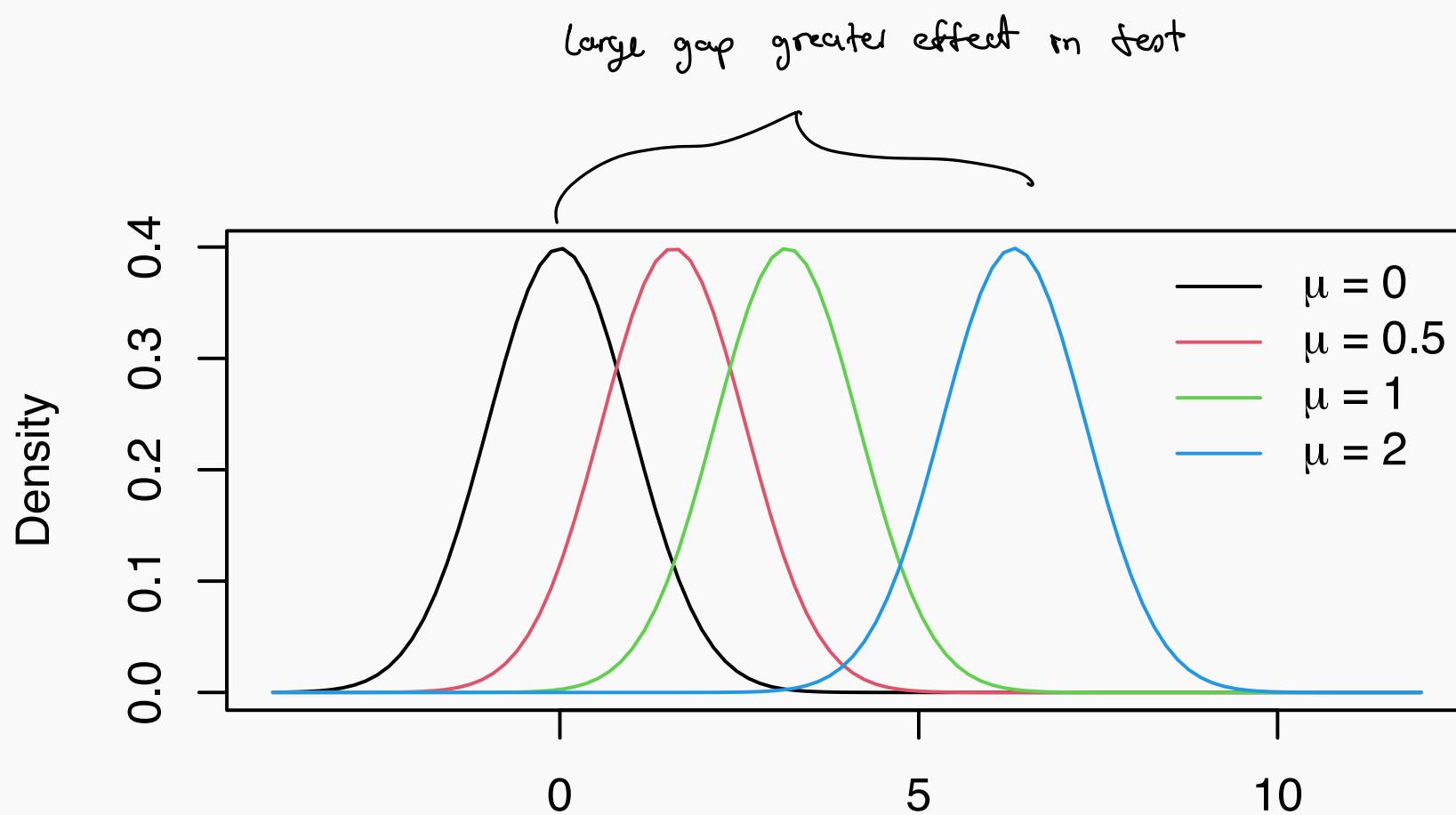
$$Z \sim \mathcal{N}(\Delta, 1)$$

where $\Delta = \sqrt{n}\mu > 0$ increases with the value of μ .

$$\begin{aligned} Z &= \sqrt{n} \bar{Y} \\ E[Z] &= E[\sqrt{n} \bar{Y}] \\ &= \sqrt{n} E[\bar{Y}] \quad \bar{Y} \rightarrow \mu \\ &= \sqrt{n} \mu \end{aligned}$$

The distributions valid under H_1 are called the **alternative distributions** of Z .

R lab: visualizing the null and alternative distributions



If $\mu \uparrow$ p will ↑ or ↓ ?

The p -value

The p -value measures the distance between the data and H_0 . Small values of it correspond to a test statistic unlikely to arise under H_0 , and suggest that H_0 and the data are inconsistent.

In the example, the idea is that any value larger than the observed z_{obs} (the value of Z computed with the observed data) would cast even greater doubt on H_0 .

The p -value is thus defined as *the probability (under H_0) of observing a value of the test statistic equal or larger than the observed one*

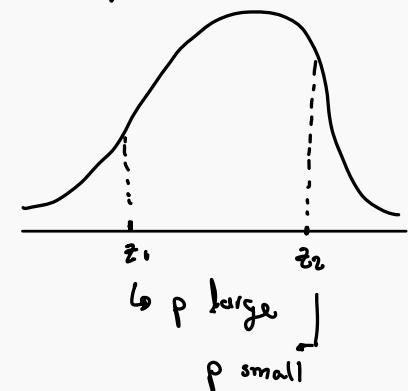
$$p = \Pr_{H_0}(Z \geq z_{obs}) = 1 - \Pr_{H_0}(Z \leq z_{obs})$$

Since under H_0 we have $Z \sim \mathcal{N}(0, 1)$, it follows that

$$p = 1 - \Phi(z_{obs})$$

p can be compared w/ the confidence level of a C.I.

$\uparrow p$ likely
 $\downarrow p$ unlikely



R lab: computing the p -value for a sample

In case the null distribution is not known, it would be possible to compute the p -value by simulation whenever it is possible to generate data under H_0 . In R:

```
set.seed(13); n <- 10; y_obs <- rnorm(n)
```

```
z_obs <- mean(y_obs) * sqrt(n)
```

```
print(z_obs)
```

```
## [1] 1.897537
```

```
M <- 100000; z.sim <- numeric(M)
```

```
for(i in 1:M) { y <- rnorm(n)
```

```
z_sim[i] <- mean(y) * sqrt(n) }
```

```
c(mean(z_sim >= z_obs), 1 - pnorm(z_obs))
```

```
## [1] 0.02877000 0.02887856
```

$p \leq 0.5 \rightarrow H_0$ is rejected

P^*
simulated

P
reg

It is likely data generated by specific values of interest
that

Other alternative hypotheses: more details

For the simple example of test on μ and the same $H_0 : \mu = 0$, other two possibilities for H_1 could then be considered.

In either case, the same test statistic Z would still be used, but the computation of the p -value would change, due to the different direction of deviation from H_0 .

For $H_1 : \mu < 0$, small values of Z would flag deviation from H_0 (that is, negative values with large absolute value), so that

Example

$$p = \Pr_{H_0}(Z \leq z_{obs}) = \Phi(z_{obs}).$$

$\begin{cases} H_0 : \pi \approx \pi_0 \\ H_1 : \pi \neq \pi_0 \end{cases}$

how far we are from the value of interest

$$\mathcal{N}(0, 1) \sim Z = \frac{\pi - \pi_0}{\sqrt{\pi_0(1-\pi_0)/n}}$$

Instead, for $H_1 : \mu \neq 0$, both directions ought to be considered, and

$$p = \Pr_{H_0}(|Z| \geq |z_{obs}|) = 2 \Pr_{H_0}(Z \geq |z_{obs}|) = 2 \{1 - \Phi(|z_{obs}|)\}.$$

Accept Null Hypothesis

$$\begin{aligned} p &= 2 [1 - \Phi(1.26)] \leftarrow Z = 1.26 \text{ (suppose)} \\ &= 2 [1 - \text{pnorm}(1.26)] \\ &= 0.208 \end{aligned}$$

Significance level

We commonly say that a the result of a test is *significant at the 5% level* whenever the p -value is smaller or equal to 0.05. Other levels of some practical interest are 1% or 0.1%.

As stated in the CS book, an often-followed convention is

Range	Evidence against the null hypothesis
$0.05 < p \leq 0.1$	<i>marginal evidence</i>
$0.01 < p \leq 0.05$	<i>evidence</i>
$0.001 < p \leq 0.01$	<i>strong evidence</i>
$p \leq 0.001$	<i>very strong evidence</i>

A test *with fixed significance level* arises when the significance level is fixed in advance, and then it is just reported whether the p -value is smaller than the fixed level. If this happens, it may be reported that H_0 is **rejected**, otherwise we may say that H_0 is **not rejected** (or **accepted**).

Rejection and acceptance regions

If we define **the sample space** as the set of the values that our available sample may take, the **rejection region** of a test with fixed significance level is the subset of the sample space corresponding to the samples that would lead to a rejection of H_0 .

The remaining part of the sample space forms instead the **acceptance region**.

Both these two regions are determined by means of a test statistic.

Rejection and acceptance regions for the example

In the simple normal example previously introduced, for $H_1: \mu > 0$, it is simple to verify that a rejection region of level α is simply

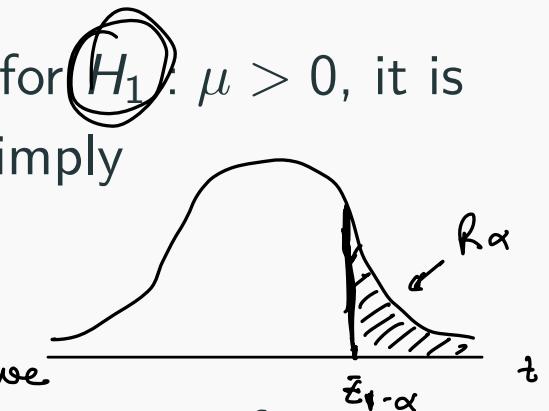
φ and α can be compared

φ and α are probabilities

$\Leftrightarrow z_\alpha$ are quantiles, i.e., $z \in \mathbb{Z}$ r.v.

$$\mathcal{R}_\alpha = \{\mathbf{y} : Z \geq z_{1-\alpha}\},$$

to critical value



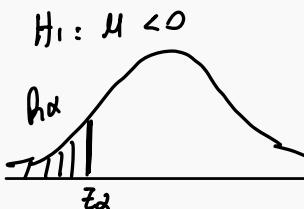
where $z_{1-\alpha}$ is the standard normal $(1 - \alpha)$ -quantile, i.e. 1.645 for $\alpha = 0.05$.

The rejection region is entirely conditioned by H_2

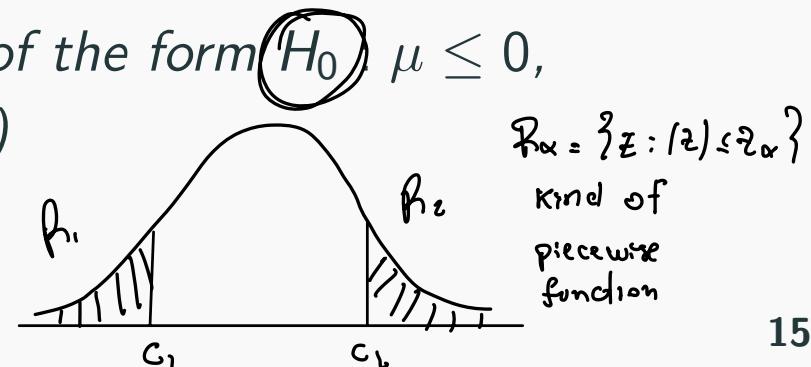
The acceptance region is just given by

$$\mathcal{A}_\alpha = \{\mathbf{y} : Z < z_{1-\alpha}\}.$$

(Note: the computation of the p-value, and of \mathcal{R}_α and \mathcal{A}_α would be exactly the same if the null hypothesis were of the form $H_0: \mu \leq 0$, maintaining the same alternative hypothesis.)



$$\mathcal{R}_\alpha = \{z : z \leq z_\alpha, z \in \mathbb{Z}\}$$



$$\mathcal{R}_\alpha = \{z : f(z) \leq q_\alpha\}$$

kind of
piecewise
function

Errors for a fixed-significance level test

Using α

Type I reject H_0 when true \rightarrow False positive / alarm

Type II accept H_0 when false \rightarrow False negative / miss

Conditioned to fail if there is an error

When we adopt a test with fixed significance level, we move from using the p -value as a measure of evidence against H_0 to using a test to decide which of H_0 and H_1 is more supported by the data.

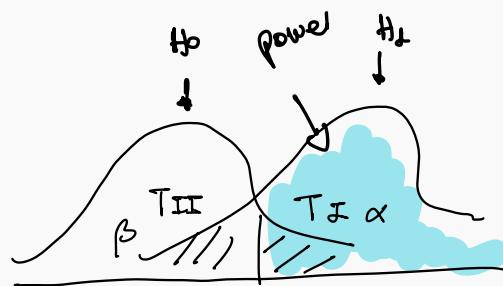
Two wrong decisions are possible. We commit a *Type I error* by rejecting H_0 when it is true, or a *Type II error* by accepting H_0 when it is false.

In the example, $\Pr_{H_0}(Y \in R_\alpha) = \alpha$, and in fact **the fixed significance level equals the probability of making a Type I error.** \rightarrow By definition

Example:

$$H_0: \mu = 0$$

$$H_1: \mu > 0$$

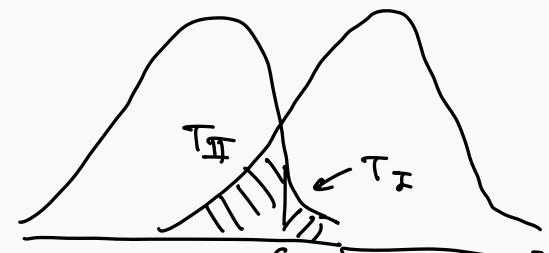


Look at plots overlaped at the beginning

$$\uparrow T_I \downarrow T_{II}$$

Errors depends on significance levels.

$$\downarrow \alpha \rightarrow \downarrow T_I \rightarrow \uparrow T_{II}$$



Power of a test

Type I error: mistake in concluding there is an effect, difference, ... when there isn't one in the population

Type II error: missing a real effect, relationship, ... in the population

For a test with fixed significance level, the power is the probability of (correctly) detecting that H_0 is false

$$\Pr_{H_1}(\mathbf{Y} \in \mathcal{R}_\alpha).$$

The power of a test can be used for comparing alternative tests for the same problem, with tests with higher power being preferable.

The power is often used for designing studies, in particular for choosing the sample size in medical or industrial studies. Indeed, for fixed significance level, the power increases with the sample size.

		Decision	
		Reject H_0	Do not reject H_0
H_0 true	Reject H_0	Type I error	Correct decision
	Do not reject H_0	Correct decision	Type II error

Power of two tests for the example

For the simple example (with $H_1 : \mu > 0$), an alternative (but silly) test statistic may be given by taking the same Z as above computed by using only half of the sample (for n even).

Fixing a significance level of 5%, the two tests have exactly the same probability of a Type I error, so for comparing them we must use their power.

The power is a function of the μ assumed under H_1 , and for a certain $\mu \geq 0$ we obtain (since $z_{0.95} = 1.645$)

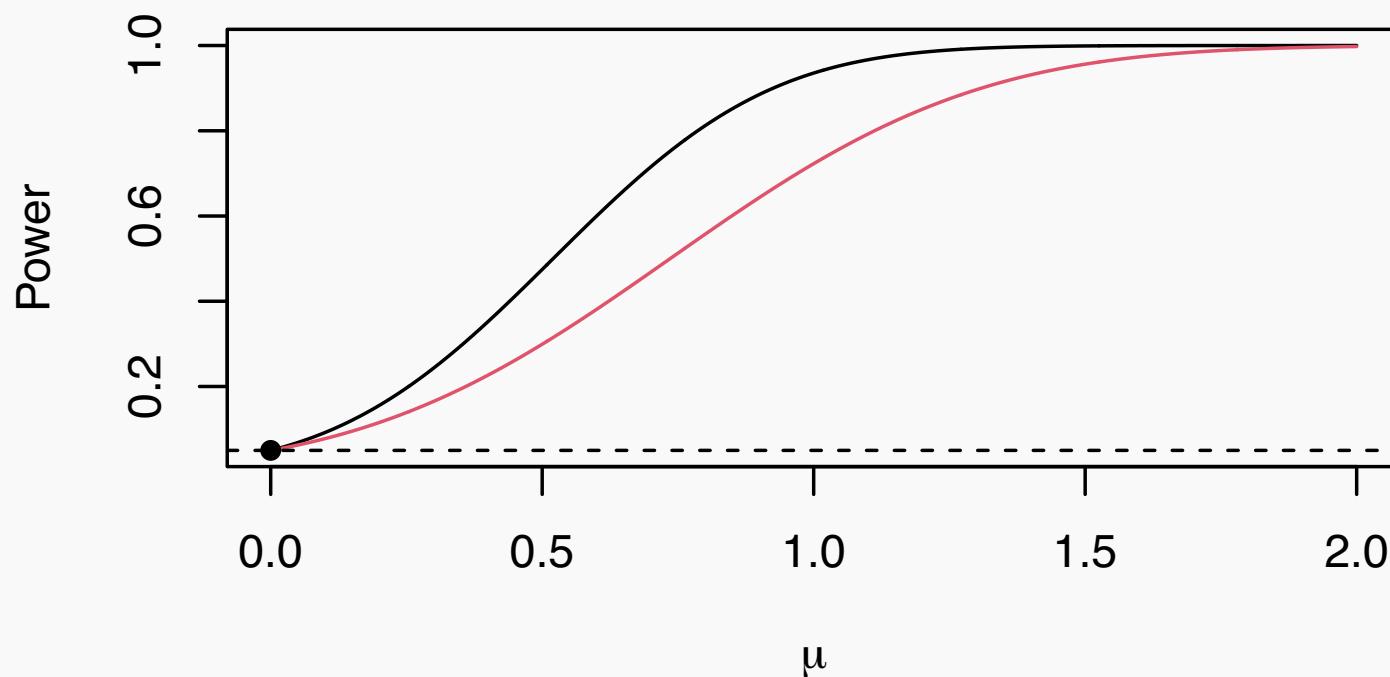
$$\Pr_\mu(Z \geq 1.645) = 1 - \Phi(1.645 - \sqrt{n}\mu)$$

$\uparrow \mu, \uparrow n \Rightarrow \downarrow \phi \Rightarrow \uparrow \text{power test}$

There is a "compromise" between μ, n and error. because $Z = \frac{\bar{Y} - \mu_0}{\sqrt{\frac{\sigma^2}{n}}}$ could lead to outliers though not correct $P \rightarrow 0$

R lab: power of two alternative tests

```
mu <- seq(0, 2 , l = 1000); n <- 10; n1 <- 5  
plot(mu, 1 - pnorm(1.645 - sqrt(n) * mu), type = "l",  
      ylab="Power", xlab = expression(mu))  
lines(mu, 1 - pnorm(1.645 - sqrt(n1) * mu), col = 2)  
abline(h=0.05, lty = 2); points(0, 0.05, pch = 16)
```



Comments on the p -value

The usage of p -values is not free of controversies, and in ending the review of the general theory on testing some comments are in order.

1. The p -value **is NOT the probability that H_0 is true**, since the latter is not even an event.
2. The results of statistical tests, and p -values in particular, should never be taken without considering context-specific knowledge. Even a small p -value may not be particularly meaningful if the alternative hypothesis is logically implausible.
3. Hypothesis testing is useful in certain contexts, but it has some important limitations. For (very) large sample sizes, even tiny deviations from the null hypothesis will lead to small p -values. For large sample sizes, there are alternative approaches which are more fruitful, and techniques based on **model selection** are often preferable to statistical tests.

$$Z = \frac{\bar{Y} - \mu_0}{\sqrt{\frac{\sigma^2}{n}}}$$

outliers and large samples sizes could make test statistic large then $p < 0$ and reject H_0 .
→ model selection.

Some commonly used tests

One-sample t test

Given a normal random sample y_1, \dots, y_n , with $Y_i \sim \mathcal{N}(\mu, \sigma^2)$, a classical testing problem on μ is of the form (for two-sided alternative, say)

$$\begin{cases} H_0 : \mu = \mu_0 & \text{parametric test: we know something about the sample} \\ H_1 : \mu \neq \mu_0 & \text{Non-param. f: we know nothing} \end{cases}$$

The test statistic is given by

$$T = \frac{\bar{Y} - \mu_0}{\sqrt{\frac{S^2}{n}}} \sim t_{n-1}, \quad \text{when } H_0 \text{ is true}$$

$\bar{Y} \sim N(\mu, \frac{\sigma^2}{n})$

definition

$$\bar{Y}_n \sim N(\mu, \frac{\sigma^2}{n})$$

with the p -value given by

$$p = \Pr_{H_0}(|T| \geq |t_{obs}|)$$

$$T = \frac{\bar{Y} - \mu_0}{\sqrt{\frac{(n-1)S^2}{n}} / \sqrt{\frac{\sigma^2}{n}}} = \frac{\bar{Y} - \mu_0}{\sqrt{\frac{(n-1)S^2}{n}} / \sqrt{\frac{\sigma^2}{n}}} \sim t_{n-1}$$

$\bar{Y} \sim N(\mu, \frac{\sigma^2}{n})$

$\sqrt{\frac{(n-1)S^2}{n}} \sim \chi_{n-1}^2$

$\sqrt{\frac{\sigma^2}{n}} \sim \frac{\sigma}{\sqrt{n}}$

which can be computed as $p = 2 \Pr_{H_0}(T \geq |t_{obs}|) = 2 \{1 - F_{t_{n-1}}(|t_{obs}|)\}$, since the t distribution is symmetric around 0.

Example

Political ideology is an ordinal scale. It is sometimes informative to treat ordinal data in a quantitative manner by assigning scores to the categories, in order to use the mean to summarize the responses

In the following we will use the GSS seven-point scale of political views ranging from 1= “extremely liberal” to 7= “extremely conservative”

$$T = \frac{\bar{Y} - \mu}{\sqrt{\frac{s^2}{n}}} \sim t_{n-1}$$

Political ideology	Race		
	Hispanic	Black	White
1. Extremely liberal	5	16	73
2. Liberal	49	52	209
3. Slightly liberal	46	42	190
4. Moderate, middle of road	155	182	705
5. Slightly conservative	50	43	260
6. Conservative	50	25	314
7. Extremely conservative	14	11	84
<i>n</i>	369	371	1835

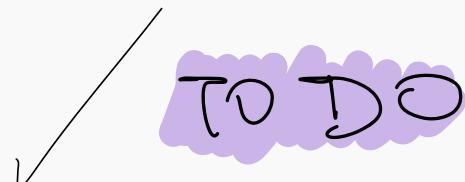
Figure 1: From Agresti, Kateri (2022, *Foundations of Statistics for Data Scientists*)

Example (cont'd)

To test whether the data show much evidence of either of these, we conduct a significance test about how the population mean μ compares to the moderate value of 4, with hypotheses

$$\begin{cases} H_0 : \mu = 4.0 \\ H_1 : \mu \neq 4.0 \end{cases}$$

```
##  
## One Sample t-test  
##  
## data: polid$ideology[polid$race == "hispanic"]  
## t = 1.2827, df = 368, p-value = 0.2004 > 0.5 → Accept H0  
## alternative hypothesis: true mean is not equal to 4  
## 95 percent confidence interval:  
## 3.952333 4.226528  
## sample estimates:  
## mean of x  
## 4.089431
```



Example

The DAAG book introduces the simple dataset pair65, about an experiment on the effect of heat on the stretchiness of elastic bands: a small sample of differences between two different conditions for 9 bands.

$$T = \frac{\bar{d} - 0}{\sqrt{\frac{s^2}{n}}} \sim t_{n-1}$$

heated	ambient	difference
244	225	19
255	247	8
253	249	4
254	253	1
251	245	6
269	259	10
248	242	6
252	255	-3
292	286	6

t because of pair observations : before and after.

Example (cont'd)

Focusing on the 9 differences on the amount of stretch, we test

$$\begin{cases} H_0 : \mu = 0 & \text{there is no difference} \\ H_1 : \mu \neq 0 & \text{there is difference} \end{cases}$$

by means of the `t.test` function, resulting in significance at 5% level

```
##  
## One Sample t-test  
##  
## data: difference  
## t = 3.1131, df = 8, p-value = 0.01438 → reject  
## alternative hypothesis: true mean is not equal to 0  
## 95 percent confidence interval:  
## 1.641939 11.024728 ↗ μ=0  
## sample estimates:  
## mean of x  
## 6.333333
```

Approximate tests

For large random samples, the Central Limit Theorem ensures that $\bar{Y} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$, being $\mu = E(Y_i)$ and $\sigma^2 = \text{var}(Y_i)$.

A test statistic for $H_0 : \mu = \mu_0$ is therefore

$$Z = \frac{\bar{Y} - \mu_0}{\sqrt{\frac{S^2}{n}}} \sim \mathcal{N}(0, 1), \quad \text{when } H_0 \text{ is true}$$

The estimator of the variance S^2 can be replaced by a more suitable one. For example, for binary data, $Y_i \sim \mathcal{B}_i(1, \pi)$, commonly used test statistics are $Z = \frac{\hat{\pi} - \pi_0}{\sqrt{\frac{\hat{\pi}(1 - \hat{\pi})}{n}}}$ or $Z = \frac{\hat{\pi} - \pi_0}{\sqrt{\frac{\pi_0(1 - \pi_0)}{n}}}$, the latter being preferable.

Tests based on the CLT are instances of **approximate tests**, for which the property concerning the Type I error level holds only approximately.

Two sample t -test

Given two **independent normal samples**, represented by

$X_i \sim \mathcal{N}(\mu_X, \sigma_X^2), i = 1, \dots, n_X$ and $Y_i \sim \mathcal{N}(\mu_Y, \sigma_Y^2), i = 1, \dots, n_Y$, the test statistic for testing the equality between the two means is

Distribution ? $\leftarrow T = \frac{\bar{X} - \bar{Y}}{\text{SE}(\bar{X} - \bar{Y})}$ $\begin{cases} H_0 : \mu_X = \mu_Y \Leftrightarrow \mu_X - \mu_Y = 0 \\ H_1 : \mu_X \neq \mu_Y \Leftrightarrow \mu_X - \mu_Y \neq 0 \end{cases}$

with $\text{SE}(\bar{X} - \bar{Y})$ estimated by $\sqrt{\frac{S_X^2}{n_X} + \frac{S_Y^2}{n_Y}}$. for $\sigma_X^2 \neq \sigma_Y^2$

A different formula is instead adopted if it is possible to assume that $\sigma_X^2 = \sigma_Y^2$. $\rightarrow \text{SE} = s \sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}$, $s = \frac{s_X + s_Y}{2}$ (pool s) $\rightarrow T \sim t_{n_1 + n_2 - 2}$

The distribution of T when H_0 is true is given by a suitable t distribution.

Like for the one-sample case, there are general formulas for large samples, employing the normal distribution.

Paired t -test

Paired observations arise whenever each unit of a random sample of size n is observed twice, under different conditions, so that we end up again with two sets of variables $X_i \sim \mathcal{N}(\mu_X, \sigma_X^2)$, $i = 1, \dots, n$ and $Y_i \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$, $i = 1, \dots, n$.

However, now the pair (X_i, Y_i) refers to the same unit, so that the two samples X_1, \dots, X_n and Y_1, \dots, Y_n are no longer independent.

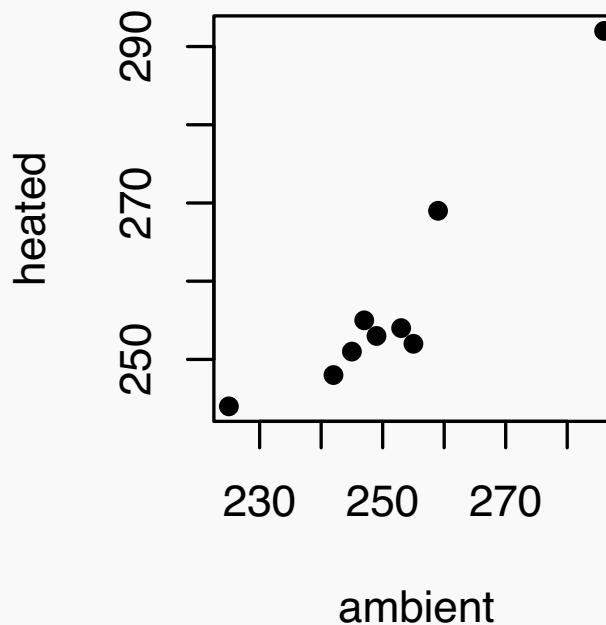
The pair65 data set is exactly of this nature. Like in that example, the resolution is to focus on the random sample of the n differences $D_i = X_i - Y_i$, for which $E(D_i) = \mu_X - \mu_Y$: for testing the equality of the two means μ_X and μ_Y we just apply the theory for the one-sample t -test, with $\mu_0 = 0$.

For the pair65 data set, the p -value of about 0.014 suggests that heat may indeed have an effect on stretchiness.

Again we reject. In either case H_0 is not supported by data

Example

Even though the pair65 data is very small, the fact that the two groups of observations are not independent is readily suggested by a scatterplot



READ BOOK

By (blindly) applying the test for independent data we would get a p -value of about 0.40, hinting at a quite different conclusion.

Example (cont'd)

We want to compare mean weight changes in conceptual populations of anorexic girls receiving cognitive behavioral therapy $n_1 = 29$ or a control $n_2 = 26$

The hypothesis system is

$$\begin{cases} H_0 : \mu_1 = \mu_2 \\ H_1 : \mu_1 \neq \mu_2 \end{cases}$$

If the therapy truly has no effect relative to the control, the weight changes for the two populations would have equal means and equal standard deviations

The sample standard deviations are $s_1 = 7.31$ for the cognitive behavioral therapy and $s_2 = 7.99$ for the control, the pooled standard deviation estimate is $s = 7.64$, and $\bar{y}_1 = 3.01$ and $\bar{y}_2 = -0.45$

Example (cont'd)

For testing $H_0 = \mu_1 = \mu_2$ the observed value of the T test statistic is

$$t = \frac{\bar{y}_1 - \bar{y}_2}{se} = 1.68$$

with $df = n_1 + n_2 - 2 = 53$

```
##                                     T =  $\frac{\bar{y}_1 - \bar{y}_2 - \mu_0}{SE_0}$ 
## Two Sample t-test
##                                     Accept H0
## data: cogbehav and control
## t = 1.676, df = 53, p-value = 0.09963
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -0.680137 7.593930
## sample estimates:
## mean of x mean of y
## 3.006897 -0.450000
```

Ex: $s_p^2 = \frac{s_1^2 + s_2^2}{2}$

pool std

Example

Without assuming the variance quality we would get

```
##  
## Welch Two Sample t-test  
##  
## data: cogbehav and control  
## t = 1.6677, df = 50.971, p-value = 0.1015 → Accept H0 again  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -0.7044632 7.6182563  
## sample estimates:  
## mean of x mean of y  
## 3.006897 -0.450000
```

Example

To compare two groups on the population proportions π_1 and π_2 having a particular outcome, we test $H_0 : \pi_1 = \pi_2$

For the difference $\pi_1 - \pi_2$ parameter this hypothesis is $H_0 : \pi_1 - \pi_2 = 0$, no difference, or no effect

We assume independent samples of sizes n_1 and n_2 , with success counts Y_1 and Y_2 having binomial distributions and sample proportion estimates $\hat{\pi}_i = y_i/n_i$ for $i = 1, 2$

The P-value is a two-tail or one-tail probability from the standard normal distribution, according to whether H_1 is two-sided, $H_1 : \pi_1 \neq \pi_2$, or one-sided, $H_1 : \pi_1 > \pi_2$ or $H_1 : \pi_1 < \pi_2$

Example (cont'd)

For the research investigation of the efficacy of prayer, data are summarized in the contingency table below

Prayer	Complications		Total
	Yes	No	
Yes	315	289	604
No	304	293	597

$$\begin{aligned} \text{Bi} \\ \downarrow n \sim 10^2 \\ Z = \frac{\hat{\pi}_1 - \hat{\pi}_2}{\text{SE}_0} \sim N(0,1) \end{aligned}$$

Figure 2: From Agresti, Kateri (2022, *Foundations of Statistics for Data Scientists*)

Accept H_0 : praying has no effect on complications

Two-sided alternative hypothesis leads to a p-value of

```
## [1] 0.6671956
```

While, for one-sided alternative hypothesis the result is halved

Example (cont'd)

Considering the square of the test statistic, we can perform a hypothesis test based on the χ^2 -distribution

For a two-sided alternative hypothesis we get Just $\& df$ because we square + rv.

```
##  
## 2-sample test for equality of proportions without continuity correction  
##  
## data: c(315, 304) out of c(604, 597)  
## X-squared = 0.18217, df = 1, p-value = 0.6695  
## alternative hypothesis: two.sided  
## 95 percent confidence interval:  
## -0.04421536 0.06883625  
## sample estimates:  
## prop 1 prop 2  
## 0.5215232 0.5092127
```

Relation between tests and confidence intervals

Main result

As displayed for the pair65 data testing, the t.test R function returns also the confidence interval for the parameter under testing, in that case the true mean of the differences in stretchiness.

This is not by chance, since there is a close connection between hypothesis testing on the value of a certain parameter and confidence intervals for that parameter.

For the case of a mean, for example, the basic idea is that
If the confidence interval for μ does not contain zero, this is equivalent to rejection of the hypothesis that the true mean is zero.

Important: the connection is between two-sided confidence intervals and two-sided alternative hypotheses. For one-sided alternative hypotheses, the connection is with one-sided confidence intervals.

try this function

More precisely

The general result is as follows, and states a perfect equivalence between the two methods:

$$\text{given } \alpha \rightarrow \text{find } p$$

1. Given a method to find a confidence interval of level $(1 - \alpha)\%$ for a certain scalar parameter θ , we can establish whether the p -value for testing $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$ is smaller than the significance level α by checking if θ_0 is included in the interval
2. Given a method to find a p -value for testing $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$, we can obtain a confidence interval of level $1 - \alpha$ by selecting all the θ_0 values that will lead to a p -value larger than α

$$\text{given } p \rightarrow \text{find } \alpha \text{ for all } \theta_0. \text{ s.t. } p > \alpha$$

Example: pair65 data

The 95% and 99% confidence intervals for the mean of the differences are, respectively

95%	1.6419	11.0247
99%	-0.4930	13.1596
98.56217%	0.0000	12.6667

The 95% confidence interval does not contain zero, while the wider 99% does, implying that the hypothesis $\mu = 0$ is rejected for $\alpha = 0.05$, but not for $\alpha = 0.01$.

Note that for a confidence interval of level $1 - p = 0.9856217$, we obtain a lower limit exactly equal to 0: the p -value, in fact, corresponds to a significance level which is borderline between rejection and non-rejection of H_0 .

Example (probability of type II error and power of a test)

An astrologer was asked which personality chart out of three options was the correct one for 116 adults, based on their horoscope.

$H_0 : \pi = 1/3$: predictions are like random guessing $H_1 : \pi = 1/2$ better than guessing
 π : prob. of correct prediction

We want to find the $P(\text{Type II error})$ for $\alpha = 0.05$ -level test

Under H_0 , the sampling distribution of $\hat{\pi}$ is

$$\hat{\pi} \sim N(1/3, 0.0019)$$

↑ suppose

$$s_e = \sqrt{\frac{\pi_0(1-\pi_0)}{n}} = 0.0438$$
$$\pi_0 = \frac{1}{3}, n = 116$$

For $\alpha = 0.05$, the rejection region is

qnorm $\rightarrow z_\alpha = 1.645$ critical value
quantile Fail to reject H_0 if $\hat{\pi} < \frac{1}{3} + z_\alpha \cdot s_e = 0.405$

Study CI.

$$\mathcal{R}_\alpha = \{\mathbf{y} : \hat{\pi} \geq 0.405\}, \text{ with } P_{H_1}$$

where k_α is the critical value, i.e. the quantile of order 0.95 of the sampling distribution under H_0 on the sample proportion scale

Example (cont'd)

The probability of type II error is the probability of failing to reject the null hypothesis when it is false

Therefore the test statistic under $H_1 : \pi = 1/2$ is

$$H_q : \pi > \frac{1}{3}$$

wrt H_0 or H_1 ? $\hat{\pi} \sim N(1/2, 0.0021)$

For $\alpha = 0.05$, the $P(\text{Type II error})$ is $P(\hat{\pi} < 0.405) = 0.02$

$$\begin{aligned} P &= P_{H_1}(0.405) \\ &= 0.66 > 0.05 \end{aligned}$$

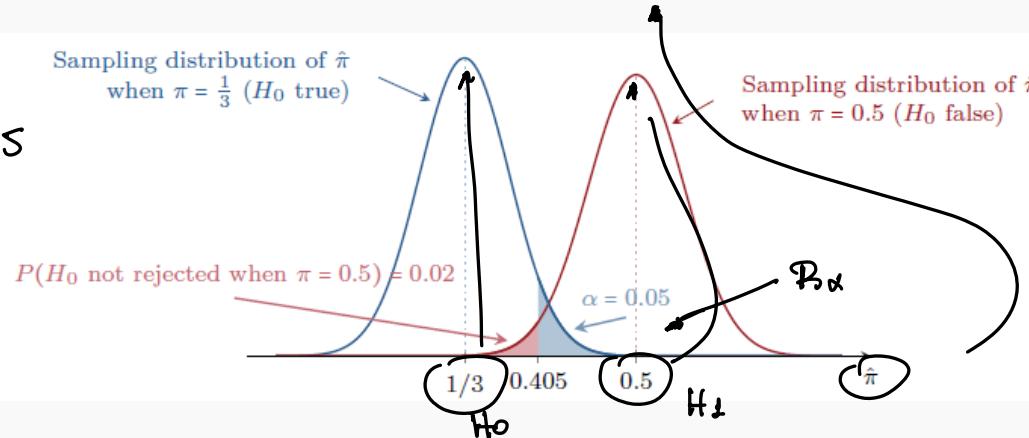


Figure 3: From Agresti, Kateri (2022, *Foundations of Statistics for Data Scientists*)

$$P_{\text{error}} \sim 1 - P(\hat{\pi} - z_{\alpha}/\sqrt{n})$$

Recall that $P(\text{Type II error})$ increases when π is closer to H_0 .

Type II	$\rightarrow P_{H_1}$
Type I	$\rightarrow P_{H_0}$

Given that we have $\pi = 0.5$
we want $P(\text{Type II})$

reject we it is false
We expect $P(\) > 0.5$
We focus on

For the normal distribution

$$z = \frac{0.405 - 0.5}{0.0464} = -2.04 \quad \text{where} \quad se = \sqrt{\frac{\pi_0(1-\pi_0)}{n}} = 0.0464$$

$$\rho = P(Z \leq z) = \Phi(z) = 0.02$$

$\beta = P(\hat{\pi} < 0.405)$ is the prob. that a std normal r.v falls below -2.04 , which is 0.02.

For $\alpha = 0.01 \rightarrow \beta = 0.02$

$$\downarrow \alpha \quad \uparrow \beta$$

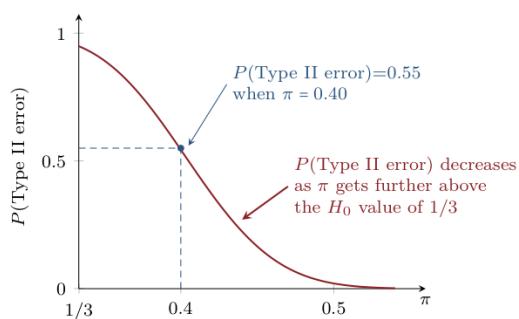
$$\alpha = P(\text{Type I error})$$

$$\beta = P(\text{Type II error})$$

If $\pi \rightarrow \pi_0$ then $\beta \uparrow$

```
> library(pwr)
> pwr.p.test(ES.h(0.5, 1/3), n=116, sig.level=0.05, alt="greater")
power = 0.978 # power of alpha=0.05 test of H0:pi=1/3 when pi=0.5
```

$$\alpha = 0.05, \pi = 0.4 \rightarrow \beta = 0.55$$



Exercise

$$\text{Power} = 1 - \beta (\text{Type II})$$

Let β denote $P(\text{Type II error})$. For an $\alpha = 0.05$ -level test of $H_0 : \mu = 0$ against $H_1 : \mu > 0$ with $n = 30$ observations, $\beta = 0.36$ at $\mu = 4$. Then:

1. At $\mu = 5$, $\beta > 0.36$. $n=5$ is even further than earlier. Since $\beta \downarrow$ when $|\mu - \mu_0| \uparrow$ then it is false
2. If $\alpha = 0.01$, then at $\mu = 4$, $\beta > 0.36$ If $\downarrow \alpha \uparrow \beta$ for fixed μ then it is true
3. If $n = 50$, then at $\mu = 4$, $\beta > 0.36$ If $\uparrow n \uparrow \beta$ for fixed μ . False
4. The power of the test is 0.64 at $\mu = 4$ True. $\text{Power} = 1 - \beta$
5. This must be false, because necessarily $\alpha + \beta = 1$

• α is the significance level of the test and $\uparrow \alpha \approx P(\text{Type I})$ error

• $\beta \approx P(\text{Type II})$

These values are related to different distributions, so their sum doesn't need to be 1.

Nonparametric tests

Main idea behind nonparametric tests

What does it mean?

Nonparametric tests specify only partially a statistical model for the data, so that they may provide more robust inferences than parametric tests with contaminated data, outliers or, more generally, in settings where model specification is hard.

This is sometimes useful, especially when only certain aspects of the data are of interest, or for checking the results obtained with a full model specification.

The details of such tests, and more generally the theory supporting their validity, would require a substantial amount of space. Here we just mention such solutions in passing, as a tool in the statistician's reservoir that at times may be a useful complement to parametric tests.

Wilcoxon rank sum and signed rank tests

The main idea of nonparametric tests is illustrated by the Wilcoxon rank sum test, which can be used to replace the t test when normality is doubtful, due to outliers or excessive rounding, for example.

The test uses the **ranks**, which are the index of each observation in the sample sorted in ascending order. For instance, for the pair65 set of differences

Difference	Ordered abs	Rank	Signed rank
19	1	1	1
8	-3	2	-2
4	4	3	3
1	6	4	5
6	6	5 → ^s 5	5
10	6	6	5
6	8	7	7
-3	10	8	8
6	19	9	9

Wilcoxon rank sum and signed rank tests

For the paired test (or one sample test), the null hypothesis is that the distribution is symmetric.

In R we obtain a result very similar to what returned by the parametric test, thus reinforcing the conclusion

```
##  
## Wilcoxon signed rank test with continuity correction  
##  
## data: obj$Difference  
## V = 43, p-value = 0.01742 → Reject H0  
## alternative hypothesis: true location is not equal to 0
```

V: \sum of + values in signed ranks.

There are also two-sample extensions, for both independent data or paired data (though the latter can be performed by considering the differences, as done here). The two-sample version (for independent samples) is known as *signed rank test* or *Mann-Whitney test*.

In a parametric test we would have obtained $p \approx 0.014$.

Example

For the study about dog petting and praise, the times that the dog interacted with the owner (in seconds) are

Petting

```
## [1] 114 203 217 254 256 284 296
```

Praise

```
## [1] 1 2 3 4 5 6 13
```

```
## [1] 4 7 24 25 48 71 294
```

RANKS	Σ RANKS
7 8 9 10 11 12 14	74
1 2 3 4 5 6 13	39

$$W = \frac{\text{largest sum} - \frac{n_1(n_1+1)}{2}}{2}$$
$$= 71 - \frac{7 \cdot 8}{2} = 43$$

For a one-sided alternative hypothesis $H_1 : Me_x - Me_y > 0$ we obtain

```
wilcox.test(Petting,Praise,alternative="greater", exact=TRUE)
```

```
##  
## Wilcoxon rank sum exact test  
##  
## data: Petting and Praise  
## W = 43, p-value = 0.008741 → Reject H0.  
## alternative hypothesis: true location shift is greater than 0
```

Pearson's chi-squared test

This is a class of test applied to sets of categorical data to evaluate whether any observed difference between the sets arose by chance. It is suitable for unpaired data from large samples.

Pearson's chi-squared test is used to assess three types of comparison:

- **goodness of fit**: establishes whether an observed frequency differs from a theoretical distribution;
- **homogeneity**: test if two or more sub-groups of a population share the same distribution of a single categorical variable;
- **independence**: determines whether two categorical variables are associated

In all the cases the distribution of test statistic is, under H_0 a Chi-square distribution with some degrees of freedom.

Goodness of fit: Darts challenge against one friend

Suppose that n observations x_1, \dots, x_n are divided among K cells. The following test statistic is then defined:

$$\chi^2 = \sum_{k=1}^K \frac{(O_k - E_k)^2}{E_k} \stackrel{H_0}{\sim} \chi^2_{K-1},$$

where

convergence in distribution under H_0

- O_k are the observed frequencies for cell k
- E_k are the expected frequencies for cell k , under H_0

But what is H_0 here? For illustration purposes, suppose you are playing darts agains another friend.



Figure 4: Darts target

Goodness of fit (cont'd)

You suspect that your friend is not a great darts player, and that his shots along the game will hit the lowest points with great probability and the highest point with low probability. Translated in probability terms, you divide the darts target into $K = 4$ zones and you assign the the following hitting probabilities:

- Zone 1 (from 1 to 3 points): $p_1 = 7/16$;
- Zone 2 (from 4 to 6 points): $p_2 = 5/16$;
- Zone 3 (from 7 to 9 points): $p_3 = 3/16$;
- Zone 4 (the highest points in the middle of the target, let's say 10,25 and 50 points): $p_4 = 1/16$.

Your null hypothesis is that, due to a moderate control on his darts skills, he has decreasing probabilities to hit the best zones:

$$H_0 : p_1 = 7/16, \quad p_2 = 5/16, \quad p_3 = 3/16, \quad p_4 = 1/16.$$

Any significative deviation from the above probability distribution, would cause the rejection of the null hypothesis.

Goodness of fit (cont'd)

For checking your assumption, you count the first $n = 50$ attempts x_1, \dots, x_n of your opponent, and you will code $x_i = k$, if the i -th shot hits the k -th zone.

These are the observed (absolute) frequencies of 50 attempts

```
## x → k
## 1 2 3 4
## 23 17 9 1 → obs   Ok
```

Performing the test by hand, we need to compute the expected (absolute) frequencies

```
p <- c( 7/16, 5/16, 3/16, 1/16)
```

```
exp <- n * p ; n=50
```

```
exp
```

```
## [1] 21.875 15.625 9.375 3.125 Ek
```

Goodness of fit (cont'd)

The observed test statistics is

```
X2 <- sum( (obs - exp)^2/exp)
```

```
X2
```

```
## [1] 1.638857    chi-square    r.v.
```

and the p-value 0.6506

```
pchisq(X2, df = K - 1, lower.tail = FALSE )  
            3
```

```
## [1] 0.6506117  ↪ Accept
```

We obtain the same result by using the **chisq.test()** function

```
chisq.test(obs, p = p)
```

```
##  
## Chi-squared test for given probabilities  
##  
## data: obs  
## X-squared = 1.6389, df = 3, p-value = 0.6506
```

What's the conclusion?

Goodness of fit (cont'd)

Your friend is hitting the target according to your probabilities but he wants to play again, and performing another challenge.

Before starting, he requires to drink an energetic drink. You are ready to count his next 50 attempts. Does the drink improve the performance?

```
## x_drink  
## 1 2 3 4  
## 12 26 7 5
```

And the test result is

```
#new test after the energetic drink  
chisq.test(new_obs_ad, p = p)
```

```
##  
## Chi-squared test for given probabilities  
##  
## data: new_obs_ad  
## X-squared = 13.074, df = 3, p-value = 0.00448
```

→ Reject H_0

H_0 : he is not a good player, wrt the probabilities defined.
Of course since he practised

What's the conclusion?

Homogeneity: Darts challenge with more friends

Suppose now that other five friends join you and other guy, for a total amount of $M = 6$ friends.

Do all of your friends share the same probabilities, with the above probabilities to hit the four zones?

The test statistic is:

$$\chi^2 = \sum_{k=1}^K \sum_{m=1}^M \frac{(O_{k,m} - E_{k,m})^2}{E_{k,m}} \stackrel{H_0}{\sim} \chi^2_{(K-1)(M-1)}$$

For each of them, you count the first 50 shots

```
chisq.test(obs, p = p)
```

```
##  
## Pearson's Chi-squared test  
##  
## data: obs  
## X-squared = 12.743, df = 15, p-value = 0.6221
```

H_0 : all are bad players wrt the probabilities defined

What's the conclusion?

Homogeneity (cont'd)

Now a great player decides to join you, we do another round and perform again the test.

```
## x_great  
## 1 2 3 4  
## 3 8 17 22
```

The observed results for all the players are

```
## [,1] [,2] [,3] [,4] [,5] [,6] [,7]  
## 1 21 23 22 21 23 25 3  
## 2 15 14 16 21 17 11 8  
## 3 9 12 7 7 6 11 17  
## 4 5 1 5 1 4 3 22
```

Homogeneity (cont'd)

The test result is

```
##  
##  Pearson's Chi-squared test  
##  
## data: obs  
## X-squared = 88.166, df = 18, p-value = 3.077e-11
```



Reject

Likelihood-based tests

The likelihood ratio test*

We saw that the likelihood ratio makes possible to choose between different parameter values. Therefore, it is not strange that the likelihood ratio can be used as test statistic, being in some sense the optimal choice, as supported by the **Neyman-Pearson lemma**.

Formally, the lemma is valid for choosing between two *simple hypotheses* $H_0 : \theta = \theta_0$ vs $H_1 : \theta = \theta_1$, for any pair of parameter values θ_0 and θ_1 .

The **likelihood ratio test statistic** is given by

$$\lambda(\mathbf{y}) = \frac{L(\theta_1)}{L(\theta_0)} = \frac{f_{\theta_1}(\mathbf{y})}{f_{\theta_0}(\mathbf{y})}$$

Prop: the Neyman-Person's lemma states that the LRT is the "most powerful" for significance level α .

with rejection region

$$\mathcal{R}_\alpha = \{\mathbf{y} : \lambda(\mathbf{y}) \geq k_\alpha\}, \quad \text{wit} \quad H_1 : \mu > 0$$

being the test's *power* $\beta(\theta_0) = \Pr_{\theta_0}\{\lambda(\mathbf{Y}) \geq k_\alpha\} = \alpha$.

Three likelihood-based tests*

We first focus on a simple one-parameter model, and on the problem of testing $H_0 : \theta = \theta_0$ vs $H_1 : \theta \neq \theta_0$.

The following three tests are available:

- The **likelihood ratio test (LRT)**

$$W(\theta_0) = 2 \{ \ell(\hat{\theta}) - \ell(\theta_0) \}$$

- The **Wald test**

$$W_e(\theta_0) = (\hat{\theta} - \theta_0)^2 J(\hat{\theta}) = \frac{(\hat{\theta} - \theta_0)^2}{\text{SE}(\hat{\theta})^2}$$

- The **score test**

$$W_u(\theta_0) = \frac{U(\theta_0)^2}{I(\theta_0)}$$

In all the three cases, we reject H_0 for large values of the statistic, so that the p -value is (for instance) $p = \Pr_{\theta_0}(W \geq w_{obs})$.

Visually*

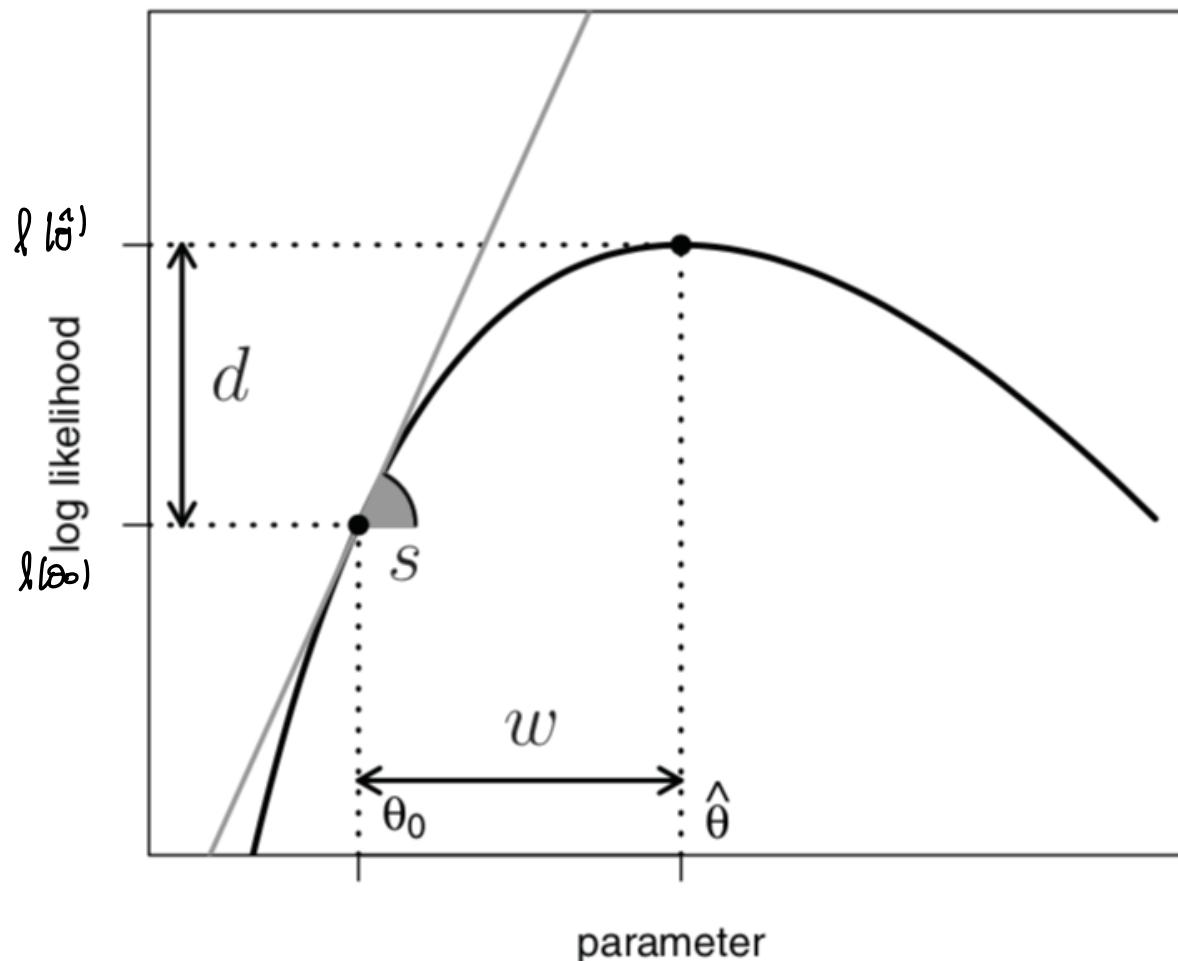


Figure 1. Comparing the three test statistics according to the traditional plot: Likelihood ratio is reported on the y scale, Wald on the x scale, and the score on the first derivative scale. The different scales do not favor understanding of the underlying connections.

Three likelihood-based tests: comments*

- Whenever available, the exact distribution of these tests can be employed.
- Which one is preferable? The likelihood ratio test is clearly an obvious choice, but for large samples the three statistics are equivalent: this fact can be proved by a Taylor expansion of $U(\hat{\theta})$ around θ_0 .
- From the asymptotic distribution of the MLE, it readily follows that the null distribution of W_e is approximately

$$W_e(\theta_0) \stackrel{d}{\sim} \chi_1^2 \quad \checkmark$$

and since the two other tests are equivalent in large samples, the same result holds also for them.

- For one-sided alternatives such as $H_1 : \theta > \theta_0$, the signed squared-root versions of the test should be used, namely (for the LRT)

$$R(\theta_0) = \text{sgn}(\hat{\theta} - \theta_0) \sqrt{W(\theta_0)}$$

and, under H_0 , $R(\theta_0) \stackrel{d}{\sim} \mathcal{N}(0, 1)$. $\sqrt{\chi_1^2} \stackrel{?}{\sim} \mathcal{N}(0, 1)$

Parameter of interest and nuisance parameters*

The three tests introduced readily generalize to hypotheses on the entire p -dimensional parameter θ . For instance, the LRT would become

$$W(\theta_0) = 2 \{ \ell(\hat{\theta}) - \ell(\theta_0) \}$$

with asymptotic null distribution given by χ_p^2 .

At any rate, the typical (and most interesting) situation is where we wish to test an hypothesis on a q -dimensional subset of θ , with $q < p$.

Following the CS book, we write $\theta^\top = (\psi^\top, \gamma^\top)$, with the null and alternative hypotheses given by $H_0 : \psi = \psi_0$ vs $H_1 : \psi \neq \psi_0$.

Here ψ is denoted as the **parameter of interest** and γ is the **nuisance parameter**.

The profile likelihood*

Likelihood theory handles nuisance parameters by introducing the **profile likelihood**.

Denoted by $\hat{\gamma}_\psi$ the MLE of γ for fixed value of ψ , namely

$$\hat{\gamma}_\psi = \underset{\gamma \in \Gamma}{\operatorname{argmax}} \ell(\psi, \gamma)$$

then we define the profile likelihood for ψ as

$$L_P(\psi) = L(\psi, \hat{\gamma}_\psi).$$

Note that the maximum of $L_P(\psi)$ is given by the MLE of ψ .

Inference based on the profile likelihood*

A crucial point is the large-sample properties of the profile likelihood are **those of a bona-fide likelihood function** for the parameter of interest only.

In particular, the profile likelihood LRT

$$W_P(\psi) = 2 \{ \ell_P(\hat{\psi}) - \ell_P(\psi_0) \} \sim \chi_q^2$$

the asymptotic null distribution is given by χ_q^2 .

Note, however, that if the dimension of γ is large, the large-sample results may be poor. In such cases, the parametric bootstrap is a more accurate route to obtain the p -value.

The t -test as a likelihood-based method*

Many noteworthy tests can be derived from the LRT based on the profile likelihood.

A very important instance is the t -test on μ for a normal random sample, $Y_i \sim \mathcal{N}(\mu, \sigma^2)$. With some simple algebra

$$\ell_P(\hat{\mu}) - \ell_P(\mu_0) = -\frac{n}{2} \log(\hat{\sigma}^2) + \frac{n}{2} \log(\hat{\sigma}_{\mu_0}^2),$$

and since $\hat{\sigma}_{\mu}^2 = \hat{\sigma}^2 + (\hat{\mu} - \mu)^2$, it follows

$$r_P(\mu_0) = \text{sgn}(\hat{\mu} - \mu_0) \sqrt{n \log \left\{ 1 + \frac{(\hat{\mu} - \mu_0)^2}{\hat{\sigma}^2} \right\}}.$$

Further simple algebra shows that $R_P(\mu_0)$ is a monotonic increasing function of the T test statistic $T(\mu_0) = (\bar{y} - \mu_0)/\sqrt{s^2/n}$, so that, for instance, $\Pr_{H_0}\{R(\mu_0) \geq r_{obs}\} = \Pr_{H_0}\{T(\mu_0) \geq t_{obs}\}$.

Other notable instances*

Several other tests can be derived as special cases of the LRT, such as the F test for one-way anova models, or exact tests employed in linear regression models.

Other famous tests are instead special cases of the score test. The most notable instance is the chi-squared test of independence for two-way contingency tables, and related tests. The underlying statistical model is the **multinomial distribution** for the observed frequencies.

Linear models

(An Introduction)

N. Torelli, G. Di Credico, V. Gioia
Fall 2023

University of Trieste



Introduction to linear models

Multiple linear model

Introduction to linear models

- Assumptions
- Estimation
- Assessment

There is not one single correct model. It depends on what we want to express.
The model must be interpretable

Linear regression model

Linear regression model is one of the basic tools for statistical analysis.

Since pioneering works of Sir Francis Galton in the late XIX century, the main aim of regression models is to study the systematic influence of

- one or more **concomitant factors** (explanatory variables, regressors, covariates) on *L_e variables observed*
- a **response variable** (dependent variable).

The main goal of regression modelling is understanding *whether and how* the response variable (the phenomenon of interest) is related to the concomitant quantities.

The basic regression model has been expanded in many directions in order to apply it in extremely complex situations and to large and complex data sets, but the basic aim remained the same.

The relation is not symmetric.

Aims of regression modelling

- prediction/forecast: regression modelling is a tool to provide a prediction of the phenomenon of interest, given the knowledge of the concomitant factors (e.g. for time reasons, costs, or because the concomitant factors are easier to measure) $x \rightarrow y$
- interpretation: which factors affect more the phenomenon of interest and how? Which is the direction of the relationship between the phenomenon of interest and a specific concomitant factor?

$$y = ax + b$$

a st. y changes

The main ingredients

- *Response variable* is the quantity of main interest (it can be quantitative or qualitative), let's denote it by Y ;
- *Explanatory variables* (also called, predictors or covariates) are the concomitant factors, let's denote them by X_1, \dots, X_{p-1} ;

A first formalization

A rough way to formalize the problem is by specifying a functional relationship:

$$Y = g(X_1, \dots, X_{p-1})$$

as an approximation of a possible “true”, yet unknown, relationship.

- $g(\cdot)$ is not known but in some case we can conjecture its shape by consideration on the nature and the characteristics of the phenomenon of interest
- or we can choose a very simple structure such as

$$Y = \beta_0 + \sum_{j=1}^{p-1} \beta_j X_j = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + X_{p-1} \beta_{p-1}$$

- usually involved variables are measured on a sample of n subjects $(x_{i1}, \dots, x_{ip-1}; y_i)$, $i = 1, \dots, n$. We want to use these data to explore possible relationship between Y and the covariates.

RHS doesn't include variability, but Y (LHS) is a r.v.

Simple linear model: a basic example

Heating consumption in a house depends on temperature?

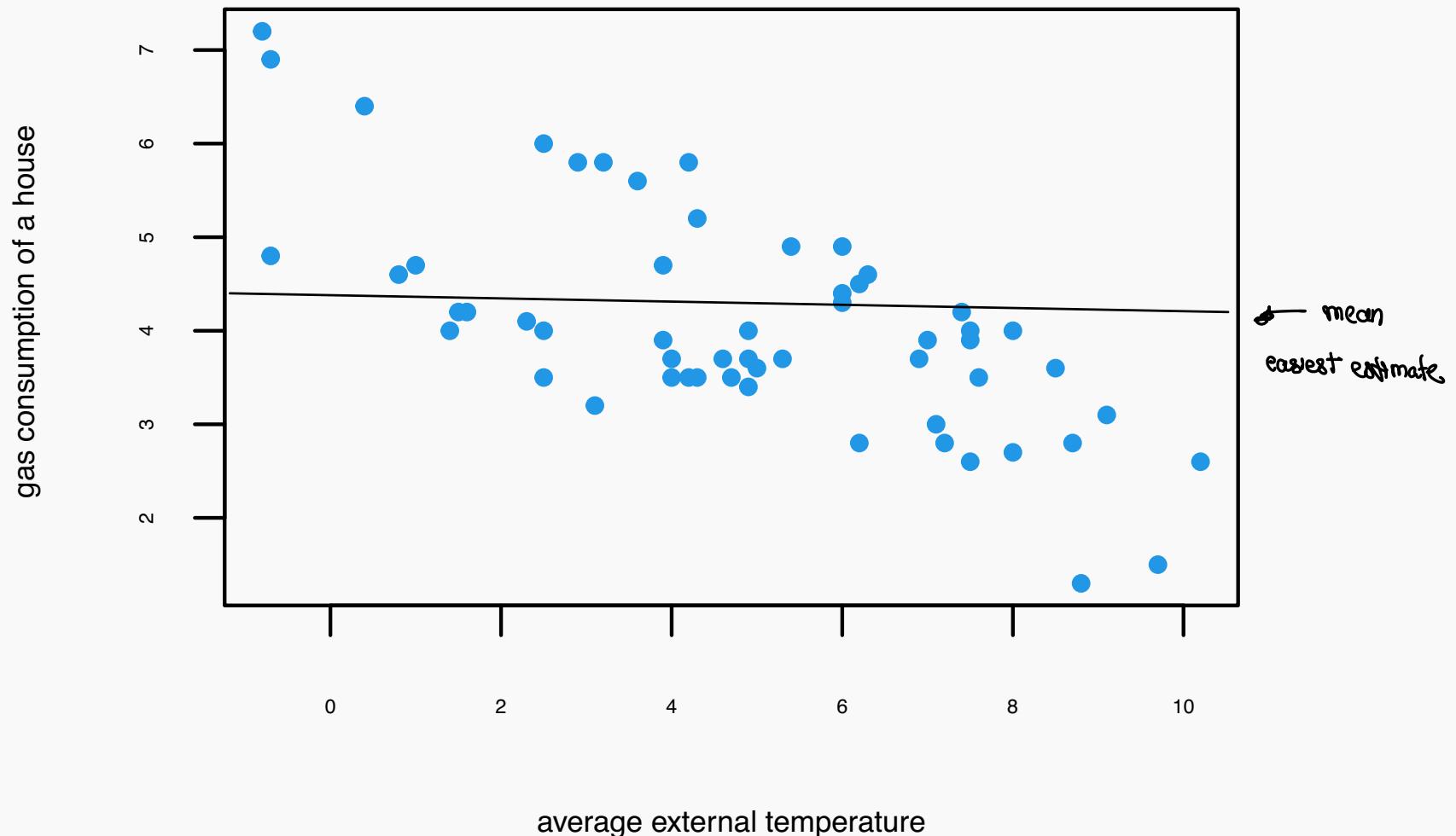
- We are interested in predicting the heating consumption in a house at some time
- We observe the weekly consumption of gas (Y , in thousands cube feet) over n weeks ($y_1, \dots, y_i, \dots, y_n$)
- A first rough prediction of the gas consumption is: $\bar{y} = \frac{\sum_{i=1}^n y_i}{n}$
- It is sensible to think that the external temperature affects the heating consumption (we expect that the gas consumption decreases as the average external temperature increases).
- In addition to y_1, \dots, y_n , we observe the average external temperature (x_i , in Celsius degrees) registered for the same weeks of observation of y_i . Thus, our sample of data is:

$$(x_1, y_1), (x_2, y_2), \dots, (x_i, y_i), \dots, (x_n, y_n)$$

Simple linear model: a basic example

Heating house's consumption for varying temperature

Heating consumption in a house for varying temperature across 56 weeks



The simple linear regression model: Model specification

- Can we better describe the conjectured relationship in order to make a more accurate prediction than \bar{y} ?

$$\begin{aligned}\text{gas consumption in week } i &= g(\text{temperature in the same week}) \\ y_i &= g(x_i)\end{aligned}$$

- Whatever g we assume, it will simply be an approximation and we should also take into account that the dependent variable is affected by a random error ϵ_i

$$\begin{aligned}\text{gas consumption in week } i &= g(\text{temperature in the same week}; \\ &\quad \text{non observed and less relevant factors}) \\ y_i &= g(x_i; \epsilon_i) \quad \begin{array}{l} \uparrow \\ \text{↳ includes "error" and non-included factors.} \end{array}\end{aligned}$$

- In fact, we may say that, conditional to a given value of x_i , *the expectation of y_i is:*

$$E[Y_i | X_i = x_i] = \mu_i = g(x_i) \quad \text{deterministic component}$$

i.e. μ_i is the population mean of y_i , conditional to the value of x_i .

The simple linear regression model: Model specification

- Data suggest that the expected heating consumption can be modeled as a linear function of the external temperature
- A simple model: the straight line

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad i = 1, \dots, n$$

or equivalently:

$$\mu_i = \beta_0 + \beta_1 x_i \quad = \text{real} - \text{predicted}$$

- The model assumes a constant growth rate of gas consumption for decreasing values of temperature: the effect on μ of a constant increase of x is the same whatever is x .

The simple linear regression model: Model assumptions

- The model is correctly specified:

$$y_i = g(x_i; \epsilon_i) = \beta_0 + \beta_1 x_i + \epsilon_i$$

= systematic component + stochastic component

$\xrightarrow{\text{constant}} \nabla \sum J = 0$

- systematic component: predictors are under the control of the researchers (non stochastic)
- stochastic component:

- the error terms have zero mean

\mapsto it does not include further systematic terms

$$\mapsto \mu_i = \beta_0 + \beta_1 x_i$$

$$\text{s.t. } E[\epsilon_i] = \mu_i$$

- the error terms have constant variance and are uncorrelated

$$\nabla \sum \epsilon_i^2 = \sigma^2 \quad \forall i$$

- (useful, albeit to some extent not necessary, the error terms are normally distributed)

$$\epsilon_i \sim \text{i.i.d. } \mathcal{N}(0, \sigma^2)$$

$$\rho(\epsilon_i, \epsilon_j) = 0$$

$$\forall i, j$$

Not necessary for y_i
but important for hypothesis testing with MLE

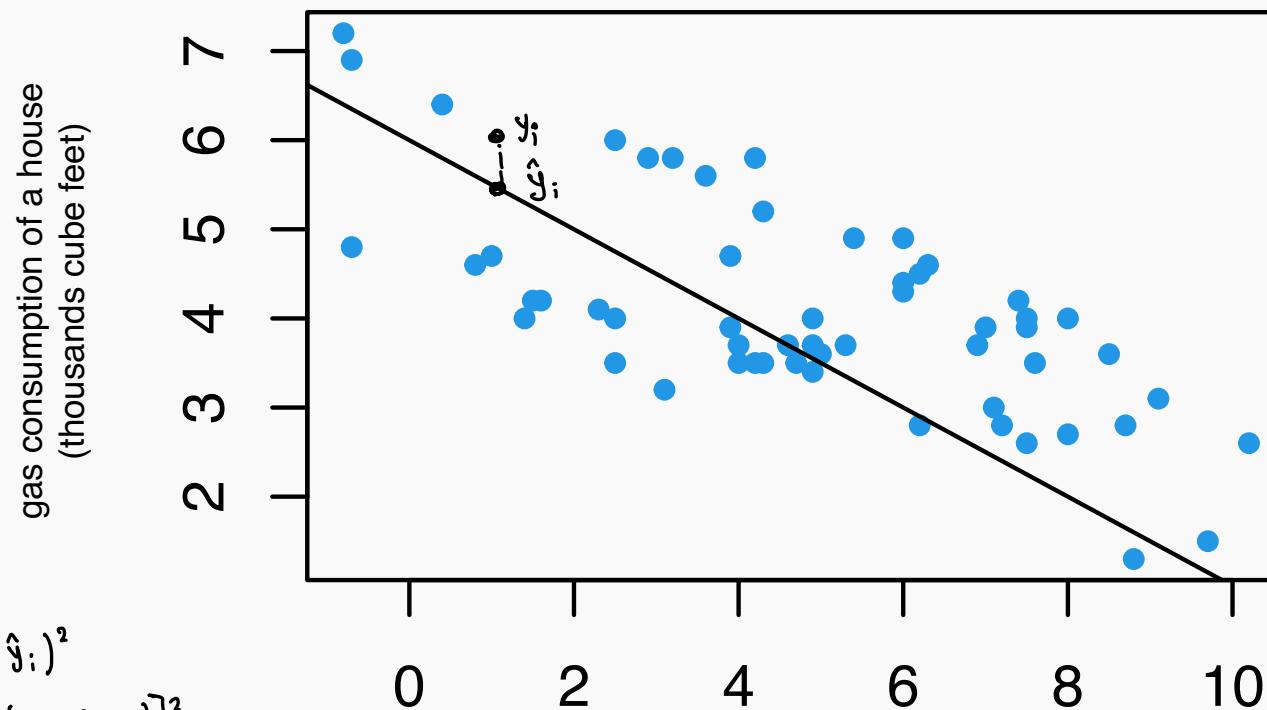
With the normality assumption

$$Y_i \sim \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2)$$

Y_i and Y_j uncorrelated $\forall i \neq j$.

The simple linear regression model: How to choose the best line?

- How to use data to estimate the unknown parameters β_0 and β_1 ?



$$S(\beta_0, \beta_1) = \sum (y_i - \hat{y}_i)^2 \\ = \sum (y_i - (\beta_0 + \beta_1 x_i))^2$$

$$\hat{\beta}_0, \hat{\beta}_1 = \min S(\beta_0, \beta_1)$$

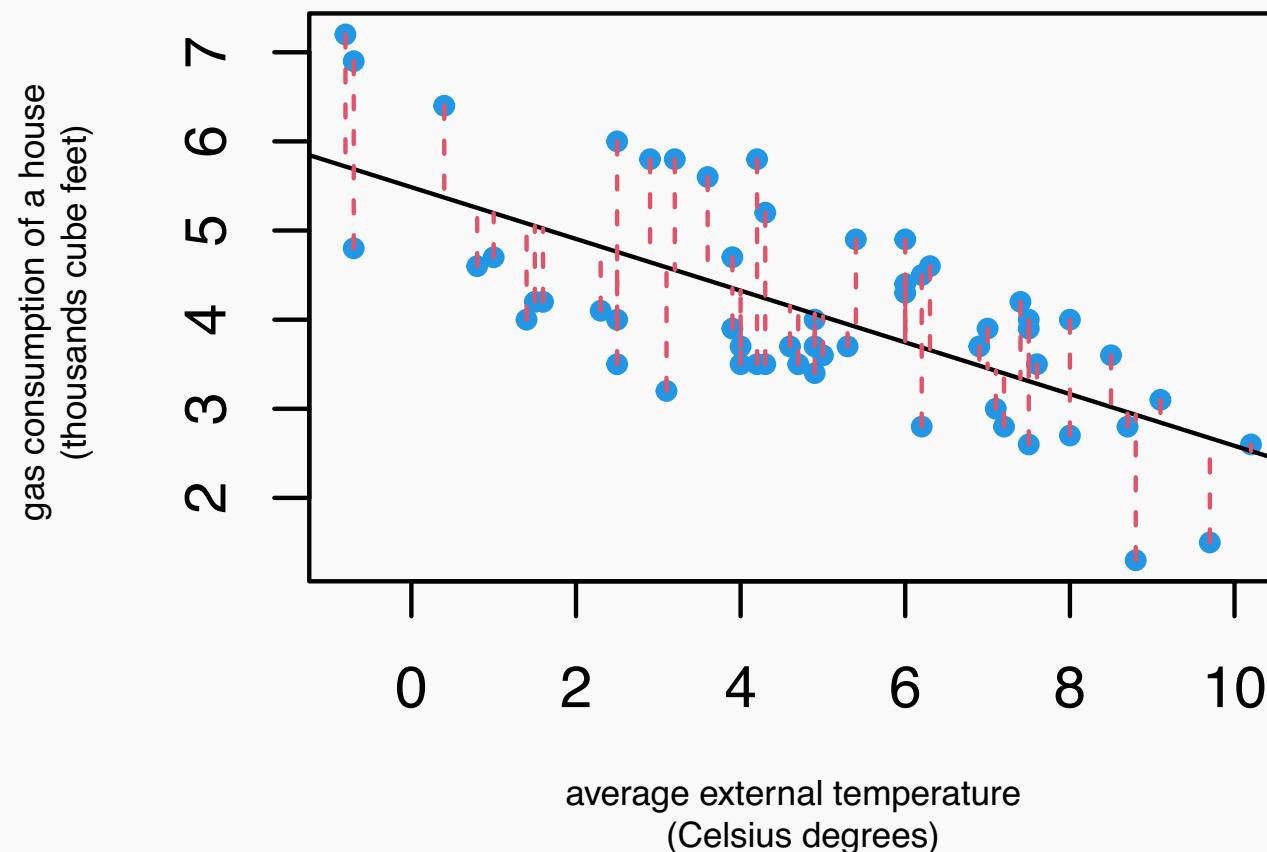
$$\frac{\partial S(\beta_0, \beta_1)}{\partial \beta_0} = 0 \rightarrow \hat{\beta}_0 = \bar{y} - \beta_1 \bar{x} \Rightarrow \hat{y} = \beta_0 + \beta_1 x$$

$$\hat{\beta}_1 = \frac{\sum x_i y_i - \bar{y} \sum x_i}{\sum (x_i^2 - \bar{x} \sum x_i)} = \frac{\sum x_i y_i - n \bar{x} \bar{y}}{\sum (x_i^2 - n \bar{x}^2)} = \frac{\text{Cov}(x, y)}{\text{Var}[x]} = g(x, y) \frac{S[y]}{S[x]}$$

sgn [slope] depends sgn [covariance]
i.e., sgn of b1 by LR

How to choose the best line? The least squares criterion

- Choose the line which minimizes the sum of squared residuals



```
## [1] 39.99487
```

How to choose the best line? The least squares criterion

- Choose the line which minimizes the sum of squared residuals:

$$\min_{\beta_0, \beta_1} \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_i))^2$$

minimize

$$L(\beta_0, \beta_1, \sigma^2) = \text{const} + \frac{1}{2\sigma^2} \sum (y_i - \mu_i)^2$$

maximizes

L
and

β_0, β_1 are the same

- The minimization problem has the following solution:

$$\hat{\beta}_1 = \frac{\text{cov}(x, y)}{\text{var}(x)} \quad (\text{slope})$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \quad (\text{intercept})$$

It is important to remind that **if we assume also that the random component of the model is normally distributed, then maximum likelihood estimation lead to the same solution** (least squares is also the maximum likelihood solution).

In the gas consumption example the two estimated coefficients are:

```
## [1] 5.4861933 -0.2902082
```

$$g_c(T=0)$$

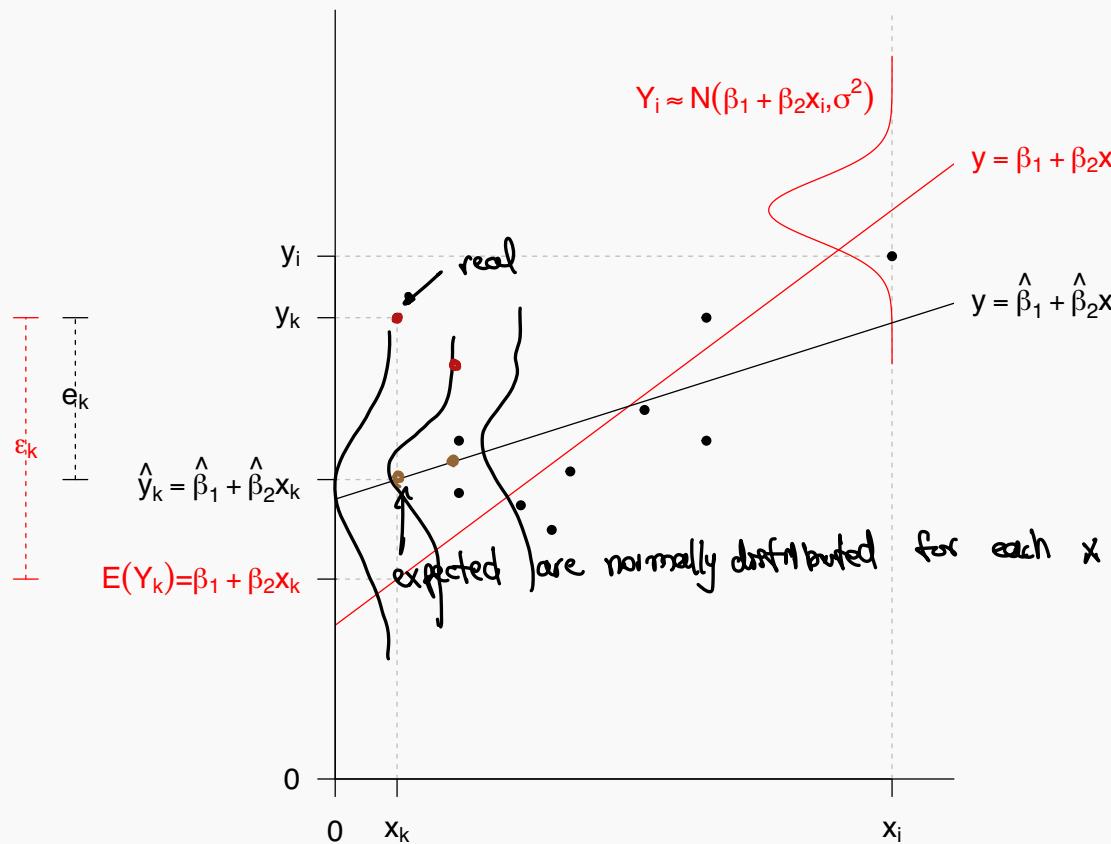
$$g_c(T=-1)$$

$$x_0 = 10$$

$$\hat{y}_0 = \hat{\beta}_0 + x_0 \cdot \hat{\beta}_1$$

$$\hat{\beta}_1 \bar{x} \Big|_{\bar{x}=-1} = \hat{\beta}_0 - \bar{y}$$

True and estimated relationship



Inference

Statistical tests allow us to draw general considerations about the model,
valid not only for the sample at hand

- Is the model useful somehow?
 - ↪ Test the usefulness of the whole model $\hat{\beta}_0 = 0$?
 - Does the explanatory variable X really affect the response variable?
 - ↪ Test the significance of a single predictor
- ↪ (in the simple linear regression model the two above are equivalent)

Estimating σ^2

ε_i : random variables
 e_i : residuals, not rvs

The estimate $\hat{\sigma}^2$ is $\frac{1}{n} \sum_1^n e_i^2$ where $e_i = (y_i - \hat{y}_i)$ are the residuals,

$$\begin{aligned}\hat{y}_i &= \hat{\beta}_0 + \hat{\beta}_1 x_i \\ \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x}\end{aligned}$$

This estimate is biased.

$$\hat{\sigma}^2 = \frac{1}{n} \sum_i^n (y_i - \bar{y})^2 - \hat{\beta}_1^2 \frac{1}{n} \sum_i^n (x_i - \bar{x})^2$$

↳ estimate of the error

observed - expected

Write code to compute it

$$\hat{\sigma}^2 = \sum_i^n y_i^2 / n - \bar{y}^2 - \hat{\beta}_1^2 \left(\sum_i^n x_i^2 / n - \bar{x}^2 \right)$$

$$\begin{aligned}\bar{y} &= \hat{\beta}_0 + \hat{\beta}_1 \bar{x} \\ \hat{y}_i &= \hat{\beta}_0 + \hat{\beta}_1 x_i\end{aligned}$$

An unbiased estimate is

$$s^2 = \frac{n}{n-2} \hat{\sigma}^2$$

$$\begin{aligned}n\hat{\sigma}^2 &= \sum e_i^2 = \sum (y_i - \hat{y}_i)^2 = \sum (y_i - \bar{y})^2 - \sum (\bar{y} - \hat{y}_i)^2 \\ &= \sum y_i^2 - n\bar{y}^2 - \sum \hat{\beta}_1 (x_i - \bar{x})^2 \\ &= \sum y_i^2 - n\bar{y}^2 - \hat{\beta}_1 (\sum x_i^2 - n\bar{x}^2)\end{aligned}$$

In general, for p covariates

$$s^2 = \frac{n}{n-p} \hat{\sigma}^2 \quad \text{is unbiased}$$

Testing usefulness of the overall model: A first useful index

- y varies in the population; its variability may be measured by:

$$\sum_{i=1}^n (y_i - \bar{y})^2 \quad \text{Total Sum of Squares (SS)}$$

Variability with respect to the mean

- The whole variability of y may be decomposed as follows:

- variability of y explained by the model:

R^2 represents the ?

$$\sum_{i=1}^n (\hat{y}_i - \bar{y})^2 \quad \text{Regression SS}$$

- residual variability (due to the chance):

$$\sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad \text{Residual SS}$$

- It can be shown that, in the linear model:

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$$

Total SS

Residual SS

Regression SS

Testing usefulness of the overall model: A first useful index

- If the model is good:

↪ Residual SS is small compared to Total SS
↪ Regression SS is the main portion of Total SS

- Coefficient of Determination:

↓
if there is not difference
between the prediction and
the system ResidualSS $\rightarrow 0$

$$R^2 = \frac{\text{RegressionSS}}{\text{TotalSS}} = 1 - \frac{\text{ResidualSS}}{\text{TotalSS}}$$

% of variability of y explained by the model

↪ $0 \leq R^2 \leq 1$
↪ The lower R^2 the worse the fitted model

Later we will penalize R^2 , reducing its value,
for each parameter of the model
Done to avoid overfitting

In the heating consumption example: Total SS, Regression SS and R^2 are respectively

```
## [1] 75.0142857 35.0194175 0.4668366
```

The model can explain about the 47% of the variability of y .

Testing usefulness of the overall model: The F test

- The quantities above may be used to build a formal statistical test:

Equivalent to:

$$\begin{cases} H_0: \beta_1 = 0 \\ H_1: \beta_1 \neq 0 \end{cases}$$

$$H_0: \mu_i = \beta_0 \quad \text{vs} \quad H_1: \mu_i = \beta_0 + \beta_1 x_i$$

We assume

$$\frac{TSS}{\sigma^2} \sim \chi^2_{n-1}$$

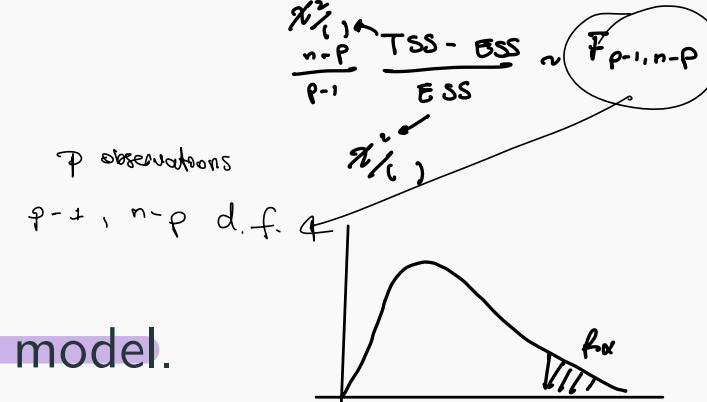
$$\frac{TSS - ESS}{\sigma^2} \sim \chi^2_{p-1}$$

$$\frac{ESS}{\sigma^2} \sim \chi^2_{n-p}$$

β_1 to "standardize" and
make variables unitless

- The F statistic is the ratio of
 - the explained variability (as reflected by R^2) and
 - the unexplained variability (as reflected by $1 - R^2$)
 suitably adjusted according to the number of observations (n) and the number of estimated parameters ($p = 2$):

$$F = \frac{R^2}{(1 - R^2)} \frac{n - p}{p - 1}$$



For extreme values
we reject H_0 and
"accept" other models
/complex models
are better.

$$E[f] = \frac{df_1}{df_1 - 2} \rightarrow \downarrow$$

don't reject H_0 .



Testing single predictors

- It is of interest to test whether the explanatory variable X really affects the response variable
- A formal statistical test may be built to check:

$$H_0 : \beta_1 = 0 \quad \text{vs} \quad H_1 : \beta_1 \neq 0$$

- The t statistic is the ratio:

$$\begin{aligned} X &\sim t = (\) / () \\ X^* &\sim F_{n,n} \end{aligned}$$

$$t = \frac{\hat{\beta}_1}{\text{standard error}(\hat{\beta}_1)}$$

- The larger the t statistic, the more evidence against H_0
- Under the assumption of gaussianity of the error term, the *probability distribution* of t is known and it allows us to define *critical values* and *p-values*.
- Note that for simple regression the t test is equivalent to the F test

Inference: The Heating consumption example

```
##  
## Call:  
## lm(formula = y ~ x)  
##  
## Residuals:  
##      Min       1Q   Median       3Q      Max  
## -1.6324 -0.7119 -0.2047  0.8187  1.5327  
##  
## Coefficients:  
##               Estimate Std. Error t value Pr(>|t|)    With these values it is possible to perform hypothesis testing  
## (Intercept) 5.4862     0.2357  23.275 < 2e-16 ***  Reject H0  
## x          -0.2902     0.0422  -6.876 6.55e-09 ***  
## ---  
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##  
## Residual standard error: 0.8606 on 54 degrees of freedom  
## Multiple R-squared: 0.4668, Adjusted R-squared: 0.457  
## F-statistic: 47.28 on 1 and 54 DF, p-value: 6.545e-09  
  

$$\left(\frac{p}{se}\right)^2$$
  

$$\uparrow \quad \quad \quad \text{F} \uparrow \quad p \rightarrow 0 \rightarrow \text{reject } H_0$$

```

Prediction

Confidence interval for the mean Y_0 (red):

$$\hat{Y}_0 \pm t_{n-2, 1-\alpha/2} \sqrt{s^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_i^n (x_i - \bar{x})^2} \right)}$$

confidence interval for.

$$\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}_0$$

mean

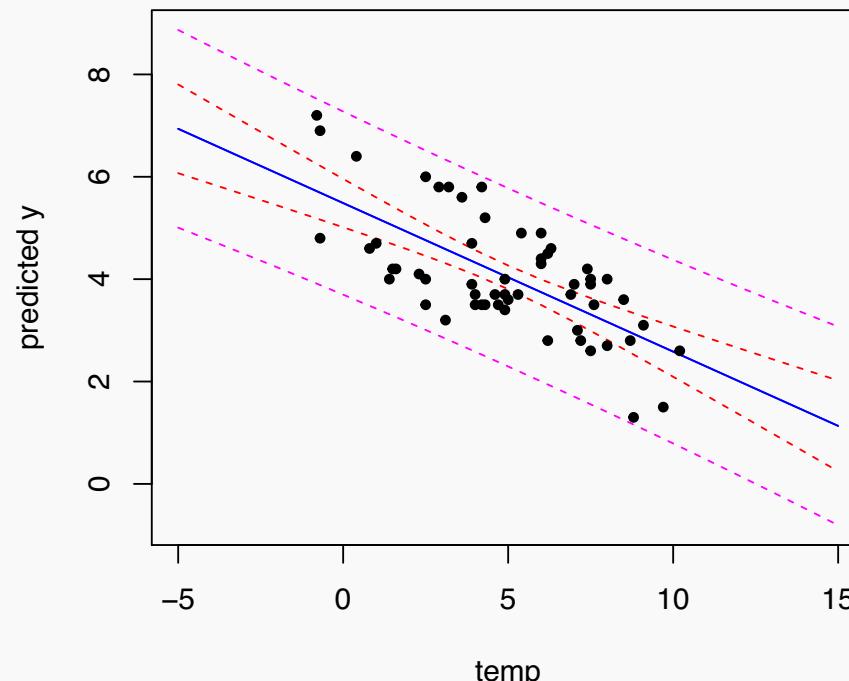
Confidence interval for prediction of a new value \hat{y}_0 (purple):

$$\hat{y}_0 \pm t_{n-2, 1-\alpha/2} \sqrt{s^2 \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_i^n (x_i - \bar{x})^2} \right)}$$

adds var. for new observation

$$\hat{\beta}_0 + \hat{\beta}_1 x_0 + \varepsilon_0$$

prediction



Model checking

- The least squares line is the line which best fits the data at hand but... *best is not (necessarily) good.*
- How to establish if the estimated model is a good one?
 - ↪ Residuals are an estimate of the error components ϵ_i

$$\begin{aligned} e_i = \hat{\epsilon}_i &= y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \\ &= y_i - \hat{\mu}_i \quad i = 1, \dots, n \end{aligned}$$

- ↪ Residual analysis allows to check if the model assumptions are met and if the model is good

Assumptions for model:
- mean error 0
- variance error const
- errors non corr.
extra error term $N(,)$

Assumptions for model

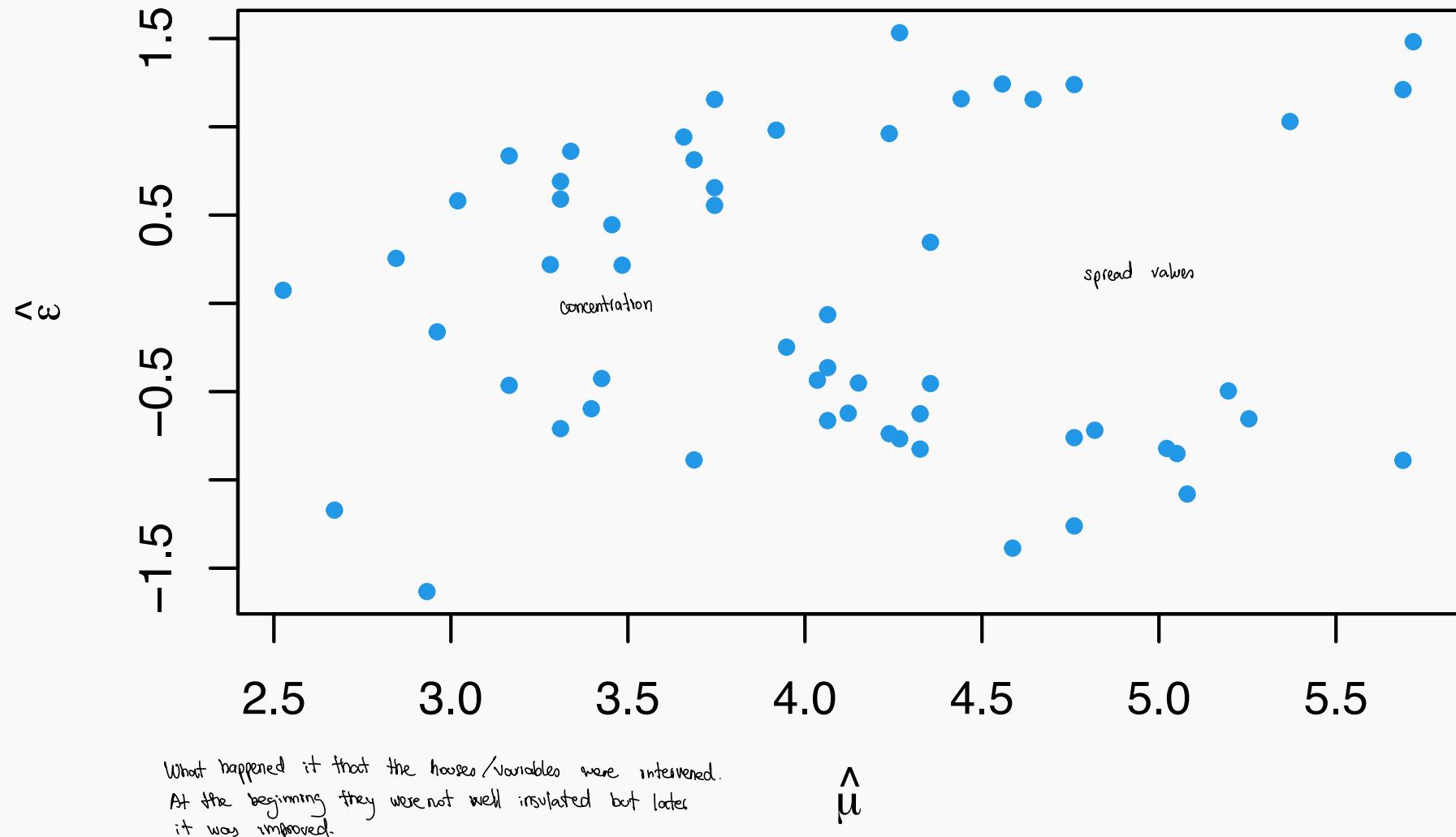
Assessment
- residuals from model follow / respect assumptions.

Model checking - Residual plots

- How should be the residuals in a good model?
 - \mapsto both positive and negative (around zero)
 - \mapsto small
 - \mapsto have constant variability
 - \mapsto scattered at random (if the model is well specified the amount of variability of y not explained by x must be due to the chance only \Rightarrow the residuals do not show any regularity)
- Residual plots: *Most Important to Assess Model's Assumptions.*
 - $\mapsto \hat{\epsilon}_i$ vs $\hat{\mu}_i$
If patterns are observed the model needs to be verified.
`par(mfrow = c(3,2))
plot(fit)
plot(x,e)`
 - $\mapsto \hat{\epsilon}_i$ vs each x_i
 - \mapsto Normal QQ-plot of $\hat{\epsilon}_i$
- Other plots based on residuals
 - \mapsto Leverages *Influence of each observation on the model*
 - \mapsto Cook distances



Model checking - Residual plots



Model checking

- Although not too bad, the residual plot suggest some problems
 - The variability of the residuals is not constant
 - There seem to be two clusters of residuals
 - ↪ Positive residuals increase for increasing $\hat{\mu}$
 - ↪ Negative residuals decrease for increasing $\hat{\mu}$
 - It seems that for two groups of observations the estimated relationship between heating consumption and external temperature is different from the estimated one (as the residuals are still some function of the temperature via $\hat{\mu}$).
- Have we forgotten anything relevant?

Multiple linear model

The (multiple) linear regression model: Introduction

- In fact, we get to know that during the observation time, there has been an insulation intervention on the house so that in the last 30 weeks the house was insulated.
- Data at hand thus become:

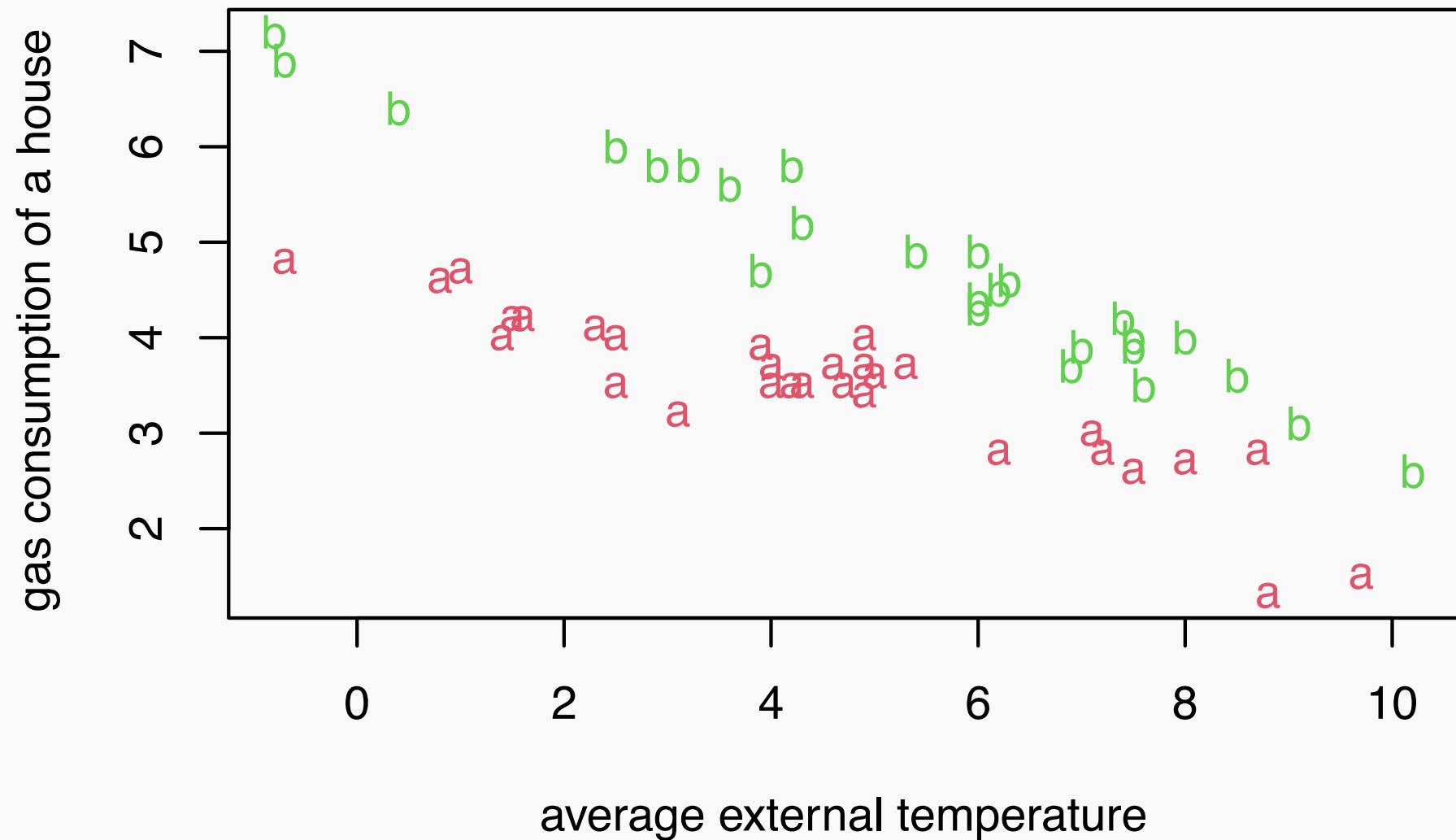
$$(x_1, z_1, y_1), (x_2, z_2, y_2), \dots, (x_i, z_i, y_i), \dots, (x_n, z_n, y_n)$$

with $z_i = \text{"before insulation"}$ for $i = 1, \dots, 26$
and $z_i = \text{"after insulation"}$ for $i = 27, \dots, 56$

$\ln(y \sim x + z)$
 z only enters in the model when it is equal to 1.

- It is sensible to expect that the isolation intervention has an impact on the mean heating consumption and that after the intervention the heating consumption is lower than before it.

The (multiple) linear regression model: Introduction



Model specification

- In the light of the availability of the additional variable z the model can be specified as follows:

gas consumption in a week = $g(\text{temperature in the same week, before/after intervention})$
non observed and less relevant factors)

$$y_i = g(x_i, z_i; \epsilon_i)$$

- The natural extension of the straight line model is the following *(multiple) linear model*

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \epsilon_i \quad (1)$$

or equivalently:

$$\mu_i = \beta_0 + \beta_1 x_i + \beta_2 z_i$$

Qualitative predictors

- The additional variable has a qualitative nature as it takes values "before intervention" and "after intervention" ⇒ the specified model does not make sense in the current form as z_i is not a number
- The standard way to overcome the problem is to introduce an auxiliary variable, an indicator variable (econometricians call it *dummy*):

$$d_i = \begin{cases} 0 & \text{if } z_i = \text{"before intervention"} \\ 1 & \text{if } z_i = \text{"after intervention"} \end{cases}$$

- The (1) then becomes:

β_2 : expected difference between two conditions

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_i + \beta_2 d_i + \epsilon_i \\ &= \beta_0 + \beta_1 x_i + \epsilon_i \quad \text{before the intervention} \\ &= (\beta_0 + \beta_2) + \beta_1 x_i + \epsilon_i \quad \text{after the intervention} \end{aligned}$$

- In other words the introduction of the indicator variable d gives rise to two parallel straight lines, one for each value of z_i

Qualitative predictors (factors)

Model estimation via least squares easily extends to the multiple linear model:

$$\min_{\beta_0, \beta_1, \beta_2} \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_i + \beta_2 d_i))^2$$

The three coefficients are respectively:

$\hat{\beta}_0$ ^{→ before int.}
 $\hat{\beta}_1$
 $\hat{\beta}_2$
[1] 6.551329 -0.336697 -1.565205

read below $\hat{\beta}_0$?

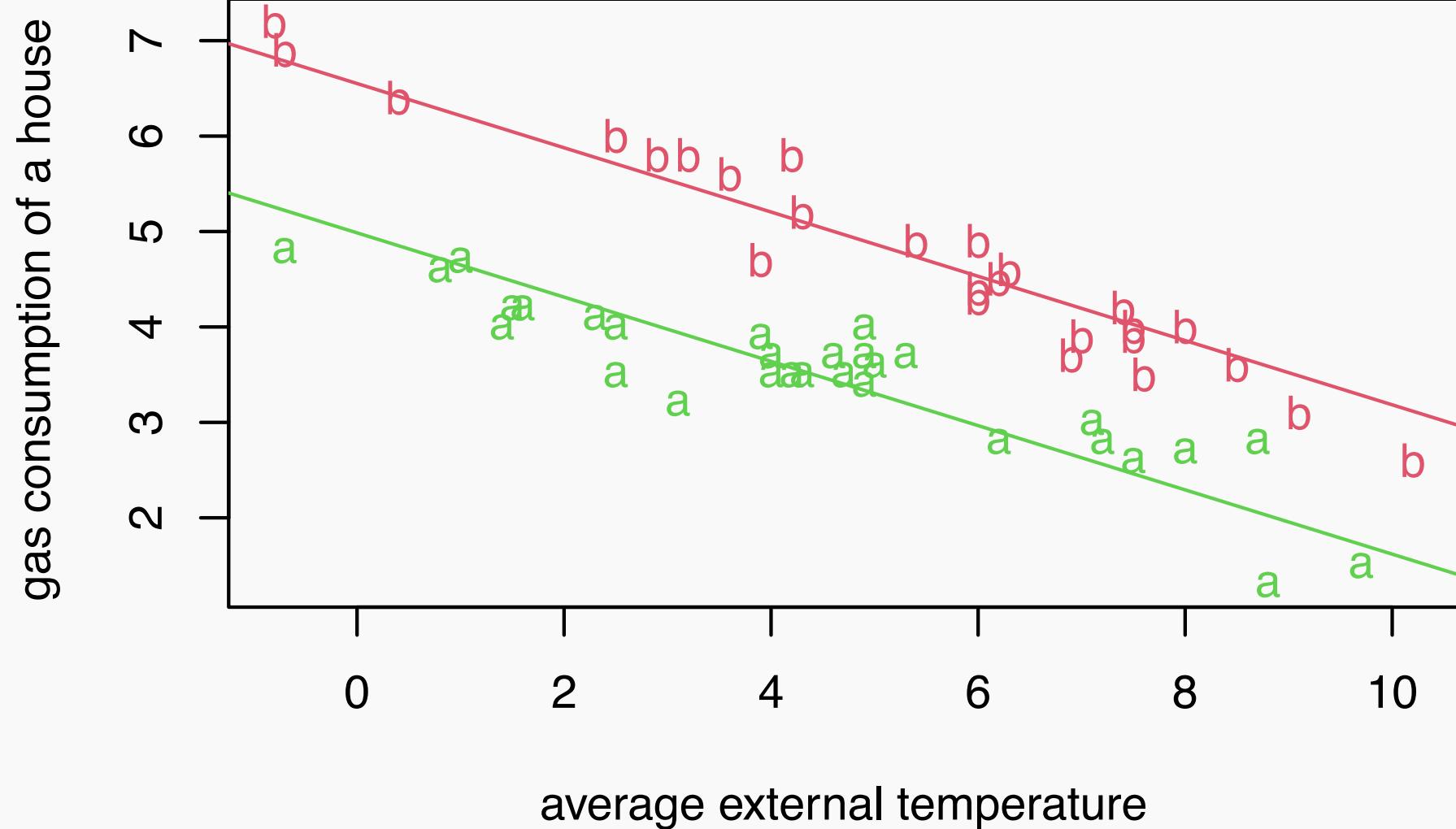
or

$\hat{\beta}_0 + \hat{\beta}_1$?



↳ after intervention

Qualitative predictors (factors)



Interpreting the model

- $\hat{\beta}_0$: expected response value when all the predictors are set at zero (if it does make sense and 0 is in the range of observed predictors)
 - ↪ If the external temperature is 0 degree, and before the isolation intervention ($d_i = 0$), the expected consumption of gas is about 6.6 thousand cube feet
- $\hat{\beta}_j$ ($j \geq 1$): expected change of y when the j -th predictor increases by 1 unit and all the other predictors are kept constant:
 - ↪ If the external temperature increases by 1 degree, the expected consumption of gas decreases by about 0.34 thousand cube feet, independently of the isolation intervention
 - ↪ If the house gets isolated (d_i passes from 0 to 1), the expected consumption of gas decreases by about 1.57 thousand cube feet independently of the external temperature

Interpreting the model

- The estimated line can be used to get a prediction of y for any value of the predictors (in the range of observed values)

$$\hat{y} = \hat{\mu} = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 d$$

→ What is the expected gas consumption when the external temperature is 5?

- if the house is not insulated:

$$\hat{\mu} = 6.551 - 0.3367 \cdot 5 = 4.8675 \text{ thousand cube feet}$$

- if the house is insulated:

$$\hat{\mu} = 6.551 - 1.565 - 0.3367 \cdot 5 = 3.3025 \text{ thousand cube feet}$$

$$\bar{y} = \hat{\beta}_0 - \hat{\beta}_1 \cdot \bar{x} \rightarrow (\hat{\beta}_0 + \hat{\beta}_2) - \hat{\beta}_1 \bar{x}$$

Inference

Statistical tests allow us to draw general considerations about the model, valid not only for the sample at hand

- Is the model useful somehow?
 - ↪ Compute the *adjusted R²*: a suitable adjustment of R^2 which penalises additional explanatory variables (descriptive)
 - ↪ Test the usefulness of the whole model: the F test
 - $H_0 : \mu_i = \beta_0 \Leftrightarrow \beta_1 = \beta_2 = 0$ vs
 - H_1 : at least one between β_1 and β_2 is not 0
- Does the j -th explanatory variable really affect the response variable?
 - ↪ Test the significance of a single predictor: the t test
 - $H_0 : \beta_j = 0$ vs $H_1 : \beta_j \neq 0$

Inference: The Heating consumption example

```
##  
## Call:  
## lm(formula = y ~ x + z)  
##  
## Residuals:  
##      Min       1Q   Median       3Q      Max  
## -0.74236 -0.22291  0.04338  0.24377  0.74314  
##  
## Coefficients:  
##             Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 6.55133   0.11809  55.48 <2e-16 ***  
## x          -0.33670   0.01776 -18.95 <2e-16 ***  
## zafter     -1.56520   0.09705 -16.13 <2e-16 ***  
## ---    ↗ this is called zafter and not z because this is a categorical variable.  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##  
## Residual standard error: 0.3574 on 53 degrees of freedom  
## Multiple R-squared:  0.9097, Adjusted R-squared:  0.9063  
## F-statistic: 267.1 on 2 and 53 DF,  p-value: < 2.2e-16
```

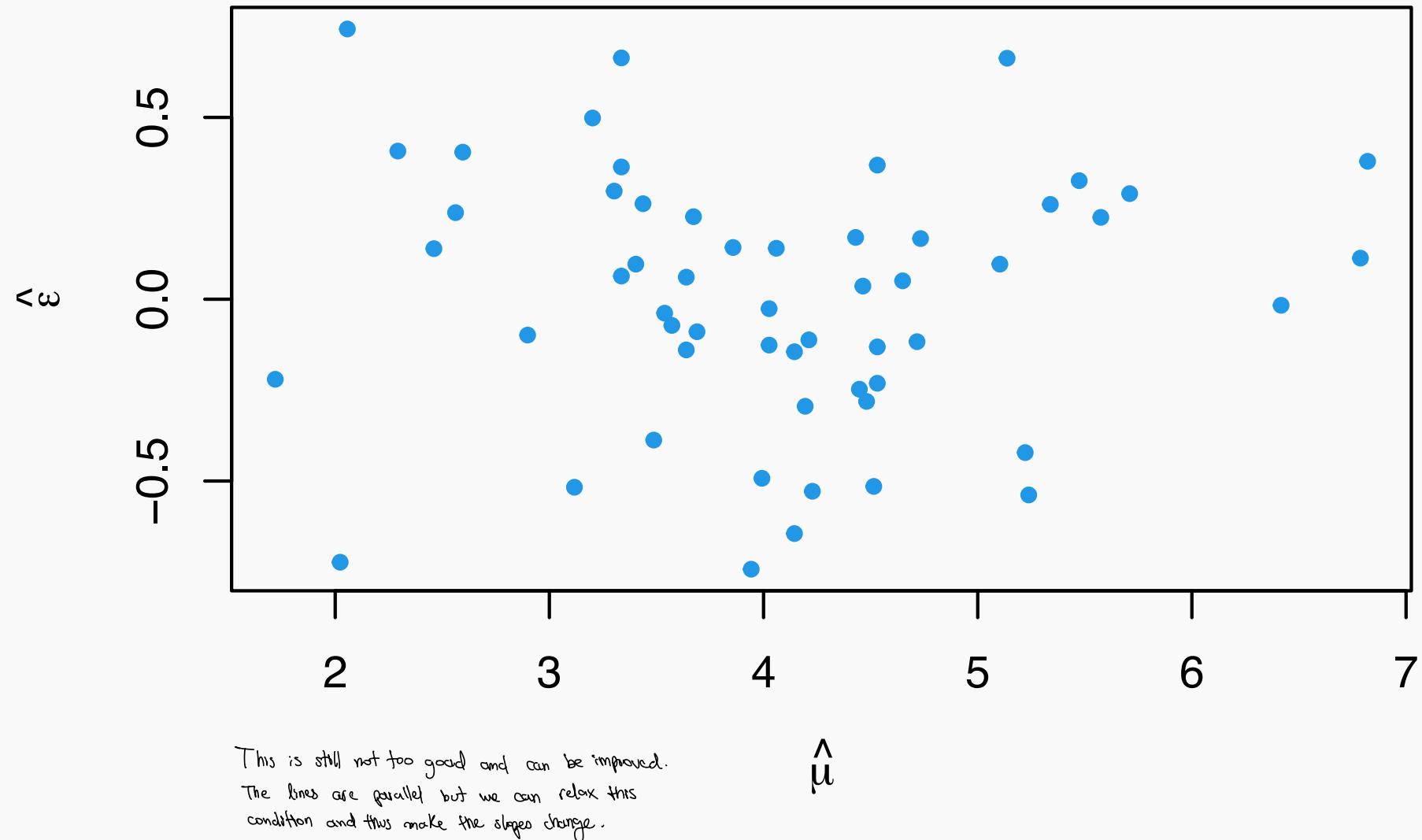
Model checking

- The least squares linear model is the estimated linear model which best fits the data at hand but... *best* is not (necessarily) *good*.
- How to establish if the estimated model is a good one?
 - ↪ Residuals are built as in the simple linear model and have the same interpretation (estimate of the error components ϵ_i)

$$\begin{aligned}\hat{\epsilon}_i &= y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\beta}_2 d_i) \\ &= y_i - \hat{\mu}_i \quad i = 1, \dots, n\end{aligned}$$

- ↪ Residual analysis allows to check if the model assumptions are met and if the model is good

Model checking - Residual plots



Model checking

- The “two clusters” problem is not present anymore
- The non-constant variability of the residuals is reduced
- The residual plot still suggests some problems: the residuals are still some function of the predictors (via $\hat{\mu}$)
 - ↪ Positive residuals for small and large $\hat{\mu}$
 - ↪ Negative residuals for intermediate values of $\hat{\mu}$
- Have we forgotten anything relevant?

The interaction term

- Going back to the scatterplot of the data...
 - ↪ the line corresponding to the observations before insulation tends to underestimate the gas consumption for small lower temperatures and overestimate it for higher temperatures
 - ↪ the line corresponding to the observations after insulation tends to overestimate the gas consumption for small lower temperatures and underestimate it for higher temperatures
- Data show that not only the intercepts but also slopes might be different before and after the insulation intervention
 - ↪ *interaction term*: different relationship between Y and X for different values of d

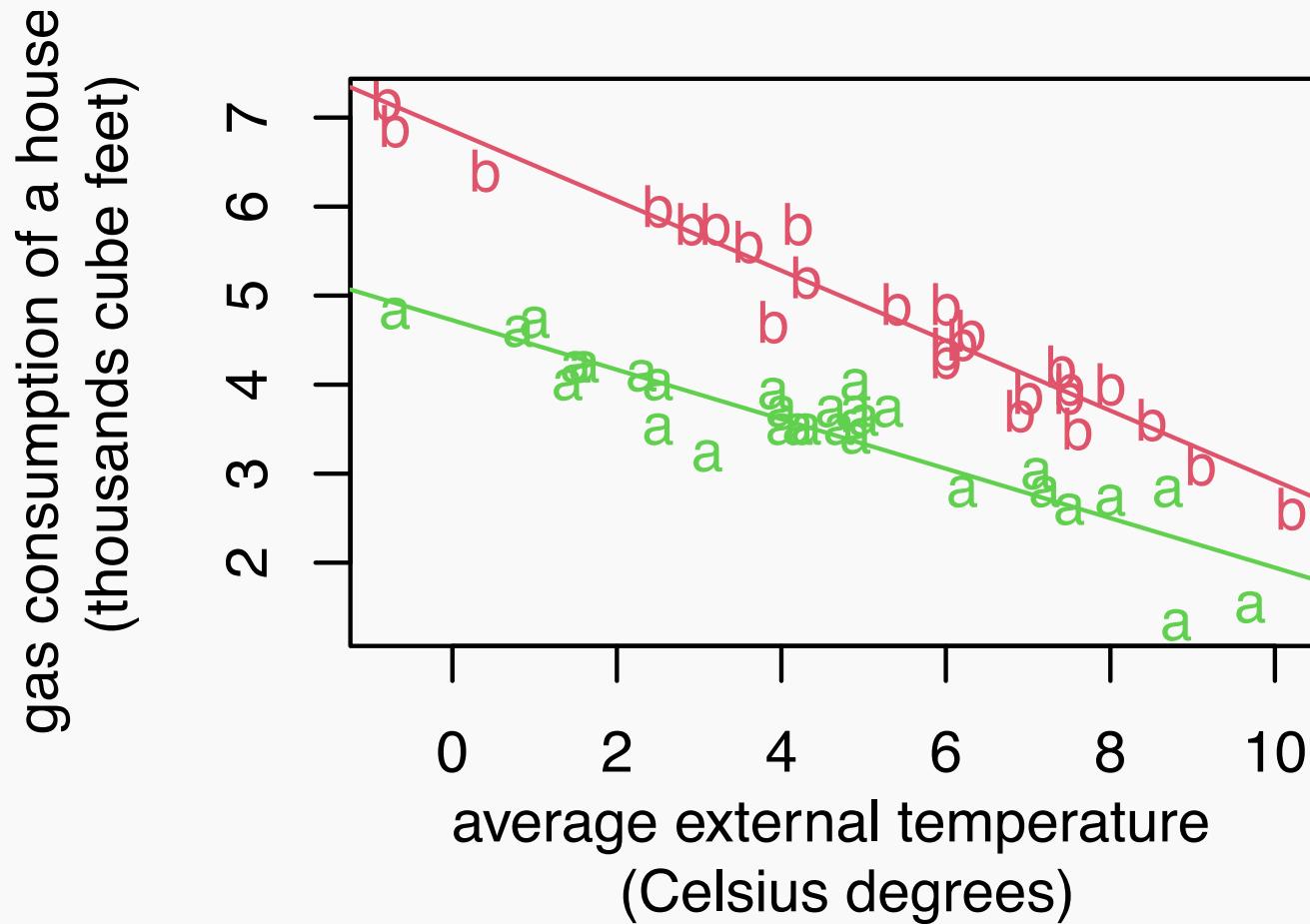
The interaction term: Formalization

- The linear model with interaction of the predictors may be formalized as follows:

$$\begin{aligned}y_i &= \beta_0 + \beta_1 x_i + \beta_2 d_i + \beta_3(x_i \cdot d_i) + \epsilon_i \\&= \beta_0 + \beta_1 x_i + \epsilon_i \quad \text{before the intervention, when } d_i = 0 \\&= (\beta_0 + \beta_2) + (\beta_1 + \beta_3)x_i + \epsilon_i \quad \text{after the intervention, when } d_i = 1\end{aligned}$$

- The model is estimated by the least squares criterion, as before

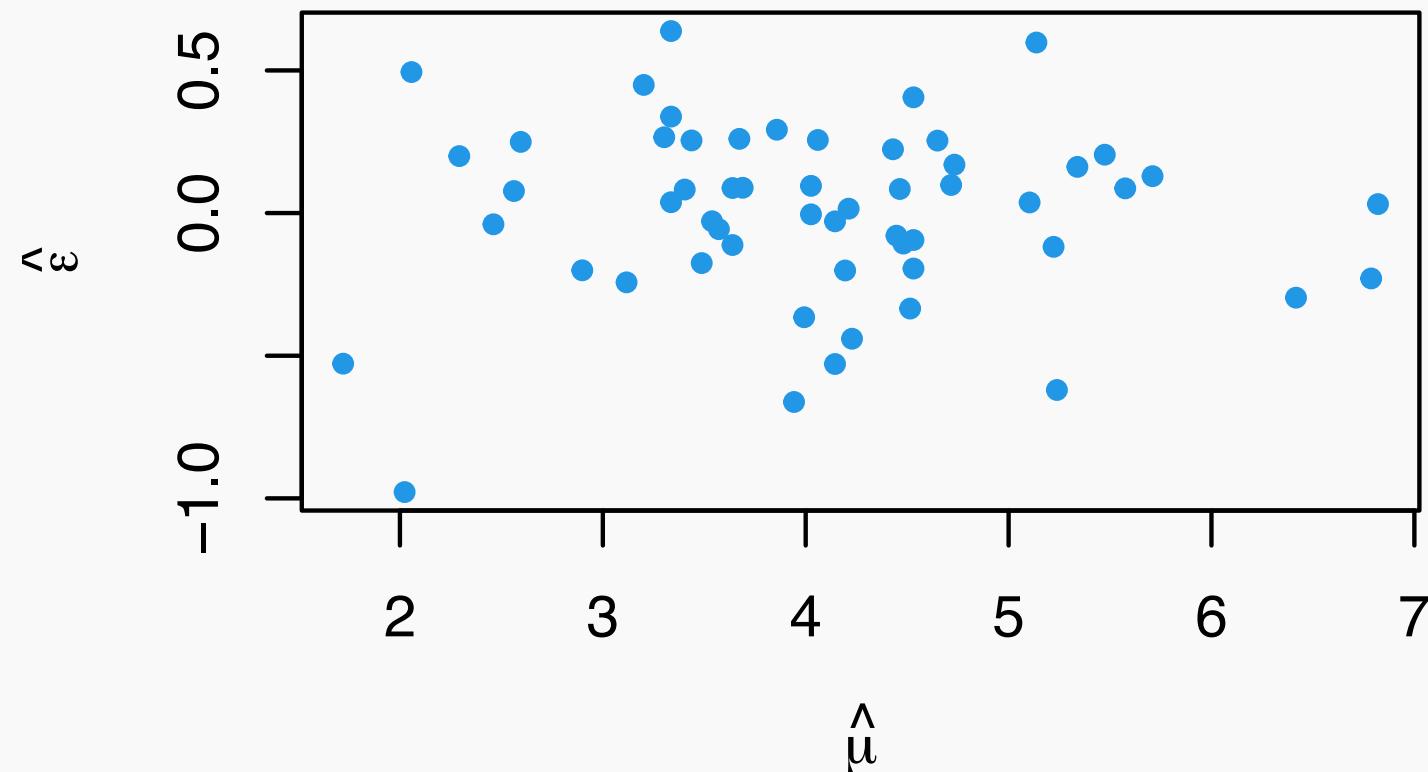
The estimated interaction model



The estimated interaction model

```
##  
## Call:  
## lm(formula = y ~ x * z)  
##  
## Residuals:  
##      Min       1Q   Median       3Q      Max  
## -0.97802 -0.18011  0.03757  0.20930  0.63803  
##  
## Coefficients:  
##                 Estimate Std. Error t value Pr(>|t|)  
## (Intercept)  6.85383   0.13596  50.409 < 2e-16 ***  
## x          -0.39324   0.02249 -17.487 < 2e-16 ***  
## zafter     -2.12998   0.18009 -11.827 2.32e-16 ***  
## x:zafter    0.11530   0.03211   3.591 0.000731 ***  
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##  
## Residual standard error: 0.323 on 52 degrees of freedom  
## Multiple R-squared:  0.9277, Adjusted R-squared:  0.9235  
## F-statistic: 222.3 on 3 and 52 DF,  p-value: < 2.2e-16
```

Model checking - Residual plots



Package lmviz: deviation from assumptions

Generalization to multiple regression model

Given a response variable Y and $p - 1$ predictors X_1, \dots, X_{p-1} , observed on a sample of n subjects, the multiple linear model is specified as follows:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ip-1} + \epsilon_i$$

- The model is correctly specified:

- systematic component:

- predictors are under the control of the researchers (non stochastic)
 - predictors are not collinear (i.e. not highly correlated)

- stochastic component:

- the error terms have zero mean
 - ↪ do not include further systematic terms
 - ↪ this implies that $E(Y_i) = \mu_i = \beta_0 + \beta_1 x_i + \dots + \beta_p x_{ip-1}$
 - the error terms have constant variance
 - the error terms are independent
 - (optional: the error terms are normally distributed)

Linear models

(Some basic results)

N. Torelli, G. Di Credico, V. Gioia

Fall 2023

University of Trieste

Matrix notation

Inference in Linear models

Model validation and model selection

Matrix notation

The linear model

- Linear models (LM) are appropriate when analyzing the relationship between a quantitative *response variable* Y and a set of *covariates* x_1, x_2, \dots, x_{p-1} .
It is assumed that a sample of n values of the response variable Y is observed as well as n values of each covariate.
- The aim is to evaluate the impact of covariates on the mean μ_i of the response variable Y_i for the i -th unit. In a linear model this is represented by the equation

$$E(Y_i) = \mu_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_{p-1} x_{ip-1} \quad (1)$$

The value y_i for the i -th unit of the sample can be written as

$$y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_{p-1} x_{ip-1} + \epsilon_i , \quad (2)$$

the model above can be also written for the set of all the n units in matrix notation

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

- $\mathbf{X}\boldsymbol{\beta}$ is the so called systematic component
- $\boldsymbol{\epsilon}$ is the stochastic component.

copy pse

Matrix notation

copy matrix

The equation for each unit in matrix notation is

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p-1} \\ \vdots & & & \vdots \\ 1 & x_{n1} & \cdots & x_{np-1} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_{p-1} \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix},$$

con:

$$(y_i) = (1 \ x_{ij}) (\beta_j) \quad j=1 \dots p$$

- ↪ \mathbf{y} = is the vector of the values of the response variable ($n \times 1$) ;
- ↪ \mathbf{X} = is a matrix ($n \times p$) which contains the values of the covariates.
- ↪ $\boldsymbol{\beta}$ = is the vector ($p \times 1$) of the regression coefficients;
- ↪ $\boldsymbol{\epsilon}$ = is the vector ($n \times 1$) of the stochastic components.

The model written for the i -th unit can be also written as $y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \epsilon_i$ where \mathbf{x}_i^T is the i -th row of the *so-called* design matrix.

Matrix notation: main assumptions

In the linear model

- The response variable Y is a quantitative variable
- The covariates X could be either:
 - ↪ quantitative (numeric) variables or
 - ↪ categorical variables (factors).
- it is usually assumed that the values in the matrix X are fixed constant (non stochastic). X is the **design matrix**
- the design matrix X is assumed to be of full rank. Since usually $n \gg p$ this means that the rank of X is p (that is $\min(p, n)$). The columns of X are linearly independent vectors.

Matrix notation: main assumptions

(copy)

The linear model is completely specified by assumptions on the stochastic components, the random variables ϵ_i

1. $E(\epsilon_i) = 0$ or equivalently $E(\epsilon) = 0$

2. $\text{Var}(\epsilon_i) = \sigma^2$ homoscedasticity

3. $E(\epsilon_i \epsilon_j) = 0$ per $i \neq j$ uncorrelation.

$$\text{Var}[\epsilon_i \epsilon_j] = E[\epsilon_i \epsilon_j]$$

\Rightarrow independence [only] for normal rv
 $\sigma = 0 \Rightarrow$ independence H rv.

The last two conditions can be more concisely expressed in matrix form as

$$\text{Cov}(\epsilon) = E(\epsilon \epsilon^T) = \sigma^2 I_n,$$

where $\text{Cov}(\epsilon)$ denotes variance-covariance matrix of the random vector ϵ .

The assumption 1-3 are called the second order assumptions (since they refer only to the first two moments of the variables).

A distributional assumption is then often added

4. $\epsilon \sim N_n(\mathbf{0}, \Sigma)$

In matrix notation

multivariate normal distribution

$$E(y) = X\beta \text{ and } y \sim N(X\beta, \sigma^2 I)$$

$\Sigma = \sigma^2 I$ reflects assumption over errors.

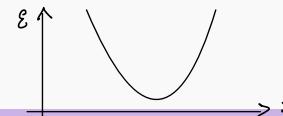
Discussion of the assumptions

copy
main
idea

- *Linearity.* The assumption about the linear effect of the covariates is actually not very restrictive. Non linear relationships can be introduced by appropriate transformations of the covariates. For instance:

- $y_i = \beta_0 + \beta_1 \log(z_i) + \epsilon_i$ introduces a logarithmic effect of z_i . But note that if one simple redefines $x_i = \log(z_i)$ then we are back to a standard linear model for the transformed variable X .
- x_i^2 introduces a parabolic effect.

can be introduced if
the errors - covariates
plot look parabolic



on F we use
 $\ln(y \sim I(x^{12}))$

It affects the
residuals plot.

- *Homoscedasticity of the random components.* This is the standard assumption.

The use of diagnostic checks can help verifying it. If possible departures from homoschedasticity are ignored it can impair quality of estimates. Possible remedies can be introduced

Variability for time series or space dependent.

- *Uncorrelated random components.*

It is assumed that all random components are mutually uncorrelated. In some context this assumption is questionable (this is the case of data that are temporally or spatially ordered). Also this assumption can be verified with diagnostic tools and remedies are available.

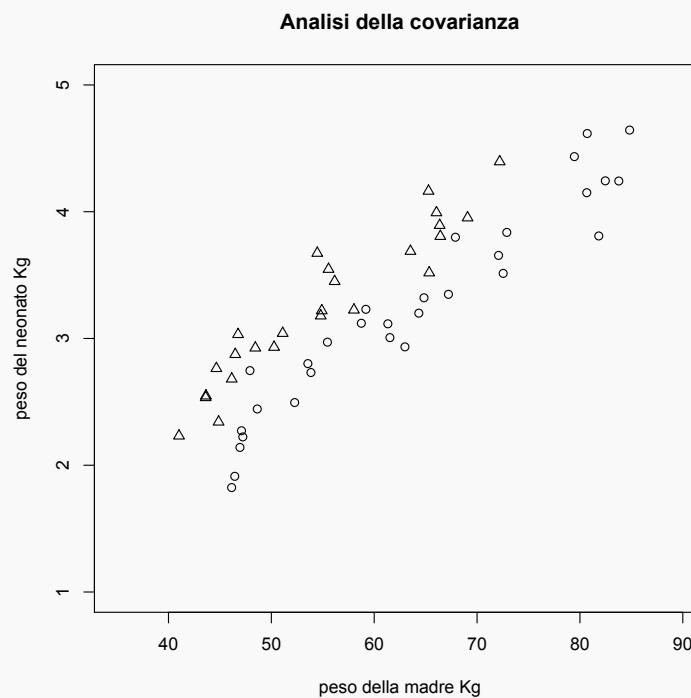
Continuous covariates, factors, interactions

copy idea

- When the covariate is a quantitative one, under the linearity assumption, then the value of the parameter associated to it represents simply the derivative of Y wrt to X .
 $x_i \text{ num}$
 $\beta_i = \frac{\partial y}{\partial x_i}$
- The effect of a categorical variable (a factor) measures the difference in the expected value of the response variable for each value of the factor wrt the reference category for the factor itself (all the other variables being equal). There will be then as many parameters as the number of levels of the factor minus one.
For categorical variables the interpretation is harder.
- Usually the interactions between two (or more) variables are introduced. Interpretation of interaction is easier when it refers to two factors or to a factor and a numeric variable.

Another example: one factor and a quantitative variable

Weight, Y , in kilograms, for a sample of newborn babies, from smoker mothers smokers (F) (in the graph different symbols are used for smokers - circles - and non smokers - triangles). For each women the pre-pregnancy weight mother weight X is observed)



As expected the weights of newborn babies is greater on the average when the mothers are non smokers. It seems that a systematic difference exists between the two groups though the relationship between weight of the mother and weight of the babies

A model with a continuos covariate and a factor

The two variables Y = “weight of the babies” and X = “weight of the mother” are both continuous numeric variables while the variable F is a factor with two levels ($F = 1$ if smoker, $F = 2$ if non smoker). The model is

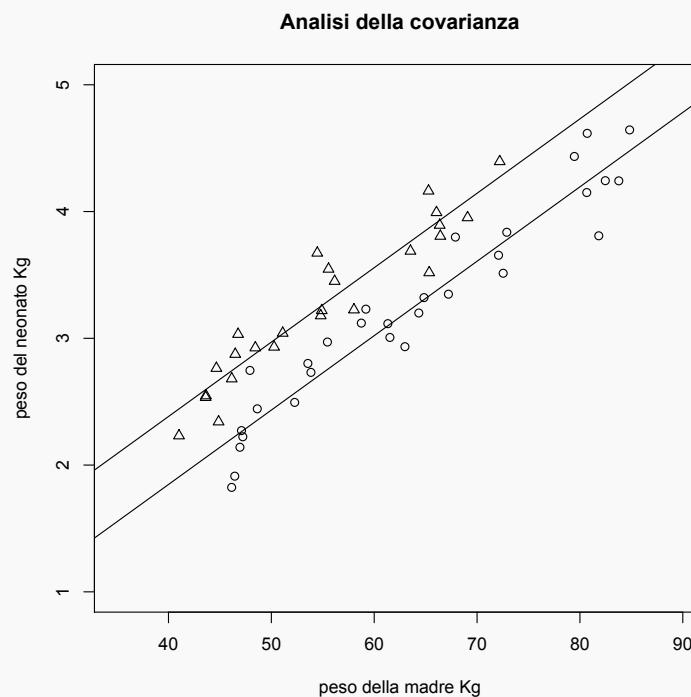
$$y_i = \beta_0 + \beta_1 x_i + \beta_2 I_{(F_i=2)} + \epsilon_i .$$

with the corresponding matrices:

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 & 0 \\ 1 & x_2 & 0 \\ \vdots & \vdots & \vdots \\ 1 & x_k & 1 \\ \vdots & \vdots & \vdots \\ 1 & x_n & 1 \end{pmatrix} \begin{matrix} \text{dummy/} \\ \text{binary} \end{matrix} \begin{matrix} \text{presence of characteristic.} \\ \text{d} = d(y, x, F) \end{matrix} + \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix} .$$

$\text{Im}(y \sim x + F, \text{data} = d)$

Interpretation of the parameters



For the given specification

- parameter β_1 measures the (linear) effect of the weight of the mother on the weight of the baby and represents the (common) slope of the two lines in the graph;
- β_0 measures the intercept of the smoker's line and β_2 is, for a given weight of the mother, the vertical distance between the two lines.

An alternative parametrization

- Any model, particularly when factors are involved, can have alternative parametrizations.

The model introduced above can be also written in the following form:

$$Y_i = \beta_1 x_i + \beta_2 I_{(F_i=1)} + \beta_3 I_{(F_i=2)} + \epsilon_i , \quad (3)$$

where $I_{(F_i=j)}$ is an indicator variable which takes on 1 if $(F_i = j)$, $j = 1, 2$, and 0 otherwise.

- The model is equivalent but interpretation of parameters changes.

In this case the matrix form is:

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} x_1 & 1 & 0 \\ x_2 & 1 & 0 \\ \vdots & & \vdots \\ x_n & 0 & 1 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix} .$$

Smokers non-smokers $\begin{pmatrix} 1 \\ 0 \\ \vdots \\ 1 \end{pmatrix}$ is not included because of linear dependence

If there are too many parameters the model can't explain the data.

$$\beta_2 = E[Y = \text{smoker}]$$

$$\beta_3 = E[Y = \text{non-smoker}]$$

Interpretation of the parameters

In this new parametrization

- parameter β_1 measures the effect of the weight of the mother on the weight of the baby
- β_2 and β_3 estimate the mean of the y , the weight of the babies, for a given weight of the mother, for smokers and non smokers respectively.
- The columns of the design matrix \mathbf{X} can be obtained from those of the previous specification by using a linear combination. The model is the same but interpretation of parameters changes.
- one should not add the intercept otherwise \mathbf{X} will become rank deficient

A model with interaction

The independent variables enter the model additively: the effect of the variable X is the same for each level of the factor F . In many case this is a too simplistic model, and the effect can change for different values of the factor.

This effect can be caught by introducing interaction between the covariates. The following regression model includes an interaction

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 I_{(F_i=2)} + \beta_3 I_{(F_i=2)} x_i + \epsilon_i .$$

Interaction implies the relationship between x (weight of the mother) and y (weight of the baby) can be different for smokers and non smokers.

$$\begin{aligned} E[y | F=1] &= \beta_0 + \beta_2 x \\ E[y | F=2] &= (\beta_0 + \beta_1) + (\beta_1 + \beta_3)x \end{aligned}$$

important to interpret model with interactions

Transforming categorical variables to numeric for fitting linear models can derive in the lack of parameters for instance β_2 and this makes the interpretation harder.
It is better to transform it to factors and in this case the interpretation is simpler.

Watch ~ 1:22

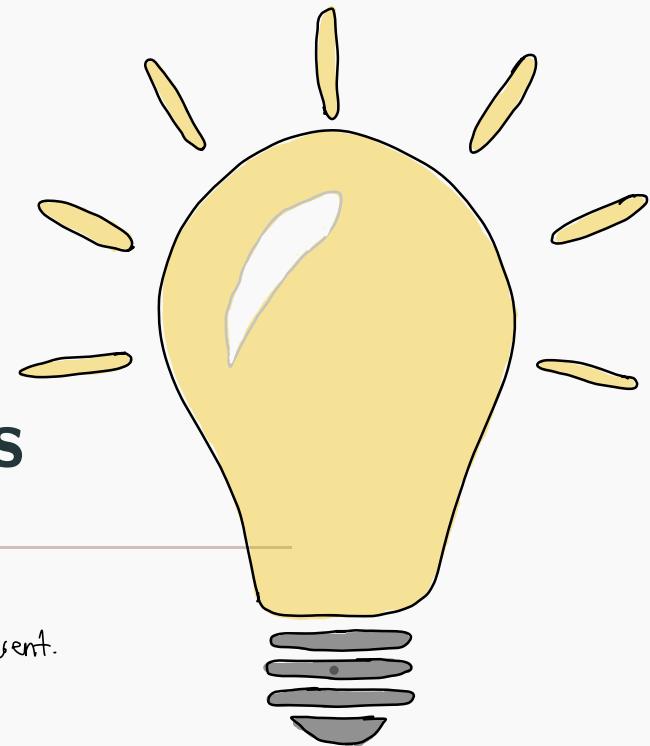
```
data <- read.csv("    ", stringsAsFactors = TRUE)  
lm(y ~ x + z + x * z) → lm(y ~ z * x)
```

Inference in Linear models

We try to analyze the dependency of y on x but causality is something different.

Interpretation of causality is not trivial.

Correlation doesn't imply causation.



Least square estimation

- A sample y_1, y_2, \dots, y_n along with the values of the vector \mathbf{x}_i for the covariates observed on i -th unit is available and the aim is to estimate the parameters of the models $(\beta_0, \beta_1, \dots, \beta_{p-1}, \sigma^2)$
- An estimator of the parameters vector β can be obtained by using the **least square method**:
 1. The least square estimator (LSE) $\hat{\beta}$ of β is the vector for which the following quantity is minimized

$$LS(\beta) = \sum_{i=1}^n (y_i - \mathbf{x}_i^T \beta)^2 = (\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta) ;$$

$$LS(\beta) = \mathbf{y}^T \mathbf{y} - \beta^T \mathbf{X}^T \mathbf{y} - \mathbf{y}^T \mathbf{X} \beta + \beta^T \mathbf{X}^T \mathbf{X} \beta$$

2. Taking the derivative $\frac{\partial LS(\beta)}{\partial \beta} = -2\mathbf{X}^T \mathbf{y} + 2\mathbf{X}^T \mathbf{X} \beta$ and then equating it to 0, the LSE is

$$\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} .$$

To invert $(\mathbf{X}^T \mathbf{X})$ we have to assume that this matrix is non singular.

This is always true if \mathbf{X} is a full rank matrix

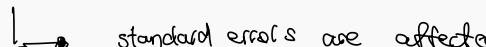
$$\Sigma \stackrel{?}{=} \text{diag}(\beta_1, \dots, \beta_n)$$

Properties of LS estimator

The LSE $\hat{\beta}$ has the following properties:

$$\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

$$\hat{e} = \text{SSE}/(n-p)$$

1. $E(\hat{\beta}) = \beta$ and $\text{var}(\hat{\beta}) = \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1}$; finite norm
2. asymptotically $\hat{\beta} \sim N_p(\beta, \sigma^2 V^{-1})$ where $V = \lim_{n \rightarrow \infty} \mathbf{X}_n^T \mathbf{X}_n$ and \mathbf{X}_n is the sequence of design matrices and V is positive definite; in most of the cases then for large n , $\hat{\beta} \sim N_p(\beta, \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1})$. This allows us to test significance of parameters and to build confidence intervals easily.
3. $\hat{\beta}$, the LSE is the *best estimator* in the sense that it has minimum variance among all linear estimators (BLUE Best Linear Unbiased Estimator - Gauss-Markov theorem).
4. When $(\mathbf{X}^T \mathbf{X})$ is not singular but its determinant is very close to 0 then estimates are very unstable. This happens if (multiple) correlation among the column of \mathbf{X} is very close to 1. Regularization could be a solution.  standard errors are affected

Obs: before fitting a model ^{some} unselected variables must be removed to avoid having redundancy.

ML estimation

- When normality is assumed for the random components and taking account of uncorrelation (which means also independence in the normal case) then β can be obtained by using **maximum likelihood estimation**.
- Choose the value β which maximizes

$$L(\beta, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2} (\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta)\right)$$

It is easy to show (by evaluating the log-likelihood and by deriving with respect to β) that the likelihood is maximized if the following quantity is minimized $(\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta)$.

- under normality assumption, MLE and LSE are equivalent**
- But to the properties already listed for the LSE now it can be added the one regarding the (exact) distribution of $\hat{\beta} \sim N(\beta, \sigma^2(\mathbf{X}^T \mathbf{X})^{-1})$

MLE is less general because assumes error normal.

but we know distribution of $\hat{\beta}$

Which one is more general?

What about in the LS?

Estimation of σ^2

- To estimate σ^2 the following estimator is usually considered

Why?

It seems that $\hat{\beta}$ and S^2 depend on each other by construction.

Their distribution are different for sure, but their definitions makes them depend on one another.

$$\hat{S}^2 = S^2 = \frac{SSE}{n - p} \quad \text{where} \quad \begin{array}{l} p \rightarrow \text{covariates and} \\ + \text{ intercept} \end{array}$$

$$\begin{aligned} SSE &= \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n e_i \\ &= (\mathbf{y} - \hat{\mathbf{y}})^T (\mathbf{y} - \hat{\mathbf{y}}) \\ &= (\mathbf{y} - \mathbf{X}\hat{\beta})^T (\mathbf{y} - \mathbf{X}\hat{\beta}) \end{aligned}$$

$$\hat{S}^2 = \frac{(\mathbf{y} - \mathbf{X}\hat{\beta})^T (\mathbf{y} - \mathbf{X}\hat{\beta})}{n - p}$$

is the *Sum of Squares of residuals* e_i . / deviance of the errors.

- Note also that, when normality holds

↳ solvable only if $\mathbf{X}^T\mathbf{X}$ is full rank $\rightarrow p$
each column / variable must be
independent

$$\hat{\beta} \sim N_p(\beta, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}) \quad \text{and} \quad \frac{(n - p)S^2}{\sigma^2} \sim \chi^2_{n-p},$$

$P \times P$ matrix

and the two estimator $\hat{\beta}$ and S^2 are independent.

- Note that for small samples if σ^2 is unknown and S^2 is used, tests and confidence intervals for a single β_j are based on student t distribution with $n - p$ df.

P. Correlated variables \rightarrow no problem \rightarrow no solution

Almost. \rightarrow problem $\nabla[\hat{\beta}] = (X^T X)^{-1} \sigma^2 \Rightarrow \nabla$
because $\det(X^T X) \rightarrow 0$

$\Rightarrow \hat{\beta} = (X^T X)^{-1} X^T Y$ is not a good estimator \rightarrow unbiased and high variance estimator.

\Rightarrow avoid correlated variables for the model.
but it should be done carefully because of
colinearity. \leftarrow reverse

colinearity is expressed through correlation of vars

Model validation and model selection

Testing a general linear hypothesis

The tests of more common interest are:

- test of significance for a single element of β

$$H_0 : \beta_j = 0 \quad \text{against} \quad H_1 : \beta_j \neq 0$$

- test on a subvector $\beta_1 = (\beta_1, \dots, \beta_r)$

$$H_0 : \beta_1 = 0 \quad \text{against} \quad H_1 : \beta_1 \neq 0$$

- test of equality of two coefficients $H_0 : \beta_j - \beta_r = 0 \quad \text{against}$

$$H_1 : \beta_j - \beta_r \neq 0 \quad \text{like categories through factors}$$

$$C = (1, 0, \dots, 0)$$

$$C = \begin{pmatrix} 1 & & & \\ & \ddots & & 0 \\ & & \ddots & \\ 0 & & & \ddots & 0 \end{pmatrix}$$

$$C = (1, -1, 0, \dots, 0)$$

All these hypotheses are special cases of the *general linear hypothesis*

$$H_0 : \mathbf{C}\beta = \mathbf{d} \quad \text{against} \quad H_1 : \mathbf{C}\beta \neq \mathbf{d}$$

\mathbf{C} is a $r \times p$ matrix with rank = $r \leq p$ and d is a $r \times 1$ vector.

If data are fit to the model under the restriction $\mathbf{C}\beta = \mathbf{d}$ the residuals of this model are $H_0 e_i$ and one can compute $SSE_{H_0} = \sum_i^n H_0 e_i^2$ and calculate the statistic

$$\frac{n-p}{r} \frac{SSE_{H_0} - SSE}{SSE} \quad \underbrace{\frac{(SSE_{H_0} - SSE)/r}{SSE/(n-p)} \sim \chi^2_{r, n-p}}_{\sim \chi^2_{n-p}} \rightarrow \chi^2_{r, n-p}$$

that, when H_0 is true, under normality assumption is a $F_{r, n-p}$ random variable.

The model w/ larger params

$$SSE < SSE_{H_0}$$

Test significance of a single coefficient

When testing significance for a single element of β

$$H_0 : \beta_j = 0 \quad \text{against} \quad H_1 : \beta_j \neq 0$$

applying general linear hypothesis (where C is a row vector of zeros with one only in position j and $d = 0$), it can be shown that

$$\frac{\hat{\beta}_j^2}{\widehat{\text{Var}}(\hat{\beta}_j)} \sim F_{1,n-p}$$

This is the square of a student t with $n - p$ df and equivalently the following test statistic is usually considered

Obs: $t = \sqrt{F}$

$$t_j = \frac{\hat{\beta}_j}{\sqrt{\widehat{\text{Var}}(\hat{\beta}_j)}}^{1/2}$$

ONLY WHEN $F_{1,n-p}$

This result can be also used to obtain for β_j the following confidence interval at level $1 - \alpha$:

$$\hat{\beta}_j \pm t_{n-p, 1-\alpha/2} (\widehat{\text{Var}}(\hat{\beta}_j))^{1/2}$$

Decomposition of Sum of Squares

- The following holds:

$$\sum (y_i - \hat{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum (y_i - \hat{y}_i)^2$$

$$SST = SSR + SSE, \quad (4)$$

regression + residuals

the Total Sum of Squares (*total deviance*) is the sum of the Regression Sum of Squares (*deviance explained by the model*) and the Residual Sum of squares (*deviance of the residuals*). Analyzing the components of (4) is of great relevance, the ratio between *SSR* and *SST* is clearly related to the quality of the model.

- Let \mathcal{F}_1 be the minimal model (the one which contains only the intercept, $p = 1$).

Let \mathcal{F}_p be the current model with p parameters and let \mathcal{F}_{p_o} be a reduced model with $1 < p_o < p$ nested in \mathcal{F}_p .

Then the variance explained by the current model \mathcal{F}_p can be partitioned as it is shown in the table that follows (Table 1), called Analysis of variance table.

The analysis of variance

Table 1: Analysis of Variance (Anova)

Source of variability	df	SS	testing models improvement
total	n	SST	
constant	1	$n\bar{Y}^2$	
total	$n - 1$	SST_{cor}	
improvement with \mathcal{F}_{p_o} with respect to \mathcal{F}_1	$p_o - 1$	SSR_{p_o}	$\frac{SSR_{p_o}/(p_o-1)}{SSE_{p_o}/(n-p_o)} \sim F_{p_o-1, n-p_o}$
improvement with \mathcal{F}_p with respect to \mathcal{F}_{p_o}	$p - p_o$	$SSR_p - SSR_{p_o}$	$\frac{(SSE_{p_o} - SSE_p)/(p - p_o)}{SSE_p/(n - p)} \sim F_{p-p_o, n-p}$
residuals $\mathcal{F}_{\tilde{p}}$	$n - p$	SSE_p	

- The fall in the fit from \mathcal{F}_{p_o} to \mathcal{F}_p can be evaluated using the statistic

$$F = \frac{(SSE_{p_o} - SSE_p)/(p - p_o)}{SSE_p/(n - p)} \sim F_{p-p_o, n-p}.$$

SSE : sum of squared residuals

tests if the $p-p_o$ parameters in H_p but not H_{p_o} are all 0.

$$\begin{cases} H_0 : H_{p_o} \\ H_1 : H_p \end{cases}$$

Coefficient of determination R^2

The coefficient of determination R^2 is defined as the proportion of total variance explained by the regression model.

It can be used as a goodness-of-fit measure for the models

$$R^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = 1 - \frac{SSE}{SST}$$

and $0 \leq R^2 \leq 1$.

This decomposition is possible if the model includes the intercept.

For nested models R^2 always increases adding covariates.

When comparing nested models the *corrected coefficient of determination \bar{R}^2* is instead used

$$\bar{R}^2 = 1 - \frac{n-1}{n-p}(1-R^2)$$

It penalizes inclusion of new variables that are non significant.

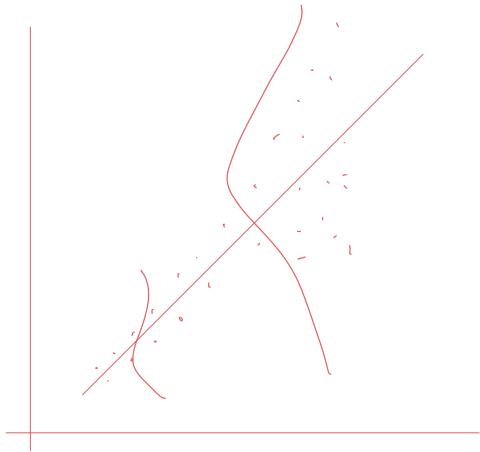
Deterministic value , the residuals must be canceled.

Residuals

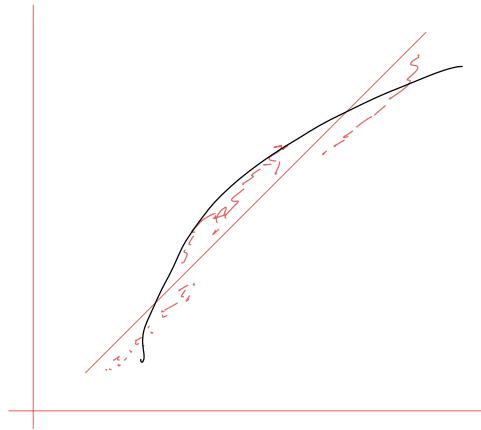
$$\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

- The mean of \mathbf{y} can be predicted once the model is estimated by $\hat{\mu}_i = \hat{\mathbf{y}} = \mathbf{X}\hat{\beta}$ and consequently $\hat{\mathbf{y}} = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} = H\mathbf{y}$
- $H = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$ is a square matrix of size n and it is called the *hat matrix* or the *projection matrix*. It has the following properties:
 - it is symmetric and idempotent $H^T = H$ and $H^2 = H$
 - $\text{rank}(H) = \text{trace}(H) = p = \sum h_{ii}$ if $h_{ii} > \frac{2}{n} \rightarrow \text{leverage}$
 - h_{ii} have values that range from $1/n$ and 1 and their sum is equal to p
 - the matrix $I - H$ is also symmetrical and idempotent with rank equal to $(n - p)$
- The residuals of the model are then $e_i = (I - H)\mathbf{y}$
- under normality assumption $e_i \sim N(\mathbf{0}, \sigma^2(I - H))$

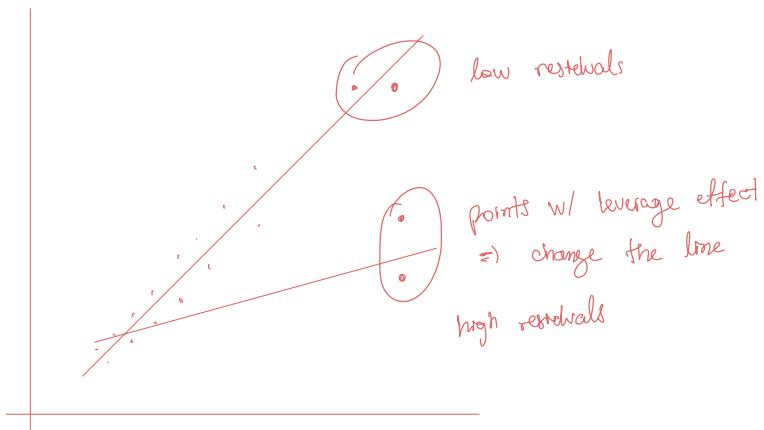
h_{ii} determines the influence of variable i on determining \hat{y}_i



non-homoscedasticity



$\sum e_i \rightarrow 0$ but relationship
might not be linear.



non-homoscedasticity

Analysis of the residuals

- The quality of the model and the validity of the assumptions can be judged by using some diagnostic tools that mainly rely upon analysis of residuals as defined above.
- In the linear models the residuals can be standardized (to take into account the fact that they have unequal variance)

$$r_i = \frac{Y_i - \hat{\mu}_i}{\sqrt{S^2(1 - h_{ii})}}, \quad (5)$$

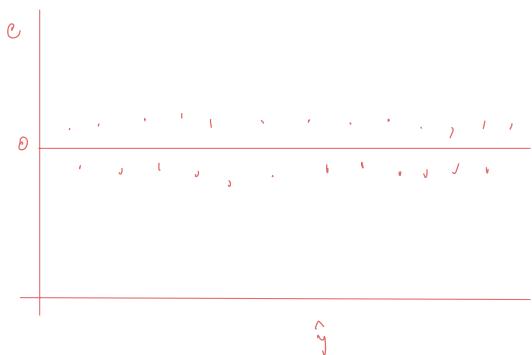
where h_{ii} is the i -th element on the diagonal of $H = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$. These values are called *leverages*. They reveal if a point has values that are far from the majority of data in the space of the xs. Suspect values have leverage $> 2p/n$

- to identify which values are outliers with respect the majority of the data points the studentized residual are introduced

$$e_i^* = \frac{e_i}{S_{(i)} \sqrt{1 - h_{ii}}}$$

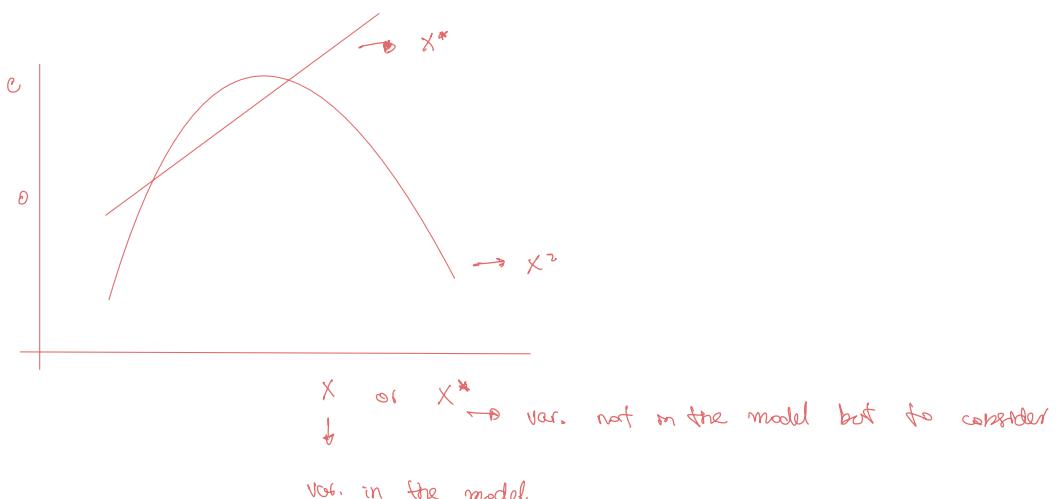
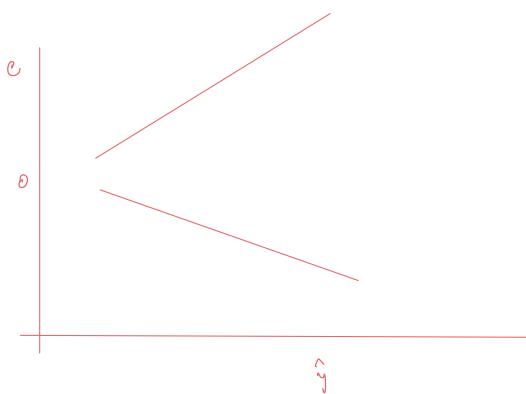
where $S_{(i)}^2$ is the variance of the residuals when the i -th observation is excluded. For a model correctly specified e_i^* follows a t with $n - p - 1$ df.

Residuals



$$e_i = y_i - \hat{y}_i$$

$$\sum e_i = 0 \quad \text{by definition}$$



Analysis of the residuals

- Cook distances are defined as

$$D_i = \frac{1}{p} r_i^2 \frac{h_{ii}}{1 - h_{ii}} = \frac{1}{pS^2} \sum_{j=1}^n (\hat{y}_j - \hat{y}_{j(i)})^2$$

residuals + leverage

- $\hat{y}_{j(i)}$ are the predicted values for the i -th observation in a model estimated without it. It reveals (when $D_i > 1$) observations that, when excluded from the analysis, will cause substantial modifications in the estimates of the parameters.
- Classical graphical tools based on (standardized) residuals are:
 - plot of residuals against the predicted values (*to reveal possible heteroscedasticity*)
 - plot of residuals against the explanatory variables (*to reveal non linearities*)
 - plot of residuals against variables not in the model (*added variable plot*)
 - Q-Q norm of residuals (*to assess normality*)
 - plot of leverages h_{ii} and of Cook distances (*to reveal outliers*)
- formal test of normality, such as Shapiro-Wilks test or Jarque-Bera test (the latter based on estimated values of third and fourth standardized moments)

Dealing with non constant variance and residual correlation

- Residual analysis can reveal that some assumptions could be questionable
- Critical assumptions are those on uncorrelation and heteroscedasticity of residuals
- The assumptions $\text{Cov}(\epsilon) = \sigma^2 I$ should be replaced by a more general $\text{Cov}(\epsilon) = \sigma^2 W^{-1}$ where W is assumed to be simply a positive definite matrix
- In this case LSE is still an unbiased estimator but the estimate of its variance covariance matrix is biased.
- If we ignore this, the main consequences are that tests or confidence intervals based on assumption of uncorrelation and homoscedasticity lead to wrong conclusions (tests tends to say a parameter is significant too often and confidence intervals appear shorter)

$$\check{\beta} = X^{-1}Y$$

$E[\check{\beta}] = \beta$ but $\text{Var}[\check{\beta}] = (X^T X)^{-1} \sigma^2$ under estimate
the variance \rightarrow change

Heteroscedasticity

- We will consider here only the case of heteroscedasticity. In this case $\text{Cov}(\epsilon) = \sigma^2 W^{-1}$ with $W^{-1} = \text{diag}(1/w_1, 1/w_2, \dots, 1/w_n)$ more weights for more reliable variables
- If each ϵ_i is multiplied by $\sqrt{w_i}$ one obtains the transformed values $\epsilon_i^* = \sqrt{w_i}\epsilon_i$ which have constant variance
- $\text{var}(\epsilon_i^*) = \text{var}(\sqrt{w_i}\epsilon_i) = \sigma^2$ and then the random components are homoscedastic.
- The model does not change if we transform also the response variable and the covariates (including the intercept) accordingly.
- We then obtain

$$y_i^* = \sqrt{w_i}y_i \text{ and}$$

$$x_{ij}^* = \sqrt{w_i}x_{ij}$$

for each of the p covariates (including the intercept) and then the model

$$y_i^* = \beta_0 x_{i0}^* + \beta_1 x_{i1}^* + \beta_2 x_{i2}^* + \cdots + \beta_{p-1} x_{i,p-1}^* + \epsilon_i^*$$

is homoscedastic and the same assumptions of a standard LM hold. These transformations, in matrix notation, are equivalent to pre-multiply all components of the model $\mathbf{y} = \mathbf{X}\beta + \epsilon$ by the matrix $W^{1/2}$

Weighted Least Squares (WLS)

- If then

$$\mathbf{y}^* = W^{1/2} \mathbf{y}, \mathbf{X}^* = W^{1/2} \mathbf{X} \text{ and } \epsilon^* = W^{1/2} \epsilon$$

we get a new model for transformed data where homoscedasticity holds and parameters can again be estimated by LSE obtaining

$$\begin{aligned}\hat{\beta} &= (\mathbf{X}^{*T} \mathbf{X}^*)^{-1} \mathbf{X}^{*T} \mathbf{y}^* \\ &= (\mathbf{X}^T W^{1/2} W^{1/2} \mathbf{X})^{-1} \mathbf{X}^T W^{1/2} W^{1/2} \mathbf{y} \\ &= (\mathbf{X}^T W \mathbf{X})^{-1} \mathbf{X}^T W \mathbf{y}\end{aligned}$$

- This estimator is the Weighted Least Square Estimator (WLS)
- Note that weights are inversely proportional to variances of ϵ (which are originally heteroscedastic)
- To units with a more erratic random component are given smaller weights
- Application of this strategy requires that the weights are known

□ More complex models introduce heteroscedasticity by default.

Model choice and variable selection

In many applications a large number of candidate predictors are available.

A **naive approach** often used is the following:

Estimate the most complex model that includes all the covariates (and possibly all the interactions). Then, remove all not significant variables from the model. (Backward selection)

This strategy is not advisable for many reasons. Let us list some of them:

- the resulting model can overfit the data and then its predictive performance for new data can decrease
- the larger the number of covariates the higher the risk of **multi-collinearity** (**correlated regressors**)
- There are many models with equivalent performances but different substantial interpretation. You are not sure that the variable which remains in your model after such backward selection strategy should be really considered the most relevant.

Other **naive** (yet often used) strategies for variable selection are:

- All subset selection (chose the best among $\sum_{j=1}^p \binom{p}{j}$ possible models)
- Forward selection
- Stepwise selection (a combination of forward and backward selection)

Model choice criteria

Since one of the principles to consider when building a model is the *Occam's razor*, criteria to select a model that has good performances and at the same time is less complex should be introduced.

When considering alternative LM we have already seen some criteria

- R^2 and corrected R^2
- F test (for nested models)
- Mallows's C_p

$$C_p = \frac{\sum_i^n (y_i - \hat{y}_{iM})^2}{\hat{\sigma}^2} - n + M$$

where M is the number of covariates in the model and \hat{y}_{iM} are the predicted values with those M covariates. The “best” model is the one with lowest C_p .

- Akaike Information Criteria (AIC)

$$AIC = -2I(\hat{\beta}_M, \hat{\sigma}^2) + 2(M + 1)$$

Better fit corresponds to smaller AIC values.

For a linear model with gaussian components and $p \beta_j$ parameters

$$AIC = n \log(\hat{\sigma}^2) + 2(p + 1).$$

Note that $\hat{\sigma}^2$ is SSE divided by n .

- Bayesian Information Criteria (BIC)

$$BIC = -2I(\hat{\beta}_M, \hat{\sigma}^2) + \log(n)(M + 1)$$

Avoiding collinearity

A diagnosis of collinearity is obtained by computing the *variance inflation factor* (VIF) associated to the j -th predictor

$$VIF_j = \frac{1}{1 - R_j^2}$$

Test if x_j can be expressed as a function of the others.
 $\uparrow R_j$.
dependent variable

where R_j^2 is the coefficient of determination when x_j is regressed on all the remaining covariates. $VIF_j > 10$ is usually taken as a symptom that the variable can cause collinearity.

Typical solutions are:

- omission of covariates
- using principal components extracted from regressors (or other combination of the regressors)
- Ridge regression: it is an alternative to LSE where

$$\hat{\beta}_{\text{Ridge}} = (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^T \mathbf{y}$$

by construction linear independent
but difficult to interpret

and λ is a chosen tuning parameter

biased estimator but
low variance
→ bias-variance trade-off
solvable using Linear Algebra

Regularization Techniques

It has been assumed so far that \mathbf{X} has full rank, and this gives a unique solution to equations

$$(\mathbf{X}^T \mathbf{X})\beta = \mathbf{X}^T \mathbf{y}$$

We have already discussed that $\mathbf{X}^T \mathbf{X}$ could be close to singularity. Regularization consists in changing the objective function by penalizing it:

$$\hat{\beta}_{PLS} = \arg \min_{\beta} [(\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta) + \lambda \text{pen}(\beta)]$$

penalizes on β for large β_i
w/ λ
 $\uparrow \rightarrow$ the more β_i are shinked

where $\text{pen}(\beta)$ is a term that measure the complexity of the model and $\lambda \geq 0$ is a smoothing parameter that reflects the weight given to the penalty. Penalty is such that it is large when many β are large.

Ridge regression is an example of penalized least square. It corresponds to introducing the following penalty

$$\text{pen}(\beta) = \sum_{j=1}^p \beta_j^2 = \beta^T \beta$$

With large λ the penalty term dominates and all (or almost all) coefficients are shrunk to 0. λ is usually chosen by k -fold cross validation.

LASSO (Least Absolute Shrinkage and Selection Operator)

Also LASSO corresponds to a penalized least square criterion and

$$\hat{\beta}_{LASSO} = \arg \min_{\beta} [(\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta) + \lambda \sum_{j=1}^p |\beta_j|]$$

shrinkage is larger

not solvable w/ Linear Algebra
by definition

- The penalization chosen with LASSO tend to shrink some of the values of the coefficients to 0. Small coefficients are more strongly shrunk to 0 compared with Ridge regression.
- Balance between fit of the data and regularization
- Note that no closed explicit solution of the minimization problem exists. Numerical optimization must be used (quadratic programming, LARS–Least Angle Regression).

$$Y = X\beta + \varepsilon$$

$$y_i = x_i^T \beta + \varepsilon_i'$$

$$y_i = [1 \ x_{1i} \ x_{2i} \ \dots \ x_{pi}] \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \varepsilon_i$$

Assumptions

- ① linear combination of x_i
- ② ε_i rv w/ $E[\varepsilon_i] = 0$, $Cov[\varepsilon_i, \varepsilon_j] = 0$
 $\Rightarrow E[\varepsilon_i] = 0 \Rightarrow \varepsilon_i \sim N(0, \sigma^2)$

Estimates

$$\hat{\beta}_{OLS} = (X^T X)^{-1} X^T y$$

Inferences

- ① Testing β_s
- ② Model testing and selection

Problems w/ linear models

- criticism w/ diagnostics like residuals
e.g. heteroscedasticity

$$\hat{\beta}_W = (X^T W X)^{-1} X^T y \rightarrow \text{Weighted Least Squares}$$

$$W = \text{diag}(1/w_1, \dots, 1/w_n)$$

$$w_i \sim \frac{1}{\text{Var}[X_i]}$$

$\uparrow \text{Var}[\cdot]$ & importance

- $\det(X^T X) \rightarrow 0 \Rightarrow (X^T X)^{-1}$ has large values

$$\text{Var}[\hat{\beta}] = \begin{bmatrix} 0 & & & \\ & \ddots & & \\ & & S & \\ & & & \text{Var}[\hat{\beta}_i] \end{bmatrix}$$

Inference

$$\frac{\hat{\beta}_i}{\sqrt{\text{Var}[\hat{\beta}_i]}}$$

$$\left. \begin{array}{ll} H_0: \beta_i = 0 & \rightarrow \text{not important to fit model} \\ H_0: \beta_i \neq 0 & \end{array} \right\}$$

\uparrow when true but also when $\det(X^T X) \rightarrow 0$

- Multicollinearity

Useful to watch to the correlation matrix of the inputs

$$\rho_{ppq} = \begin{bmatrix} 1 & & & & \\ & 1 & * & & \\ & & 1 & & \\ & & & 1 & \\ & & & & 1 \end{bmatrix}^j$$

o \uparrow \leftrightarrow correlation

o \uparrow $\perp \nparallel$ correlation $x_i \approx w_1 x_j + w_2 x_k$
 must be avoided.

Logistic regression

Dichotomous response

N. Torelli, G. Di Credico, V. Gioia

2023

University of Trieste

Introduction

Regression for dichotomous response: Logistic regression

Parameters interpretation

Inference for logistic regression parameters

Alternative specification of the response function

Estimation issues

Introduction

GLM: introduction and basic ideas

- GLMs allow to extend classical normal linear models in many directions:
 - response variables can be assumed non-normal (*including discrete distributions or distributions with support $[0, \infty)$*);
 - The mean and the variance of the response are assumed to vary according to values of observed covariates
 - The impact of covariates on the mean of the response is specified according to a (possibly) *non-linear* function of a linear combination of the covariates
- Main advantages are:
 - Unification of seemingly different models: it makes easy to use, understand and teach the techniques. Many of the standard ways of thinking LM carry over to GLMs;
 - Normal LMs, probit and logit models, log-linear models for contingency tables, Poisson regression, some survival analysis models are GLMs;
 - A single general theory and a single general computational algorithm can be developed for inference.

Regression for dichotomous response: Logistic regression

Dichotomous response: Some examples

- In many cases the variable of interest is not a quantitative (numeric) variable.
- The simplest, yet interesting, case is the one where the response variable is dichotomous. Very often we observe for a sample of units whether an event occurred or not. Of interest in financial or actuarial applications could be:
 - whether a person prefer to use an electric vehicle
 - whether a person purchases an item
 - whether a person decides to change Adsl provider
 - whether a individual decides to retire o to continue to work in a given year
 - whether a firm becomes insolvent
 - whether a individual has a defined disease

Binary dependent variable

- As in the case of quantitative response variables we are interested in building a statistical model that allows us to predict whether a specific event occurs (or if a unit belongs to one class).
- Exactly like in standard linear regression model we aim to explain (predict) a dependent variable y_i by using observed characteristics of the i -th unit such as their age, sex, education, income, etc..
- A dichotomous dependent variable y_i can in general take on two values denoted by 0 or 1. Generally it is assumed that the variables take on the value 1 if an event of interest occurred.
- For instance if the response variable reports whether a unit decided or not to buy a new car, we could put for the i -th unit

$y_i = 0$ if the car have not been purchased

$y_i = 1$ if the car have been purchased

Bernoulli variables

- Variables like the one introduced above are characterized by a Bernoulli probability distribution

Y	$Pr(Y = y)$
0	$1 - p$
1	p

"First Ingredient": distribution of data

$$Y_i \sim \text{Be}(p_i)$$

$$\begin{aligned} E[Y_i] &= p_i = \mu_i \\ V[Y_i] &= p_i(1-p_i) \end{aligned} \quad \xrightarrow{\text{mean-Variance dependency}}$$

p_i depends on x

"Second ingredient": linear combination / prediction

$$\begin{aligned} \eta_i &= \beta_0 + \beta_1 x_i \\ \text{"Third ingredient": mean-covariates relationship} \\ \mu_i &= r(\eta_i) \end{aligned}$$

link between predictor and mean

$$Y \sim \text{Bi}(1, p) \quad \leftarrow \text{dichotomous var.}$$

$$Y \sim \text{Be}(p), \quad Y = \{0, 1\}$$

$$P(Y, p) = P^y (1-p)^{1-y}$$

$$E[Y] = p, \quad \text{Var}[Y] = p(1-p)$$

for linear normal model

$$\eta_i = E[\mu_i] = \mu_i = \mu_i(x)$$

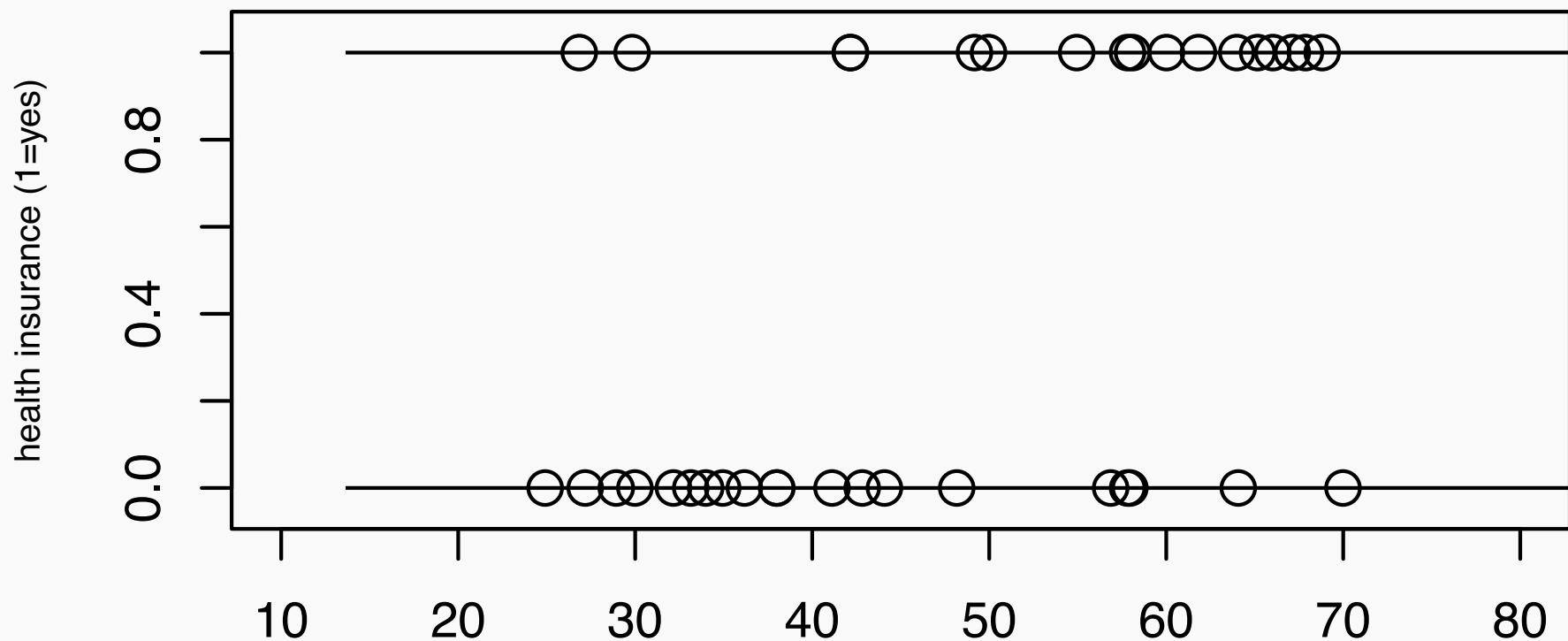
for the $\text{Be}(p)$ we have that $\mu_i = p_i \in [0, 1]$
but η_i can take values outside $[0, 1]$

$$\Rightarrow p_i = r(\eta_i) \in [0, 1]$$

for suitable r .

- p is a probability and then varies between 0 and 1.
- We expect that the probability p that a given event occurs varies according to the values of some covariates x_i .
- p is also the mean of the variable Y and so we are trying to understand if (and possibly how) the mean of the response variable varies as a function of a set of covariates.

A first example: Health Insurance coverage



For a sample of 37 individuals we observe the age of any sample unit and whether he/she owns a private health insurance.

It seems that older units are more likely to own a health insurance. For these data response variable Y can be assumed Bernoulli

1. $Y_i \sim \text{Bernoulli}(h(x_i))$.
2. and a possibly non linear model can be specified for $h(\cdot) \rightarrow [0, 1]$.

Logistic regression: Choosing an appropriate curve

- Just like in the case of simple linear regression, our model aims to represent the mean μ_i of the dependent variable Y_i as a function of a covariate x_i ;
- In this case since the Y_i s are drawn from a Bernoulli (or more generally Binomial) random variables, its mean is a probability.
- As we have seen an appropriate curve is not a straight line (in fact, curves that are S shaped seem more appropriate).
- There are many curves (functions) that could be considered. A possible function is the following

$$\begin{aligned} \lim_{z \rightarrow -\infty} r(z) &= 0 \\ \lim_{z \rightarrow \infty} r(z) &= 1 \end{aligned} \quad r(z) = \frac{e^z}{1 + e^z} \quad \Leftrightarrow \eta = z = \log\left(\frac{\rho}{1-\rho}\right) = r^{-1}(\rho) = g(\rho)$$

logit function
↓ odds

- This function, called the **response function**, is monotone increasing in z , exhibits an S shaped behaviour and takes on values in the interval $[0, 1]$
- Moreover, if we have a single covariate x_i we can assume that this covariate enters the function linearly, i.e.,

$$p(x_i) = r(\beta_0 + \beta_1 x_i) = \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}} \in [0, 1]$$

not very choice but
good one

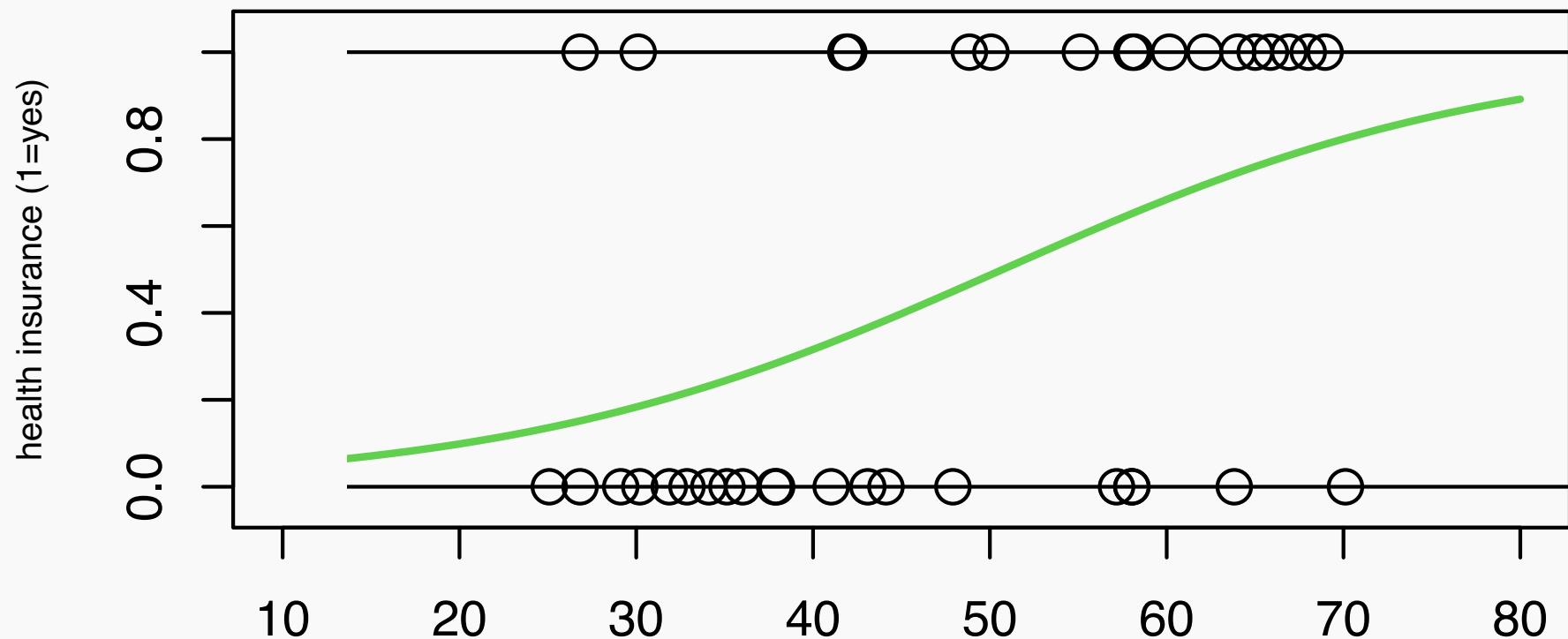
$$\beta_0 + \beta_1 x_i = \log \frac{p(x_i)}{1 - p(x_i)}$$

R session

$\text{mod} = \text{lm}(Y \sim x_1 + \dots + x_n)$ normal distribution by default

GLM $\text{mod} = \text{glm}(Y \sim X_0 + \dots + X_p, \text{family} = \text{binomial}, (\text{link} = \text{logit}))$
Ingredients \longrightarrow ① ② ③

A first example: Health Insurance coverage



The green line is the curve

$$p(\text{eta}) = g(-3.653 + 0.072\text{eta}) = \frac{e^{-3.653+0.072\text{eta}}}{1 + e^{-3.653+0.072\text{eta}}}$$

Logistic regression: Finding a “good” function

- The model defined above is the **logistic regression model**.
- We want to find the parameters β_0 and β_1 that define a curve that give a better description of the data.
- Note that in this case criteria like minimization of the sum of least squares do not provide simple solutions given the non linear nature of the function r .
- But we have assumptions about which probability distribution has generated the data, more precisely we assume that:
 - for a given value of x_i we observe $y_i = 1$ with probability $p(x_i)$
 - $p(x_i) = \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}}$
 - data are independent (i.e., derived from a simple random sample of n units from a population)

Logistic regression: Maximum likelihood estimation

- Under the assumptions stated above, once data (y_i, x_i) are observed, we can evaluate what is the probability $L(\beta_0, \beta_1)$ that the observed data are generated for each possible pair of values β_0, β_1 .
- The probability $L(\beta_0, \beta_1)$ is called the likelihood function and takes on different values for any possible couple (β_0, β_1) .
- We could then choose that couple $\hat{\beta}_0, \hat{\beta}_1$ which corresponds to the maximum probability (**maximum likelihood estimation**). This couple is the **maximum likelihood estimate**.
- Finding the maximum likelihood estimates $\hat{\beta}_0, \hat{\beta}_1$ usually requires the use of an iterative algorithm.

The solution obtained have "good" statistical properties especially if the sample is large

Multiple logistic regression: Extending the model

- The previous example has shown how the model can be easily extended to include more explanatory variables (in fact, we added gender).
- We can simply extend to the case where the log-odds depend linearly from a set of explanatory variables.
- This is similar to the multiple linear regression model. Then for the *i*-th unit in the sample we can write

$$E[Y_i] = p_i \quad \log\left(\frac{p_i}{1 - p_i}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_j x_{ij} = \mathbf{x}_i^T \boldsymbol{\beta}$$

- The covariates can be quantitative variables or indicator variables that account for qualitative factors. Also Interactions can be considered.

Structure of the model

Note that the model has a structure which is similar to the linear model

1. We specify a distributional assumption fro the response Y_i : a Bernolli variable in this case. Then $E(Y_i) = p_i$
2. We specify the way the inputs (the covariates) are combined in order to measure their impact on the expected value p_i : it i a linear combination

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_j x_{ij} = \mathbf{x}_i^T \boldsymbol{\beta}$$

3. We specify how the linear combination η_i is related to p_i . In the case of logistic regression $r(\eta_i) = \frac{e^{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_j x_{ij}}}{1 + e^{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_j x_{ij}}} = p_i$ so that the inverse function, called the link function is also defined

$$g(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = \mathbf{x}_i^T \boldsymbol{\beta}$$

Maximum likelihood in details

We have a precise idea of the distribution of the response variable and we will also assume that a random sample of size n is available.

The log-likelihood $\log(L(\beta)) = I(\beta)$ is

$$\begin{aligned} L(\beta) &= \prod_{i=1}^n P(y_i) \\ &= \prod_{i=1}^n \pi^{y_i} (1 - \pi_i)^{1-y_i} \\ &\quad \ell(\beta) = \sum_{i=1}^n [y_i \log(\pi_i) - y_i \log(1 - \pi_i) + \log(1 - \pi_i)] \\ &= \sum_{i=1}^n \left[y_i \log\left(\frac{\pi_i}{1 - \pi_i}\right) + \log(1 - \pi_i) \right] \end{aligned}$$

Logistic regression implies $\log\left(\frac{\pi_i}{1 - \pi_i}\right) = \mathbf{x}_i^T \beta$ and then

$$\ell(\beta) = \sum_{i=1}^n [y_i \mathbf{x}_i^T \beta - \log(1 + \exp(\mathbf{x}_i^T \beta))] = \sum_{i=1}^n [y_i \eta_i - \log(1 + \exp(\eta_i))]$$

Equating to 0 the first derivative of $\ell(\beta)$ we obtain the likelihood equations

$$s(\beta) = \frac{\partial I(\beta)}{\partial \beta} = \sum_{i=1}^n \mathbf{x}_i (y_i - \pi_i) = 0$$

It is a system of n non linear equations whose solution requires numerical methods.

$$\begin{aligned} P(y_i) &= \pi_i^{y_i} (1 - \pi_i)^{1-y_i} \\ \pi_i &= \frac{e^{\eta_i}}{1 + e^{\eta_i}}, \quad \eta_i = \mathbf{x}_i^T \beta = \logit(\pi_i) \end{aligned}$$

$$\eta_i = \log\left(\frac{\pi_i}{1 - \pi_i}\right)$$

$$\begin{aligned} \pi_i &= \frac{1}{1 + e^{-\eta_i}} \\ 1 - \pi_i &= \frac{e^{-\eta_i}}{1 + e^{-\eta_i}} = \frac{1}{1 + e^{\eta_i}} \\ &= \left(1 + e^{\eta_i}\right)^{-1} \end{aligned}$$

- Comparing probabilities:
- Difference in probabilities is not too useful because it depends on the starting point
 - Instead, we should focus on the odds, because comparing ratios is more meaningful
 - ↳ then difference of logarithms gave the same interpretation that ratios of odds

Parameters interpretation

Logistic regression

- The interpretation of the parameters of a logistic regression model is slightly different compared with linear regression. Let us consider a simple regression model with just one variable
- The intercept, β_0 , is meaningful only if $x = 0$ makes sense in the context considered. There is not simple interpretation
- In the simple model here introduced, the important parameter is the one associated with the covariate: β_1
 - if β_1 is positive the larger is x the higher will be the probability that the event occurs
 - if β_1 is negative for large values of x the probability that the event occurs will be lower
 - $\beta_1 = 0$ implies no effect of X on the probability of the event

$$\beta_0 = \log\left(\frac{p}{1-p}\right)$$

Logistic regression

- The parameter β_1 of a logistic regression, unlike linear regression, cannot be interpreted as the variation in the probability corresponding a variation of 1 unit in X
- In fact the slope of the curve is different for different values of X (since the relationship is not linear)
- but
 - we can consider the inverse of relationship
$$P(y=1) = p(x_i) = g(\beta_0 + \beta_1 x_i) = \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}}$$

β_1 act linearly on the logit of P_i

 - in this case we obtain $\log \frac{p(x_i)}{1-p(x_i)} = \beta_0 + \beta_1 x_i$
 - $\log \frac{p}{1-p}$ is the so called **logit transform** of a probability p
- In the logistic regression model (or logit model) we assume that X affects linearly the logit $\log \frac{p}{1-p}$

$$\beta_1 = E\left[\log \frac{P_1}{1-P_1} \mid X = x+1\right] - E\left[\log \frac{P_0}{1-P_0} \mid X = x\right] = E\left[\log \frac{\frac{P_1}{1-P_1}}{\frac{P_0}{1-P_0}} \mid X = x\right]$$

A second example: A dose-response analysis

- Consider the data in the table below

dose	1.66	1.74	1.75	1.76	1.78	1.80	1.86	1.88
n. positive	3	9	23	30	46	54	59	58
n. of patients	59	60	62	56	63	59	62	60
proportion	0.051	0.150	0.371	0.536	0.730	0.915	0.951	0.967

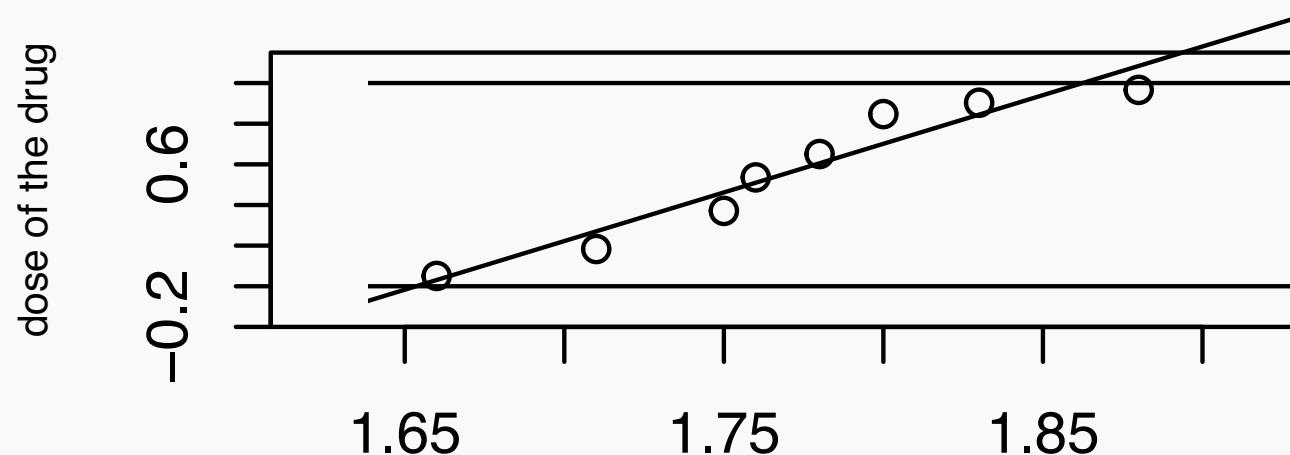
\uparrow dose \uparrow proportion , $n \approx \text{const.}$

$\hat{\pi}$
estimate
of sample
proportion

This is a binomial distribution

- The data refer to 481 individuals who received a drug. For each dose of the drug it has been observed if the individual had a positive response or not.
- Since only 8 different doses have been considered we can obtain the proportion positive responses for each dose.

Binomial response



- The plot shows that the proportion of positive responses out of m_i on trial, increases with the dose of the drug.
- A linear relationship is patently inappropriate. The data are proportions and their values should lie in the $[0,1]$ range
- $Y_i \sim \text{Binomial}(m_i, h(x_i))$. Specify a non linear model for $h(\cdot) \rightarrow [0, 1]$.

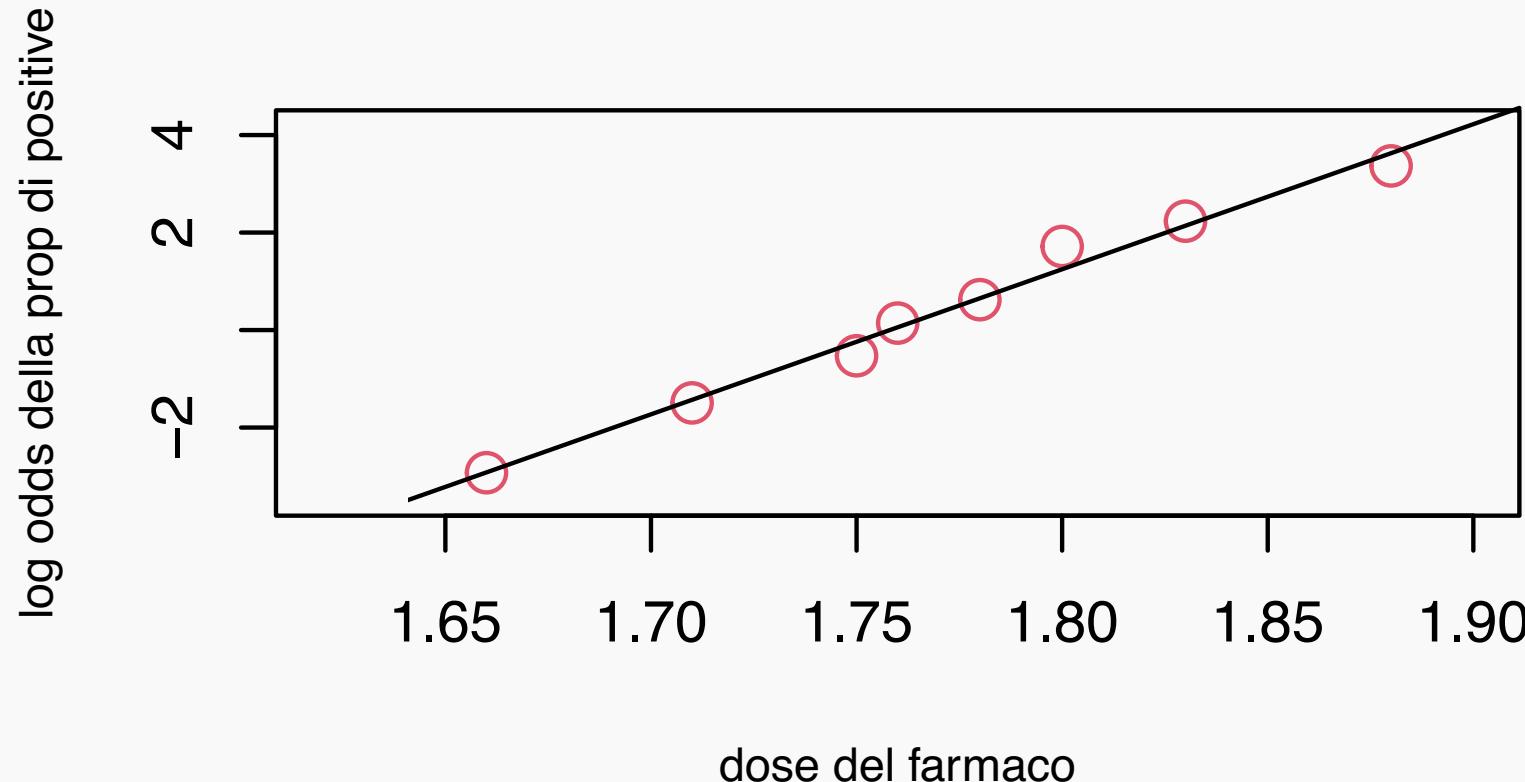
Logistic regression: The logit transform

Let us consider again the data about the proportion of positive responses to the drug.

dose	1.66	1.74	1.75	1.76	1.78	1.80	1.86	1.88
n. positive	3	9	23	30	46	54	59	58
n. of patients	59	60	62	56	63	59	62	60
proportion (p)	0.051	0.150	0.371	0.536	0.730	0.915	0.951	0.967
$p/(1 - p)$	0.05	.177	0.59	1.15	2.71	10.80	19.67	29.00
<i>interpretation in terms of differences.</i>	-2.92	-1.73	-0.53	0.14	0.99	2.38	2.98	3.36

- $\frac{p}{1-p}$ are the odds. Odds provide an alternative way to describe the probability of an event. They take on values between 0 and ∞

Logistic regression: Alternative representation of the dose response model



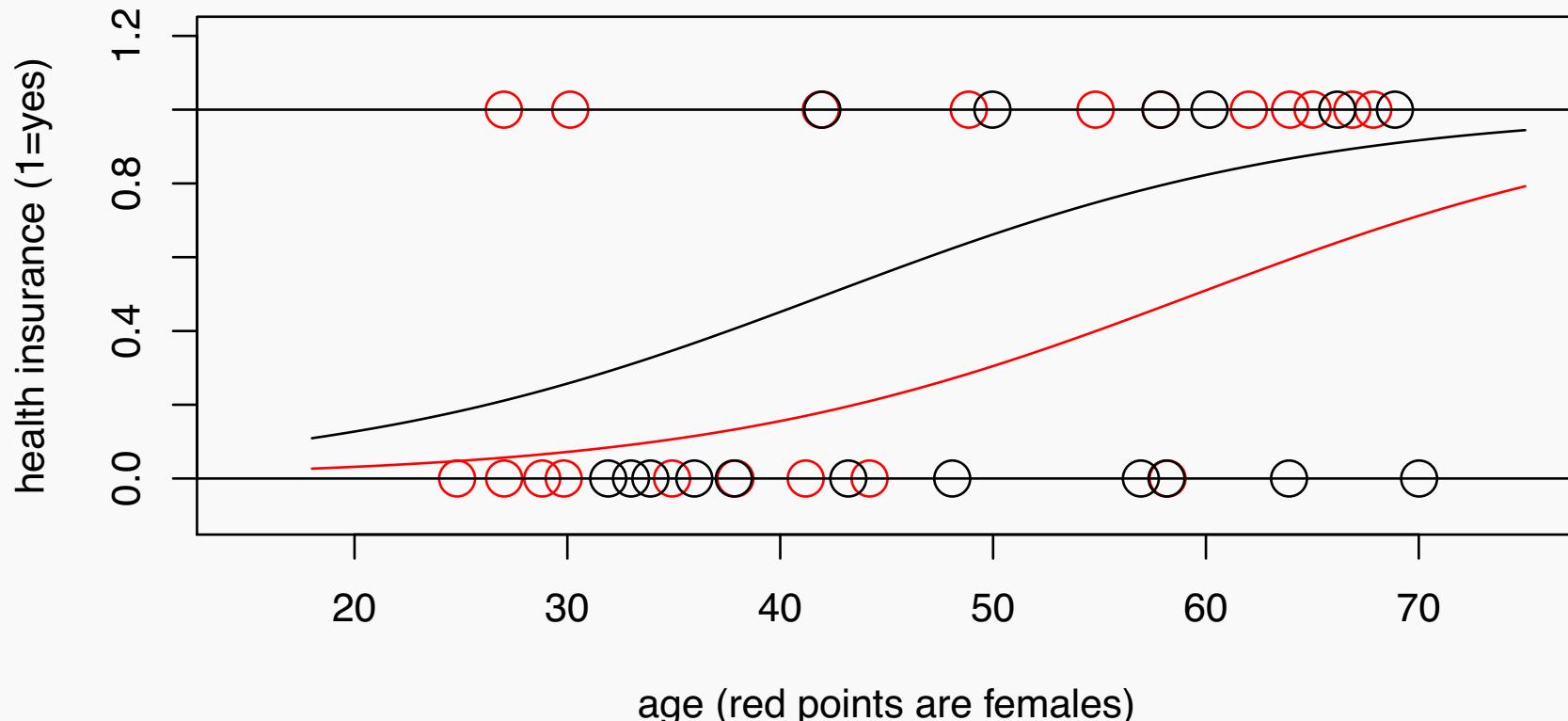
- The relationship between dose and log–odds of the proportion is linear!
- This means that a unit increase of the dose will cause an increase of β_1 in the log–odds of the proportions

Logistic regression: Odds and log-odds

- Bernoulli random variables are completely defined by the value of p , the probability of a “success”. The odds defined as $\frac{p}{1-p}$, obtained by a simple transformation of p , have an important interpretation.
- Suppose p indicates whether a given football team wins the next match. If $p = 0.2$ than the odds of the team winning are $0.2/(1-0.2)=1/4$ and we may say that the odds of winning are 1 on 4.
- This means that if we bet 1 euro on the team winning, in a fair game, if the team wins we get the euro back plus 4 euros. If the team does not win, we lose our euro.
- The odds provides the important information in this context (bet of 1 and winning of 4) and in fact when betting the information provided are simply the odds.
- If we know the odds we can calculate the probability p and vice versa.
- The odds can take on any positive value and the odds are 1 when an event has probability $p = 0.5$.
- The logarithm of the odds is often used, it can take any value and it is equal to 0 if the probability $p = 1/2$.
- As we have noted β_1 in our simple logistic regression model is the proportional variation we observe in the log-odds if the covariate X is increased by a unit.

Interpretation of a dichotomous covariate: Health Insurance coverage continued

- Let us consider again the data on private health insurance and assume we know observe the gender of the respondents



Logistic regression with a dichotomous covariate: Health Insurance coverage continued

- This is the result for a more complex logistic regression model

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x + \beta_2 \text{sex}$$

sex can take on only two values 0 (if female) or 1 (if male)

- The maximum likelihood estimates of the coefficients are
(Intercept) eta sex
-5.152 0.087 1.496
- Probability of owing a health insurance is higher for males and increases with age

Logistic regression with a dichotomous covariate: Health Insurance coverage continued

- If we evaluate the difference in the log-odds of the probability of health insurance (at a given age) for males, p_{male} , and females, p_{female} , this will be simply equal to 1.496
- $\log \frac{p_{male}}{1-p_{male}} - \log \frac{p_{female}}{1-p_{female}} = 1.496$
- or equivalently $\log \frac{\frac{p_{male}}{1-p_{male}}}{\frac{p_{female}}{1-p_{female}}} = 1.496$
- The estimated coefficient $\beta_2 = 1.496$ represents the so called log-odds ratio
- And $e^{1.496}$ is the odds ratio
- Odds ratio is 1 if the two odds (or) the two probabilities are the same for males and female

$$\begin{aligned} \frac{\frac{1}{p_m} - 1}{\frac{1}{p_f} - 1} &= k \\ \frac{\frac{1}{p_m} - 1}{\frac{1}{p_f} - 1} &\approx k \times \left(\frac{\frac{1}{p_m} - 1}{\frac{1}{p_m} - 1} \right) \\ p_f &= \left[e^k \left(\frac{1}{p_m} - 1 \right) + 1 \right]^{-1} \end{aligned}$$

Logistic regression with a dichotomous covariate: Health Insurance coverage continued

- Log-odds ratio is 0 if the two probabilities are the same ...
- and when the probability of a health insurance is the same for males and females then having or not a health insurance policy do not depend on the gender.
- In this case the value of $\beta_2 = 1.496$ indicates a seemingly not negligible change in the log-odd ratio and it means that probability is different for males and female.
- The odds ratio $e^{\beta_2} = e^{1.496} = 4.464$ indicates that the odds of having a health insurance for a male are more than 4 times the same odds for a female.

e^{β_2} gives the expected change in the odds for pairs

Males are about 4.5 times more likely to have a health insurance policy than females.

Inference for logistic regression parameters

Testing parameters significance

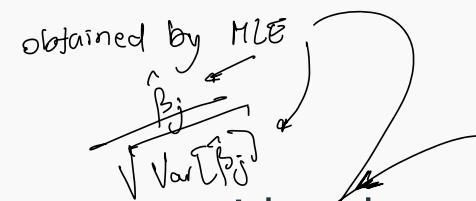
- Maximum likelihood method provides good estimates of the β s.
- For the j -th variable X_j we want to state if the data convey enough evidence to draw the conclusion that this variable is relevant to predict the response variable.
- Maximum likelihood methods provides also estimates of the standard errors of the estimated parameters.
- For (moderately) large sample we are able to answer to the question:
"is a given parameter significantly different from zero?"

Testing parameters significance

Testing β_j

$$\hat{\beta}_j \sim N(\beta_j, \text{Var}(\hat{\beta}_j))$$

then we compute the p-value : prob. of obtaining a value larger than z .



- As in the linear regression case we can consider the ratio

This test statistic directly specifies the null hypothesis $\beta_j = 0 \Leftrightarrow z = \frac{\hat{\beta}_j - \beta_j}{\text{s.e.}(\hat{\beta}_j)} \sim N(0, 1)$ for large enough sample if H_0 is true

- If absolute value of z is “large” then data support the hypothesis that the parameter is significantly different from zero
- to decide when “large” is really large, one can give a look to the associated p-values. This is the probability that we obtain a z even larger than the one observed when the parameter is actually equal to 0.
- since p-values are probabilities they lies between 0 and 1. And usually one judge the j -th variable relevant if the p-value associated to its estimate is (possibly much) smaller than 0.05

Testing parameters significance

1. The result above follows from the asymptotic properties of MLE: for large n we know that $\hat{\beta}_{ML} \stackrel{\text{approximately}}{\sim} \mathcal{N}(\beta, I(\beta)^{-1})$ where $I(\beta)$ is the expected information matrix, which in the case of a Bernoulli model is

$$X^T W X = I(\beta) = \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^T \pi_i (1 - \pi_i)$$

↳ problem w/ the inverse

Regularization techniques
can also be applied.

where $\pi_i = r(\mathbf{x}_i^T \beta)$.

2. This matrix depends on the unknown quantities β but a consistent estimate is obtained by substituting to β its estimate $\hat{\beta}$.
3. The element on the diagonal of $I(\beta)^{-1}_{jj}$ is an estimate of the variance of $\hat{\beta}_j$.
 $\text{Var}[\hat{\beta}_j]$
4. For this reason the ratio $\frac{\hat{\beta}_j}{\sqrt{I(\hat{\beta})^{-1}_{jj}}}$ evaluated, is asymptotically distributed as a Standard Gaussian assuming $H_0 : \beta_j = 0$.

Inference for logistic regression parameters: Judging the overall performance of the model

- For the logistic regression model it is not possible to obtain a quantity that has the same interpretation of R^2 in the linear model.
- It is possible to measure the difference between the value of the likelihood for the estimated parameters $L_{\hat{\beta}} = L(\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_j)$ and the value of the likelihood we would obtain in other cases.
- Two relevant cases are
 - the likelihood L_{max} one could achieve if considers as many parameters as available data (thus achieving a perfect fit)
 - the likelihood L_0 one obtains in a null model , i.e., a model with only the intercept β_0 (this means that no covariate has a significant effect on the response).
- Comparing those likelihoods helps to judge whether the model is useful to predict the response variable

Deviance of residuals to assess model was used before

$$DEVf = \sum (y_i - \hat{y}_i)^2$$

Inference for logistic regression parameters: Judging the overall performance of the model

- It is possible to look at the ratio between $L_{\hat{\beta}}$ and L_0 or at the difference between $\log L_{\hat{\beta}}$ and $\log L_0$:
if the latter difference is small then the model is not supported by the data
- It is also possible to consider the difference between the $\log L_{max}$ and $\log L_{\hat{\beta}}$. This difference should be small for good models.
- The value $D = 2(\log L_{max} - \log L_{\hat{\beta}})$ is called the deviance.
- It behaves like the deviance in the linear model: is large for bad models and decreases as we improve the model for instance by adding more significant explanatory variables.
- Comparing the deviances of two alternative models that differ only because a simpler model is obtained by setting some parameters equal to 0 (i.e. excluding some potential covariates) helps to decide which one among the two models should be preferred.

Logistic regression results: Health Insurance coverage

```
mod1<-glm(formula = sani ~ eta + sex, family = binomial(link=logit))  
summary(mod1)
```

Logistic regression: Predicting the response variable

We would probably keep the model w/ 3 variables because we have few clusters
= **only**

- Remind that in a logistic regression model we assume that

$$p_i = \frac{e^{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_j x_{ij}}}{1 + e^{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_j x_{ij}}}$$

- We can simply estimate the probabilities p_i by substituting the estimated values to the β s

invariant of MLE

$$\hat{p}_i = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \dots + \hat{\beta}_j x_{ij}}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \dots + \hat{\beta}_j x_{ij}}}$$

- These predicted probabilities are used when this model is used for classification. Simply define a threshold $c \in (0, 1)$ and predict
- $$Y_i = 1 \quad \text{if} \quad \hat{p}_i > c$$

To test improvement in models we can have nested models and then compare the Sum Squared of Residuals

$$\hat{y}_i = \mathbf{H}\mathbf{y} = \mathbf{x}_i^T \boldsymbol{\beta} = (\mathbf{x}_i^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}_i^T) \mathbf{y}$$

↙ reduction considering less covariates for \mathbf{X}

$$SSR_R = \sum (y_i - \hat{y}_i)^2 \quad \text{by construction it is smaller than the larger model}$$

$$SSR_L = \sum (y_i - \hat{y}_i)^2$$

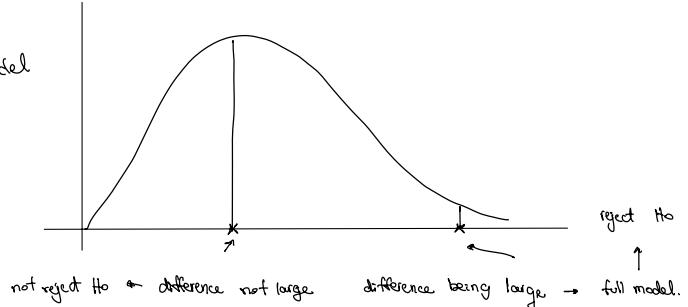
The SSR metric is related w/ the F test and ANOVA to assess the models.

We define the deviance for small deviance

$$-2 \left(l(\boldsymbol{\beta}_0) - l(\hat{\boldsymbol{\beta}}_F) \right) \downarrow \chi^2 \rightarrow \text{full model}$$

full model / saturated m.

for large values we keep the reduced model.



There exist another model to make classification called Probit model based on the CDF of the standard normal ϕ . This model "requires" proper random and balanced datasets to work. However, this is not always the case. Additionally the estimators based on the Probit model are biased.

In the logistic regression model the dataset needs not to be balanced and the estimators are not biased apart of β_0 , which is not too problematic because it is not related to any variable.

$$\Rightarrow r(\hat{\eta}_i) = \hat{\mu}_i + \text{bias}$$

Alternative specification of the response function

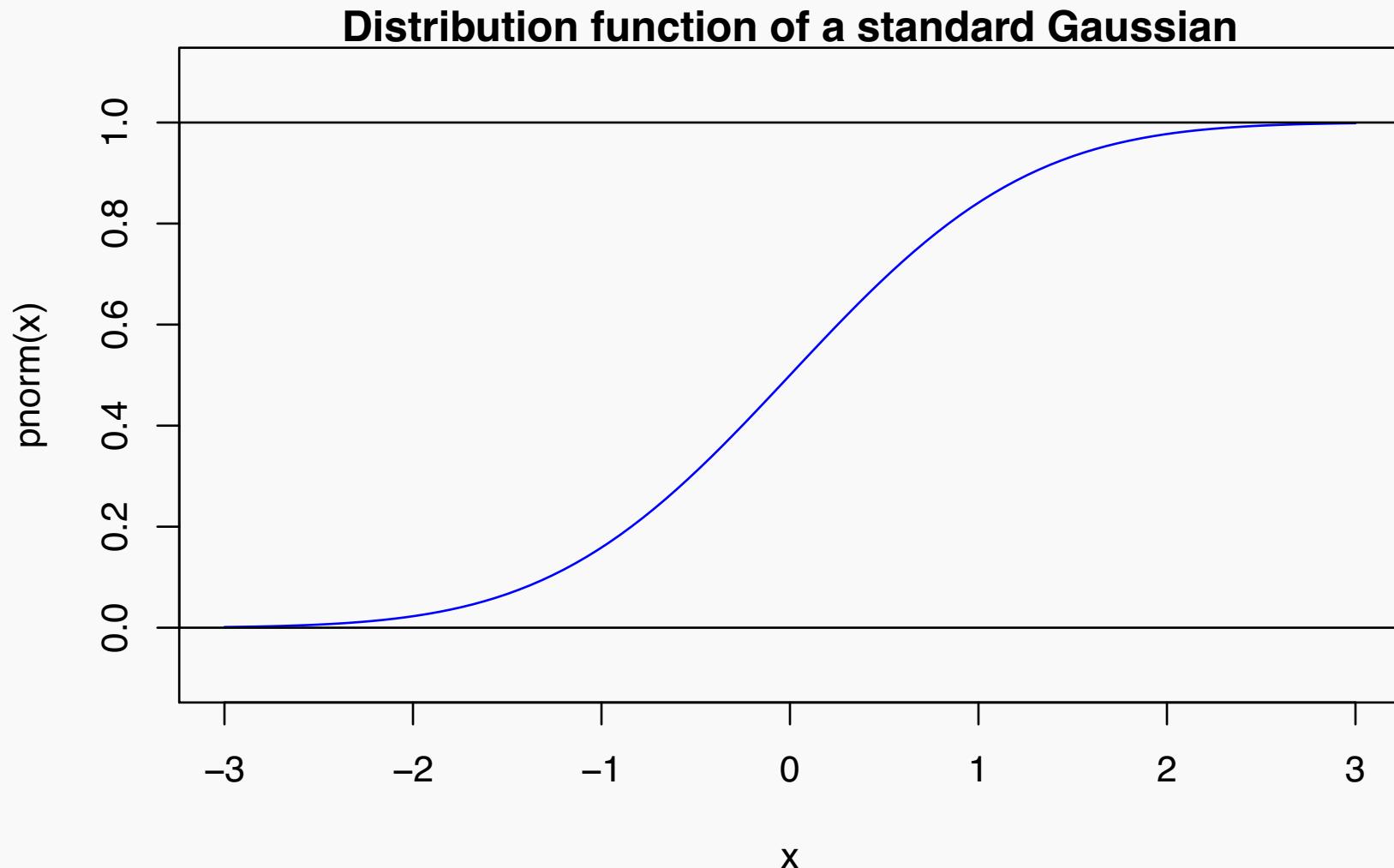
Probit regression

- We justified the choice of the response function $g(z)$ that gave rise to logistic regression by saying that we needed a S shaped function that lies within the $[0, 1]$ range since we want it to represent probabilities.
- But there are many function that we could choose. For instance a function that could work well is the distribution function of the standard Gaussian
- In fact we could write

$$p_i = \Phi(\beta_0 + \beta_1 x_i)$$

where the function Φ is the distribution function of the standard Gaussian random variable

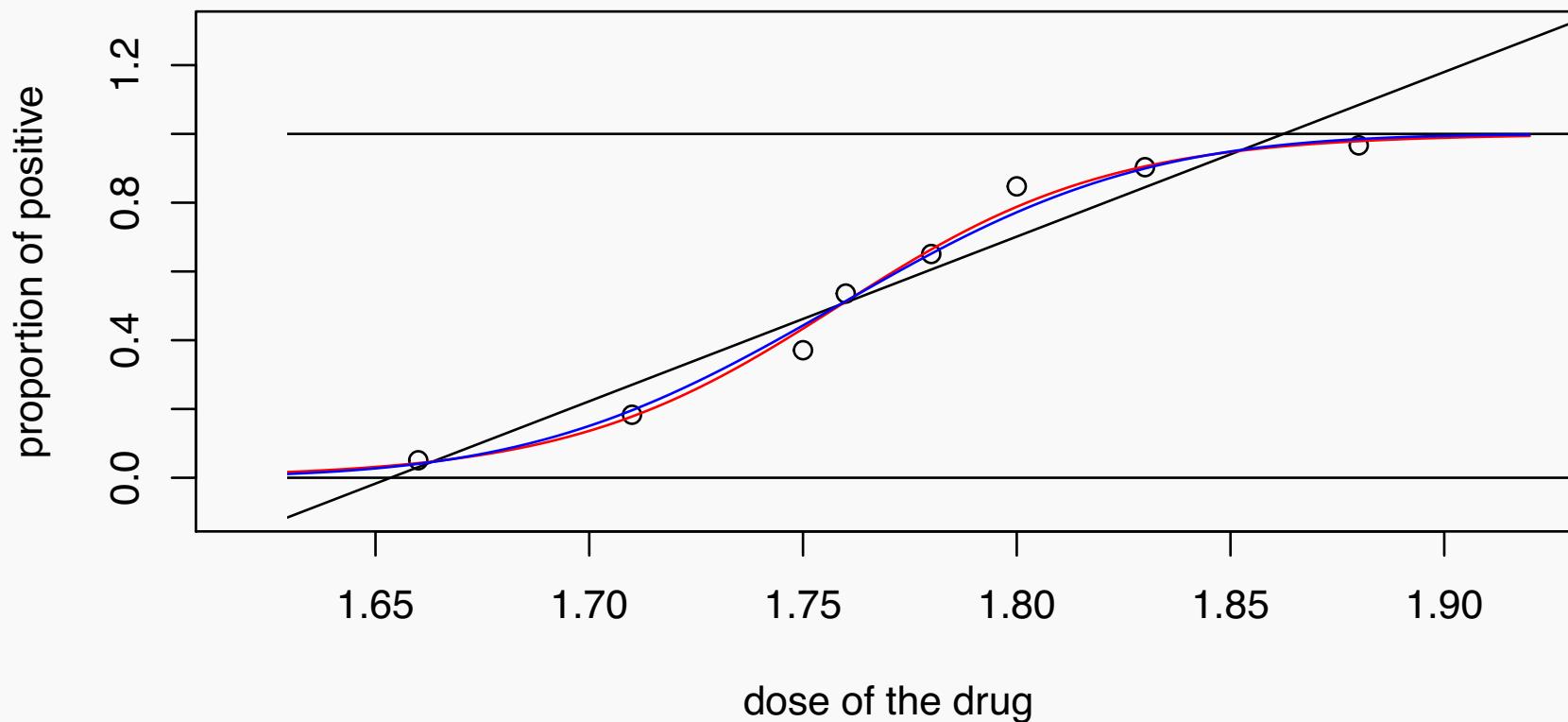
Probit regression



- This choice of the response function defines the **probit regression model**
- Probit regression model is also very popular
- Other choices are also possible for $g(\cdot)$

Probit vs logistic regression

- Actually probit regression gives results that are very similar to those obtained with logistic regression



the blue curve represents prediction by a probit regression model

Estimation issues

The case of perfect separation

- The maximum likelihood estimates for a binomial model are generally easily found using efficient numerical algorithms
- However, there may be convergence problems if it exist a function of the covariates that perfectly separates $y_i = 1$ and $y_i = 0$. Or if for some categories defined by a covariate, one observes y only 0 or only 1.
- In this case the likelihood function does not have a maximum and as a results the estimates provided are highly unstable.
- The main symptom is therefore given by a message that says “the algorithm has not reached convergence” and that “probability predictions have been obtained which are numerically equal to 1 or 0”. Another symptom is that the values of the standard errors of the estimates are very high.
- There are several solutions. One possibility is the one that uses a penalized likelihood.
- This solution can be obtained by considering a likelihood to which a term is added to eliminate the bias in ML estimates for logistic regression (for example by using the {R brglm} package)

↑

bias reduction

Why logit link?

We want to relate $E[Y_i] = \mu_i$ w/ $\eta_i = x_i^T \beta$

using $g(\mu_i) = \eta_i \Leftrightarrow r(\eta_i) = \mu_i$

such that $g^{-1} = r$ monotonic function

Reasons:

- logit is numerically efficient
- sample can be unbalanced
- β_i except β_0 are unbiased but β_0 is not related to any variable

↳ it shifts the probability function just as ξ shifts

$$(x + \xi)^2$$

Different linkers change the probability

Here we change + or - ingredients

Regression for count data

Poisson regression and other models for counts

N. Torelli, G. Di Credico, V. Gioia

Fall 2020

University of Trieste

Introduction

Poisson regression

Inference

Overdispersion

Poisson regression... and beyond

Introduction

Count variables

Let us now consider the case of a response variable Y that is a count.

The variable Y can take on values of 0, 1, 2 and so on.

Relevant examples of count dependent variables are:

- the number of car accidents
- the number of phone calls over a day for an assistance service
- the number of items bought on a sales portal
- the number of cases of a disease in a given territory
- the number of those accessing a web site

Usually the counts refer to all events occurred within a specified time interval (or a space interval)

Counts as response variable in a statistical model

- Also in this case we want to build a statistical model and the goal is to predict the number of events.
- We try to explain/predict the counts y_i by using observed characteristics of the i -th unit (such as age, sex, education, income, etc..) .
- We expect that the distribution of the counts varies as the other covariates vary and possibly we can try to summarize how the distribution of the counts varies by describing how the mean number of the counts vary.
- It is important to remind that there are more models that can be used for the distribution of a count variable.
- The simplest model, and possibly the best known, for this specific case is the Poisson one.

Poisson regression

The Poisson distribution

- A count variable Y is a variable that takes on 0 or any positive integer number, i.e., 0, 1, 2, ...
- Its probability function is the sequence of probabilities $Pr(Y = 0), Pr(Y = 1), Pr(Y = 2), \dots$
- It could be represented in a table like this

y	$Pr(Y = y)$
0	$e^{-\mu}$
1	$\mu e^{-\mu}$
2	$\frac{\mu^2 e^{-\mu}}{2!}$
3	$\frac{\mu^3 e^{-\mu}}{3!}$
...	...

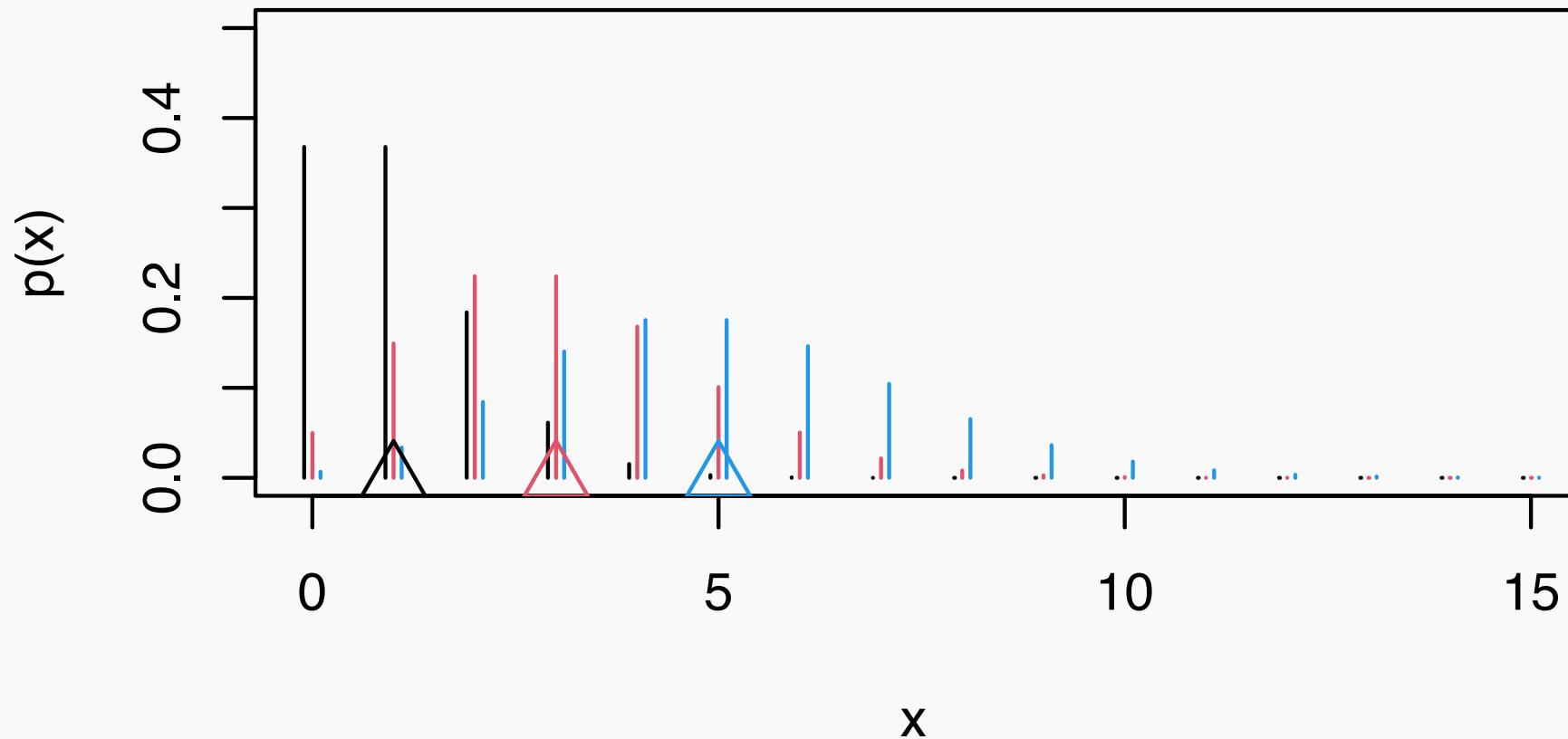
- The Poisson random variable then defines the probabilities of any value $Y = 0, 1, 2, 3, \dots$ as follows

$$Pr(Y = y) = \frac{\mu^y e^{-\mu}}{y!}$$

- the parameter μ is the expected value of the counts Y distributed according to a Poisson random variable. μ is greater than 0

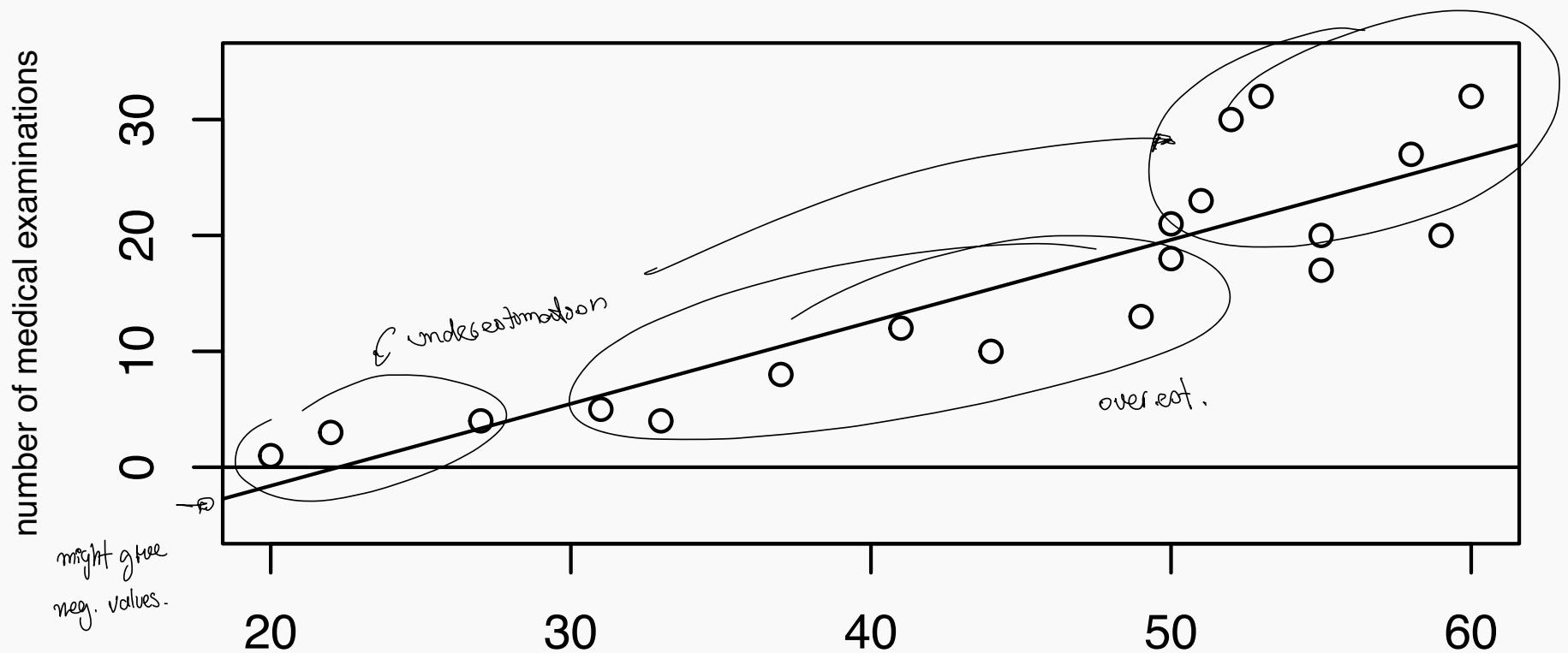
The Poisson distribution: some examples

Poisson rv for different values of the parameter



The graph shows the Poisson distribution for different values of the parameter, i.e., the mean ($\mu = 1$ black, $\mu = 3$ red, $\mu = 5$ blue)

Count data: an example with medical examinations



- Y_i (number of examinations) can be assumed Poisson: $Y_i \sim \text{Poisson}(\mu_i)$.
- We can assume that $\mu_i = h(x)$, i.e., it is a function of the covariate x .
- A linear specification is clearly inappropriate (also because it will predict negative values). We should choose among functions that $h(\cdot) \rightarrow [0, \infty)$.

Poisson regression: Basic framework

- We assume, like in other regression models, that for any value of the covariate X the curve represents the mean μ of the dependent variable Y .
- In the example, we assume that for a given age x_i the number of medical examinations has a Poisson distribution whose parameter (mean) is μ_i .
- The mean μ_i of the variable Y_i is assumed to lie on a curve (a function) like the one depicted in the previous slide.
- More specifically, we assume:

$$\mu_i = E(y_i) = e^{\beta_0 + \beta_1 x_i} = e^{\beta_0} (e^{\beta_1 x_i})$$

multiplicative factor.

- The linear component $\beta_0 + \beta_1 x_i$ is transformed by the exponential function and as a result it takes on only positive values.
- Remind that the mean of a count variable can be only positive
- This will define the Poisson Regression

Poisson regression: Interpretation of the parameters

$$E[Y_i] = \mu_i = e^{\beta_0 + \beta_1 x}$$

$$\Delta E = e^{\beta_0 + \beta_1 x} \cdot e^{\beta_1} - e^{\beta_0 + \beta_1 x} = e^{\beta_0 + \beta_1 x} (e^{\beta_1} - 1)$$

$$\frac{\Delta E}{E_0} = e^{\beta_1} - 1$$

$$\log\left(1 + \frac{\Delta E}{E_0}\right) = \beta_1$$

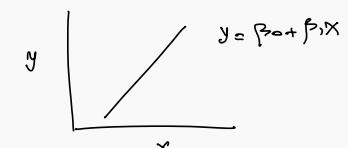
- The linear specification within the exponential function ease interpretation.

$$\log(1+x) \approx \left|_{x=0} x - \frac{x^2}{2} + O(x^3)\right.$$

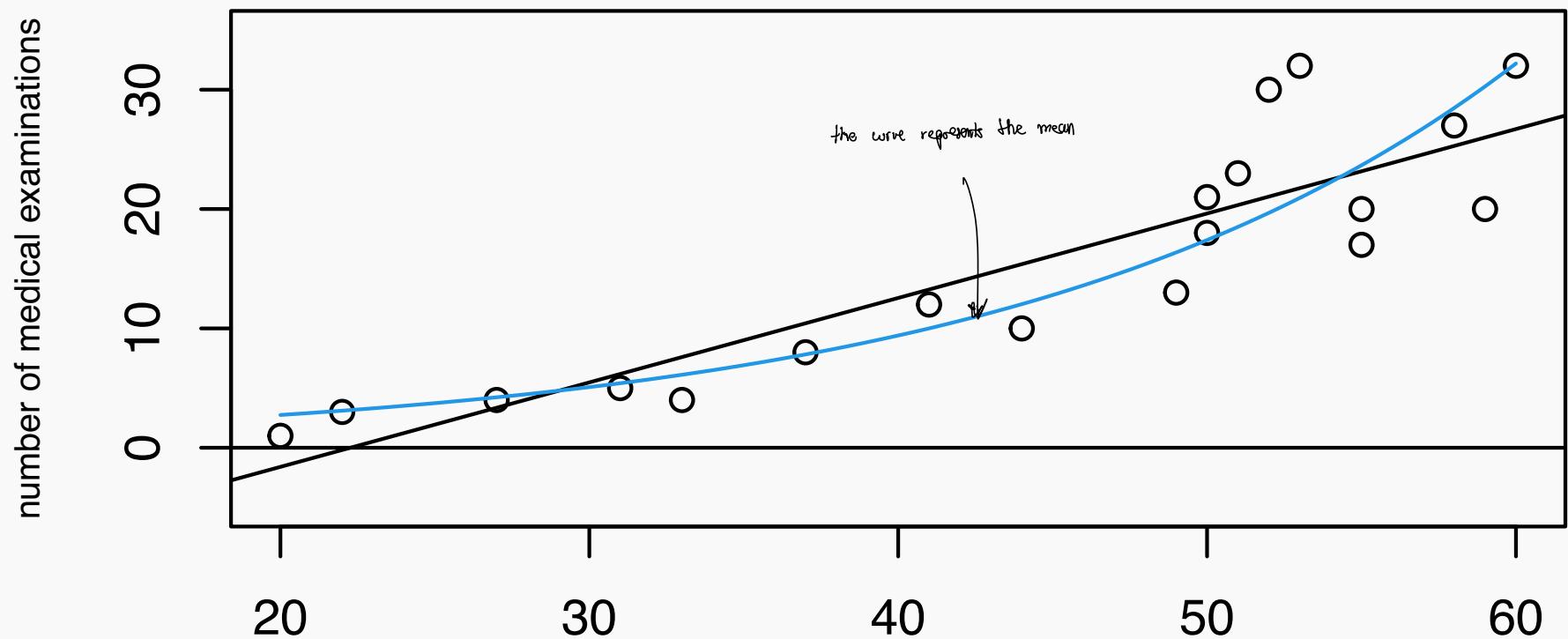
- For this simple model the most interesting parameter is β_1 that is associated to the covariate X .
- β_1 can be interpreted as the proportional change in the mean corresponding to a unit change in X . We can multiply it by 100 and interpret it as the percentage variation in Y .
- In the example presented above we obtained $\beta_0 = -0.220$ and $\beta_1 = 0.062$. It means that the model predicts that if we add one year to age the mean number of medical examinations raises of about 6.2 percent.

$$\mu_i = e^{\beta_0 + \beta_1 x}$$

$$= e^{\beta_0} e^{\beta_1 x} \approx_{\beta_1} e^{\beta_0} \left(1 + \beta_1 x\right)$$



Count data: an example with medical examinations



This new curve is more appropriate for the mean μ of the number of examinations as a function of age.

Poisson regression: Estimation of the parameters

- The non linear specification adopted makes a bit more difficult to detect the curve that approximate the data points. There are many possible criteria to find it.
- Since we have postulated a data generating mechanism based on the Poisson distribution we can again use the maximum likelihood method.
- If we observe a random sample of data (y_i, x_i) than we can evaluate the probability of obtaining those data.
- More specifically, it is assumed that each data point is drawn from a Poisson distribution with mean $\mu_i = e^{\beta_0 + \beta_1 x_i}$.

Then using independence assumption (implied by random sampling) we can evaluate the probability $L(\beta_0, \beta_1)$ of observing that specific set of data points for each possible couple (β_0, β_1)

The MLE is the value $\hat{\Theta}_{ML}$ the maximizes the Likelihood function, which translates to maximizing the probability of observing the data given the model.

It is not a probability but the set of parameters that maximize the prob. distib. given the data

- The maximum likelihood estimate is the couple $(\hat{\beta}_0, \hat{\beta}_1)$ that corresponds to the highest value of $L(\beta_0, \beta_1)$
- locating the maximum likelihood estimates $(\hat{\beta}_0, \hat{\beta}_1)$ is not straightforward and requires the use of an iterative procedure.

Multiple Poisson regression

- The Poisson regression model defined in the example is very simple. It can be easily extended to include more covariates (quantitative variables or qualitative factors).
- A more general specification for the model is then + Y_i has a Poisson distribution with mean μ_i + $\mu_i = e^{\beta_0 + \beta_{i1}x_{i1} + \beta_{i2}x_{i2} + \cdots + \beta_{ip}x_{ip}}$
- This is a log-linear model since a linear regression model is assumed for the logarithm of μ_i :
$$\log(\mu_i) = \beta_0 + \beta_{i1}x_{i1} + \beta_{i2}x_{i2} + \cdots + \beta_{ip}x_{ip} = \mathbf{x}_i^T \boldsymbol{\beta}$$

Inference

Deviance for Normal LR

$$D = -2 \log \left(\frac{L_{\text{max}}}{L_{\hat{\mu}}} \right) = -2(L_{\text{max}} - L_{\hat{\mu}})$$

$$L_{\hat{\mu}} = L(\hat{\mu}, y) = -n \log \sqrt{2\pi\sigma^2} - \frac{\sum (y_i - \hat{\mu}_i)^2}{2\sigma^2}$$

$$L_{\text{max}} = L(\mu, y) = -n \log \sqrt{2\pi\sigma^2}$$

$$D = \frac{-2 \sum (y_i - \hat{\mu}_i)^2}{\sigma^2}$$

Poisson regression: parameters estimates

- The Poisson assumption for Y_i allows us to use maximum likelihood for parameters estimation.
- log-likelihood, assuming random sampling, is:

$$\log(L(\beta)) = \ell(\beta) = \sum_{i=1}^n (y_i \log(\mu_i) - \mu_i) + \text{const}(y_i)$$

the constant $-n\log(y_i!)$ is omitted since it does not depend on β

- using the link function $\log(\mu_i) = \eta_i$ we obtain

$$\ell(\beta) = \sum_{i=1}^n \ell_i(\beta) = \sum_{i=1}^n y_i \mathbf{x}_i^T \beta - \exp(\mathbf{x}_i^T \beta) = \sum_{i=1}^n (y_i \eta_i - \exp(\eta_i))$$

we can evaluate the score function to obtain the likelihood equations

$$s(\beta) = \sum_{i=1}^n \frac{\partial \ell_i(\beta)}{\partial \beta} = \sum_{i=1}^n \mathbf{x}_i (y_i - \exp(\eta_i)) = \sum_{i=1}^n \mathbf{x}_i (y_i - \mu_i) \quad \text{and equate it to 0} \quad s(\beta) =$$

expected Fisher information $i(\beta)$ can be also obtained

$$i(\beta) = E(s(\beta)s(\beta)^T) = \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^T \mu_i$$

- solving likelihood equations is not straightforward and numerical solutions are necessary (e.g. Newton-Raphson)

R command : `glm(Y ~ x1 + ... + xn, family = Poisson(link = log))`

Inference in Poisson regression: testing significance of single β -s

Definition & inference

- Once the parameters of the model are estimated we are interested in deciding if a given covariate is relevant or not to predict the response variable.
- Also for this model, for large samples, maximum likelihood method provides also good estimates of the standard errors of the β s.
- We can then evaluate if a given β_j associated to the covariate X_j is large enough by looking at the ratio $\frac{\hat{\beta}_j}{s.e.(\hat{\beta}_j)}$.
- If the absolute value of this ratio is large there is evidence that the variable X_j affects the mean of Y . For large samples, the ratios above are well approximated by a standard Gaussian distribution when the parameter is actually 0. This allow us to judge if the ratio is large enough.
- The rule of thumb is that large means greater than 2. But it is always a good idea to look at p-values associated to the estimated parameters.
Because in this case the p-values for the normal distribution are less than 0.05

Inference for Poisson regression models: Judging the overall performance of the model

- Also in this case, just like in the logistic regression, one can measure the difference between the value of the likelihood for the estimated parameters $L_{\hat{\beta}} = L(\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p)$ and the value of the likelihood we would obtain when considering:
 - as many parameters as the available data (that would give a perfect fit) L_{max}
 - the null model with only the intercept β_0 , i.e., L_0
 - an alternative model that is equal to the one estimated but with some parameters set to 0 ($L(\hat{\beta}_R) = L(\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k, 0, \dots, 0, 0)$). In the last expression the last $p - k$ parameters are set to 0
- Comparing those likelihoods (or their logarithm) helps to judge the quality of the model

Inference for Poisson regression models: Judging the overall performance of the model

- The difference between $\log L_{\hat{\beta}}$ and $\log L_0$ give a first indication: if the latter difference is small then the model is not supported by the data
- The difference between the $\log L_{max}$ and $\log L_{\hat{\beta}}$ should be small for good models.
- Twice this difference is called the deviance and is defined as $D_{\hat{\beta}} = 2(\log L_{max} - \log L_{\hat{\beta}})$.
- A way to compare two models is to compare their deviances. We could compare the deviance $D_{\hat{\beta}_R}$ where the likelihood is evaluated for a reduced model $L(\hat{\beta})_R$.
- If the difference between the two deviances is small we keep the simpler model.
- To decide when “small” is small enough we can use statistical criteria (comparing the difference with the value of a appropriate χ^2 distribution with $p - k$ degrees of freedom)

$$D_{\hat{\beta}_1} = 2(\ell(y, y) - \ell(\hat{\mu}_1, y))$$

$$D_{\hat{\beta}_0} = 2(\ell(y, y) - \ell(\hat{\mu}_0, y))$$

$$\Delta D = 2 \left\{ 2\ell(y, y) - [\ell(\hat{\mu}_1, y) - \ell(\hat{\mu}_0, y)] \right\}$$

$$= D_{\hat{\beta}_1} - D_{\hat{\beta}_0} \quad \text{not}$$

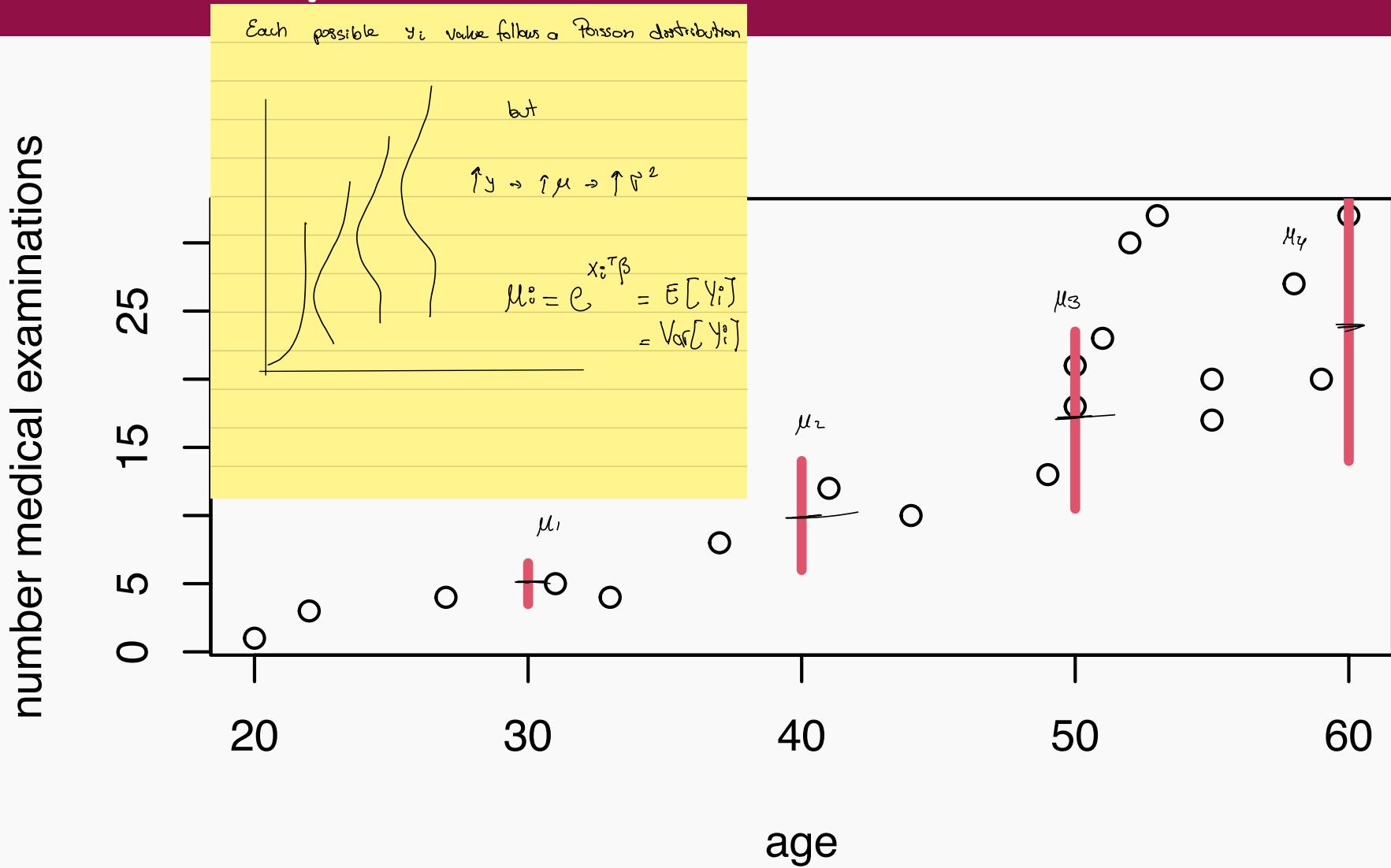
$$\Delta D = 2(\ell(\hat{\mu}_1, y) - \ell(\hat{\mu}_0, y)) = 2(\ell(\hat{\mu}_1, y) - \ell(y, y) + \ell(y, y) - \ell(\hat{\mu}_0, y)) = D_{\hat{\beta}_0} - D_{\hat{\beta}_1}$$

yes

Overdispersion

Variance increases more than expected

Again the example with medical examinations



The vertical red strips illustrate how dispersion of the number of medical examinations increases with the mean. Can the model accommodate this?

Overdispersed count data

- Poisson regression models in a GLM context imply that the variance function then functionally related to the mean function (it is actually the same).
- In fact, a striking characteristic of a Poisson distribution is that its mean is equal to its variance.
- A Poisson model states that the mean of our response variable varies according to the model. This implies that the variance of Y_i varies accordingly.
- The Poisson regression model implicitly introduces a form of heteroschedasticity.
- This characteristic makes the Poisson model very peculiar and in many cases reduces its flexibility and its ability to describe real situations.
- A simple way to check appropriateness of the model is to verify if data reflect the specific requirement $\text{mean}(Y_i) = \text{variance}(Y_i)$
- Considering a model for counts with overdispersion will be more realistic in many practical cases

Poisson regression: Residual checks

- In the Normal linear regression models residuals checks are a powerful tool for assessing model adequacy
- We can evaluate residuals also for Poisson regression
- First we can predict μ by using our model. More specifically:

$$\hat{\mu}_i = e^{\hat{\beta}_0 + \hat{\beta}_{i1}x_{i1} + \hat{\beta}_{i2}x_{i2} + \dots + \hat{\beta}_{ip}x_{ip}}$$

- We can now obtain residuals by comparing the predictions with the observed values and dividing them by the estimated standard deviations (remind that the model is heteroschedastic)

$$r_i = \frac{\hat{\mu}_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i}} \sim N(0, 1)$$

only if appropriate it holds.

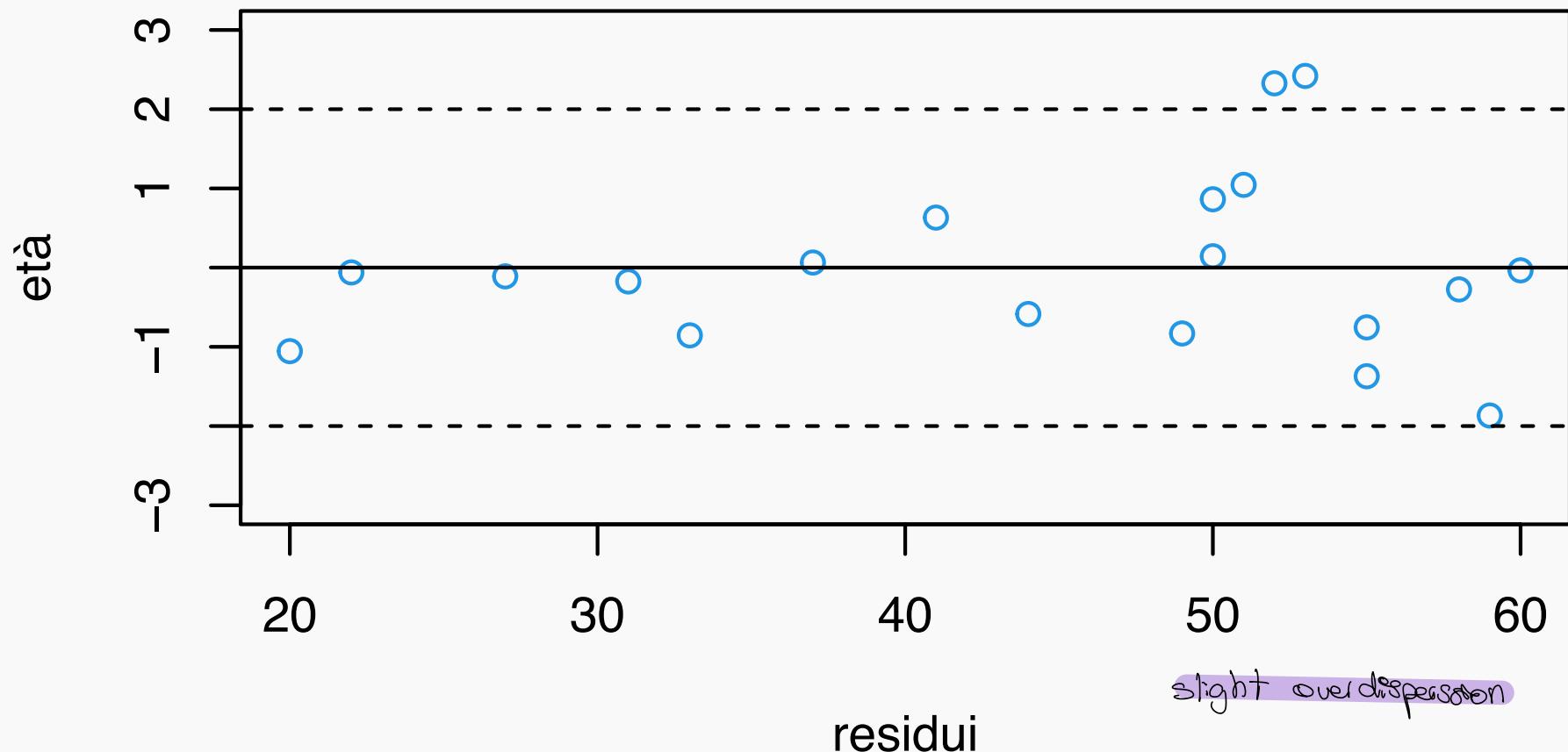
$$\Rightarrow r_i \in [-2, 2]$$

Poisson regression: Residual checks

- A look at residuals plots can help detecting if our assumptions are reasonable
- Residuals move around 0, they should be reasonably small and should not show any specific pattern.
- For large samples they are approximately equivalent to draws from a standard Gaussian. This means that the large majority (about 95%) of them have a value between 2 and -2. 
- A large number of residuals whose absolute value is greater than 2 could be a symptom that $\text{variance}(Y_i) > \text{Mean}(Y_i)$.
- This situation, called overdispersion, indicates that a Poisson model could be not appropriate

Data an medical examination

residuals graph



Note that two residuals are outside the interval $[-2, +2]$

Overdispersed count data

- Poisson regression models in a GLM context imply that the variance function then functionally related to the mean function (it is actually the same).
- For a Poisson models the standardized residuals are

$$z_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i}}$$

where $\hat{\mu}_i = \exp^{x_i^T \hat{\beta}}$

- If the Poisson model holds the z_i are approximately independent and will have mean equal to 0 and variance equal to 1. Approximately $\sum_{i=1}^n z_i^2$ is a χ^2_{n-p} distribution if the model holds. This can be used for detecting overdispersion.
- Considering a model for counts with overdispersion will be more realistic in many practical cases

Dealing with overdispersion

- For LMs the method of LS allows to obtain estimates of the regression parameters without the specification of a probabilistic model.
- The method of LS requires only the specification of the relation between the expected value of the response variable and the linear predictor, and the specification of the variance of the error term,
- Also for the GLMs it is possible to specify only these two relations (assuming that the variance function $V(\mu_i)$ is known).
- In other words, this means that the parametric assumption $Y_i \sim EF(\cdot, \phi)$ could not even be satisfied. Only the assumption about expectations is essential: $\mu_i = E(Y_i) = g^{-1}(\eta_i)$
- the only distributional feature that must be known in order to calculate the estimating equation is the variance function $V(\mu)$.

Quasi-likelihood model

We get rid of the family assumption

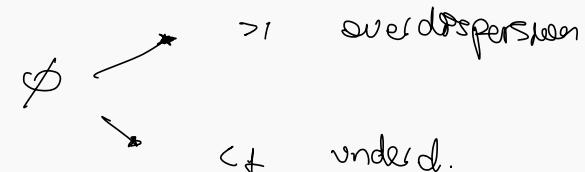
$$Y_i \sim \text{Po}(\mu_i)$$

Keep	$E[Y_i] = \mu_i$
change	$V[Y_i] = \phi V^2 = \phi \mu_i$
Keep	$g(\mu_i) = \eta_i = X_i^\top \beta$

`glm(, family(quasipoisson))`

but it works well "only"
for large samples

- Under suitable regularity conditions, the likelihood equations for a GLM give estimates for the coefficients β which maintain several properties, also if the parametric assumptions of Y_i are substituted with weaker assumptions:
 - $g(\mu_i) = g(E(Y_i)) = \eta_i, \quad i = 1, \dots, n,$
 - $\text{var}(Y_i) = \phi V(\mu_i), \quad i = 1, \dots, n,$
 - $\text{cov}(Y_i, Y_j) = 0, \text{ if } i \neq j.$
- The semi-parametric statistical model specified by assumptions 1–3 is called **quasi-likelihood model**.



Quasi-likelihood model

- The assumptions 1–3 above offer an increase in flexibility with respect to the usual parametric specifications based, respectively, on the Poisson, binomial or exponential distributions.
- In general, the quasi-likelihood approach allows to deal with *overdispersion problems*: it is possible to specify $\text{var}(Y_i)$ so that there is more variability with respect to the exponential family.
- The case of *underdispersion*, i.e. $\phi < 1$, is less important in applications, but can be dealt with under the quasi-likelihood model as well.

Using quasi-likelihood in `glm`

- When estimating a GLM by using quasi-likelihood one can use the same variance function derived from a Binomial or from a Poisson model and using the canonical link for those models. In R this leads to a specification of the `family` that is called `quasibinomial` or `quasipoisson`.
- Estimates of the β are the same since the estimating equations do not change
- But standard errors of estimates will change since a value different from 1 is estimated for ϕ . In `quasipoisson` one should take into account that variance is modelled as $\text{Var}(y_i) = \phi\mu_i$
- The parameter ϕ can be also estimated as

$$\hat{\phi} = \frac{1}{n-p} \sum \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i}$$

- In those cases also the Deviance of the model has to be corrected because it is computed assuming $\phi = 1$. The deviance reported has to be divided by $\hat{\phi}$
- Also the standardized residuals are different. E.g., for the Poisson:

$$z_i^{QL} \rightarrow z_i^{QL} = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\phi}\hat{\mu}_i}} \quad vs \quad z_i^{GLM} = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i}}$$

95% $-2 < z_i^{QL} < 2$

$-2 \leq z_i^{GLM} \leq 2$

Poisson regression... and beyond

Overdispersion occurs when data is not collected in the same conditions/situations
sample in different cities w/ different populations.

Poisson regression: Taking into account exposure

- The basic Poisson model can be extended to take into account the fact that counts can arise under different conditions.
- Assume we want to model the average number of accidents Y_i on some roads. Obviously, this number depends on how many vehicles in a given period have been on the road. The set of those exposed to the risk of accident is different for each data unit.
- The number of vehicles is an exposure variable e_i and it should be taken into account.
- It could be sensible to model the rates instead of the counts, i.e., we could write for $\mu_i = E(y_i)$
- $\log\left(\frac{\mu_i}{e_i}\right) = \beta_0 + \beta_{i1}x_{i1} + \beta_{i2}x_{i2} + \cdots + \beta_{ip}x_{ip}$
- But this is equivalent to put
$$\log(\mu_i) = \beta_0 + \beta_{i1}x_{i1} + \beta_{i2}x_{i2} + \cdots + \beta_{ip}x_{ip} + \log(e_i)$$
- This means that we insert an additional variable $\log(e_i)$ among the covariates but imposing its coefficient to be equal to 1
- The special covariate $\log(e_i)$ is called **offset**

glm(... , offset = log e)

Other models for count data

- The basic Poisson model is the simplest one to use when modelling counts.
- Poisson random variables are not the only ones that can be used to describe the distribution of counts.
- Other slightly more complex models, for instance Negative Binomial random variables, could be more flexible to cope with situations (such as overdispersion) frequently encountered in practise.
- Poisson models could be also refined to face non standard and more complex data patterns, such as:
 - + The case of counts that can never take on the value 0 (truncated Poisson)
 - + The case where counts can be observed only for a portion of the sample. For the remaning portion we know that only 0 is possible (zero inflated models). This can occur in automobile claims data because insured are reluctant to report claims fearing that this will result in higher future insurance premiums.
- More complex model are also needed to accommodate situations where the theoretical conditions that give raise to Poisson counts are not met.

Negative binomial regression

- It is an alternative model that can be considered when data exhibits overdispersion. Its probability function is

$$Pr(Z = z) = \binom{z + k - 1}{z} p^k (1 - p)^{z-k} \quad z = 0, 1, \dots$$

where $E(Z) = k(1 - p)/p$ and $Var(Z) = k(1 - p)/p^2$.

- Interpretation: probability to observe z *failures* until the pre-specified number of *successes* k is observed.
- Compared with Poisson
 - since it has an extra parameter it proves to be more flexible
 - mean is larger than variance and then it accommodates overdispersion
 - Poisson is a limiting case of negative binomial (if $p \rightarrow 1$ and $k \rightarrow 0$ then $kp \rightarrow \lambda$)
- Recall that negative binomial emerges as a mixture of Poisson when each unit Y is Poisson with mean λ and λ are drawn from a *Gamma* distribution.

Negative Binomial regression

- When building a model for Negative Binomial a different parametrization is more appropriate, by defining $Y = Z - k$ and

$$p = \frac{1}{1+\alpha}$$

- $$Pr(Y = y) = \binom{y + k - 1}{k - 1} \frac{\alpha^y}{(1 + \alpha)^{y+k}} \quad y = 0, 1, \dots$$
- Then
 - $E(Y) = \mu = k\alpha$
 - $Var(Y) = k\alpha + k\alpha^2 = \mu + \mu^2/k$
- and the following link can be used $\log \frac{\alpha}{1+\alpha} = \log \frac{\mu}{k+\mu}$
- In R a specific function has to be used: `glm.nb(...)` included in the package MASS

Zero inflated Poisson

- Zero inflation means that we have far more zeros than what would be expected for a Poisson or BiN distribution
- Ignoring zero inflation can have two consequences:
 - the estimated parameters and standard errors may be biased
 - the excessive number of zeros can cause overdispersion
- A possible model hypothesizes that the observed counts derive from a mixture of two populations:
 - for a part of the population (with probability p) Y can only be 0
 - for the remaining part (with probability $1 - p$) Y is distributed as a Poisson or a BiN.
- Distribution of counts is then, in case of Poisson

$$P(y_i = 0) = p_i + (1 - p_i)e^{-\mu_i}$$
$$P(y_i = y_i | y_i > 0) = (1 - p_i) \frac{\mu_i^{y_i} e^{\mu_i}}{y_i!}$$

- Covariates can be introduced, like in GLM, for modelling p_i and μ_i

Statistical Methods - II partial test

November 29, 2023 - 1 hour

To allow evaluation, you must submit the text with name and surname.

Name and Surname Luis Fernando Palauca Flores	Student ID	38
--	------------	----

1 The coefficient of determination R^2 is generally used

- to evaluate the coefficients significance
- to compare non-nested models *
- to evaluate model misspecification *
- to evaluate the explained variability by the model ✓

Not completely true

$$\rho = P(T \geq |t|)$$

2 Code 1 reports the summary of a linear model that describes the professor evaluation eval using the beauty score of the professor (beauty, continuous) and the gender (female, dichotomous).

With reference to the interaction coefficient, which of the following statement is correct?

- The probability of observing a value larger than -0.113 is 0.0789 ✗
- The interaction coefficient is statistically significant at 0.15 ✗
- The interaction coefficient is statistically significant at 0.05 ✗
- The probability of observing a value larger than -1.761 is 0.0789 ? ✗

This time I was lucky

$$\rho = P(T > |t|)$$

3 Figure 1 shows the residuals $e_i = y_i - \hat{y}_i$ vs the predictor x of a simple linear regression model. absolute value

Which of the following assumptions of the linear model seems to be violated?

- Normality
- Linearity
- Homoscedasticity *
- None

4 Consider a normal linear model $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$, $i = 1, \dots, n$ with $\varepsilon_i \stackrel{iid}{\sim} \mathcal{N}(0, \sigma^2)$.

Which of the following statements on the residuals $e_i = y_i - \hat{y}_i$ is false?

- The residual sum of squares allow us to evaluate the proportion of variability not explained by the regression model → The errors ε_i are homoscedastic and uncorrelated
- The residuals are homoscedastic / For the residuals we have $e = y - \hat{y} = (I - H)y$ ↗ correlated
- The residuals sum up to 0 ↗ heteroscedastic
- The sum of the squares of the residuals is the minimum over all possible $(\beta_0, \beta_1, \beta_2)$ ↗ ✗

5 A Poisson GLM relates the number of yearly shark attack with the yearly average global surface temperature (GST). The intercept of the model is $\hat{\beta}_0 = 2.997$, and the predictor coefficient is $\hat{\beta}_1 = 0.005$. Which of the following statements is false?

- When the GST increases of 1 degree, there is a multiplicative effect of 1.005 on the expected number of shark attack
- The expected number of shark attack for a GST of 25 is 22.692 ✓ $\mu_0 = e^{\beta_0} e^{\beta_1 x} \approx \beta_1 x_0 e^{\beta_0} (1 + \beta_1 x)$
- For a unit increase of the GST, the average number of shark attacks raises of about 0.5% ✓
- When the GST is of 0 degrees, the expected number of shark attack is 2.997 ✗

6 Code 2 reports the summary of a logistic model on the variable type, representing diabetic subjects, to evaluate the probability of having diabetes (type=1). The covariates describe the plasma glucose concentration (glu, continuous), the diabetes pedigree function (ped, continuous), the age (age, continuous), and the obesity condition (obese, dichotomous). Construct a 99% approximate confidence interval for the coefficient glu.

- 0.013; 0.105
- 0.014; 0.048 ✓
- 0.259; 3.65
- 0.052; 2.349

7 Code 1 reports the summary of a linear model that describes the professor evaluation eval using the beauty score of the professor (beauty, continuous) and the gender (female, dichotomous).

Select the correct value of the residual sum of squares, i.e. $\sum_i (\text{eval}_i - \widehat{\text{eval}}_i)^2$.

- 0.073
- 131.918 ✓
- 246.07
- 0.927

Generalized Linear Models (GLM)

(Extending the linear model)

N. Torelli, G. Di Credico, V. Gioia

2023

University of Trieste



Probably not in the test

Introduction

Generalized Linear Models: Basic ideas

Inference = how we do estimate the parameters ?

Solution of the likelihood equations

Model Evaluation Deviance

Quasi-likelihood

GLM: Extensions and recent development

Introduction

We try to summarise information w/ the mean because it minimizes the loss / squared error
For each observation the response r.v Y_i depends on covariates X_i

$$E(Y_i | X_i) = h(X_i)$$

\uparrow
regression function

We could use another function to summarize the information of the data but the theory would be more difficult.

Introduction

Generalized linear models (GLMs) encompass the statistical models reviewed so far

- The response variable in a GLM can be a quantitative variables (also considering the case when the response takes on only positive values), a dichotomous variable, a count variable.
- GLMs recognizes that for many models the idea that covariates effects can be summarized by a linear combination is useful and flexible enough.
- And it is recognized that the aim is to study how this linear combination of the covariates affects the expected value of the response variable.
- Usually the main interest is in estimating the unknown coefficients of the linear combination (the β parameters).
- Linear regression model, logistic and probit regression, Poisson regression (and in fact many other models) have this common structure

From LM to GLM

- Recall that Normal LMs, in matrix notation, are defined by

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where

We change this ingred.

- $Y_i \sim N(\mu_i, \sigma^2)$, independent, where
 $\mu_i = \mathbf{x}_i^T \boldsymbol{\beta}$ and
 \mathbf{x}_i^T is i -th row of \mathbf{X} , $i = 1, 2, \dots, n$;
- The density of $Y_i \sim N(\eta_i, \sigma^2)$ and covariates \mathbf{x}_i appear through the **linear predictor**:

$$\eta_i = \sum_{j=1}^p x_{ij} \beta_j = \mathbf{x}_i^T \boldsymbol{\beta}$$

;

- $\boldsymbol{\beta}$ e σ^2 are unknown parameters.

Introducing GLMs

GLMs generalize LMs by:

- Y_i are assumed to be (independent) measurements from a distribution with density (probability) function from the **exponential dispersion family**
- Existence of the mean $E(Y) = \mu$ is assumed and μ is determined by η that is related to it by a suitable function

$$g(E(Y_i)) = g(\mu_i) = \eta_i = \mathbf{x}_i^T \boldsymbol{\beta}$$

$g(\cdot)$ is called the **link function**.

- in principle f could be any suitable density (or probability) function, but a family of distribution plays a key role:

Y_i are assumed to be (independent) measurements from a distribution with density (probability) function from the **exponential dispersion family**

The exponential (dispersion) family

- A random variable Y belongs to exponential (dispersion) family if its density (probability) function can be written as

$$f(Y; \theta, \phi) = \exp \left\{ \frac{Y\theta - b(\theta)}{\phi} + c(Y, \phi) \right\}, \quad (1)$$

θ e ϕ are unknown scalar parameters, $b(\cdot)$ and $c(\cdot) > 0$ are known functions and domain of Y do not depend on θ or ϕ .

We will denote this by $Y \sim EF(b(\theta), \phi)$.

↗ link function

- θ is called the *natural or canonical parameter* of the exponential family.
- ϕ is called the *dispersion parameter*. It can be known in some cases. When it is unknown, the family is more properly called the *exponential dispersion family*.
- Many of the most common continuous and discrete distributions belong to this family (i.e. Normal, Gamma, Poisson, Binomial, etc)

θ : not directly what we would expect from the known distribution

Bernoulli 0 FP

Example: Poisson

- As we already noted it is the basic choice when modelling count data
- if $Y \sim \text{Poisson}(\lambda)$, its probability function is

$$\begin{aligned} f(Y; \lambda) &= \frac{e^{-\lambda} \lambda^Y}{Y!} \\ &= \exp\{Y \log \lambda - \lambda - \log Y!\}, \end{aligned}$$

for $Y = 0, 1, \dots,$

link function

- This shows that it is a member of (1) where $\theta = \log \lambda$ is the natural parameter, $\phi = 1$, $b(\theta) = \lambda = e^\theta$ and $c(Y, \phi) = -\log Y!$.
- We can write $Y \sim EF(e^\theta, 1)$.

Example: Binomial

- Standard distribution when modelling binary responses
- If $Y \sim \text{Bin}(n, \pi)$, its probability function is

$$\begin{aligned}f(Y; \pi) &= \binom{n}{Y} \pi^Y (1 - \pi)^{n-Y} \\&= \exp\{\log \binom{n}{Y} + Y \log \pi + (n - Y) \log(1 - \pi)\} \\&= \exp\left\{Y \log \frac{\pi}{1 - \pi} + n \log(1 - \pi) + \log \binom{n}{Y}\right\},\end{aligned}$$

for $Y = 0, 1, \dots, n$. log odds is the link function

- It belongs to (1) where $\theta = \log \frac{\pi}{1 - \pi}$ natural parameter, $\phi = 1$,

$$b(\theta) = -n \log(1 - \pi)|_{\pi=\frac{e^\theta}{1+e^\theta}} = n \log(1 + e^\theta)$$

and $c(Y, \phi) = \log \binom{n}{Y}$. *For members of the exponential dispersion family have dispersion*

$$\phi = 1$$

- $Y \sim EF(n \log(1 + e^\theta), 1)$

Normal Distribution

natural parameters

vectors

sufficient statistic

$$f(y; \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left\{ -\frac{(y-\mu)^2}{2\sigma^2} \right\} = h(\bar{x}) \exp \left\{ \bar{\eta}(\theta) \cdot \bar{T}(\bar{x}) - A(\theta) \right\}$$

$$= \exp \left\{ \frac{\mu x}{\sigma^2} - \frac{\mu^2}{2\sigma^2} - \left(\frac{x^2}{2\sigma^2} + \frac{1}{2} \log 2\pi\sigma^2 \right) \right\}$$

for μ, σ^2 known

$$\theta = \mu/\sigma^2 \Rightarrow \mu = \theta\sigma^2$$

$$b(\theta) = \mu^2/2\sigma^2$$

$$g(b(\theta)) = \theta$$

$$X \sim EF \left(\mu^2/2\sigma^2, 1 \right) = EF \left(\theta^2\sigma^2/2, 1 \right)$$

$$b'(\theta) = \theta\sigma^2 = \mu$$

$$g(\mu) = \mu$$

$$= \exp \left\{ \frac{\mu x}{\sigma^2} - \frac{x^2}{2\sigma^2} - \frac{\mu^2}{2\sigma^2} - \frac{1}{2} \log 2\pi\sigma^2 \right\}$$

$$= \exp_{\text{unknown}} \left\{ \begin{pmatrix} \mu/\sigma^2 \\ -1/2\sigma^2 \end{pmatrix} \begin{pmatrix} x & x^2 \end{pmatrix} - \left(\frac{\mu^2}{2\sigma^2} + \log \sigma^2 \right) \right\} \times \frac{1}{\sqrt{2\pi}}$$

$$\eta(\theta) = \begin{pmatrix} \mu/\sigma^2 \\ -1/2\sigma^2 \end{pmatrix} = (\theta)$$

$$b(\theta) = -\frac{\mu^2}{2\sigma^2} + \log \sigma^2$$

Generalized Linear Models: Basic ideas

The structure of GLMs

- A general theory has been defined for GLMs also because this allowed to implement a single general procedure for estimating the parameters, checking their significance, evaluating the goodness of fit of the model, selecting the “best” model, obtaining predictions and computing residuals.
- Most software packages have in fact been implemented to this aim.
- In GLMs the response variables are assumed to be distributed according to a more general family of random variables, the **exponential** dispersion family.
- This family includes, among the others, the Binomial, the Poisson, the Gamma and the Normal families.
- Consequently linear regression models (where the response are usually assumed to be Normal), logistic regression (where the response is binomial), Poisson regression (where the response is a count) can be written as GLMs.

The ingredients of GLMs

Components of GLMs are the following:

1. a **response** Y_i distributed as a member of a quite comprehensive family of distributions the **exponential dispersion family**
 $Y_i \sim EF(b(\theta_i), \phi)$ where $E(Y_i) = \mu_i$ and whose variance is
 $V(Y_i) = \phi V(\mu_i)$. Variance in general heteroscedastic
2. a **linear predictor** $\eta_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi}$
3. a **link function** $g()$ assumed to be monotone and that relates the linear predictor η_i to μ_i so that $g(\mu_i) = \eta_i \Leftrightarrow \mu_i = g^{-1}(\eta_i)$,
 $i = 1, 2, \dots, n$ and then

$$E(Y_i) = \mu_i = g^{-1}(\beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi})$$

and g^{-1} is called the response function

Example: Poisson

- As we already noted it is the basic choice when modelling count data
- if $Y \sim \text{Poisson}(\lambda)$, its probability function is

$$\begin{aligned} f(Y; \lambda) &= \frac{e^{-\lambda} \lambda^Y}{Y!} \\ &= \exp\{Y \log \lambda - \lambda - \log Y!\}, \end{aligned}$$

for $Y = 0, 1, \dots,$

- This shows that it is a member of (1) where $\theta = \log \lambda$ is the natural parameter, $\phi = 1$, $b(\theta) = \lambda = e^\theta$ and $c(Y, \phi) = -\log Y!$.
- We can write $Y \sim EF(e^\theta, 1)$.

Copy equation

Example: Binomial

- Standard distribution when modelling binary responses
- If $Y \sim \text{Bin}(n, \pi)$, its probability function is

$$\begin{aligned}f(Y; \pi) &= \binom{n}{Y} \pi^Y (1 - \pi)^{n-Y} \\&= \exp\{\log \binom{n}{Y} + Y \log \pi + (n - Y) \log(1 - \pi)\} \\&= \exp\left\{Y \log \frac{\pi}{1 - \pi} + n \log(1 - \pi) + \log \binom{n}{Y}\right\},\end{aligned}$$

for $Y = 0, 1, \dots, n$.

- It belongs to (1) where $\theta = \log \frac{\pi}{1 - \pi}$ natural parameter, $\phi = 1$,

$$b(\theta) = -n \log(1 - \pi)|_{\pi=\frac{e^\theta}{1+e^\theta}} = n \log(1 + e^\theta)$$

and $c(Y, \phi) = \log \binom{n}{Y}$.

- $Y \sim EF(n \log(1 + e^\theta), 1)$.

Copy equation

Mean and variance for Exponential family

- The function $b(\cdot)$ is called the *cumulant function* and it is important in evaluating and interpreting first moments of the distribution.
- by using identities related to derivatives of log-likelihood function:

$$E(\ell_*(\theta)) = E\left(\frac{d}{d\theta}\ell(\theta; Y)\right) = 0$$

and

$$i(\theta) = \text{var}(\ell_*(\theta)) = E(-\ell_{**}(\theta)) = E\left(-\frac{d^2}{d\theta^2}\ell(\theta; Y)\right), = E\left[\left(\frac{\partial}{\partial\theta}\ell(\theta; y)\right)^2\right]$$

under usual regularity assumptions.

If Y is a r.v. member of the exponential family, log-likelihood for θ it follows that:

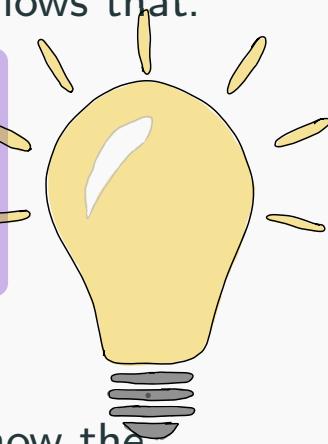
$$E\left(\frac{Y - b'(\theta)}{\phi}\right) = 0 \quad \text{and}$$

$$\text{var}\left(\frac{Y - b'(\theta)}{\phi}\right) = \frac{b''(\theta)}{\phi} \Rightarrow$$

$$\boxed{E(Y) = \mu = b'(\theta)}$$

$$\boxed{\text{var}(Y) = \phi b''(\theta)}$$

Denote $V(\mu) = b''(\theta)$, we can write $\boxed{\text{var}(Y) = \phi V(\mu)}$



- The function $V(\mu)$ is the so called *variance function* since it indicates how the variance depends on the mean of Y (GLM can be heteroscedastic). This becomes clear if we recall that μ is related to θ , i.e., $\mu = b'(\theta)$.

Some relevant member of the exponential family and their moments

Poisson

We have for a Poisson with mean λ

$$b(\theta) = e^\theta \quad \text{and} \quad \phi = 1 \quad \text{and} \quad E(Y) = b'(\theta) = e^\theta = \lambda .$$

$$\text{var}(Y) = b''(\theta) = e^\theta = \lambda \quad \text{then} \quad V(\mu) = \mu$$

Binomial

We have for a Binomial with parameters (n, π)

$$b(\theta) = n \log(1+e^\theta), \quad \phi = 1 \quad \text{then} \quad E(Y) = \mu = b'(\theta) = n \frac{e^\theta}{1+e^\theta} = n\pi .$$

$$= n \frac{\mu}{n} \left(1 - \frac{\mu}{n} \right) = \mu \left(1 - \frac{\mu}{n} \right) \xrightarrow[n \rightarrow \infty]{\mu}$$

$$\text{var}(Y) = b''(\theta) = n \frac{e^\theta}{(1+e^\theta)^2} = n\pi(1-\pi) \quad \text{and} \quad V(\mu) = \mu(1-\mu)/n .$$

\uparrow
 $\cancel{\phi = 1}$

The link function

- The second important step in specifying a GLM is the definition of the function relating μ_i and the linear predictor η_i .
- It is assumed that the link between μ_i , the mean of Y_i , and \mathbf{x}_i^T , the covariate vector, is

$$g(\mu_i) = \eta_i \quad \text{and} \quad \eta_i = \mathbf{x}_i^T \boldsymbol{\beta} .$$

- $g(\cdot)$ is a known monotone and differentiable function. The function $g(\cdot)$ is the *link function* between μ_i and η_i .
- the inverse function $g(\cdot)^{-1} = r(\cdot)$ is also called the response function
- Covariates enter into the model by the linear predictor $\mathbf{x}_i^T \boldsymbol{\beta}$, but the μ_i and η_i are generally non linearly related.
- Appropriate choices of the link function are such that $\mu_i = g^{-1}(\eta_i)$ takes on values on the appropriate range.

The canonical link

- A typical choice is to write directly the natural parameter θ as a linear function of the covariates Formally,

natural parameter is equal to the linear predictor which relates to the mean of the random variable

$$\eta = g(\mu) = g(b'(\theta)) = \theta , \theta = \theta(\mu)$$

$g(\cdot)$ is then the inverse function of $b'(\cdot)$. This choice of the link function is called *canonical link*. Default choice

- Some interesting properties derives from choosing a canonical link. Moreover the canonical link is the default link used in many softwares for estimation of GLMs (including R).

Inference

Estimation of the parameters

- ML can be used since distributional assumptions on parameters are available (for the normal LM it coincides with LS).
- A property of the exponential families is that they satisfy enough regularity conditions to ensure that the MLE is given uniquely by the solution of the likelihood equations.

- Let us recall some important features of GLM:

- $g(\mu_i) = \eta_i = \mathbf{x}_i^T \beta \Leftrightarrow \mu_i = g^{-1}(\mathbf{x}_i^T \beta);$

- $\mu_i = b'(\theta_i) \Leftrightarrow \theta_i = (b')^{-1}(\mu_i) = (b')^{-1}(g^{-1}(\eta_i)) = (b')^{-1}(g^{-1}(\mathbf{x}_i^T \beta))$

- $\text{var}(Y_i) = \phi V(\mu_i)$, with $V(\mu_i) = b''(\theta_i)$.

- Assuming independence of (y_1, \dots, y_n) , the log-likelihood $\ell(\beta, \phi)$ is simply given by

$$\ell(\beta, \phi) = \sum_{i=1}^n \log f(y_i; \theta_i, \phi) = \sum_{i=1}^n \ell_i(\beta, \phi) = \sum_{i=1}^n \left\{ \frac{y_i \theta_i - b(\theta_i)}{\phi} + c(y_i, \phi) \right\}$$

where θ_i is a function of β through

$$g(\mu_i) = g(b'(\theta_i)) = \eta_i = \mathbf{x}_i^T \beta .$$

$$\partial_{\beta_j} \ell_i = \partial_{\theta_j} \ell_i \partial_{\beta_j} \theta_j = \partial_{\theta_j} \ell_i \partial_{\eta_j} \partial_{\beta_j} \eta_j = \partial_{\theta_j} \ell_i \partial_{\eta_j} \mu_j \partial_{\mu_j} \partial_{\beta_j} \eta_j = \partial_{\theta_j} \ell_i (\partial_{\theta_j} \mu_j)^{-1} (\partial_{\mu_j} \eta_j)^{-1} \partial_{\beta_j} \eta_j$$

Likelihood equations

- To obtain the MLE of β it is necessary to solve the *likelihood equations*:

$$\frac{\partial \ell}{\partial \beta_j} = \sum_{i=1}^n \frac{\partial \ell_i}{\partial \beta_j} = 0 \quad \text{for } j = 1, 2, \dots, p.$$

- Let us compute

$$\begin{aligned} \frac{\partial \ell_i}{\partial \beta_j} &= \frac{\partial \ell_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j} = \frac{\partial \ell_i}{\partial \theta_i} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \beta_j} \\ &= \frac{\partial \ell_i}{\partial \theta_i} \left(\frac{\partial \mu_i}{\partial \theta_i} \right)^{-1} \left(\frac{\partial \eta_i}{\partial \mu_i} \right)^{-1} \frac{\partial \eta_i}{\partial \beta_j}, \end{aligned}$$

- where the terms can be written as

$$\begin{aligned} \frac{\partial \ell_i}{\partial \theta_i} &= \frac{y_i - b'(\theta_i)}{\phi} = \frac{y_i - \mu_i}{\phi}, \\ \frac{\partial \mu_i}{\partial \theta_i} &= b''(\theta_i) = \frac{\text{var}(Y_i)}{\phi}, \\ \frac{\partial \eta_i}{\partial \mu_i} &= g'(\mu_i), \\ \frac{\partial \eta_i}{\partial \beta_j} &= x_{ij}. \end{aligned}$$

Likelihood equations

- Thus, we have

$$\begin{aligned}\frac{\partial \ell_i}{\partial \beta_j} &= \frac{y_i - \mu_i}{\phi} \frac{\phi}{\text{var}(Y_i)} \frac{1}{g'(\mu_i)} x_{ij} \\ &= \frac{(y_i - \mu_i)x_{ij}}{\phi V(\mu_i)g'(\mu_i)}.\end{aligned}$$

- The likelihood equations for β are then

$$\partial_{\beta_j} \ell(y_i; \beta_j) = \sum_{i=1}^n \frac{(y_i - \mu_i)}{V(\mu_i)g'(\mu_i)} x_{ij} = 0,$$

$j = 1, 2, \dots, p$, where $\mu_i = g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})$.

$V(\mu_i) \rightarrow$ the model can be heteroscedastic.

NOT NEED OF LINK MLE
FUNCTION COMING FROM

- Note that the MLE of $\boldsymbol{\beta}$ for a fixed value of ϕ , does not depend on ϕ and coincides with the unconstrained MLE.

$$\frac{d}{d\theta} g(\mu_i) = g'(\mu_i) b''(\theta) \xrightarrow{\mu_i = b'(\theta)} \frac{d}{d\theta} = + \Rightarrow g'(\mu_i) = \frac{1}{b''(\theta)} = \frac{1}{\text{Var}(\mu_i)}$$

Canonical link

- The use of the *canonical link* ($\eta_i = g(\mu_i) = g(b'(\theta_i)) = \theta_i$) produces some simplifications in the inference based on the log-likelihood $\ell(\beta, \phi)$.
- With the canonical link, we have $g'(\mu_i) = 1/V(\mu_i)$ and the first derivative reduces to

$$\sum_{i=1}^n \frac{\partial \ell_i}{\partial \beta_j} = \sum_{i=1}^n \frac{(y_i - \mu_i)x_{ij}}{\phi V(\mu_i)g'(\mu_i)} = \sum_{i=1}^n \frac{(y_i - \mu_i)x_{ij}}{\phi}.$$

- This result implies that the likelihood equations simplify and take the form

non-linear eq.s in general

$$\sum_{i=1}^n y_i x_{ij} = \sum_{i=1}^n \mu_i x_{ij} .$$

$\eta_i = x_i^\top \beta = \theta_i(\mu)$

Using matrix notation, $X^T y = X^T \mu$.

Normal: $\hat{\beta} = (X X^\top)^{-1} X^\top \bar{y}$

- These equations agree with the general structure of the likelihood equations in exponential families: the observed value of the minimal sufficient statistic is equated to its expectation.
- As regards the existence and uniqueness of the MLE of β , if the link is the canonical one, the theory of exponential families applies.
- In general the likelihood equations for β are nonlinear and must be solved with iterative methods. To this end, the expected Fisher information for β is useful.

Fisher information

- Since β and ϕ are orthogonal, we can proceed as if ϕ were known and we can focus only on β .
- Let us consider the second derivatives of ℓ_i :

$$\begin{aligned}
 -E\left(\frac{\partial^2 \ell_i}{\partial \beta_j \partial \beta_k}\right) &= E\left(\frac{\partial \ell_i}{\partial \beta_j} \frac{\partial \ell_i}{\partial \beta_k}\right) \\
 &= E\left(\left(\frac{(Y_i - \mu_i)x_{ij}}{\phi V(\mu_i)g'(\mu_i)}\right) \left(\frac{(Y_i - \mu_i)x_{ik}}{\phi V(\mu_i)g'(\mu_i)}\right)\right) \\
 &= \frac{x_{ij}x_{ik}}{\phi^2(V(\mu_i))^2(g'(\mu_i))^2} E((Y_i - \mu_i)^2) \\
 &= \frac{x_{ij}x_{ik}}{\phi V(\mu_i)(g'(\mu_i))^2},
 \end{aligned}$$

which gives the (j, k) -element of the Fisher information matrix for β . Using matrix notation,

$$i(\beta) = \frac{X^T W X}{\phi}, \quad \text{s.e. } \hat{\beta} = (i(\beta))^{-1/2}$$

Variance considers automatically heteroscedasticity

$$\hat{\beta} = (i(\beta))^{-1/2}$$

$\uparrow V(\mu_i) \downarrow \text{weight}$

with $W = \text{diag}(w_1, \dots, w_n)$ and

$$w_i = \frac{1}{V(\mu_i)(g'(\mu_i))^2},$$

\uparrow
 \downarrow

for canonical link
i.e., known family

What does $X' = W^{1/2}X$ do?

Does X' follow the same distribution as X ?

and X is the matrix of the explanatory variables.

Fisher information

- With the **canonical link**, the observed and the expected informations coincide and have (j, k) -element

$$\frac{x_{ij}x_{ik}V(\mu_i)}{\phi}.$$

In matrix form,

$$i(\beta) = j(\beta) = \frac{X^T V X}{\phi},$$

with $V = \text{diag}(V(\mu_i))$.

- Asymptotic normality of the MLE gives

$$\hat{\beta} \sim N_p(\beta, \phi(X^T W X)^{-1}),$$

for large n .

- Therefore, a consistent estimate of the covariance matrix of β is $i(\hat{\beta}) = \phi(X^T \hat{W} X)^{-1}$, where \hat{W} is the matrix W evaluated at $\hat{\beta}$.
- If ϕ is unknown, it should be replaced by a consistent estimator, such as the MLE or the estimator based on the method of moments.
- For normal distribution with identity link we have $g(\mu) = \mu$, so that $g'(\mu) = 1$. Moreover, $V(\mu) = 1$, $\phi = \sigma^2$ and $\mu_i = x_i^T \beta$. The likelihood equations are

$$\sum_{i=1}^n \frac{(y_i - x_i^T \beta)x_{ij}}{\sigma^2} = 0 \text{ that leads to usual LSE}$$

Some models

Normal Linear Model

We have $g(\mu) = \mu$, so that $g'(\mu) = 1$. Moreover, $V(\mu) = 1$, $\phi = \sigma^2$ and $\mu_i = \mathbf{x}_i^T \beta$. The likelihood equations are

$$\sum_{i=1}^n \frac{(y_i - \mathbf{x}_i^T \beta)x_{ij}}{\sigma^2} = 0 \quad j = 1, 2, \dots, p.$$

Simplifying σ^2 and using matrix notation, the above equations reduce to the usual LS equations: $X^T(\mathbf{y} - X\beta) = 0$ or, equivalently,

$$X\beta = \mathbf{y} \quad \rightarrow \quad X^T X\beta = X^T \mathbf{y} \quad \text{that leads to} \quad \hat{\beta} = (X^T X)^{-1} X^T \mathbf{y}$$

Poisson regression

We have $g(\mu) = \log \mu$, so that $g'(\mu) = 1/\mu$. Moreover, $V(\mu) = \mu$, $\phi = 1$ and $\mu_i = \mathbf{x}_i^T \beta$. The likelihood equations are

$$\sum_{i=1}^n (y_i - e^{\mathbf{x}_i^T \beta})x_{ij} = 0 ,$$

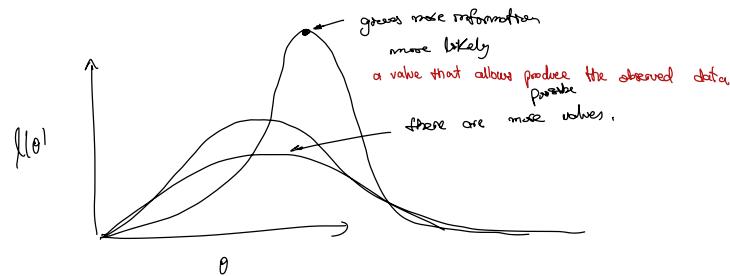
which are generally nonlinear in β . In view of this, an explicit solution does not exist in general.

Approaches for general data

$$1. \quad Y_i \begin{cases} \text{cont.} & \text{Normal or Gamma} \\ \text{discrete} & (\text{count}) \text{ Poisson} \\ \text{Binomial} & (\text{Bernoulli}) \end{cases} \Rightarrow \text{EF [dispersion]} \quad w/ \quad f(y; \theta, \phi) = \exp\left(\frac{y\theta - b(\theta)}{\phi} + c(y, \phi)\right)$$

Fisher information

\uparrow curvature \uparrow more information



Solution of the likelihood equations

An iterative algorithm

- Likelihood equations for GLMs do not usually have explicit solutions. They should be solved by iterative methods.
- For the GLM there exists the possibility to use a simple algorithm for the solution of the likelihood equations: the MLEs of the parameter β in the linear predictor can be obtained by iterative weighted least squares.
- Starting with appropriate initial value $\hat{\beta}^{(0)}$ and obtaining a sequence $\hat{\beta}^{(1)}, \hat{\beta}^{(2)}, \dots$, using a rule to update $\hat{\beta}^{(t+1)}$ with $\hat{\beta}^{(t)}$, until that the value of

$$\|\hat{\beta}^{(t+1)} - \hat{\beta}^{(t)}\|$$

is sufficiently small ($< \epsilon$).

Newton-Raphson and Fisher scoring

- Let

$$X^T \beta - \eta = g(\mu) = g(E[Y|X])$$

$$\ell_* = \left(\frac{\partial \ell}{\partial \beta_1}, \dots, \frac{\partial \ell}{\partial \beta_p} \right)^T$$

be the *score* vector. We want to solve the equation

$$\ell_* = \ell_*(\beta) = 0 .$$

$\mathbf{I} \rightarrow \mathbf{x}$
 $\mathbf{U} = \mathbf{J} \beta \rightarrow \mathbf{\hat{x}}$
 $\mathbf{J} = \mathbf{J}^2 \beta \rightarrow \mathbf{\hat{x}}$

$\frac{\mathbf{I}}{\mathbf{J}} \rightarrow \begin{bmatrix} \mathbf{\hat{x}} \\ \mathbf{\hat{x}} \end{bmatrix} = \frac{\mathbf{m}_S}{\mathbf{m}_S} = \mathbf{s}^2 \mathbf{I}$

- The Newton-Raphson method is based on the updating rule at the $(t + 1)$ -th iteration

observed inf. matrix

$$\hat{\beta}^{(t+1)} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}^{(t)}$$

$$\ell^{(t+1)} = \ell^{(t)}(\beta^{(t)}, \mathbf{y}^{(t)}) \quad (2)$$

with $\ell_*^{(t)} = \ell_*(\hat{\beta}^{(t)})$.

Maybe useful

..

- The observed information can be replaced by the expected Fisher information $i(\beta)$. This algorithm takes the name of *Fisher scoring* method. This maintains the convergence of the algorithm and simplifies the expressions (if the canonical link function is used, the two expressions coincide).

Developing the algorithm

Expression (2) is equivalent to

$$\text{expected mle. matrix } i(\hat{\beta}^{(t)})\hat{\beta}^{(t+1)} = i(\hat{\beta}^{(t)})\hat{\beta}^{(t)} + \ell_*^{(t)} .$$

Remember that the (j, k) -th element of $i(\beta)$ is

$$\sum_{i=1}^n \frac{x_{ij}x_{ik}}{\text{var}(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_j} \right)^2 ,$$

which gives $i(\beta) = \frac{X^T W X}{\phi}$, with $w_{ii} = \frac{1}{V(\mu_i)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2$.

In view of this, the right hand term can be written as

$$\begin{aligned} & (i^{(t)})\hat{\beta}^{(t)} + \ell_*^{(t)} \\ &= \sum_{k=1}^p \sum_{i=1}^n \frac{x_{ij}x_{ik}}{\text{var}(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_k} \right)^2 \hat{\beta}_k^{(t)} + \sum_{i=1}^n \frac{(y_i - \mu_i)x_{ij}}{\text{var}(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_j} \right) \\ &= X^T W^{(t)} s^{(t)} , \end{aligned}$$

Weighted Least Squares

- where $\mathbf{s}^{(t)}$ is a vector with elements

$$i(\beta) = \frac{x^T W x}{\phi}, \text{ with } w_{ii} = \frac{1}{V(\mu_i)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2.$$

$$s_i^{(t)} = \sum_{k=1}^p x_{ik} \hat{\beta}_k^{(t)} + (y_i - \mu_i) \left(\frac{\partial \mu_i}{\partial \eta_i} \right),$$

and all the involved quantities are evaluated at $\hat{\beta}$.

- Therefore, it is possible to arrive to the expression

$$X^T W^{(t)} X \hat{\beta}^{(t+1)} = X^T W^{(t)} \mathbf{s}^{(t)}. \quad (3)$$

- Clearly, the parameter ϕ simplifies.
- The above expression has the form of the normal equations for a LM obtained with weighted least squares, except that the equation above has to be solved iteratively because in general \mathbf{s} and W depend on β .

Iterative Weighted Least Squares (IWLS)

- Indeed, the Newton-Raphson iteration is

$W^{(t)}$ and $s^{(t)}$ depend on $\hat{\beta}^{(t)}$

$$\hat{\beta}^{(t+1)} = (X^T W^{(t)} X)^{-1} X^T W^{(t)} s^{(t)}. \quad (4)$$

- Each iteration of the algorithm is equivalent to a weighted least squares estimate, in which the adjusted dependent variable and the weights depend on the fitted values, for which only current estimates are available.
- The algorithm has two main steps:
 - Given $\hat{\beta}^{(t)}$, compute $s^{(t)}$ and $W^{(t)}$;
 - Obtain $\hat{\beta}^{(t+1)}$ through (4).

To start the algorithm a simple and convenient choice of the starting values is $s^{(0)} = g(Y_i)$ and $W^{(0)}$ equals to the identity matrix.

Estimating the dispersion parameter ϕ

- For the LM, the estimation of β is independent from the value of the variance σ^2 . A similar situation holds for the dispersion parameter ϕ in GLMs.
- Obviously, the MLE of ϕ , with β replaced by $\hat{\beta}$, could be used.
- Also estimators based on the method of moments are often used for ϕ .
- Since $\text{var}(Y_i) = \phi V(\mu_i)$ or, equivalently, since $\frac{E((Y_i - \mu_i)^2)}{V(\mu_i)} = \phi$ if β is known, an unbiased estimator of ϕ is

$$\hat{\phi} = \frac{1}{n} \sum_{i=1}^n \frac{(y_i - \mu_i)^2}{V(\mu_i)} .$$

If the expected values μ_i are replaced with their estimates based on $\hat{\beta}$, then the following adjusted consistent estimator is obtained

$$\hat{\phi} = \frac{1}{n-p} \sum_{i=1}^n \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)},$$

ML invariant and
df for variance

where

$$\hat{\mu}_i = g^{-1}(\mathbf{x}_i^T \hat{\beta}) .$$

Exploiting asymptotic normality of $\hat{\beta}$

- For n large, the asymptotic distribution of the MLE is

$$\hat{\beta} \sim N_p(\beta, [i(\hat{\beta})]^{-1}) \quad \text{where} \quad i(\hat{\beta}) = \frac{X^T \hat{W} X}{\phi}$$

We do still have high variance for collinear vars than can be tackled w/ regularization

↳ other algorithms.

with \hat{W} computed at $\hat{\beta}$. The estimated asymptotic variances are the diagonal elements of the matrix $(X^T \hat{W} X)^{-1} \phi$.

- Using the asymptotic distribution of $\hat{\beta}$, a confidence interval for β_j with approximate level $1 - \alpha$ is

$$\hat{\beta}_j \pm z_{1-\alpha/2} \sqrt{\phi[(X^T \hat{W} X)^{-1}]_{j,j}} .$$

First we have collinear ~~not~~ vars
are not collinear

→ V.I.F.

and anova test for nested models

- and the statistic $\frac{\hat{\beta}_j}{\sqrt{\phi[(X^T \hat{W} X)^{-1}]_{j,j}}}$ can be used to test significance of a single β_j

Model Evaluation



Comparing nested models

- Let us start by considering two nested GLMs. Let denote the models by M_C and M_R , such that $M_R \subset M_C$. Specifically, the current model M_C contains p parameters and the reduced model M_R contains p_0 parameters, where $p > p_0$.
- Consider the following partition of $\beta = (\beta_{MR}, \beta_{MC})$, where $\beta_{MR} = (\beta_1, \dots, \beta_{p_0})$ and $\beta_{MC} = (\beta_{p_0+1}, \dots, \beta_p)$. Suppose we want to test the following hypothesis

$$H_0 : \beta_{MC} = 0 \quad \text{against} \quad H_1 : \beta_{MC} \neq 0 .$$

- The criterion we will adopt to compare M_C and M_R is the likelihood ratio

$$W = 2\{\ell(\hat{\beta}) - \ell(\hat{\beta}_{MR})\} . \quad \begin{aligned} &= 2 \log \frac{\mathcal{L}(\hat{\beta})}{\mathcal{L}(\hat{\beta}_{MR})} \end{aligned}$$

The deviance in LMs

- In normal LMs, with σ^2 known, the likelihood ratio is a function of the deviance (sum of square of residuals) $D = SSE = \sum_i(y_i - \hat{y}_i)^2$ of the two models. When comparing two nested models ($M_R \subset M_C$), the likelihood ratio criterion will lead to rejection of H_0 for large values of the following statistic

$$W = 2\{\ell(\hat{\beta}) - \ell(\hat{\beta}_{MR})\} = \frac{D_{MR} - D}{\sigma^2},$$

where $D_{MR} = SSE_{H_0}$ and $D = SSE$ are sums of square of residuals in the reduced and current models respectively.

- When H_0 holds this statistic has a $\chi^2_{p-p_0}$ distribution.

What distribution follows W when H_0 don't hold?

LR test

- Like Normal LMs, we look for an interpretation of (log-)likelihood ratio in GLMs so that the relationship between the two classes of models is clear. It will help if we can define an analogous quantity as deviance in LMs.
- Log-likelihood for a GLM is

$$\ell(\beta) = \sum_{i=1}^n \ell_i(\beta) ,$$

where

$$\ell_i(\beta) = \frac{y_i \theta_i - b(\theta_i)}{\phi} + c(y_i, \phi) .$$

- With nested GLM, the statistic

$$W = 2\{\ell(\hat{\beta}) - \ell(\hat{\beta}_{MR})\}$$

is asymptotically distributed as a $\chi^2_{p-p_0}$ when H_0 holds.

The saturated model

- Analogy with Normal LM can be kept by introducing likelihood associated to a model where there are as many parameters as observations. This model will be denoted as **saturated or full**.
- At the other extreme there is a model as simple as possible, *i.e.*, a model where a single parameter represents a common μ for all the y_i

A ““good”” model usually stands between these two extremes since a saturated model is uninformative being unable to summarize data: it just repeats them in full, and a null model is usually too simple to be useful. We should seek a balance between conflicting goals of parsimony and goodness of fit.

- Saturated model is defined as:
 - ↪ a GLM having the same distribution and link function of the current model;
 - ↪ but a number of parameter equal to n (or to the number of different groups sharing the same x vector).
- We can evaluate likelihood function for the saturated model and the current model at the value of the MLE obtained in both cases ($\tilde{\theta}$ and $\hat{\theta}$ respectively). If the current model fits the data, $\ell(\tilde{\theta})$ should be very similar to $\ell(\hat{\theta})$. In case of a poor fit then $\ell(\hat{\theta})$ should be much smaller than $\ell(\tilde{\theta})$.

The deviance in GLMs

- Formally, the quantity

$$D(y; \hat{\theta}) = 2\phi\{\ell(\tilde{\theta}) - \ell(\hat{\theta})\} = \phi \sum_{i=1}^n D_i$$

with $D_i = 2\{y_i(\tilde{\theta}_i - \hat{\theta}_i) - b(\tilde{\theta}_i) + b(\hat{\theta}_i)\}$, is called *deviance function* of the model and

$$\frac{D(y; \hat{\theta})}{\phi} = \sum_{i=1}^n D_i \tag{5}$$

is the *scaled deviance*: note that it is always non negative.

This quantity is small for good models and is large when the current model gives a poor fit. Behaviour of deviance is equivalent to that of SSE in LMs.

- $\ell(\tilde{\theta})$ is the log-likelihood obtained by letting $\mu_i = b'(\theta_i) = y_i (\Leftrightarrow (\partial \ell_i / \partial \theta_i) = 0)$, so the saturated model has $p = n$ parameters.
- The saturated model is useless but $\ell(\tilde{\theta})$ provides a benchmark to compare log-likelihood of the current model.

Example: Normal regression model

Since Normal LMs are GLMs with identity link functions we can show that calculating the above defined deviance we give in this case the same result obtained by standard theory for goodness of fit evaluation in Normal LMs.

- $Y_i \sim N(\mu_i, \sigma^2)$, $b'(\theta) = \frac{\theta^2}{2}$, $\theta = \mu = b'(\theta)$ and $\phi = \sigma^2$.
- $\ell(\theta) = -\frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mu_i)^2$
- For the saturated model $\tilde{\mu}_i = y_i$, and

$$\ell(\tilde{\theta}) = -\frac{n}{2} \log \sigma^2 .$$

- For the current model $\hat{\mu}_i = \mathbf{x}_i^T \hat{\beta}$, and

$$\ell(\hat{\theta}) = -\frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \hat{\mu}_i)^2$$

- Scaled deviance is

$$D(y; \hat{\theta}) = \sum_{i=1}^n \frac{(y_i - \hat{\mu}_i)^2}{\sigma^2}$$

Poisson

- $Y_i \sim Poisson(\mu_i)$, $b(\theta_i) = e^{\mu_i} = b'(\theta_i)$, $\phi = 1$, $\log \mu_i = \mathbf{x}_i^T \boldsymbol{\beta}$
- $\ell(\theta) = \sum_{i=1}^n y_i \log \mu_i - \sum_{i=1}^n \mu_i$
- For the saturated model $\tilde{\mu}_i = y_i$, and

$$\ell(\tilde{\theta}) = \sum_{i=1}^n y_i \log y_i - \sum_{i=1}^n y_i .$$

- For the current model $\log \hat{\mu}_i = \mathbf{x}_i^T \hat{\boldsymbol{\beta}}$, and

$$\ell(\hat{\theta}) = \sum_{i=1}^n y_i \log \hat{\mu}_i - \sum_{i=1}^n \hat{\mu}_i .$$

- So deviance is $D(y; \hat{\theta}) = 2 \left(\sum_{i=1}^n y_i \log \frac{y_i}{\hat{\mu}_i} - \sum_{i=1}^n y_i + \sum_{i=1}^n \hat{\mu}_i \right)$

Binomial

- $Y_i \sim Bin(1, \pi_i)$, con $\pi_i = Pr(Y_i = 1) = E(Y_i) = \mu_i$
- $\ell(\theta) = \sum_{i=1}^n (y_i \log \pi_i + (1 - y_i) \log(1 - \pi_i))$
- For the saturated model $\tilde{\mu}_i = y_i$ and

$$\ell(\tilde{\theta}) = \sum_{i=1}^n (y_i \log y_i + (1 - y_i) \log(1 - y_i)) .$$

- For the current model $logit(\hat{\mu}_i) = \mathbf{x}_i^T \hat{\beta}$ and

$$\ell(\hat{\theta}) = \sum_{i=1}^n (y_i \log \hat{\pi}_i + (1 - y_i) \log(1 - \hat{\pi}_i)) .$$

- The deviance is

$$D(y; \hat{\theta}) = 2 \sum_{i=1}^n \left(y_i \log \frac{y_i}{\hat{\pi}_i} + (1 - y_i) \log \frac{1 - y_i}{1 - \hat{\pi}_i} \right) .$$

Comparing nested models

- Considering two nested models M_C and M_R , likelihood ratio test is

$$\begin{aligned} W &= 2 \left\{ \ell(\hat{\beta}) - \ell(\hat{\beta}_{MR}) \right\} \\ &= \frac{D(Y, \hat{\theta}_{MR}) - D(Y, \hat{\theta})}{\phi}, \end{aligned}$$

as $n \rightarrow \infty$ it is distributed $\chi^2_{p-p_0}$ when H_0 holds.

- So to test if reduced model can be accepted we can compare

$$W = \frac{D(Y, \hat{\theta}_{MR}) - D(Y, \hat{\theta})}{\phi}$$

with the quantiles of the distribution $\chi^2_{p-p_0}$. We reject H_0 for large values of the statistic (or for a small *p-value*).

Residual Deviance

- It is important to note that since deviance is defined as a function of the difference arising from a log-likelihood ratio of two nested model one is tempted to use the same criteria for evaluating if deviance of the current model is significantly small. One can look if value of deviance is not large enough when compared to a χ^2_{n-p} .
- In this last case standard asymptotic theory could not work when the number of parameter in the saturated model is not fixed as n goes to infinity.

Nonetheless the criterion could work when the number of parameters is fixed: this is, for instance, the case of a binomial model for grouped data or a Poisson model with factors as the only covariates (as it happens in log linear model from contingency tables).

- In some cases (the most notable being binomial and Poisson) the dispersion parameter is fixed to 1.
- When dispersion parameter ϕ is not known another consistent estimate of it must be considered

$$\hat{\phi} = \frac{D(Y, \hat{\theta})}{(n - p)}$$

and under mild conditions the result stated above still works.

Model selection

- Model selection strategies can exploit the tools defined above to explore which combination of explanatory variables leads to a satisfactory model.
- So one can consider a stepwise backward search by starting with a model that includes all the covariates and then consider a set of reduced sub models obtained by removing certain variables (backward selection). In order to choose among models, one can consider the sub-model obtained by deleting variables with a large p -value.
- A forward search starts from the null model (usually the one including only the intercept) and (groups of) variables are included if the p -values associated are small.
- A combination of the two strategies can also be considered.
- To compare models also the well known criteria AIC and BIC can be used. For instance, in this case $AIC = -2\ell(\hat{\theta}) + 2p$ where p is the number of parameters of the model (when dispersion parameter is known) and one chooses the model where AIC is smaller.

Residuals in GLM

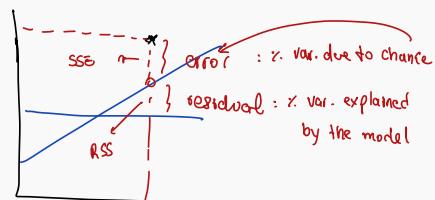
- Let us recall the basic ideas in using residual analysis in LMs:
 - residuals are easily defined as the difference between the observed datum and the estimated systematic part of the model: this step is less natural in GLM.
 - residuals tell us if there are symptoms of systematic differences between observed and fitted values (i.e. plot of residuals against fitted values, or against covariates)
 - residuals help us recognizing discrepancies between few data and the rest (ouliers detection, evaluation of leverage: hat matrix, case deletion measures -Cook's distance-, jackknife residuals, etc.)
- Some of these ideas can be generalized in GLMs.
- A straight extension of the concept of **standardized residual** is given by

$$r_{Pi} = \frac{Y_i - \hat{\mu}_i}{\sqrt{\hat{\phi}V(\hat{\mu}_i)}}, \quad (6)$$

called *Pearson residuals*. The definition (6) resembles that for residuals in LMs based on the estimation of the error term ϵ_i .

Deviance residuals

- Recall that in GLMs ϵ_i does not exist in general, so we can measure the contribution of each observation to deviance. This is analogous to LMs where SSE is defined as



$$SSE = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (Y_i - \mathbf{x}_i^T \hat{\beta})^2 ,$$

$$\begin{aligned} Y_i &= \mathbf{x}_i^T \beta + \varepsilon_i \neq \hat{Y}_i = \mathbf{x}_i^T \hat{\beta} + e_i \\ e_i &= (\hat{Y}_i - \mathbf{x}_i^T \hat{\beta}) \end{aligned}$$

"error"

while in GLMs a similar quantity is the deviance. Recall that deviance is defined as

$$D(y, \hat{\theta}) = \sum_{i=1}^n D_i .$$

Large individual contributions to total deviance D_i reflect data that are not properly reproduced by the model. Let us define

$$r_{Di} = \operatorname{sgn}(y_i - \hat{\mu}_i) \sqrt{D_i} ,$$

that is called *deviance residual* of the model.

For large n it is possible to show that $r_{Pi} \approx r_{Di}$.

Other residuals, such as Anscombe residuals, are also defined for GLMs.

Residual analysis

- Actually if the model is valid, residuals of any type, possibly scaled by $\hat{\phi}$, will have a distribution that can be (loosely) approximated by a $N(0, 1)$. This suggest to use standard graphical tools, like
 - ↪ normal probability plot of the residuals;
 - ↪ plot of residuals against the fitted values \hat{Y}_i ;
 - ↪ plot of residuals against explanatory variablesto check assumptions.

- It is also possible to generalize the Hat matrix H to check influence and leverage of residuals. Recall that H in LMs is such that $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$ and
$$H = X(X^T X)^{-1}X^T.$$
- Generalized hat matrix is similarly obtained as
$$H = W^{\frac{1}{2}}X(X^T W X)^{-1}X^T W^{\frac{1}{2}}$$
where W is substituted by \hat{W} .
- A generalization of the Cook's distances is also possible.

Quasi-likelihood

More on quasi-likelihood

- For LMs the method of LS allows to obtain estimates of the regression parameters without the specification of a probabilistic model.
- The method of LS requires only the specification of the relation between the expected value of the response variable and the linear predictor, and the specification of the variance of the error term, which is not related to the expected value:

$$E(Y_i) = \mu_i = \eta_i \quad \text{var}(Y_i) = \sigma^2$$

- Also for the GLMs it is possible to specify only these two relations (assuming that the variance function $V(\mu_i)$ is known).
- Indeed, the likelihood equation for β

$$\sum_{i=1}^n \frac{y_i - \mu_i}{V(\mu_i)g'(\mu_i)} x_{ij} = 0 , \quad j = 1, \dots, p ,$$

is an unbiased estimating equation provided that $E(Y_i) = \mu_i = g^{-1}(\eta_i)$.

- In other words, this means that the parametric assumption $Y_i \sim EF(\cdot, \phi)$ could not even be satisfied. Only the assumption about expectations is essential:
 $\mu_i = E(Y_i) = g^{-1}(\eta_i)$
- The only distributional feature that must be known in order to calculate the estimating equation is the variance function $V(\mu)$.

Quasi-likelihood model

- Under suitable regularity conditions, the likelihood equations for a GLM give estimates for the coefficients β which maintain several properties, also if the parametric assumptions of Y_i are substituted with weaker **second order assumptions**:
 1. $g(\mu_i) = g(E(Y_i)) = \eta_i, \quad i = 1, \dots, n$
 2. $\text{var}(Y_i) = \phi V(\mu_i), \quad i = 1, \dots, n$
 3. $\text{cov}(Y_i, Y_j) = 0, \text{ if } i \neq j.$
- The semi-parametric statistical model specified by assumptions 1–3 is called **quasi-likelihood model**.
- If $V(\mu) = 1$ and $g(\mu) = \mu$, the assumptions 1–3 match the usual second order assumptions of the classical LM.
- On the other hand, if $V(\mu) = \mu^2$ we obtain a multiplicative model, $Y_i = \mu_i \epsilon_i$, with $E(\epsilon_i) = 1$ and $\text{var}(\epsilon_i) = \phi$.

Quasi-likelihood equations

- Gauss-Markov (BLUE) optimality of LS extends to quasi-likelihood estimates and it has minimum asymptotic variance among estimating equations that are linear (in Y) and unbiased
- Indeed, the likelihood equation for β

$$q(y; \beta) = \sum_{i=1}^n q(y_i; \beta) = \sum_{i=1}^n \frac{y_i - \mu_i}{V(\mu_i)g'(\mu_i)} x_{ij} = 0 , \quad j = 1, \dots, p ,$$

behaves like a score vector. Specifically:

$$E(q(Y; \beta)) = 0, \quad \text{and} \quad \text{var}(q(Y; \beta)) = -E(\partial q(Y; \beta)/\partial \beta) .$$

- Quasi likelihood estimators shares many properties of a proper likelihood: the quasi-MLE β is asymptotically normal, the quasi-likelihood ratio statistic has a null chi-squared distribution.

Quasi-likelihood and overdispersion

- The assumptions 1–3 offer an increase in flexibility with respect to the usual parametric specifications based, respectively, on the Poisson, binomial or exponential distributions.
- In practice, there are situations in which the dispersion parameter does not agree with the assumed exponential family.
- For example, for the binomial or Poisson distributions we have $\phi = 1$, but data could show agreement with $\phi > 1$.
- In this case we have *overdispersion*, i.e. the variance of Y is greater than its theoretical value, and it is more plausible to assume $\text{var}(Y_i) = \phi V(\mu_i)$, with $\phi > 1$. For example, for proportions, it can be assumed that $\text{var}(Y) = \phi n\pi(1 - \pi) > n\pi(1 - \pi)$, with $\phi > 1$, where $n\pi(1 - \pi)$ is the variance of a binomial distribution.
- In general, the quasi-likelihood approach allows to deal with *overdispersion problems*: it is possible to specify $\text{var}(Y_i)$ so that there is more variability with respect to the exponential family.

GLM: Extensions and recent development

GLM: Extensions and recent development

- Generalized linear models and the relevant theory have been introduced in the “80ies.
- They have been extended in many directions to take into account more complex data structures
- It has been introduced the use of quasi-likelihood estimators that allows to specify only the mean and the variance of the response Y . This leaves more flexibility and is a simple strategy to cope with overdispersion.
- It has been extended to multivariate responses.
- it has been estended to take into account nested structure of the data and the lack of independence that can arise in those cases (Generalized Linear Mixed Models GLMM).
- Procedures for regularization, such as the LASSO, can be adopted also in case of GLMs
- it has been estended to take into account non linear functions of the covariates and to estimate this functions non parametrically (Generalized additive models GAMs).
- Use of a different inferential approach, such as Bayesian inference, have been considered.

Bootstrap Methods

(An introduction)

N. Torelli, G. Di Credico, V. Gioia

2023

University of Trieste

For inference we assume that the sample comes from a population w/ some distribution.

$$Y \sim N(\mu, \sigma^2) \quad \begin{matrix} \text{assumption} \\ \downarrow \text{unknown parameters} \end{matrix}$$

We approximate θ by $\hat{\theta} = g(y_1, \dots, y_n)$

$$\text{We assess it w/ } \sqrt{E(\hat{\theta} - \theta)^2} = \sqrt{V(\hat{\theta})} \rightarrow \text{s.e. } \hat{\theta}$$

$\hat{\theta}$ unbiased

Parametrization is not always a good assumption because we don't know the distribution of the test statistics.

$$\text{e.g., not } \bar{Y} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

and thus we can't obtain $\hat{\theta} \pm z_{1-\frac{\alpha}{2}} \text{s.e. } \hat{\theta}$ because we have no easy way to compute $z_{1-\frac{\alpha}{2}}$

We could use simulations to make approximations

Jackknife: reuse data to analyze quality of statistics. \Rightarrow limited to understand s.e. $\hat{\theta}$

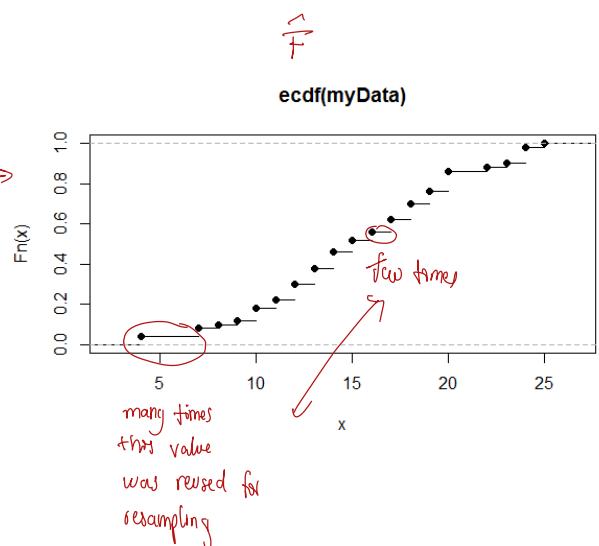
Bootstrap: reuse samples in scheme.

Why? Theoretical distribution might be complicated or unknown \times

By resampling we can obtain a sampling distribution \hat{F}

The empirical cumulative allows understand how \hat{F} is obtained \rightarrow

Comments: data points must be iid but can be extended, e.g., time series



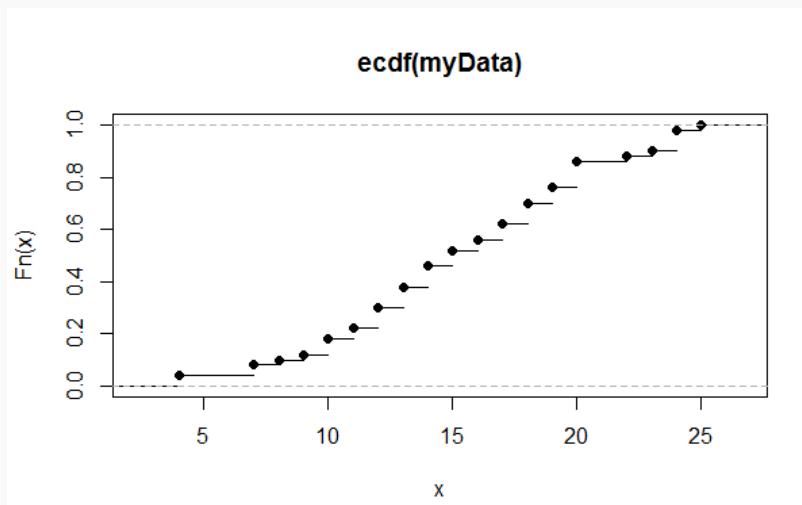
Resampling methods

The nonparametric bootstrap

The parametric bootstrap

Bootstrap-based confidence intervals

Resampling methods



Using the original dataset to resample implies that some values will be repeated and this makes some values "more probable" than others.

Each selection is independent and probabilities remain the same for each draw.

Consider drawing balls from a box w/ or w/out replacement

w/ rep.: the prob. of drawing a ball is high than in the case of

w/out rep.: because the number of elements to choose from reduces.

The idea of resampling methods

- Resampling methods are **computer-intensive methods** that employ simulation to carry out inferential conclusions for the data available.

In some sense, they replace mathematical formulas with computer simulation, though proving their validity requires quite sophisticated mathematics.

There are several such methods, but by far the most important are **bootstrap methods**. They are relatively modern, but their initial development predates the modern computer age!

(*Note: this lecture follows in particular the CASI book*)

The jackknife: introduction

The **jackknife** is, so to speak, the ancestor of the bootstrap. Its main usage is to obtain a nonparametric estimate of the standard error of an estimate, resulting in a simpler alternative to the delta method for complex functions of model parameters.

Let us consider a random sample y_1, \dots, y_n , with $Y_i \sim F$, for some distribution F .

We are interested in a real-valued statistic (parameter estimate) $\hat{\psi} = s(\mathbf{y})$, where $s(\cdot)$ is a given function of the n observations.

The jackknife: details

Let $\mathbf{y}_{(i)}$ be the sample without the i -th observation y_i

$$\mathbf{y}_{(i)} = (y_1, y_2, \dots, y_{i-1}, y_{i+1}, \dots, y_n)$$

so that $\hat{\psi}_{(i)} = s(\mathbf{y}_{(i)})$ is the corresponding statistic of interest.



The jackknife estimate of standard error for $\hat{\psi}$ is

$$\widehat{\text{SE}}_{\text{jack}} = \left[\frac{n-1}{n} \sum_{i=1}^n \left(\hat{\psi}_{(i)} - \hat{\psi}_{(\cdot)} \right)^2 \right]^{1/2}, \quad \text{with} \quad \hat{\psi}_{(\cdot)} = \frac{1}{n} \sum_{i=1}^n \hat{\psi}_{(i)}.$$

Note: the formula works also when each observation y_i is multidimensional (e.g. for regression models).

The jackknife: comments

1. When $\hat{\psi} = \bar{y}$, with some algebra we obtain that $\widehat{SE}_{jack} = s/\sqrt{n}$, the usual estimated standard error of the mean. (This is the reason to introduce the factor $(n - 1)/n$ in the definition).
 $\widehat{SE}_{JN} > S.E.$
2. The jackknife standard error is upwardly biased as an estimate of the true standard error. The bias disappears with larger n .
3. The important property of the procedure is that the definition can be applied to any statistic of interest, even very complex ones.

We only need an algorithm to compute $s(y)$: **computer power replaces the theoretical Taylor series calculations of the delta method.**

An example

Let us consider the standard error of the correlation coefficient for a random sample of bivariate normal data $(x_1, y_1)^\top, \dots, (x_n, y_n)^\top$:

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

This is an estimate of the corresponding parameter $\text{cov}(X_i, Y_i)/(\sigma_x \sigma_y)$, and we can compute its standard error by the multidimensional version of the delta method.

The related formula is not exactly friendly (from CASI book, page 157)

$$\widehat{\text{se}}_{\text{taylor}} = \left\{ \frac{\hat{\theta}^2}{4n} \left[\frac{\hat{\mu}_{40}}{\hat{\mu}_{20}^2} + \frac{\hat{\mu}_{04}}{\hat{\mu}_{02}^2} + \frac{2\hat{\mu}_{22}}{\hat{\mu}_{20}\hat{\mu}_{02}} + \frac{4\hat{\mu}_{22}}{\hat{\mu}_{11}^2} - \frac{4\hat{\mu}_{31}}{\hat{\mu}_{11}\hat{\mu}_{20}} - \frac{4\hat{\mu}_{13}}{\hat{\mu}_{11}\hat{\mu}_{02}} \right] \right\}^{1/2} \quad (10.10)$$

where

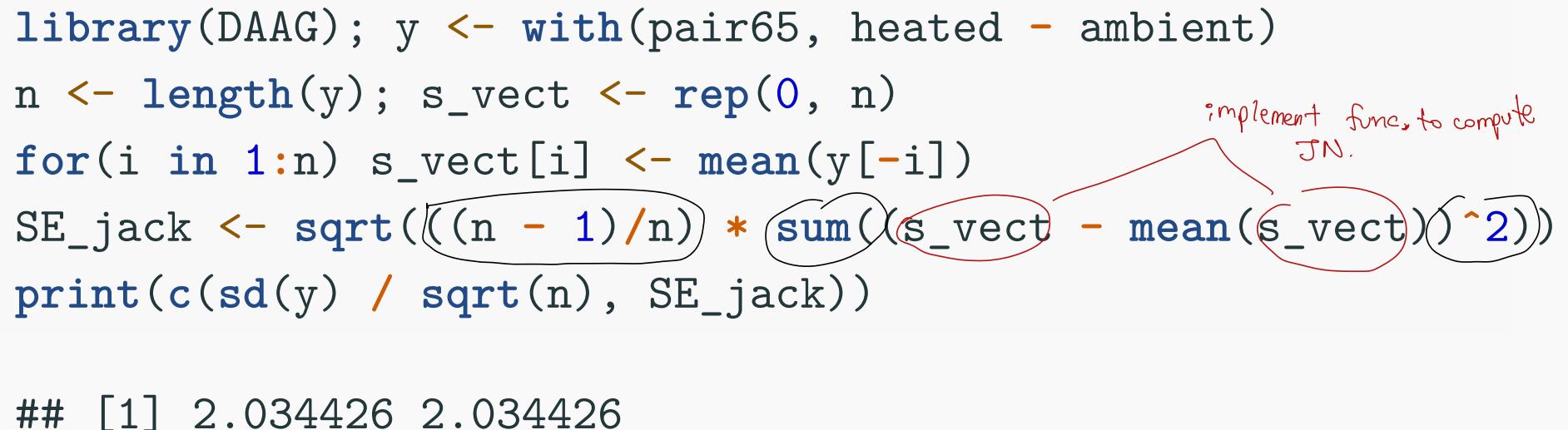
$$\hat{\mu}_{hk} = \sum_{i=1}^n (x_i - \bar{x})^h (y_i - \bar{y})^k / n. \quad (10.11)$$

R lab: the jackknife at work I

As a first example, let us consider the case of the sample mean.

```
library(DAAG); y <- with(pair65, heated - ambient)
n <- length(y); s_vect <- rep(0, n)
for(i in 1:n) s_vect[i] <- mean(y[-i])
SE_jack <- sqrt(((n - 1)/n) * sum((s_vect - mean(s_vect))^2))
print(c(sd(y) / sqrt(n), SE_jack))

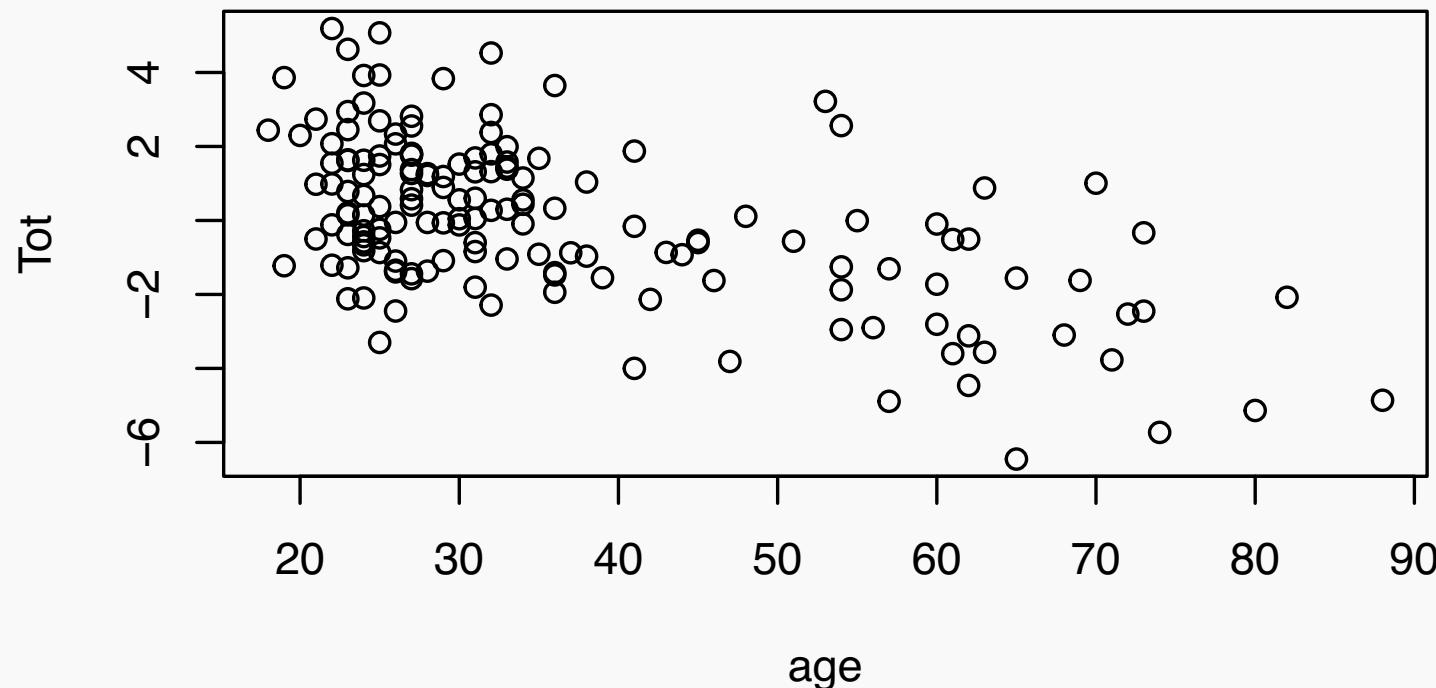
## [1] 2.034426 2.034426
```



R lab: the jackknife at work II

The second example concerns the correlation coefficient, and we use the same data set of the CASI book, the `kidneydata` dataset; here `Tot` is a composite measure of kidney overall function.

```
load("kidneydata.RData")
with(as.data.frame(kidneydata), plot(Tot ~ age))
```



R lab: the jackknife at work II

The standard error based on the delta method is stored in the variable `se_delta`, and it is taken from the CASI book. We note that `SE_jack` is slightly larger.

```
SE_delta <- 0.057
n <- nrow(kidneydata)
s_vect <- rep(0, n)
for(i in 1:n) s_vect[i] <- cor(kidneydata[-i, ])[1, 2]
SE_jack <- sqrt(((n - 1)/n) * sum((s_vect - mean(s_vect))^2))

print(c(SE_delta, SE_jack))
## [1] 0.05700000 0.05820618

$$\hat{S.E.}_{\delta} < \hat{S.E.}_{JN}$$

```

The nonparametric bootstrap

Introduction

As reported in the CASI book, the jackknife lies *between classical methodology and a full-throated use of electronic computation*, whereas the bootstrap is an undisputed *computer-intensive* statistical method.

Another important difference is that the bootstrap has a rather wide scope of application, while instead the jackknife is mainly used for standard errors.

A legendary beginning

The two methods are indeed related, as testified by the paper that introduced the bootstrap.

The Annals of Statistics
1979, Vol. 7, No. 1, 1–26

THE 1977 RIETZ LECTURE

BOOTSTRAP METHODS: ANOTHER LOOK AT THE JACKKNIFE

BY B. EFRON

Stanford University

We discuss the following problem: given a random sample $\mathbf{X} = (X_1, X_2, \dots, X_n)$ from an unknown probability distribution F , estimate the sampling distribution of some prespecified random variable $R(\mathbf{X}, F)$, on the basis of the observed data \mathbf{x} . (Standard jackknife theory gives an approximate mean and variance in the case $R(\mathbf{X}, F) = \theta(\hat{F}) - \theta(F)$, θ some parameter of interest.) A general method, called the “bootstrap,” is introduced, and shown to work satisfactorily on a variety of estimation problems. The jackknife is shown to be a linear approximation method for the bootstrap. The exposition proceeds by a series of examples: variance of the sample median, error rates in a linear discriminant analysis, ratio estimation, estimating regression parameters, etc.

The bootstrap idea

The bootstrap idea is very simple, and to illustrate it we start from the same problem introduced for the jackknife, namely the estimation of the standard error of $\hat{\psi} = s(\mathbf{y})$.

The standard error requires the computation of $\text{var}(\hat{\psi})$, something computable by drawing a large number of independent random samples from the true model F .

This is impossible, since F is unknown, so the bootstrap uses instead an estimate \hat{F} in place of F , and then it proceeds with the simulation.

In particular, when \hat{F} is the empirical distribution function (we met it in the very first class) a single simulated sample is obtained by **random selection with replacement** from the observed sample.

We have data \rightarrow we compute ecdf \hat{F} \rightarrow we resample w/ replacement



An example (from Boos and Stefanski, 2010, *Significance*)

Table 1. Random sample of 25 yearly incomes in thousands of dollars (ordered from lowest to highest)

1	4	6	12	13	14	18	19	20	22	23	24	26
31	34	37	46	47	56	61	63	65	70	97	385	

Figure 2: $n = 25$ adult male yearly incomes in a fictitious county

The data were actually generated from a known distribution, namely

$$Y_i \sim 30 \exp(Z_i), \quad Z_i \sim N(0, 1) \quad i = 1, \dots, 25$$

so that in this case we know the true distribution of the data (the population).

Example: two bootstrap samples

Nonparametric bootstrap treats the data of the previous table as the population and draws samples of size $n = 25$ (with replacement) from it.

```
y <- c(1, 4, 6, 12, 13, 14, 18, 19, 20, 22, 23, 24, 26, 31, 34,  
      37, 46, 47, 56, 61, 63, 65, 70, 97, 385)  
n <- length(y); set.seed(1989); B <- 10^4  
boot.sample <- matrix(NA, nrow = B, ncol = n) B samples of n size  
boot.sample[1,] <- sample(y, n, replace = TRUE)  
boot.sample[2,] <- sample(y, n, replace = TRUE)  
kable(boot.sample[1:2, 1:15])
```

22	20	4	34	70	13	24	70	13	63	18	12	46	6	23
65	31	24	4	34	65	37	19	34	4	70	70	1	97	97

The bootstrap at work



The bootstrap samples can be used to obtain an estimate of the standard error: denoted by $\hat{\psi}^{*b}$, $b = 1, \dots, B$ the statistic of interest for each bootstrap sample, we get

mean of bootstrap estimates

$$\widehat{SE}_{boot} = \left[\frac{1}{B-1} \sum_{b=1}^B \left(\hat{\psi}^{*b} - \hat{\psi}^{*\cdot} \right)^2 \right]^{1/2}, \quad \text{with} \quad \hat{\psi}^{*\cdot} = \frac{1}{B} \sum_{b=1}^B \hat{\psi}^{*b}.$$

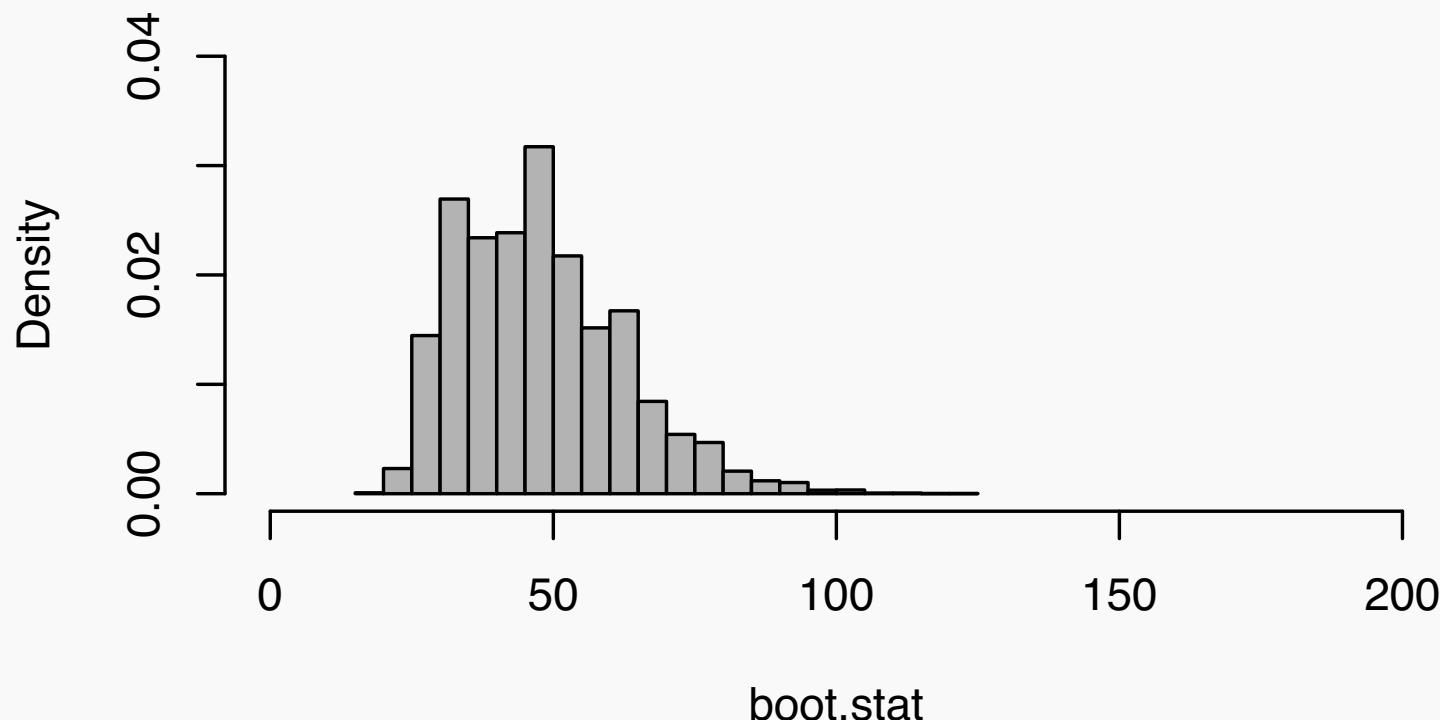
We can surely go beyond the computation of standard errors, since the set of bootstrap estimates $\hat{\psi}^{*1}, \dots, \hat{\psi}^{*B}$ can be used to approximate the distribution of $\hat{\psi}$.

What is $\hat{\psi}^{*b}$? b-th sample = $f(y_1, \dots, y_n)$ vector of size n containing
0, 1 or more than 1 element among y_1, \dots, y_n

The application of a function to the b-th bootstrap sample.

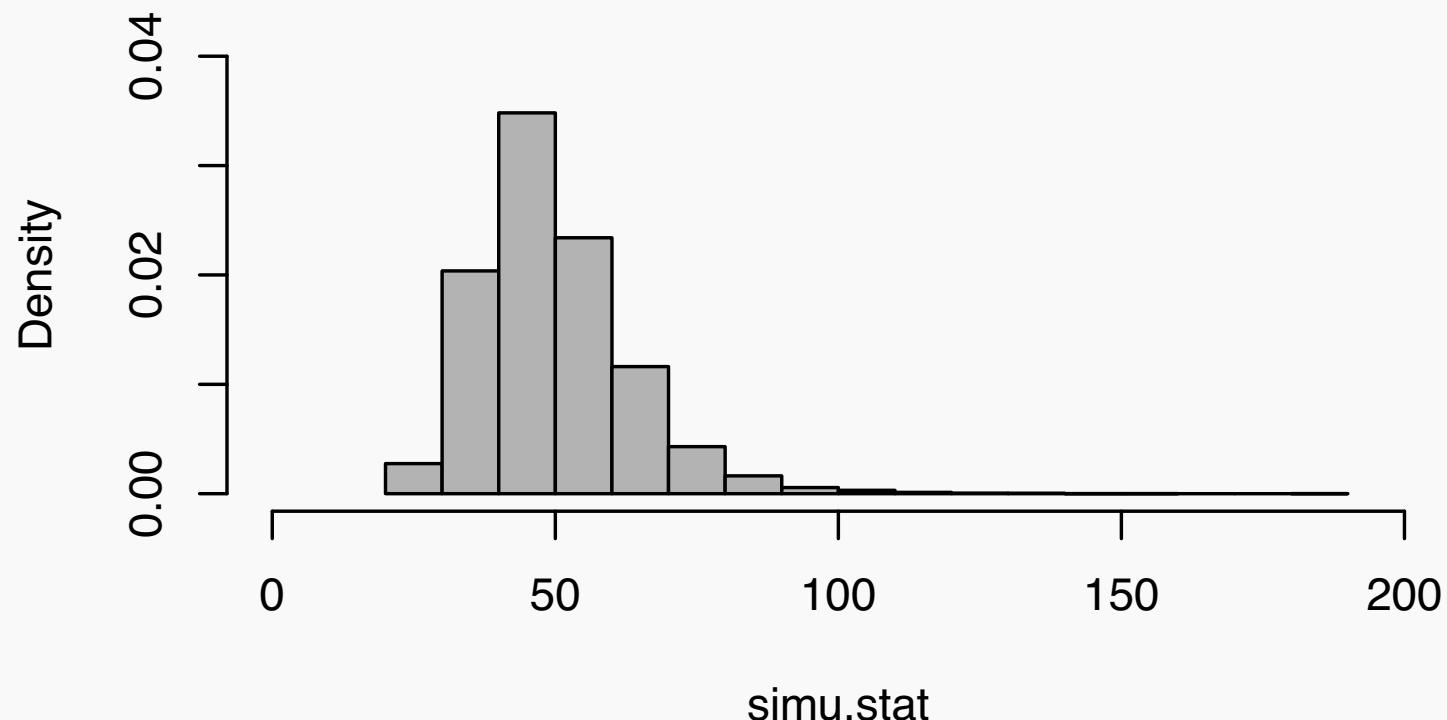
R lab: bootstrap distribution of $\hat{\psi} = \bar{Y}$

```
B <- 10^4; boot.sample <- matrix(NA, nrow = B, ncol = n)
for(i in 1:B) boot.sample[i,] <- sample(y, n, replace = TRUE)
boot.stat <- rowMeans(boot.sample)      Calculates the mean of several rows of a matrix
hist(boot.stat, main="", breaks=20, prob=TRUE, col=gray(0.7),
     xlim=c(0, 200), ylim=c(0, 0.04))
```



R lab: comparison with the true distribution

```
B <- 10^4; simu.sample <- matrix(NA, nrow = B, ncol = n)
for(i in 1:B) simu.sample[i,] <- mean(30 * exp(rnorm(n)))
simu.stat <- rowMeans(simu.sample)
hist(simu.stat, main="", breaks=20, prob=TRUE, col=gray(0.7),
      xlim=c(0, 200), ylim=c(0, 0.04))
```



Back to the standard error computation



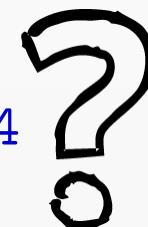
Provide B is large enough, the bootstrap-based standard error is unbiased, thus outperforming the jackknife.

```
n <- nrow(kidneydata); B <- 10^4  
s_vect <- rep(0, B)  
for(i in 1:B) {ind <- sample(1:n, n, replace = TRUE)  
    s_vect[i] <- cor(kidneydata[ind,])[1, 2]}
```

```
SE_boot <- sd(s_vect)
```

```
print(c(SE_delta, SE_jack, SE_boot))
```

```
## [1] 0.05700000 0.05820618 0.05820597
```



What is this doing?

$$\text{Corr}_{xy} = \frac{\text{cov}(x,y)}{\sqrt{x}\sqrt{y}}$$

$$\text{Corr}_{xx} = \frac{\text{cov}(x,x)}{\sqrt{x}^2} = \frac{E[\text{Var}[x]]}{\sqrt{x}^2} = 1$$

yes

U

1

=

1

U

1

=

1

U

1

=

1

U

1

=

1

U

1

=

1

U

1

=

1

U

1

=

1

U

1

=

1

U

1

=

1

U

1

=

1

U

1

=

1

U

1

=

1

U

1

=

1

U

1

=

1

U

1

=

1

U

1

=

1

U

1

=

1

U

1

=

1

U

1

=

1

U

1

=

1

U

1

=

1

U

1

=

1

U

1

=

1

U

1

=

1

U

1

=

1

U

1

=

1

U

1

=

1

U

1

=

1

U

1

=

1

U

1

=

1

U

1

=

1

U

1

=

1

U

1

=

1

U

1

=

1

U

1

=

1

U

1

=

1

U

1

=

1

U

1

=

1

U

1

=

1

U

1

=

1

U

1

=

1

U

1

=

1

U

1

=

1

U

1

=

1

U

1

=

1

U

1

=

1

U

1

=

1

U

1

=

1

U

1

=

1

U

1

=

1

U

1

=

1

U

1

=

1

U

1

=

1

U

1

=

1

U

1

=

1

U

1

=

1

U

1

=

1

U

1

=

1

U

1

=

1

U

1

=

1

U

1

=

1

U

1

=

1

U

1

=

1

U

1

=

1

U

1

=

1

U

1

=

1

U

1

=

1

U

1

=

1

U

1

=

1

U

1

=

1

U

1

=

1

U

1

=

1

U

1

=

1

U

1

=

1

U

1

=

1

U

More on the bootstrap idea

The bootstrap idea can be appreciated by noticing the parallel interpretation existing for the statistical model for the sample data

$$F \xrightarrow{\text{i.i.d.}} \mathbf{y} \xrightarrow{s(\cdot)} \hat{\psi}$$

and the bootstrap mechanism

$$\hat{F} \xrightarrow{\text{i.i.d.}} \mathbf{y}^* \xrightarrow{s(\cdot)} \hat{\psi}^*.$$

The link between the two representations is given by the fact that \hat{F} **approaches the true F when $n \rightarrow \infty$** , which is the key fact.

Comments on nonparametric bootstrap

1. It is completely automatic! The underlying math is not simple, but it has been rigorously carried out.
2. It is large-sample method, since its accuracy increases with n .
3. Can be extended to any statistic of interest, not just estimated standard errors.
4. Can be extended to more complex settings, including some models with dependent data.
5. It also has some limitations, like being not appropriate for sample extremes, such as the minimum or maximum value of the observed data. (For these latter problems, there are some specific adjustments, but they are not simple).

The parametric bootstrap

Parametric bootstrap

$$\text{e.g. } F = N(\mu, \sigma^2) \xrightarrow{\substack{\mu \text{ known} \\ \sigma^2 \text{ unknown}}} \hat{F} = N(\hat{\mu}, \hat{\sigma}^2)$$

Going back to the bootstrap mechanism, there is actually no need \hat{F} be the nonparametric estimate of F (the empirical cdf).

Another alternative is to assume a parametric statistical model $f_{\theta}(\mathbf{y})$ for the data, and simulate the bootstrap samples from $f_{\hat{\theta}}$, where as usual $\hat{\theta}$ is a point estimate of θ .

Take \hat{F} as known but its parameters unknown

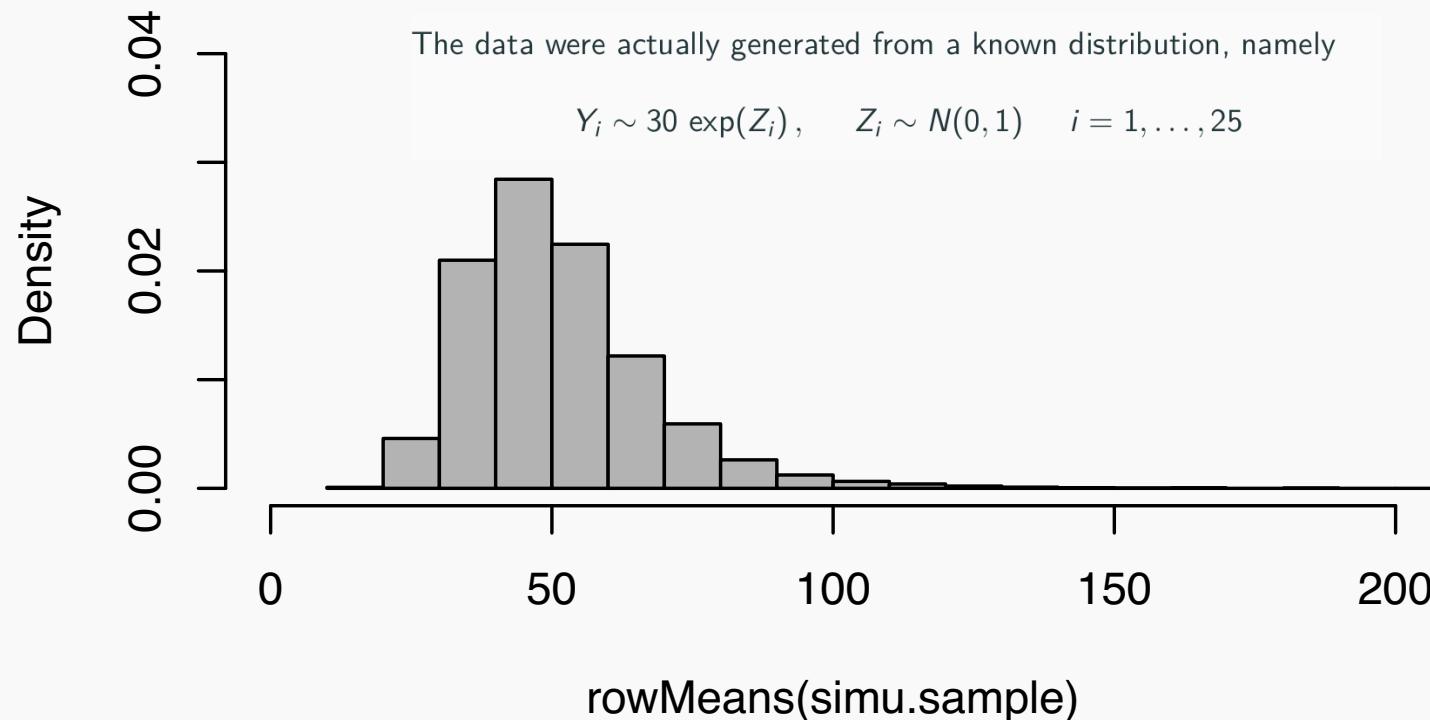
The mechanism becomes

$$f_{\hat{\theta}} \longrightarrow \mathbf{y}^* \xrightarrow{s(\cdot)} \hat{\psi}^*$$

Extensions to (very) complex models become easier, but a realistic model is required.

R lab: parametric bootstrap for $\hat{\psi} = \bar{Y}$

```
n <- length(y); mu <- mean(log(y)); sigma <- sd(log(y))  
simu.sample <- matrix(NA, nrow = B, ncol = n)  
for(i in 1:B) simu.sample[i,] <- mean(exp(rnorm(n, mu, sigma)))  
hist(rowMeans(simu.sample) , main="", breaks=25, prob=TRUE,  
col=gray(0.7), xlim=c(0, 200), ylim=c(0, 0.04))
```



Application to hypothesis testing

For parametric models we have parameters and we can make inferences about them given that we know F .

Parametric bootstrap can be employed for obtaining p -values by simulation, also for those cases when the model has some parameters that have to be estimated also under H_0 .

For example, we can obtain a fairly good approximation to the exact p -value for the classic one sample t -test.

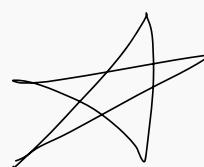
We only need to keep in mind that the bootstrap samples must be generated from the model estimated under H_0 . This means that for testing

$$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu \neq \mu_0 \end{cases}$$

condition
↓

for the usual i.i.d. normal model, we need to generate data with $\mu = \mu_0$ and $\sigma^2 = \hat{\sigma}_0^2 = \sum_i (y_i - \mu_0)^2 / n$.

maybe implement.



R lab: t-test by parametric bootstrap

```
library(DAAG); y <- with(pair65, heated - ambient)
n <- length(y)
z_obs <- mean(y) / sqrt(var(y) / n)
s0 <- sqrt(mean((y - 0)^2))
B <- 10000; z_sim <- numeric(B)
for(i in 1:B) { ys <- rnorm(n, m = 0, s = s0)
  z_sim[i] <- mean(ys) / sqrt(var(ys) / n)}
c(t.test(y)$p.val, mean(abs(z_sim) >= abs(z_obs)))
```

$$z_{\text{obs}} = \frac{\bar{y} - 0}{\text{s.e. } \bar{y}}$$

$$\text{s.e. } \bar{y} = \sqrt{\frac{1}{n} \sum (y_i - 0)^2}$$

$$z_{\text{sim}} = \frac{\bar{y}_s - 0}{\text{s.e. } \bar{y}_s}$$

[1] 0.01437832 0.01310000

Remainder of p-value definition: $P = P(T \geq |t|)$

Probabilities are calculated for r.v.s and not for fixed values, e.g., μ is fixed and not a r.v.

For confidence intervals we don't calculate the prob. of μ being in the C.I. but the C.I. is a r.v. itself and not μ . We calculate the prob. of C.I. containing μ , which is slightly different.

If $\hat{\beta}_j \sim N(\beta_j, V(\beta))$ ~ $N(0,1)$

$$P\left(\frac{z_{\alpha/2}}{2} < \frac{\hat{\beta}_j - \beta_j}{s.e.\hat{\beta}} < z_{1-\frac{\alpha}{2}}\right) = 1-\alpha$$

$$C.I.: \hat{\beta}_j \pm z_{1-\frac{\alpha}{2}} \times s.e.\hat{\beta}$$

On average C.I. contains $(1-\alpha)\%$ β

Bootstrap-based confidence intervals

The bootstrap automation of confidence intervals

We mentioned the standard approximate Wald-type 95% confidence interval for a parameter of interest ψ , in a model with parameter θ :

$$\hat{\psi} \pm 1.96 \text{SE}(\hat{\psi})$$

*↳ bootstrap
for instance*



This is widely used, but it has two shortcomings:

1. It requires the estimated standard error $\text{SE}(\hat{\psi})$, which may be hard to compute.
2. It is symmetric around the point estimate, and sometimes this leads to inaccuracy, since the finite sample distribution of $\hat{\psi}$ (and hence of the related pivot) is often asymmetric.

The first point is solved by $\widehat{\text{SE}}_{\text{boot}}$, but the bootstrap provides some further, more satisfactory solutions for confidence intervals.

Bootstrap-based confidence intervals

There are several available methods, and an extensive literature. Here we focus on the main ones (the approach is inspired by the MASS book), which are

1. The **percentile** method.
2. The **basic** method.
3. The **studentized** method.

These three methods work both for nonparametric and parametric bootstrap.

Further methods exist, such as the BC_a method, but they are used less often in practice.

Running example for confidence intervals

We use the student score dataset of the CASI book as a running example. It concerns the score of 22 students in 5 tests:

```
score <- read.table("figs/student_score.txt", header = TRUE)
print(cor(score))
```

##	mech	vecs	alg	analy	stat
## mech	1.0000000	0.4978075	0.7560364	0.6534763	0.5357744
## vech	0.4978075	1.0000000	0.5922624	0.5071353	0.3786038
## alg	0.7560364	0.5922624	1.0000000	0.7627546	0.6698255
## analy	0.6534763	0.5071353	0.7627546	1.0000000	0.7376712
## stat	0.5357744	0.3786038	0.6698255	0.7376712	1.0000000

The parameter of interest is the *eigenratio* statistic for the above correlation matrix, namely $\psi = \text{largest eigenvalue} / \text{sum eigenvalues}$.

fst step: eigen v and \tilde{v} ↳ how relevant is first component.

R lab: bootstrap (nonparametric) standard error for the student score data

```
psi_fun <- function(data) {eig <- eigen(cor(data))$values  
                           return(max(eig) / sum(eig))}  
  
psi_obs <- psi_fun(score)  
n <- nrow(score); B <- 10^4  
s_vect <- rep(0, B)  
for(i in 1:B) {ind <- sample(1:n, n, replace = TRUE)  
               s_vect[i] <- psi_fun(score[ind,])}  
  
SE_boot <- sd(s_vect)  
psi_obs + c(-1, 1) * 1.96 * SE_boot
```

```
## [1] 0.5448847 0.8401859
```

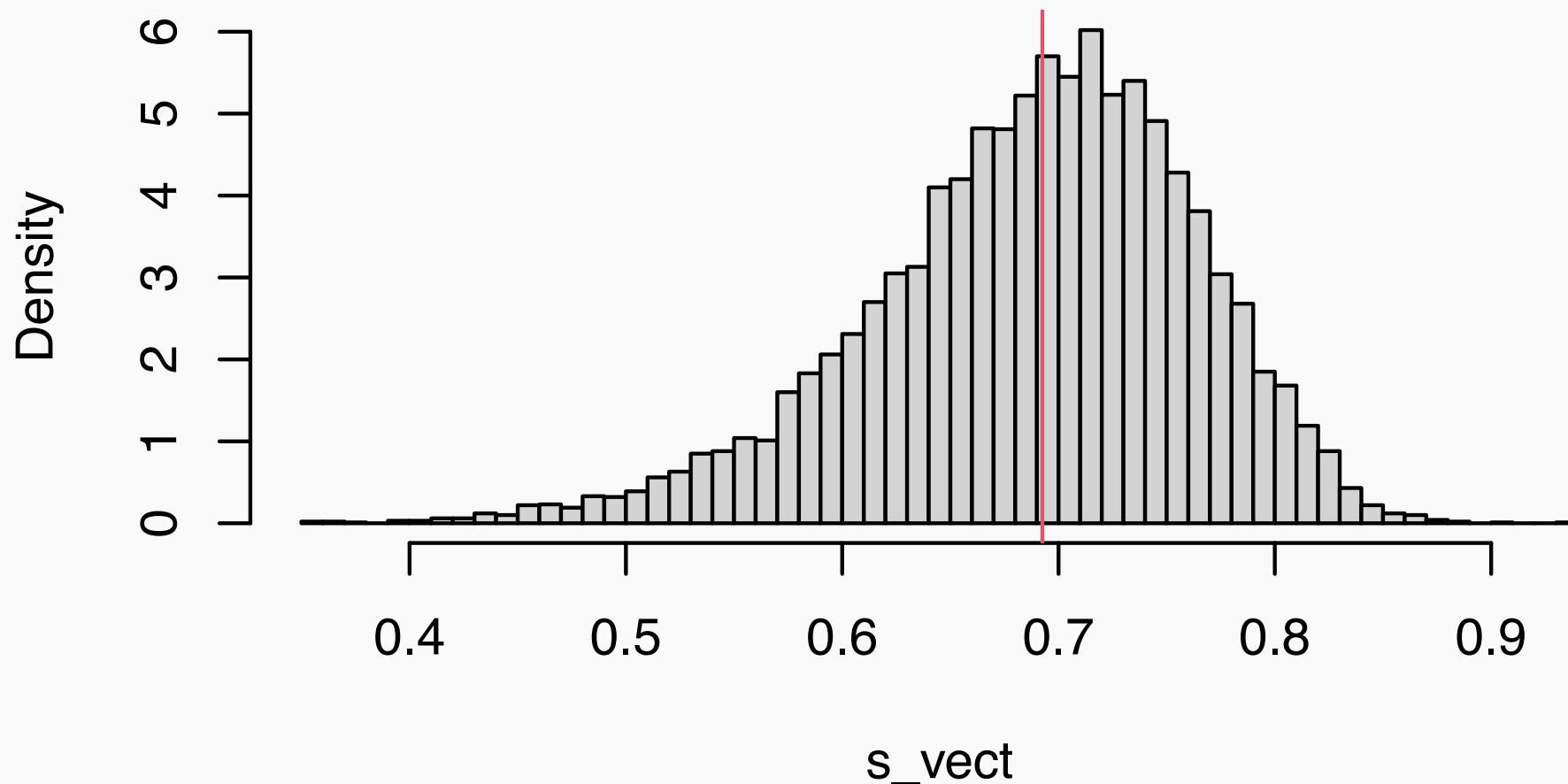
Wald C.I.



What is this?

R lab: bootstrap distribution

```
hist.scott(s_vect, main = "")  
abline(v = psi_obs, col = 2)
```



The percentile method

It simply uses the quantiles of the bootstrap distribution $\hat{\psi}^{*1}, \dots, \hat{\psi}^{*B}$.

In the example, we get

```
perc_ci <- quantile(s_vect, prob=c(0.025, 0.975))
```

```
attr(perc_ci, "names") <- NULL
```

```
perc_ci
```

$$10000 \times \frac{25}{1000}$$

```
## [1] 0.5175641 0.8136610
```

Compared to the above Wald-type interval, and taking the point estimate as reference, the percentile confidence interval is wider on the left side and shorter on the right side.

Shift to the left.

The basic method

The deviations of $\hat{\psi}^*$ from $\hat{\psi}$ in the Bootstrap should approximate the deviations of $\hat{\psi}$ from ψ in the real world.

The basic intervals are based on the idea that the distribution of $\hat{\psi}^* - \hat{\psi}$ mimics that of $\hat{\psi} - \psi$. If this is the case, we would get

$$0.95 = \Pr(L \leq \hat{\psi} - \psi \leq U) \approx \Pr(L \leq \hat{\psi}^* - \hat{\psi} \leq U)$$

Using the first probability we obtain that a confidence interval for ψ is $(\hat{\psi} - U, \hat{\psi} - L)$, and then we use the second probability to obtain that $L + \hat{\psi}$ and $U + \hat{\psi}$ are estimated by the 2.5% and 97.5% bootstrap quantiles, respectively (here denoted by $q_{0.025}^*$ and $q_{0.975}^*$).

Putting the two things together we get the **basic bootstrap confidence interval**

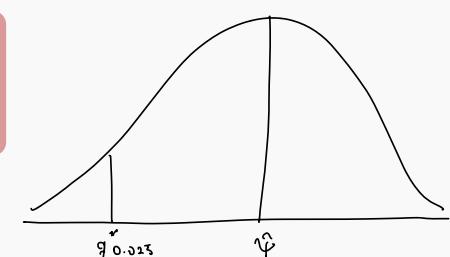
$$(\hat{\psi} - U, \hat{\psi} - L) = (2\hat{\psi} - q_{0.975}^*, 2\hat{\psi} - q_{0.025}^*)$$

$$L + \hat{\psi} = q_{0.025}^* \Leftrightarrow L = q_{0.025}^* - \hat{\psi}$$

$$\hat{\psi} - L = 2\hat{\psi} - q_{0.025}^*$$

$$\hat{\psi} - U = 2\hat{\psi} - q_{0.975}^*$$

Observed



$$L = \hat{\psi} - Z_{\alpha/2} s.e. \hat{\psi}$$

R lab: basic confidence interval

```
basic_ci <- 2 * psi_obs - quantile(s_vect, prob=c(0.975, 0.025))
attr(basic_ci, "names") <- NULL
basic_ci

## [1] 0.5714096 0.8675066
```

Since the result is essentially the percentile interval reflected about $\hat{\psi}$, the basic confidence interval is shorter on the left side and wider on the right side.

For asymmetric distributions of $\hat{\psi}$ the basic confidence interval may have coverage probability closer to the target value than the percentile one. On the other hand, the percentile interval is invariant to monotonic transformations of ψ , and this is perhaps more important.

The studentized method

The last method is perhaps the most reliable of all the methods. but it requires a standard error estimate $\text{SE}(\hat{\psi}^*)$ from each bootstrap sample

Denoting by $z_{0.025}^*$ and $z_{0.975}^*$ the bootstrap quantiles of z^{*1}, \dots, z^{*B} , where $z^{*b} = (\hat{\psi}^{*b} - \hat{\psi})/\text{SE}(\hat{\psi}^{*b})$, the **studentized bootstrap confidence interval** is given by

SKIP

$$(\hat{\psi} - \text{SE}(\hat{\psi}) z_{0.975}^*, \hat{\psi} - \text{SE}(\hat{\psi}) z_{0.025}^*)$$

This is perhaps too challenging for the running example, since explicit estimates of $\text{SE}(\hat{\psi}^*)$ would be very hard. We could employ the jackknife within each bootstrap sample, or a *double bootstrap scheme*, though ...

Non parametric smoothing

(An introduction)

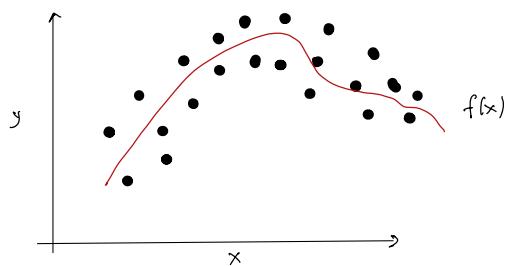
N. Torelli, G. Di Credico, V. Gioia

2023

University of Trieste

We want to relax the relationship between y and the regressors. x_s , that must be numerical variables,

f is a free-function / no parameters.



Question: What do you mean by parameters?

In physics we estimate the parameters of regression functions like this:

$$y = a_0 + a_1 x + a_2 x^2$$

$\rightarrow (a_0, a_1, a_2)$ might be relevant physical quantities.

In some cases LM or GLM will perform better than scatter smoothing methods

The point is to estimate relationships directly from data and not w/ parameters

Nonlinear regression and scatterplot smoothing



Polynomial Regression

Step functions

Kernel Smoothing

Regression splines

Smoothing splines

Nonlinear regression and scatterplot smoothing

Intro: the limitations of linearity

- Models that are built upon the linear effects of predictors, such as **linear models** or **generalized linear models**, play a crucial and non-replaceable role in the applications of statistics.
- Linearity is always an approximation but in many cases it is simple, reasonable, and it leads to very sensible results.
- Yet, there are instances where linearity is too strong a limitation, preventing the development of realistic models.
- That's why **nonlinear models** are important in statistics.
- This set of slides is based on ch.7 of the book **An Introduction to Statistical Learning** by James, Wittem Hastie and Tibshirani (freely downloadable from <https://www.statlearning.com/>).

Classes of nonlinear models

- Whereas linear models (including also GLMs) are easy to characterize, nonlinear models may be of several different types.
- We will not consider here the case of models which are *nonlinear in the parameters*. They include **nonlinear regression models**, often based on some biological or physical model, but also **Neural networks** and their extension (such as the models used in *deep learning*). These models have their own peculiarities and would deserve a specific treatment.
- A case of great interest attains to models which are *nonlinear in the predictors*.
- They belong to the class of **semiparametric regression models** (often the term nonparametric is also used)



βx

linear

x^β

non-linear (on the parameters)



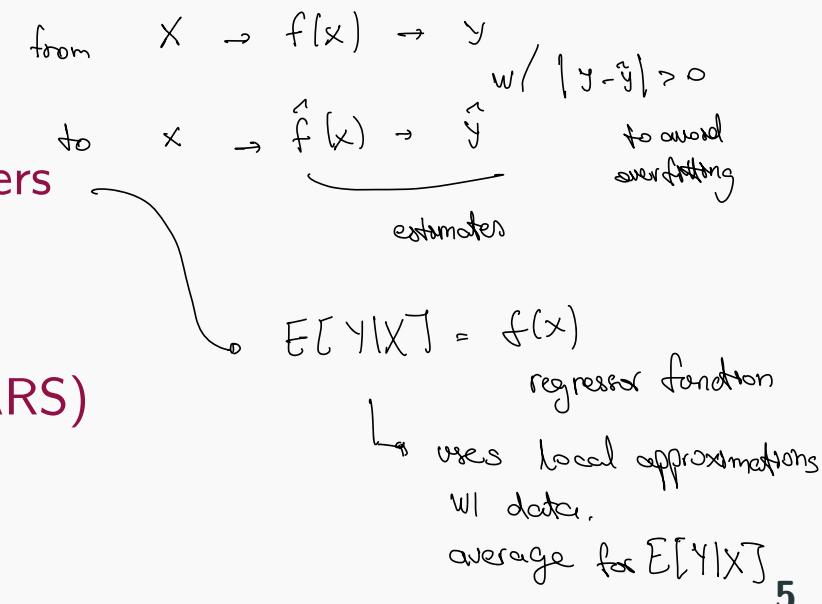
$f(x, \beta^*)$

not thus one

Semi-parametric regression models

- **Semi-parametric** regression modelling keeps the usual specification:
 $y = g(x_1, \dots, x_{p-1}, \epsilon)$ but relaxes the assumption of linear combination of predictors, and replaces it with a much weaker assumption of a smooth g
- Pro's and con's
 - greater flexibility and potentially more accurate estimate of g than LM
 - greater computation and sometimes more difficult-to-interpret results
- The more popular possible solutions are:

- ↪ Polynomial Regression
- ↪ Step functions
- ↪ Kernel and local-polynomial smoothers
- ↪ Regression and Smoothing splines
- ↪ (Generalized) Additive models
- ↪ Decision (regression) trees (and MARS)



Polynomial Regression

CART: Classification and regression tree

Polynomial Regression

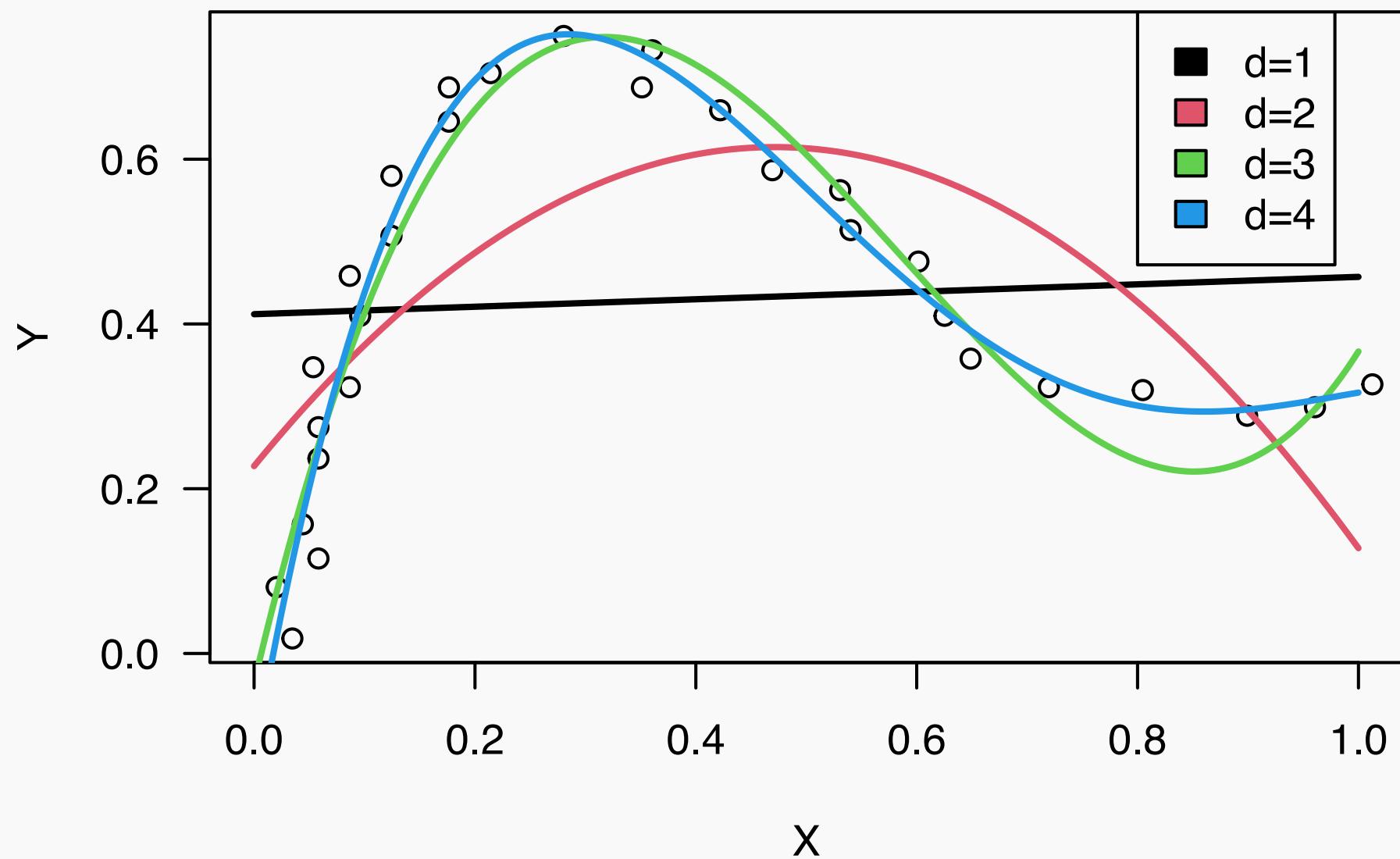
Non-parametric because we are not trying to estimate a parameter of the data, like the population mean, but we are just trying to represent the data by a suitable function.

- A simple, yet typical, way to extend linear regression for a single covariate (input variable) is to consider a polynomial of degree d within the classical linear model:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \dots + \beta_d x_i^d + \epsilon_i,$$

- For a moderately large d it fits to a scatterplot a curve which exhibits evident non-linearities
- The model is linear in the coefficients $\beta_0, \beta_1, \dots, \beta_d$ and they can be estimated by using ordinary least squares.
- It is actually a standard linear model with predictors $x_i, x_i^2, x_i^3, \dots, x_i^d$.
- It is unusual to take d larger than 3 or 4, to avoid overfitting and possible shapes of the curve which are very strange within the support of x .

Example



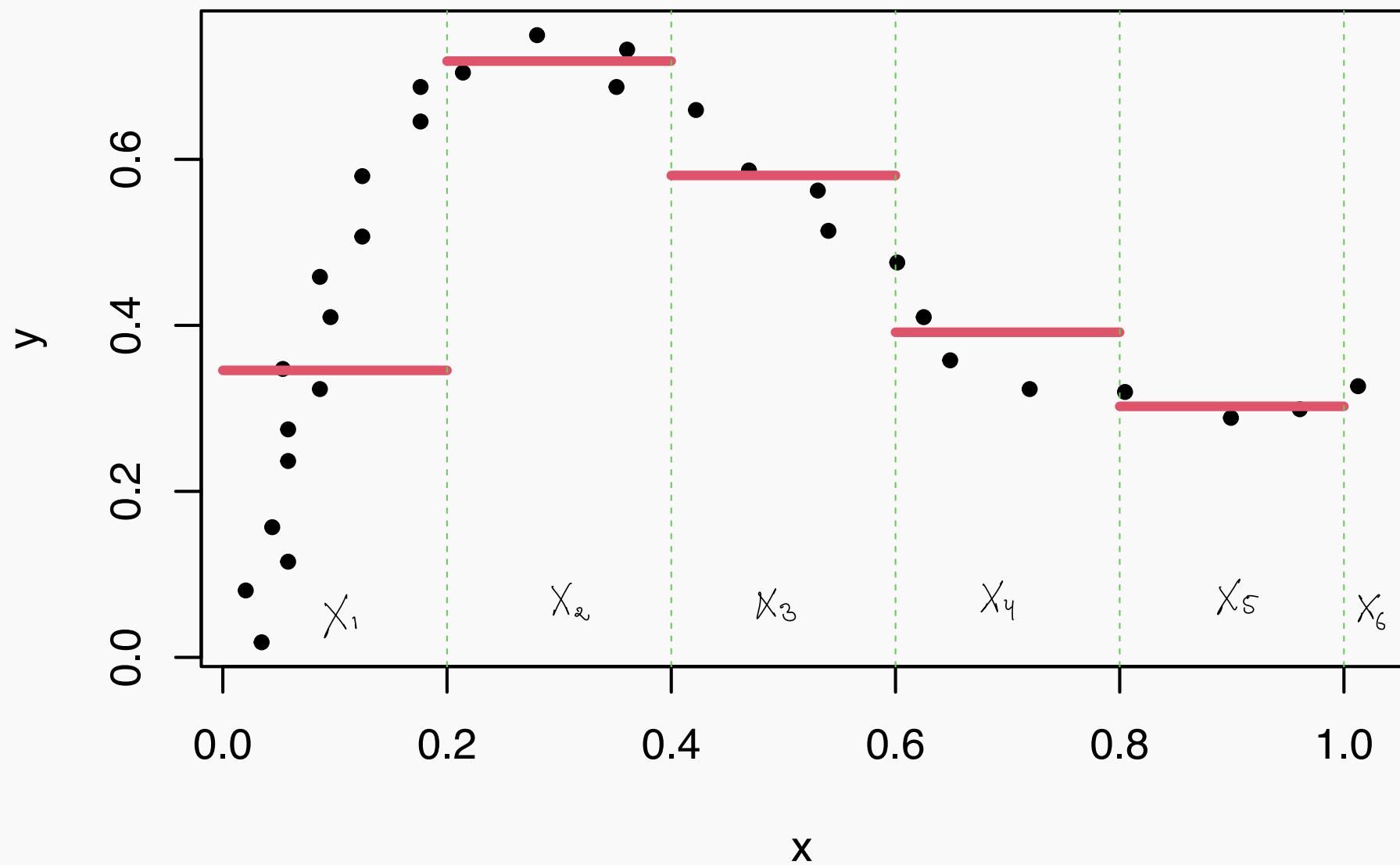
Step functions

Step functions

- A simple, yet effective, way to approximate a generic function $f(x)$ is to use a step function, that is, a piecewise constant function
- The values of the X_4 variable is subdivided into K disjoint classes $[x_0 - x_1], [x_1 - x_2], \dots [x_{K-1} - x_K]$
- In this case the value of the constant will be simply estimated by the average of the Y coordinates for those points within each interval (this is the least square solution)
- The level of the mean of the Y variable within each class k is then the predicted value for $X \in [x_{k-1} - k_k]$
- Fitting step functions is straightforward: it implies a linear regression model with a factor whose levels are the classes adopted to split the X variable.

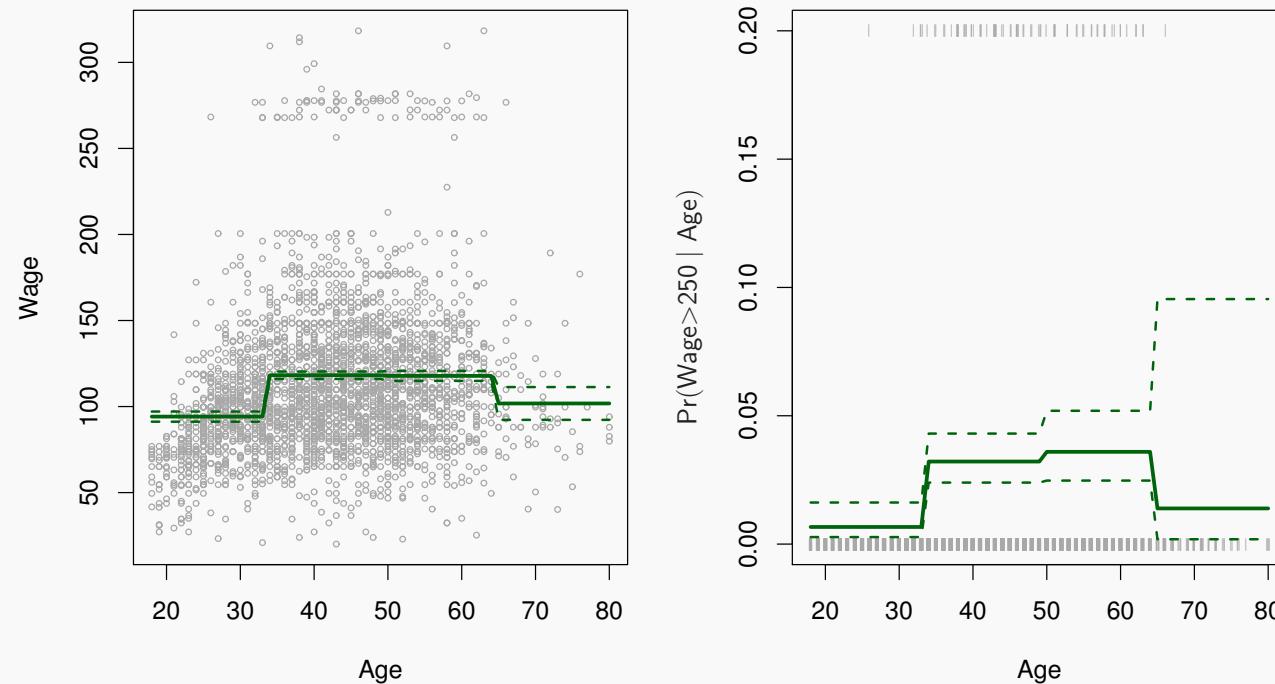
J There is no need for the classes to have the same "width".
But the density of points in each class must be the same.

Example



Another example (with salary data)

Piecewise Constant



The right panel represent a step function to estimate a model for a dicotomous variable.

Kernel Smoothing

Goals of smoothing

- The goal is again estimation of a regression function:

$$f(x) = E(y|x)$$

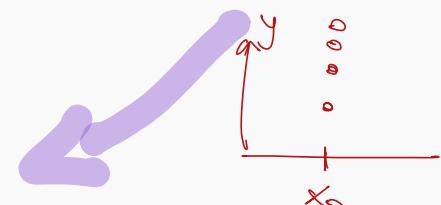
Recall: we use the mean because it minimizes MSE.

through a model $y = f(x) + \epsilon$

$$\Leftrightarrow E[\epsilon] = 0$$

- data on y and x are available $(x_i, y_i), i = 1, 2, \dots, n$
- a straightforward simple solution would be the following:
 - to predict y at $x = x_0$ gather all the pairs (x_i, y_i) having $x_i = x_0$, then
 - estimate $f(x_0)$ as the mean of the y_i values:

$$\hat{f}(x_0) = \text{Average}(y|x=x_0)$$



- typically available data do not include observations having exactly

$$x_i = x_0$$



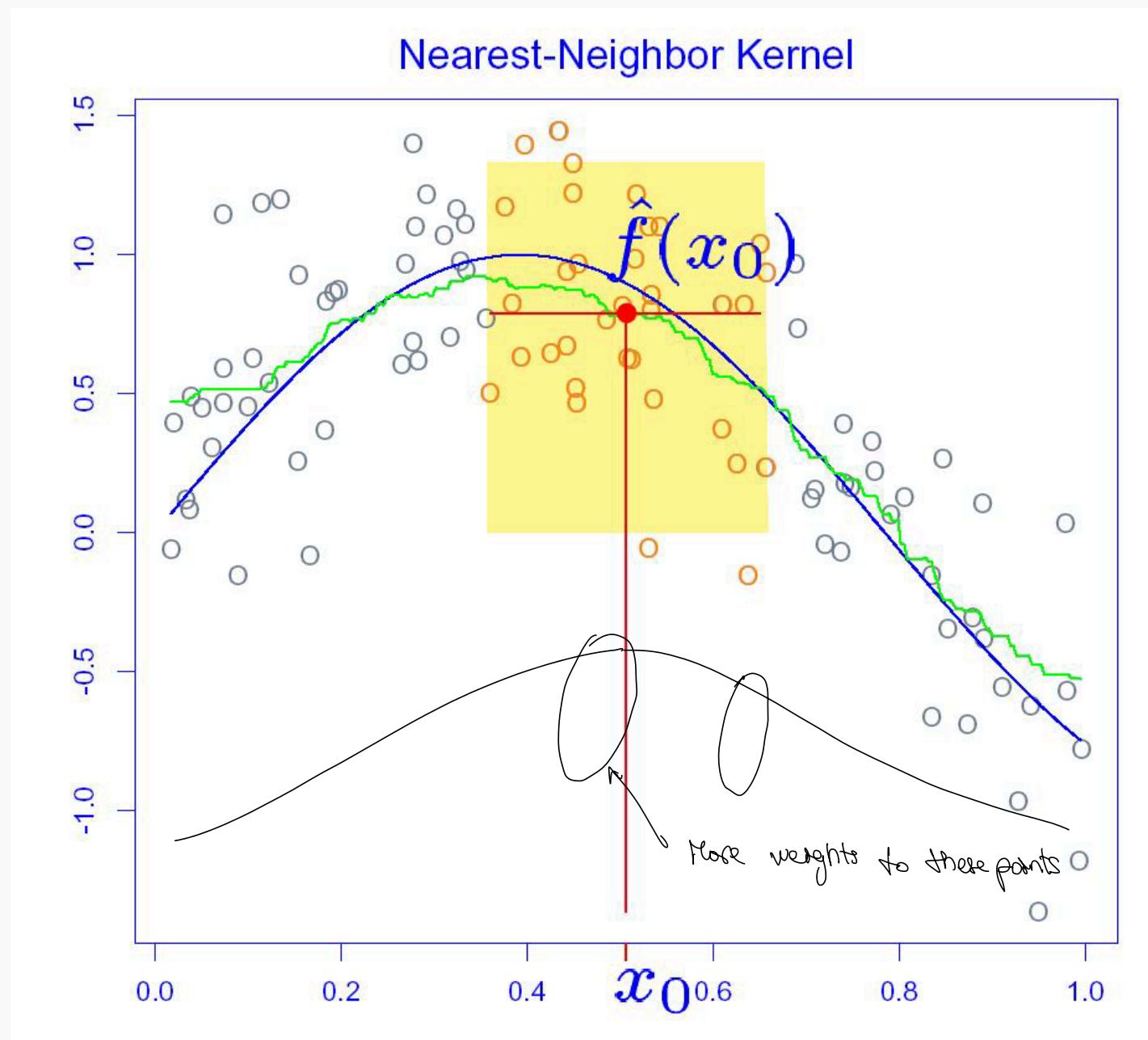
Nearest Neighbour Averaging

- an alternative solution could be to estimate $E(y|x = x_0)$ by averaging those y_i whose x_i are in a neighbourhood of $x = x_0$
- e.g. define the neighbourhood to be the set of k observations having values x_i closest to x_0 in euclidean distance $\|x_i - x_0\|$
- (in the univariate case this is simply the absolute value $|x_i - x_0|$).

This method is called nearest neighbour

A neighbourhood is a concept complicated to define since it depends on the dimension they are being considered. Sometimes, projecting 2 points in a lower dimensional space gives smaller distances.

Nearest neighbour



Choosing k

- Small k implies that we use only the k points which are closer to target x (low bias), but averages when based on a small sample have high variance.
bias-variance
trade off
w/ high/low
 k -value
- Large k includes points far from x (high bias), but they have smaller variance.
- selecting a “good” value for k depends on how smooth the true function $f(x)$ is, and how noisy y is.
- One could try different values of k on a validation dataset, and pick the one with the best prediction performance.
- cross-validation can be also used.
- An alternative is to penalize the fitting criterion (in this case is least square).

Local regression (kernel smoothers)

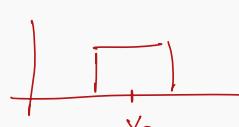
- A different solution could be to consider the weighted least squares by weighting observations x_i with their distance from x_0 :

$$\min_{\alpha, \beta} \sum_n e_i^2(x_0) \rightarrow w_h(x_i - x_0) \quad \dots \quad \text{It generalizes to many parameters } \beta_j$$

$$f(x) = \min_{\alpha, \beta} \sum_{i=1} [y_i - \beta_0 - \beta_1(x_i - x_0)]^2 w_h(x_i - x_0)$$

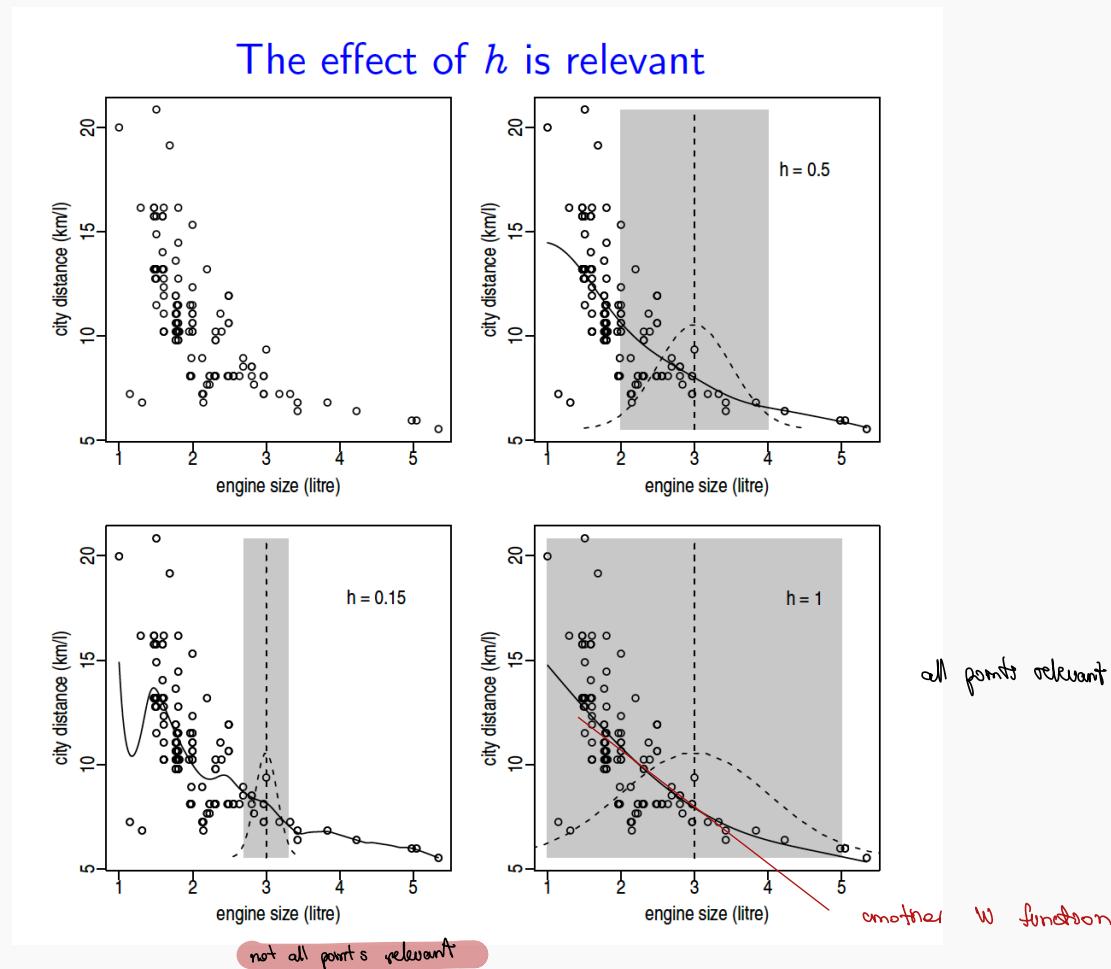
↳ Kernel: in general symmetric functions,
but not symmetric can also be used

- $h > 0$ is a scale factor, called bandwidth or smoothing parameter, and
 - $w_h(\cdot)$ is a symmetric density function around 0, said **kernel** whose variance depends on h .
 - By varying x_0 , we obtain a whole estimated curve $\hat{f}(x)$.
 - The most important ingredient is the **smoothing parameter** h , which regulates the smoothness of the curve, while the choice of the kernel w is less relevant.
 - w is often taken to be the density of the normal distribution $\mathcal{N}(0, h^2)$
- the R function ksmooth can be used to obtain a simple solution using local averages

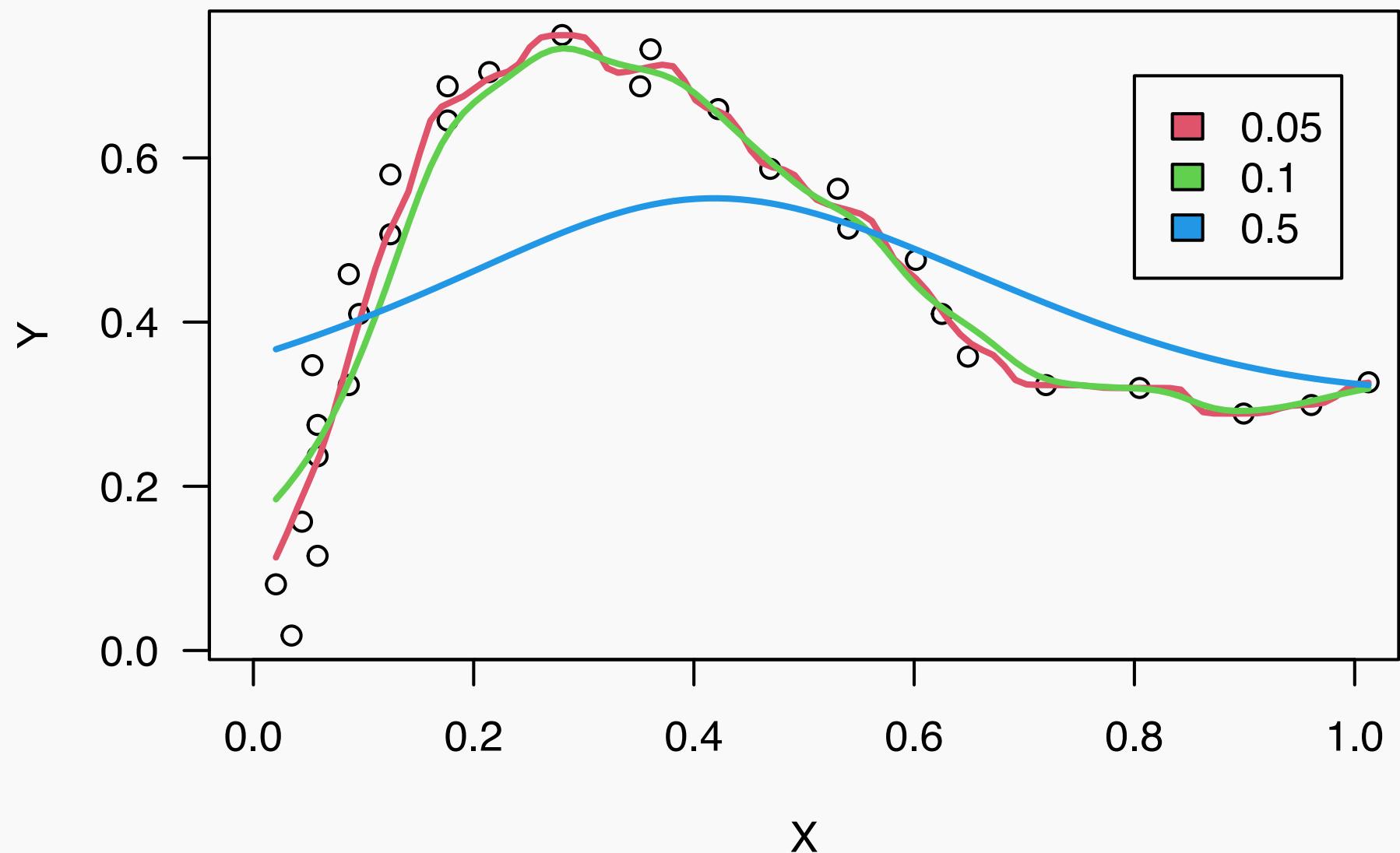
 the kernel implies the step function



The effect of h



Some kernel smoothers



Variable bandwidth and the loess

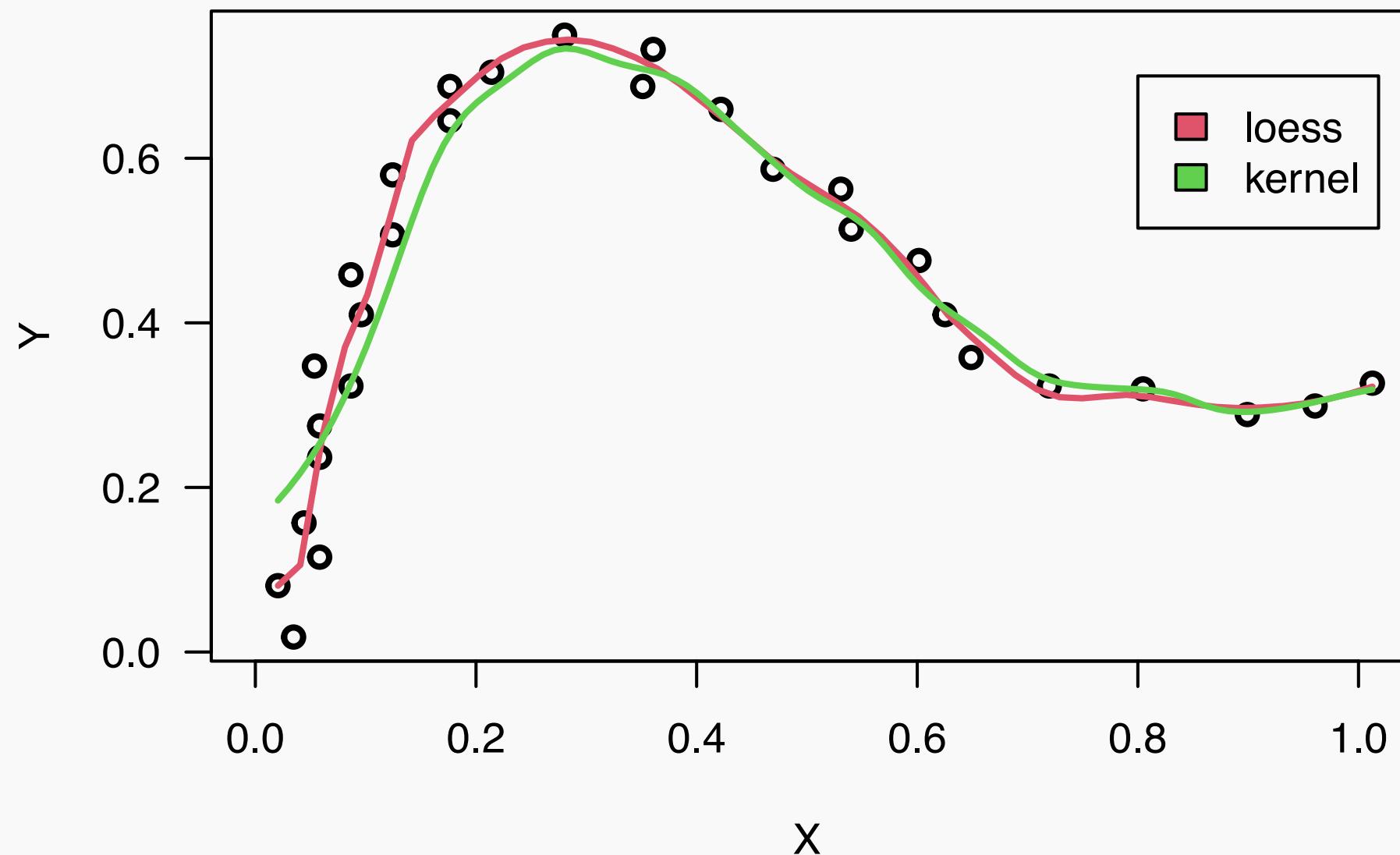
- In many cases, there is an advantage in using a nonconstant bandwidth along the x -axis, according it to the level of sparseness of observed points
- variable bandwidth: it is reasonable to use larger values of h when x_i are more scattered
- Good idea! but how do we modify h ? Locally Estimated Scatterplot Smoothing
- a popular solution is given by **loess**: it expresses the smoothing parameter by defining the fraction of effective observations for estimating $f(x)$ at a certain point x_0 on the x -axis;
 \uparrow variance $\rightarrow 0$ few points
 \uparrow bias $\rightarrow \downarrow$ many points
- this fraction is kept constant $\in [0, 1]$ to represent the variability in data
- this imply automatically a setting of the bandwidth related to the sparsity of data
- in addition, this idea is combined with the use of robust estimation
- loess is a very popular technique for smoothing a scatterplot (see the R function loess)

$$h = \alpha x (\max(x) - \min(x)) \quad , \quad \alpha \text{ span}$$
$$w_i(x_0) = \begin{cases} \left(1 - \left(\frac{x_i - x_0}{h}\right)^3\right)^3 & , \quad |x_i - x_0| \leq h \\ 0 & , \quad \text{otherwise} \end{cases}$$

(tricubic Kernel)

Although Gaussian Kernel is generally used

Loess vs kernel smoothinng



$$\begin{aligned}
 p(x) &= a_0 + a_1 x + a_2 x^2 + a_3 x^3 \\
 &= a_0 + x (a_1 + a_2 x + a_3 x^2) \\
 &= a_0 + x (a_1 + x (a_2 + a_3 x)) \\
 &= a_0 + x (a_1 + a_2' x) \\
 &= a_0 + a_1' x \\
 &= a_0'
 \end{aligned}$$

Regression splines

Basis function

- Polynomial and piecewise-constant regression models are in fact special cases of a **basis function**
- The idea is to have at hand a family of functions or transformations that can be applied to a variable X , $b_1(X), b_2(X), \dots, b_K(X)$.
Instead of a linear model in X let us fit the model:

S.t.  $y_i = \beta_0 + \beta_1 b_1(x_i) + \beta_2 b_2(x_i) + \dots + \beta_K b_K(x_i) + \epsilon_i,$

- basis functions $b_1(\cdot), b_2(\cdot), \dots, b_K(\cdot)$ are fixed and known.
 - In polynomial regression $b_j(x_i) = x_i^j$
 - with step function $b_j(x_i) = I(c_j \leq x_i < c_{j+1})$.
- Least squares can be used in both cases to estimate the β parameters.

Step function is a spline!!!

Piecewise regression

- Instead of fitting a high-degree polynomial over the entire range of X , **piecewise polynomial regression** involves fitting separate low-degree polynomials over disjoint regions of X separated by points called **knots**.
- For example, a piecewise cubic polynomial with a single knot at c has the form:

$$y_i = \begin{cases} \beta_{01} + \beta_{11}x_i + \beta_{21}x_i^2 + \beta_{31}x_i^3 + \epsilon_i & \text{if } x_i < c; \\ \beta_{02} + \beta_{12}x_i + \beta_{22}x_i^2 + \beta_{32}x_i^3 + \epsilon_i & \text{if } x_i \geq c. \end{cases}$$

- Note that 8 degrees of freedom are needed in fitting this model

Splines

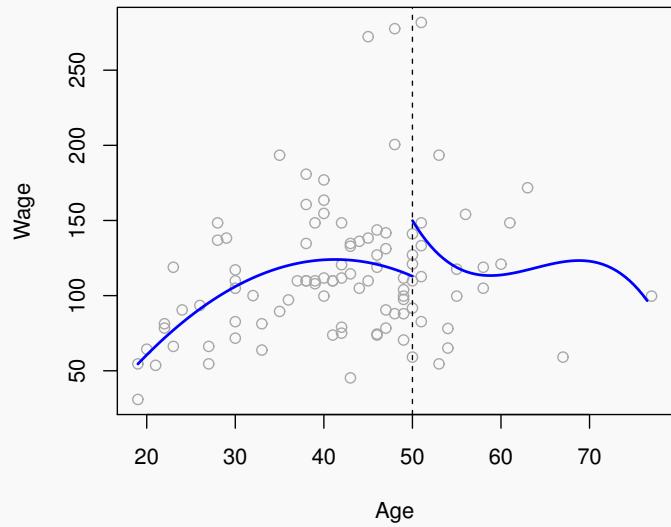
- Unlike polynomials, where flexibility is gained by using higher powers, splines introduce flexibility by increasing the number of knots keeping low the degree of the polynomials: this second option seems to be more feasible.
- With K knots, $K + 1$ distinct polynomials are defined (over each interval defined by the sequence of knots). Knots don't include the boundary points.
- Using polynomial with lower degree is possible, but one concern is continuity of the function in the knots
- Some constraints, such as continuity of the function in the knots (possibly also of the first and second derivatives) can be imposed.
- This leads to reduction of complexity of the curve and then frees up degrees of freedom. $1 \text{ knot} \Rightarrow df 4 + 4 - 3 = 4 + + \rightarrow$
- The cubic with one knot in the example seen before will need 5 degrees of freedom (continuity up to the second derivatives frees 3 degrees of freedom). A cubics spline with K knots needs $4 + K$ degrees of freedom
- Continuity can be also imposed when piecewise linear function are used.

K knots w/ wrie splines $\Rightarrow 4 \times (K+2)$ d.f. Constraints $3 \times K$	Total df $4K + 4 - 3K$ $K+4$
---	------------------------------------

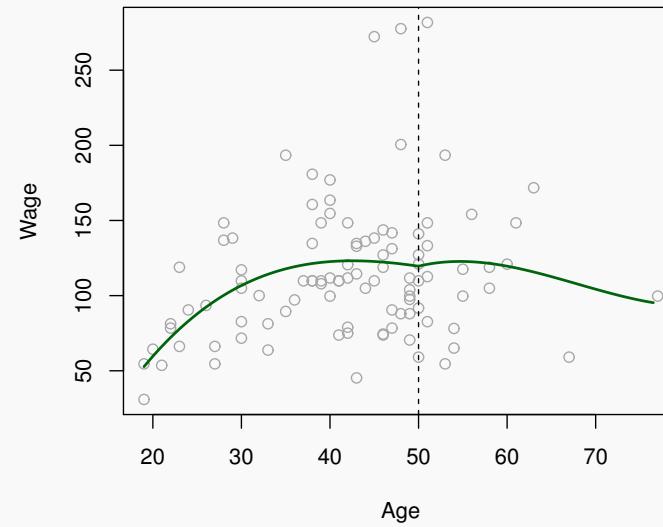
$$K+4 = 4 + (3) \\ K=3. \quad \uparrow \\ 2 \rightarrow 3 \rightarrow K+1 = \frac{4}{2} \rightarrow$$

Credit dataset

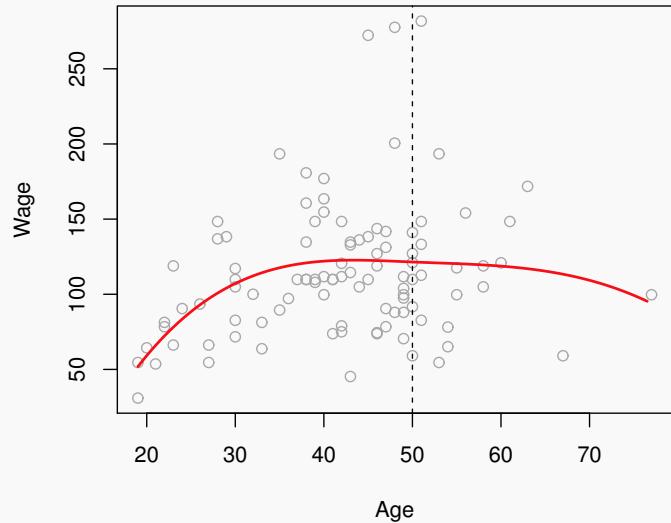
Piecewise Cubic



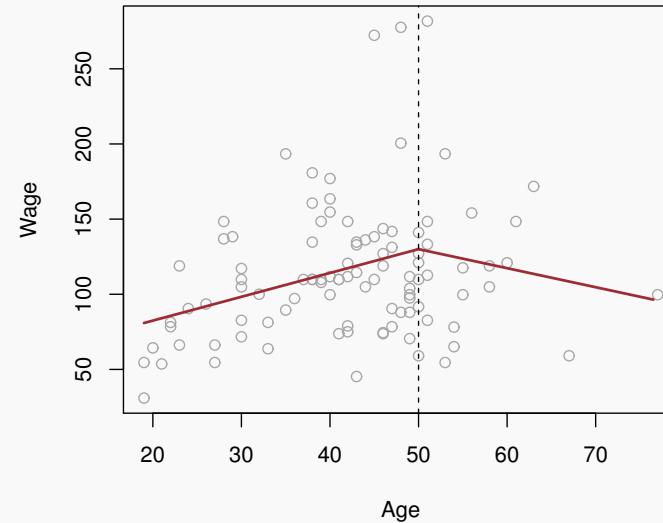
Continuous Piecewise Cubic



Cubic Spline



Linear Spline



Linear splines through basis functions

a linear spline with knots in ξ_k , $k = 1, 2, \dots, K$ is a piecewise linear function continuous in the knots

It can be represented as

$$y_i = \beta_0 + \beta_1 b_1(x_i) + \beta_2 b_2(x_i) + \dots + \beta_{K+1} b_{K+1}(x_i) + \epsilon_i,$$

where b_k are *basis functions*:

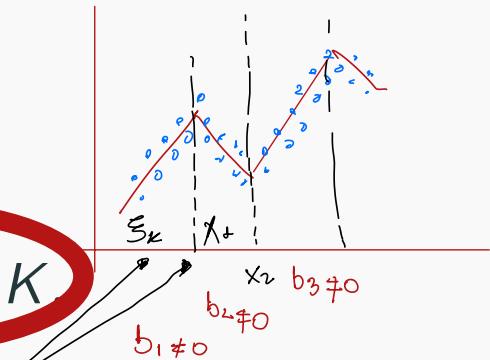
$$b_1(x_i) = x_i$$

$$b_{k+1}(x_i) = (x_i - \xi_k)_+, \quad k = 1, \dots, K$$

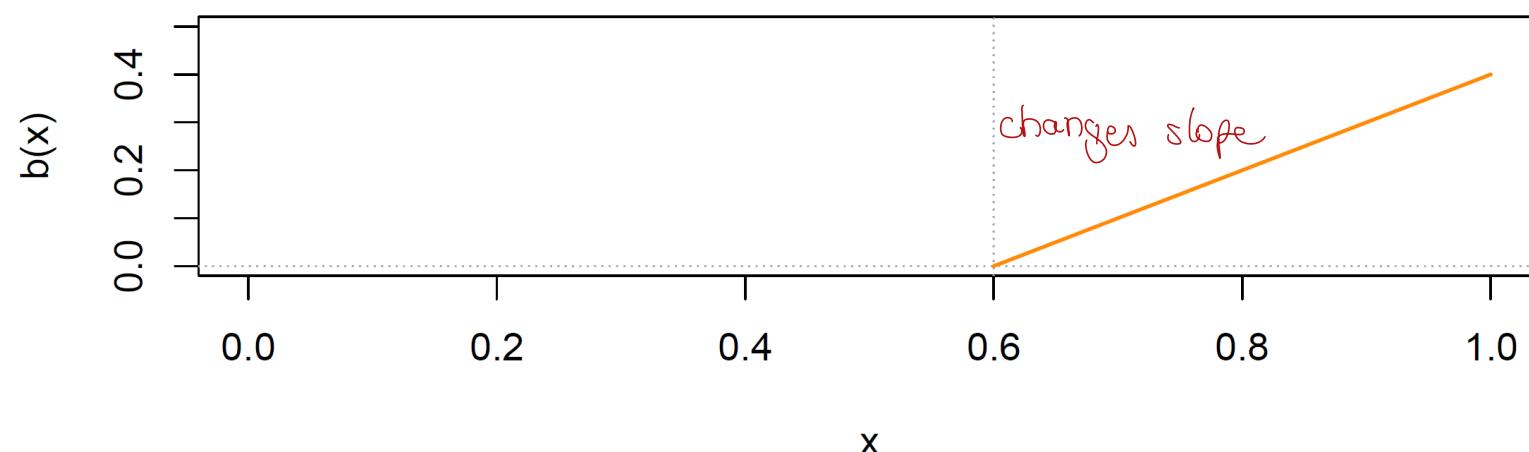
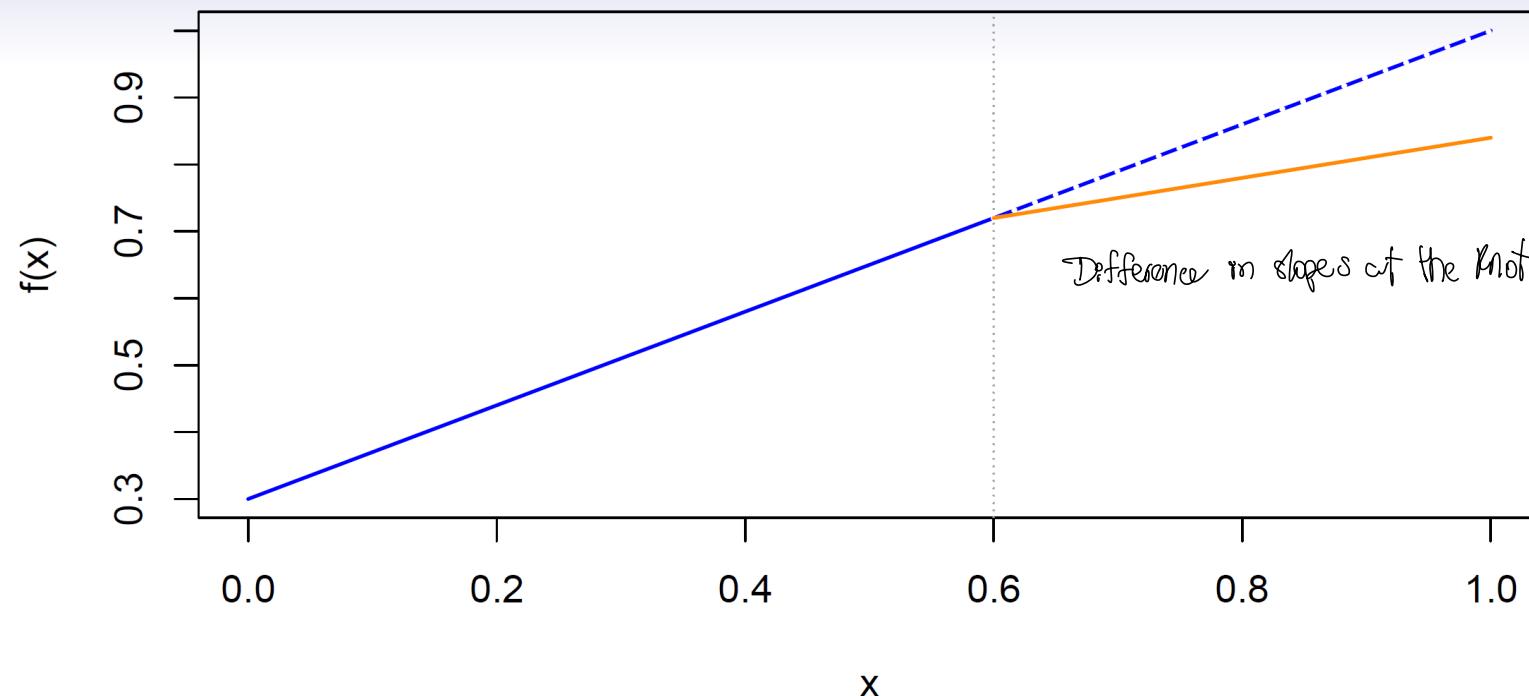
$z, 3, \dots, K+1$

The symbol $(\cdot)_+$ denotes the *positive part*:

$$(x_i - \xi_k)_+ = \begin{cases} x_i - \xi_k & \text{if } x_i > \xi_k \\ 0 & \text{otherwise} \end{cases}$$



The positive part



Cubic splines through basis splines

Later where splines is defined through a basis of functions defined in $[0,1]$ and the final spline is "expanded on the entire range of fit"

A cubic spline with K knots in ξ_k , $k = 1, 2, \dots, K$ is a piecewise cubic polynomial with continuous derivatives up to order two

It turns out that we can use the basis functions representation for a spline. A cubic spline with K knots can be modeled as:

$$y_i = \beta_0 + \beta_1 b_1(x_i) + \beta_2 b_2(x_i) + \dots + \beta_{K+3} b_{K+3}(x_i) + \epsilon_i,$$
$$b_1(x_i) = x_i$$
$$b_2(x_i) = x_i^2$$
$$b_3(x_i) = x_i^3$$
$$b_{k+3}(x_i) = (x_i - \xi_k)_+^3, \quad k = 1, \dots, K,$$

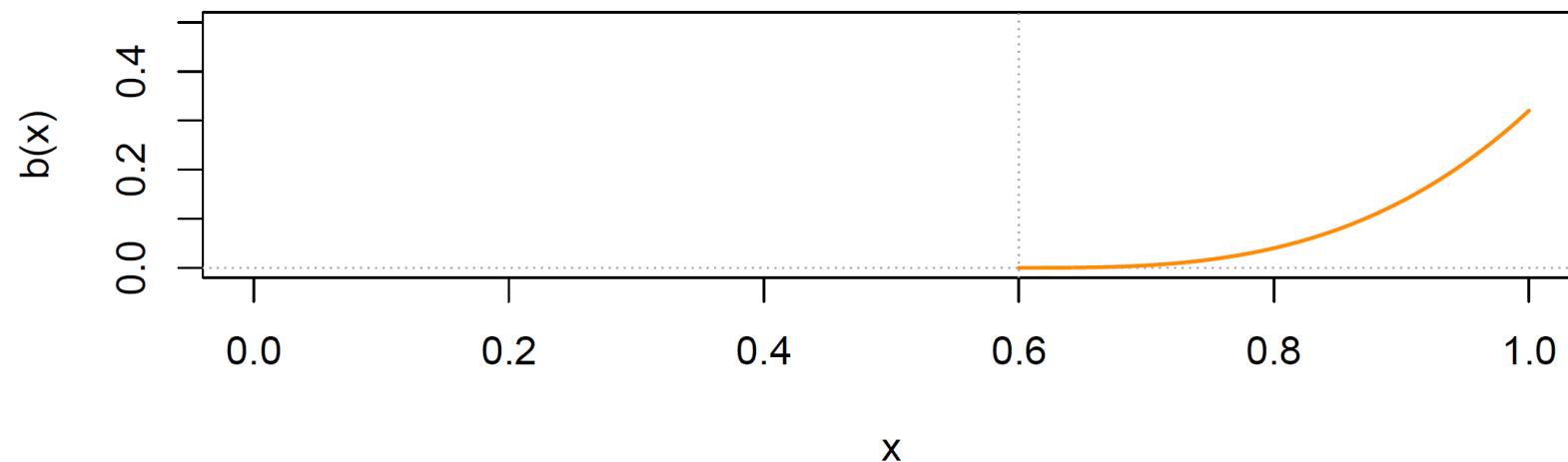
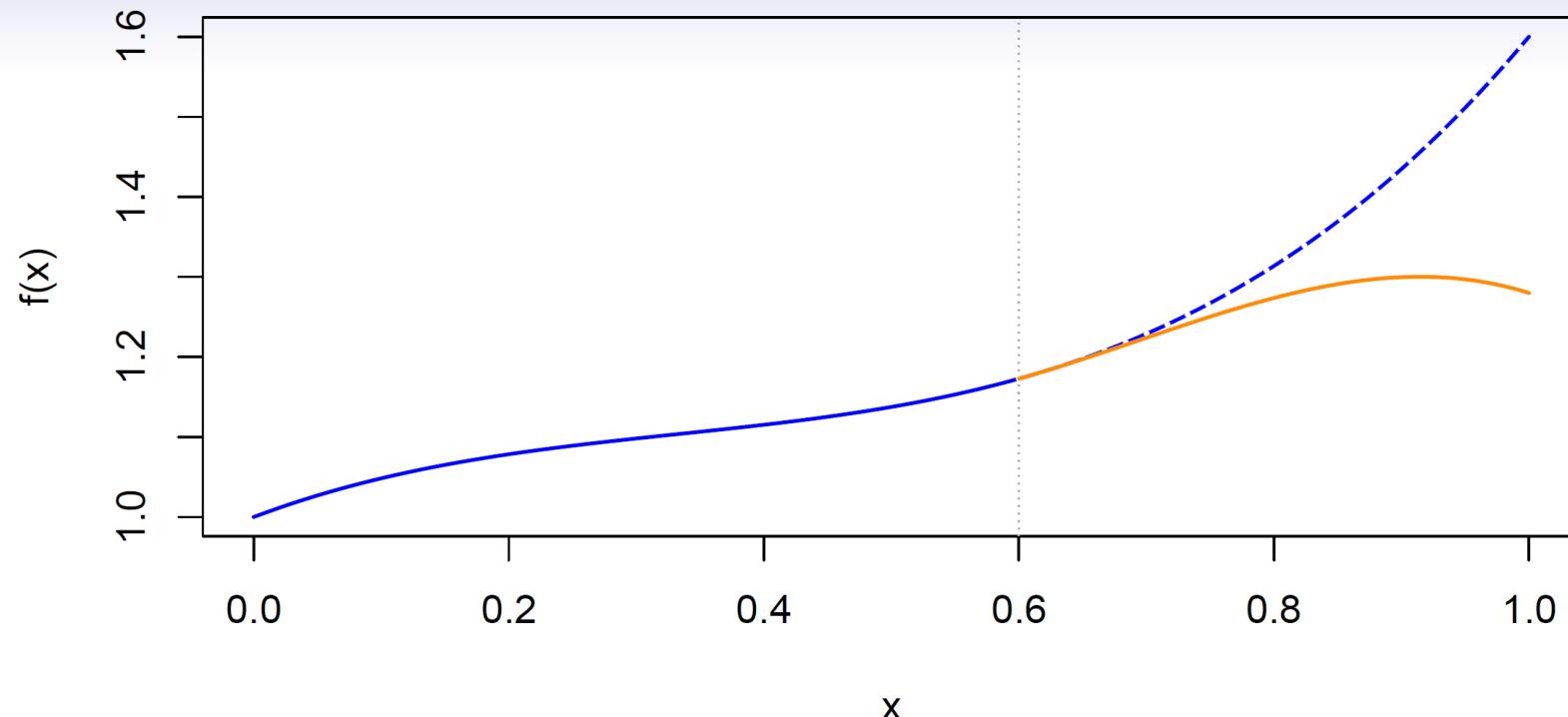
where a truncated power basis function is added for each knot. A truncated power basis function is defined as:

$$(x_i - \xi_k)_+^3 = \begin{cases} (x_i - \xi_k)^3 & \text{if } x_i > \xi_k \\ 0 & \text{otherwise} \end{cases}$$

- To fit a cubic spline with K knots to a data set we can use least squares regression with an intercept and $3+K$ predictors, of the form $X, X^2, X^3, h(X, \xi_1), h(X, \xi_2), \dots, h(X, \xi_K)$, where ξ_1, \dots, ξ_K are the knots. This amounts to estimating a total of $K + 4$ regression coefficients.

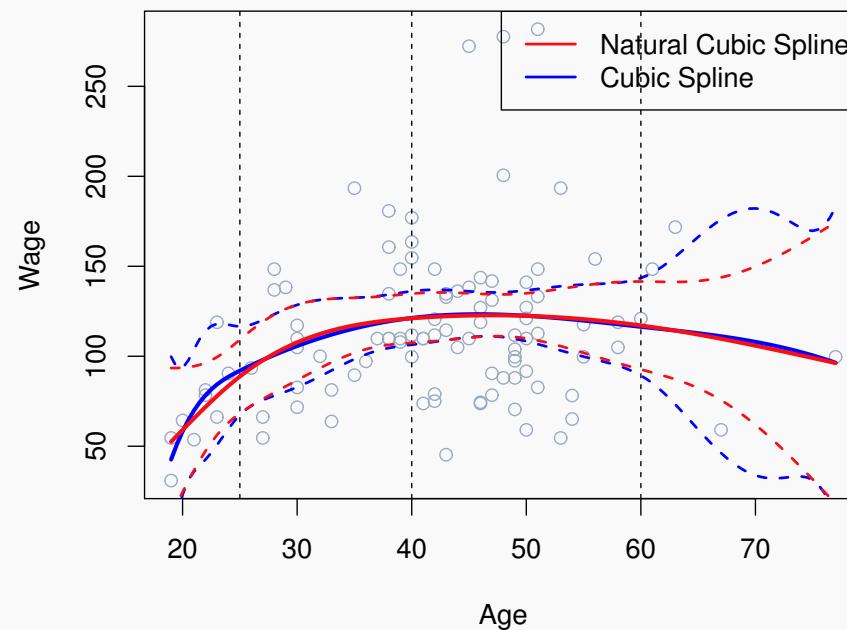
↳ degrees of freedom

The positive part (cubic spline basis)



Natural cubic splines

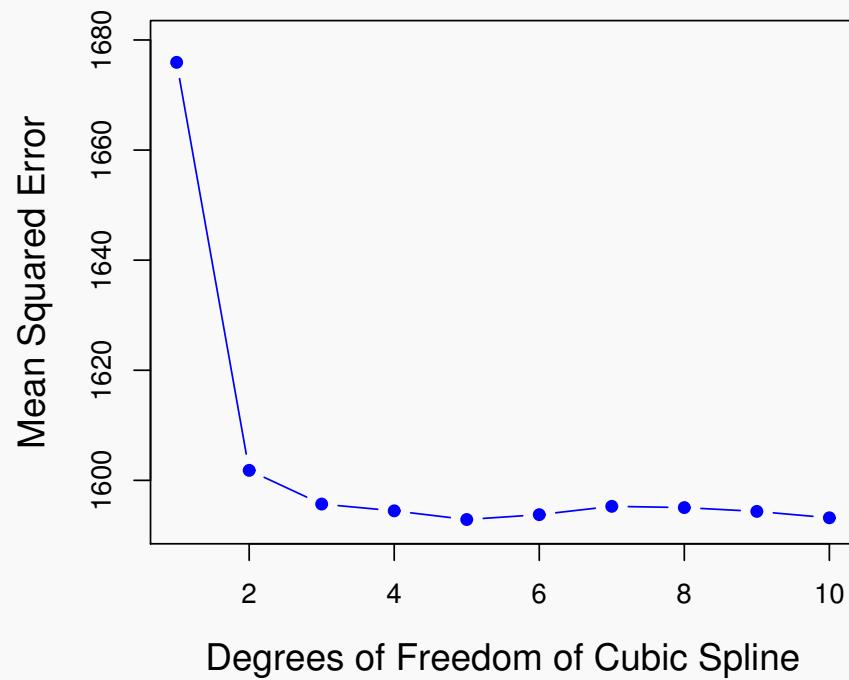
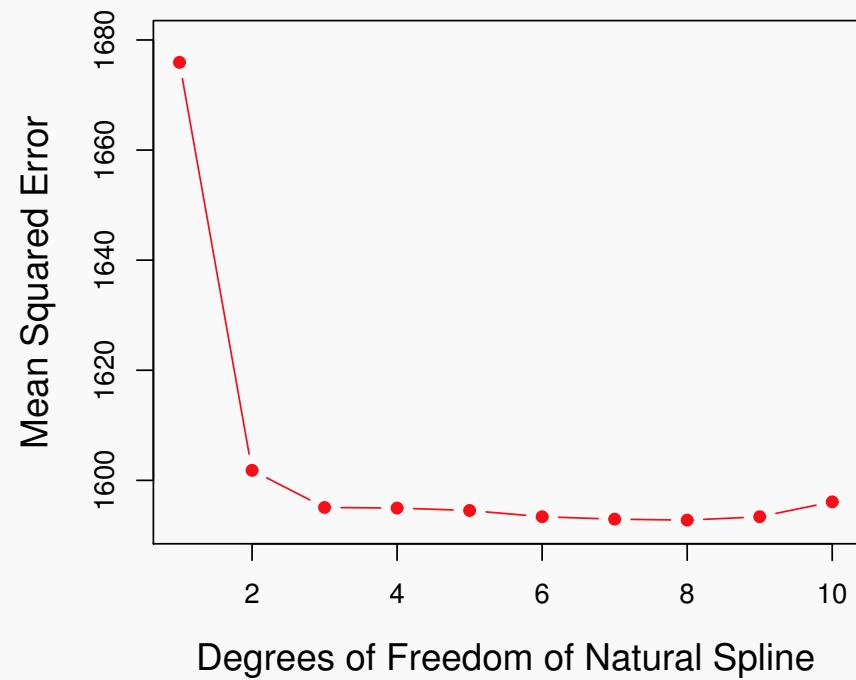
A *natural cubic spline* A natural spline is a regression spline with additional boundary constraints: the natural function is required to be linear at the boundary (in the region where X is smaller than the smallest knot, or larger than the largest knot).



Number and Locations of the Knots

- Once the degree of the polynomial is decided, where should we place the knots? And how many knots should we consider?
- The number of knots can be fixed according to the desired degrees of freedom and the desired flexibility of the curve.
- For a given number of knots a possible choice is to position them uniformly over the range of X .
- A more sensible choice could be to set more knots where the function changes more rapidly. *There are many ways to decide the location of knot, for instance*
- Another strategy is to put the knots in a sequence of quantiles of X . 
- Cross-validation can be also used:
 - use a portion of the data (say 10 %) and fit a spline with a certain number of knots to the remaining data, and then use the spline to make predictions for the held-out portion
 - repeat this process multiple times and then compute the overall cross-validated RSS.
 - repeated for different numbers of knots K . Then the value of K corresponding to the smallest RSS is chosen.

Choosing K with CV



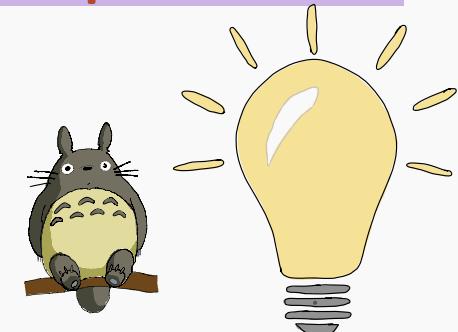
Smoothing splines

Smoothing splines

- The splines introduced so far are called **regression splines**. Their use implies:
 - choosing knots (number and position)
 - consider appropriate basis functions
 - use least squares to estimate coefficients
- When detecting a function $g(\cdot)$ to represent the relationship between a input variable X and a output Y two conflicting goals are:
minimizing an appropriate loss measure, such as RSS, and obtaining a simple, smooth, curve.
- According to this we can find *smooth* function $g(x)$ which minimizes:

$$\min_{g \in S} \sum_{i=1}^n (y_i - g(x_i))^2 + \lambda \int g''(t)^2 dt.$$

no wiggly function



Smoothing splines

■

$$\min_{g \in \mathcal{S}} \sum_{i=1}^n (y_i - g(x_i))^2 + \lambda \int g''(t)^2 dt$$

- The first term is the Residuals Sum of Squares RSS
- the second term is a **penalty term** that penalizes the variability in $g(\cdot)$ (roughness penalty) whose relevance depends on λ . In other words, $\int g''(t)^2 dt$ is simply a measure of the total change in the function $g'(t)$, over its entire range.
- $\lambda \geq 0$ is a **tuning parameter** : the larger the value of λ , the smoother the function will be.
 - When $\lambda = 0$, the penalty term has no effect, and the function $g(\cdot)$ will exactly **interpolate** the observations.
 - When $\lambda \rightarrow \infty$, the function $g(\cdot)$ is linear.
- For an intermediate value of λ the function can lead to a reasonable fit of the data but will be somewhat smooth. λ controls the bias-variance trade-off of the smoothing spline.

to minimize the equation $g''(x) = 0$

$$g(x) = ax + b \Rightarrow g'(x) = a \Rightarrow g''(x) = 0$$

\Rightarrow g is linear

Smoothing splines

- The function $g(x)$ that minimizes the penalized least squares, can be shown to have some special properties:
 - it is a piecewise cubic polynomial with knots at the unique values of the observed values of X and continuous first and second derivatives at each knot.
 - it is linear in the region outside of the extreme knots (natural cubic splines).
- However, it is not the same (natural) cubic spline that one would get if one applied the basis function approach described for regression splines: it is a regularized version of such a (natural) cubic spline, where the value of the tuning parameter controls the level of regularization.

Choosing the Smoothing Parameter

- It might seem that a smoothing spline will have too many degrees of freedom, since a knot at each data point allows excessive flexibility.
- the tuning parameter λ controls the roughness of the smoothing spline, and hence the effective degrees of freedom.
- It is possible to show that as λ increases from 0 to ∞ , the effective degrees of freedom, df_λ , decrease from n to 2. Actually, df_λ can be specified instead of λ .
- For instance, in R: `smooth.spline(age, wage)`.
- In fitting a smoothing spline, we do not need to select the number or location of the knots. Instead, we have another problem: we need to choose the value of λ .

Using cross-validation

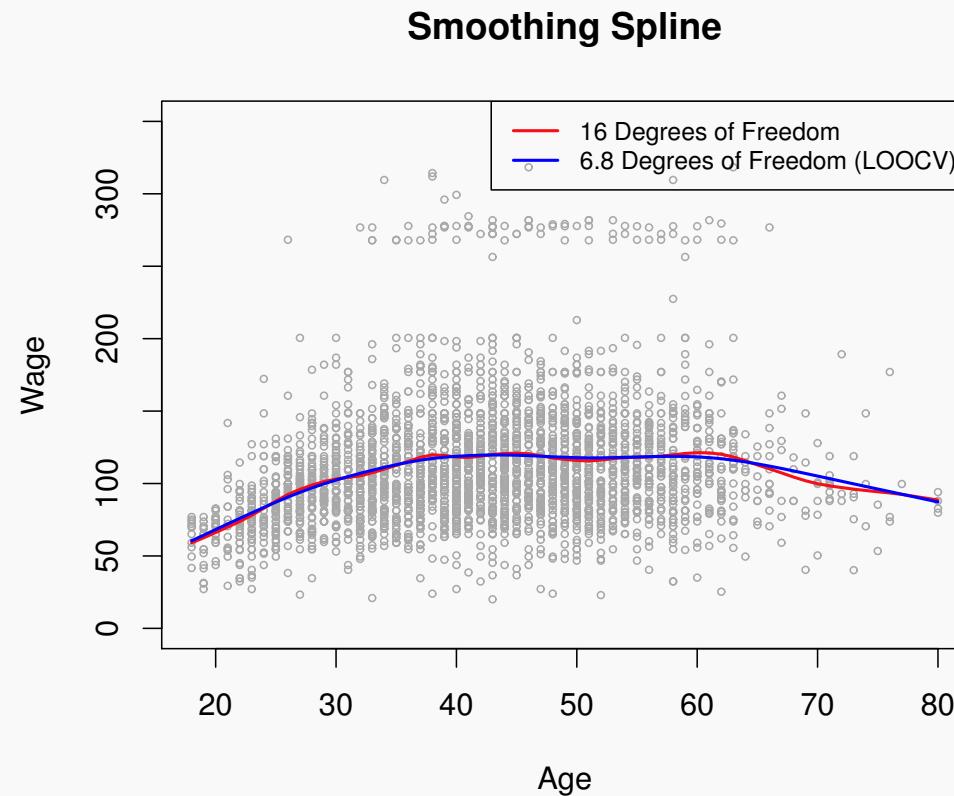
- We can find the value of λ that makes the cross-validated RSS as small as possible
- The leave-one-out cross-validation error (LOOCV) can be computed very efficiently for smoothing splines,

$$\text{RSS}_{cv}(\lambda) = \sum_{i=1}^n (y_i - \hat{g}_\lambda^{(-i)}(x_i))^2 = \sum_{i=1}^n \left[\frac{y_i - \hat{g}_\lambda(x_i)}{1 - \{\mathbf{S}_\lambda\}_{ii}} \right]^2,$$

What is \hat{g}_λ here?

- where $\hat{g}_\lambda^{(-i)}(x_i)$ is the fitted value evaluated at x_i , using all the data except the i -th, (x_i, y_i)
- \mathbf{S}_λ (whose formal definition is not detailed here) can be thought to be equivalent of the H matrix in linear models (such that $\hat{\mathbf{g}}_\lambda = \mathbf{S}_\lambda \mathbf{y}$ is the solution of penalized least squares for a given λ).
- the effective degrees of freedom corresponds to the trace of the matrix \mathbf{S}_λ .

Smoothing splines for the credit dataset



- The red curve results from specifying 16 effective degrees of freedom. For the blue curve, λ was found automatically by leave-one-out cross-validation, which resulted in 6.8 effective degrees of freedom.

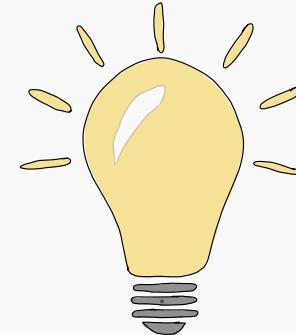
GAM only must be used
w/ numerical variables

Generalization:

$$Y = \beta_0 + S(x_1) + \beta_1 x_2 + S(x_3)$$

↳ the entire sphere

Splines for factors make no sense.



Generalized Additive Models

(An introduction)

N. Torelli, G. Di Credico, V. Gioia

2023

University of Trieste

Semiparametric regression: an introductory example

Some theory

Generalized Additive Models (GAMs)

Semiparametric regression: an introductory example

An actuarial example: Automobile Bodily Injury Claims

Dataframe AutoBi in the R package insuranceData:

The data contains information about the claimant, attorney involvement and the economic loss (LOSS, in thousands), among other variables. We consider here a sample of $n = 1,340$ losses.

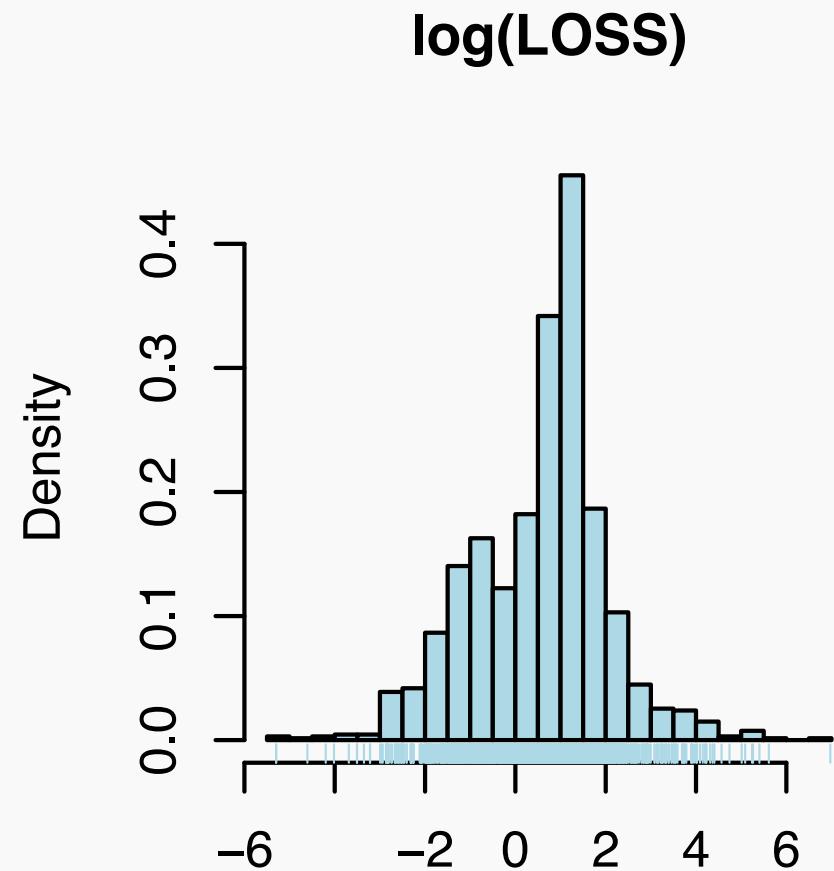
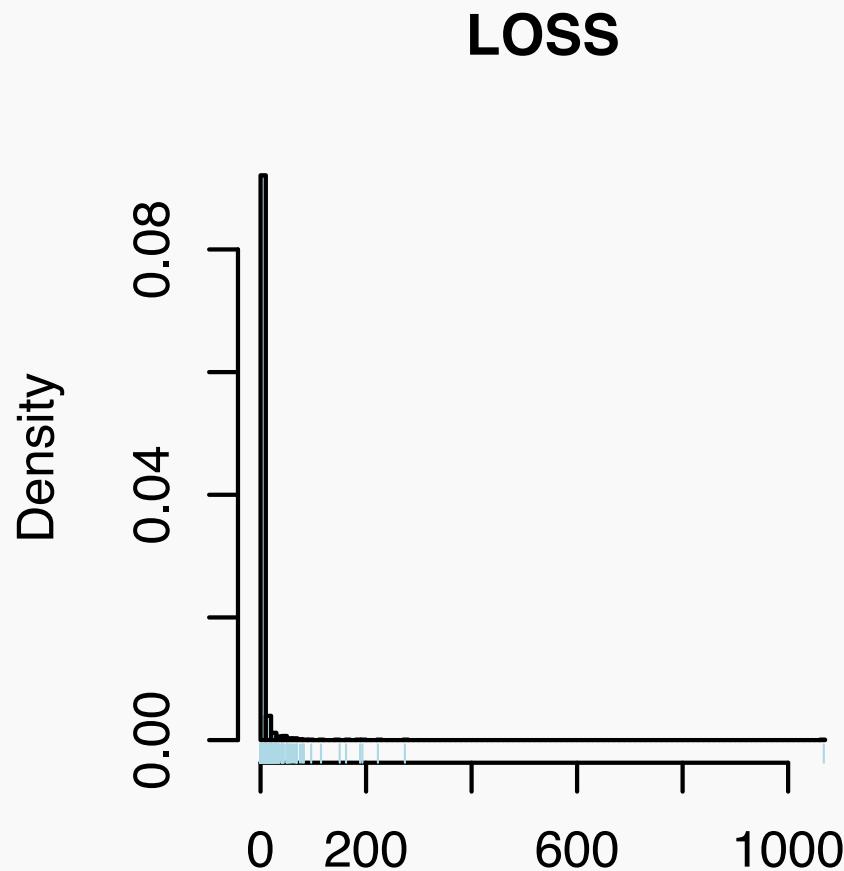
Main variables (after some transformations):

- ATTORNEY : The claimant is represented by an attorney (yes/no)
- CLMSEX : male/female
- MARITAL : married(M)/single (S)/widowed (W)/divorced (D)
- CLMINSUR : The driver of the claimant's vehicle uninsured (yes/no)
- SEATBELT : The claimant was wearing a seatbelt/child restraint (yes/no)
- CLMAGE : Claimant's age
- AGECLASS : Claimant's age split into five classes: (-18] / (18,26] / (26,36] / (36,47] / (47+)
- LOSS : Claimant's total economic loss (in thousands)

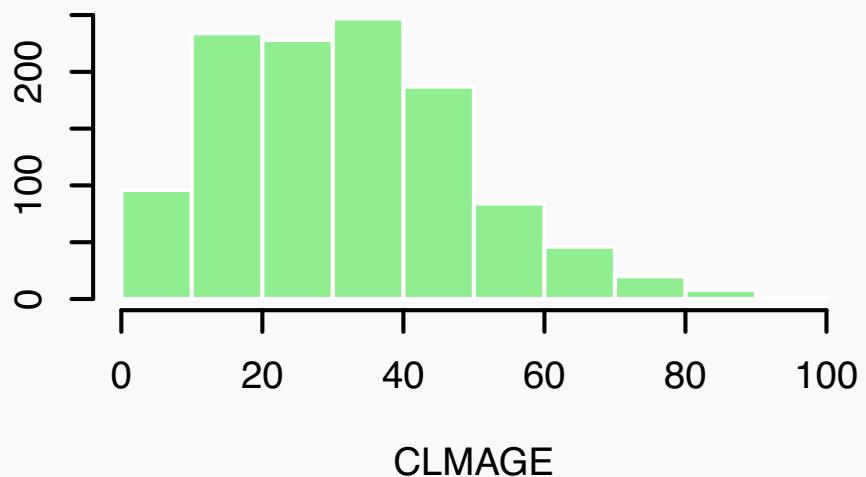
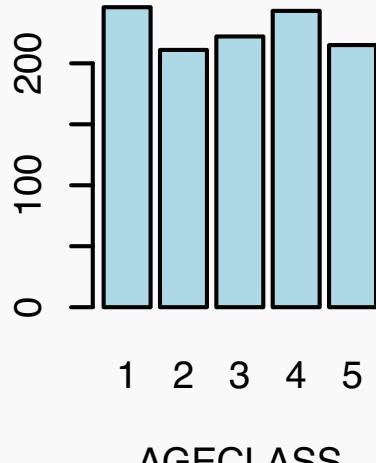
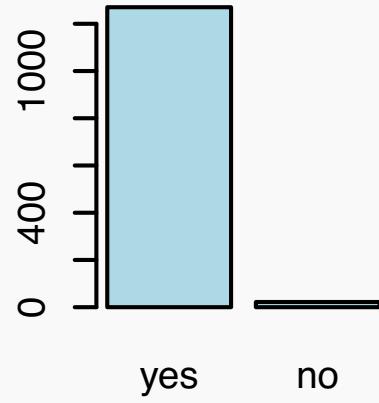
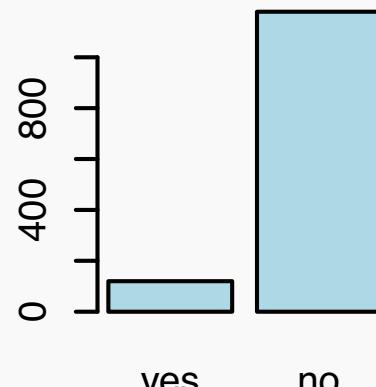
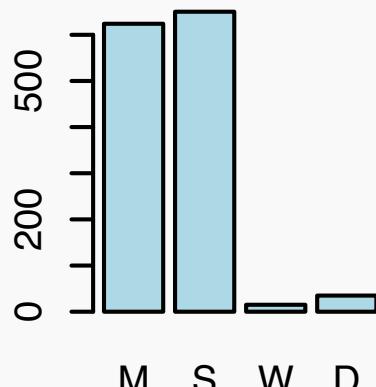
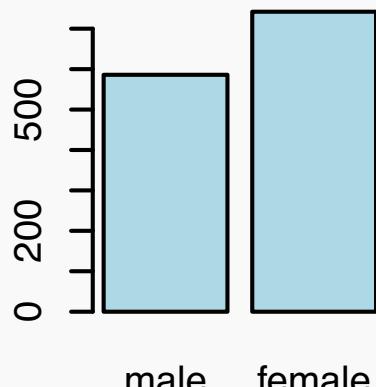
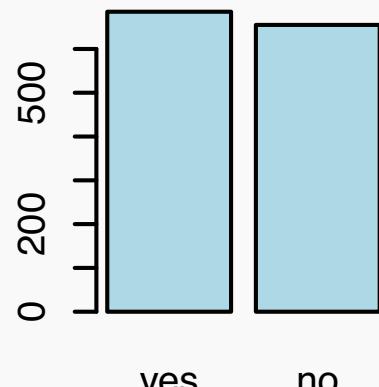
not categ.

AutoBi: Distribution of LOSS

- Severity refers to the amount of a claim.
- It is of interest to build a statistical model for predicting claim amount in future policies based on a sample of claim amounts of past policies.

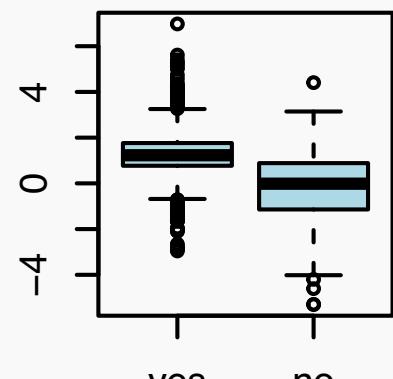


AutoBi: Information about predictors

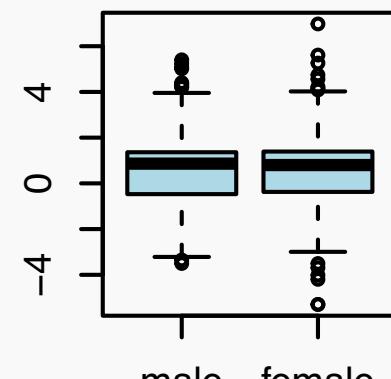


slightly skewed

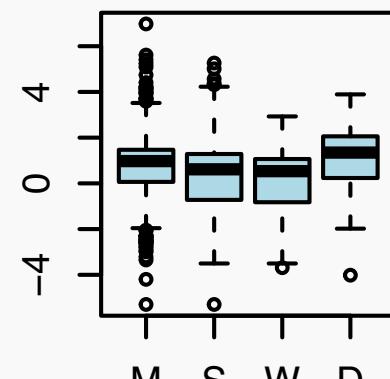
AutoBi: Relation between predictors and log(LOSS)



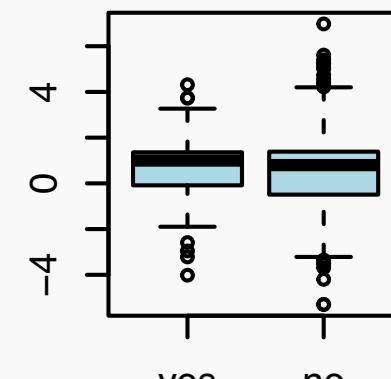
ATTORNEY



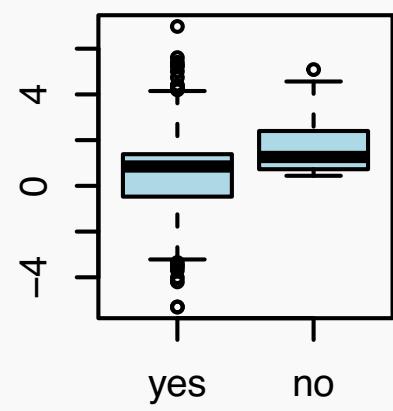
CLMSEX



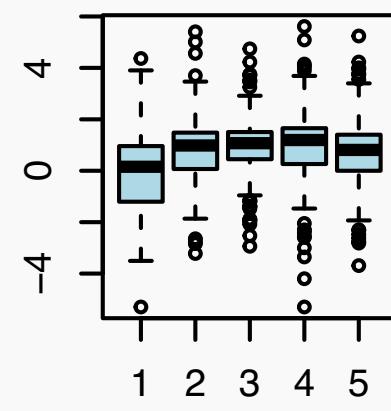
MARITAL



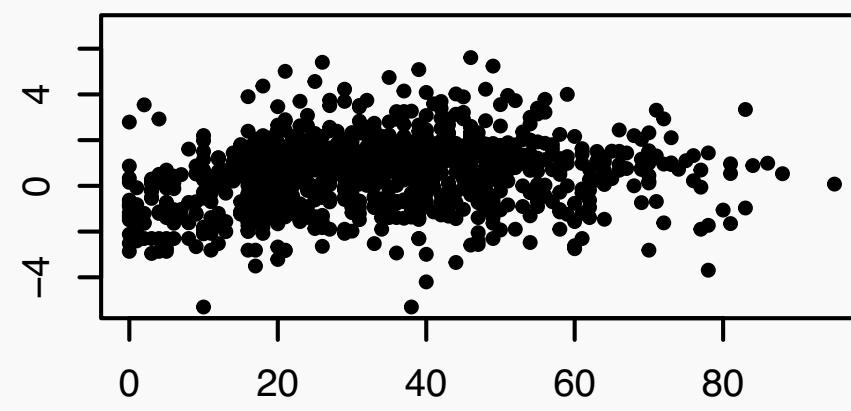
CLMINSUR



SEATBELT



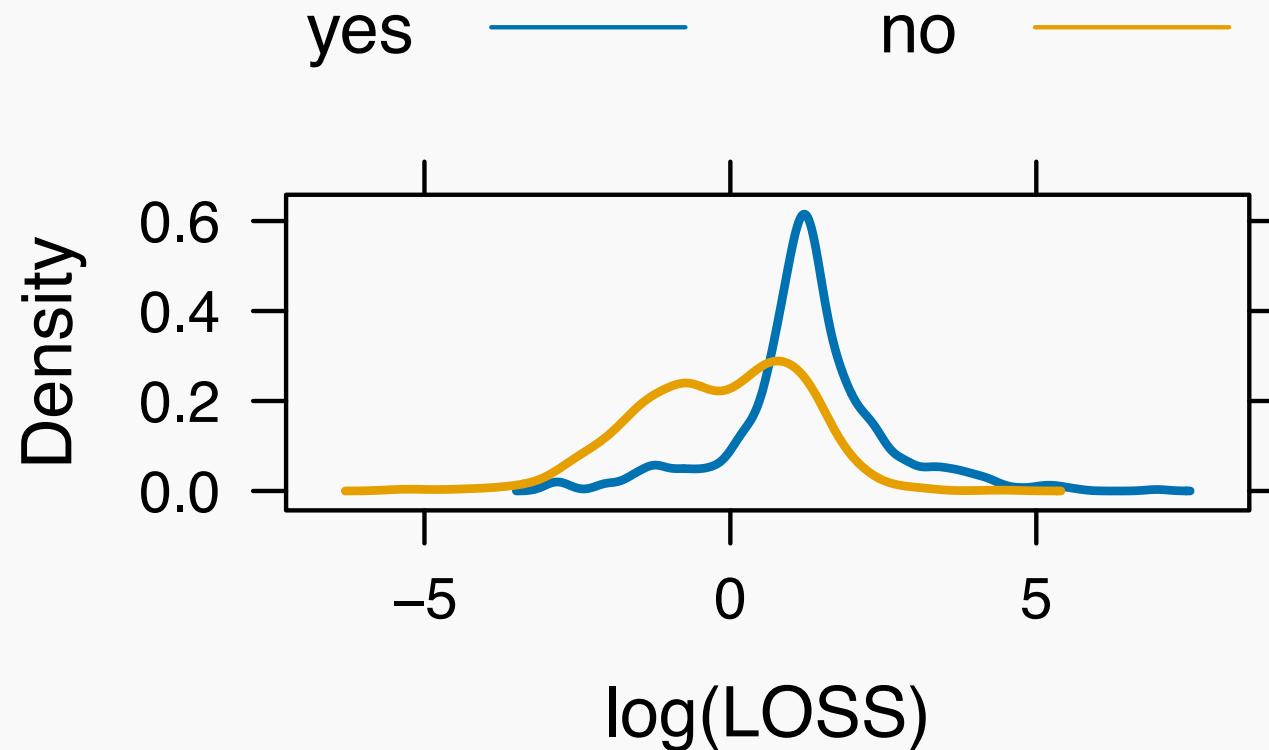
AGECLASS *step function*



CLMAGE

AutoBi: Relation between predictors and log(LOSS)

Being represented by an attorney may be important



Also other variables may matter, such as SEATBELT, MARITAL and AGECLASS. On the contrary CLMSEX, CLMINSUR seem not to have a *marginal* effect. The effect of CLMAGE is not clear.

AutoBi: linear models

Based on the AIC, the following model is selected

 to compensate skewness

Table 1: Fitting linear model: $\log(\text{LOSS}) \sim \text{ATTORNEY} + \text{CLMAGE} + \text{I}(\text{CLMAGE}^2) + \text{SEATBELT}$

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-0.22	0.14	-1.6	0.1
ATTORNEYno	-1.4	0.072	-19	1.6e-67
CLMAGE	0.083	0.0075	11	6.1e-27
I(CLMAGE^2)	-0.00091	9.5e-05	-9.5	1.1e-20
SEATBELTno	0.92	0.27	3.4	0.00059

The model has an R^2 value of around 0.32, and the diagnostic plots (not shown) do not highlight any serious flaw.

More on the effect of age

CLMAGE (in what follows denoted as z) was introduced in the model with three different specifications

1. As a linear term

$$y_i = \beta_0 + \beta z_i + \text{other covariates} + \varepsilon_i$$

2. As a quadratic curve

$$y_i = \beta_0 + \beta_1 z_i + \beta_2 z_i^2 + \text{other covariates} + \varepsilon_i$$

Other specifications could have been considered, such as

$$y_i = \beta_0 + \beta \log z_i + \text{other covariates} + \varepsilon_i$$

$$y_i = \beta_0 + \beta \sqrt{z_i} + \text{other covariates} + \varepsilon_i$$

or any other function $h(z)$:

3. as a discretized version, yet a different number of levels and/or different boundaries for the categories could have been chosen.

Different choices could lead to different results and we can reasonably explore a very restricted number of alternatives.

A different solution: nonlinear (semi-parametric) regression

We would like to specify a model for CLMAGE which is at the same time

- ↪ simple, i.e. depending on a restricted number of parameters;
- ↪ flexible, i.e. capable of modelling a wide range of shapes.

One intuitive (*naive*) idea may be

$$y_i = \beta_0 + \beta_1 z_i + \beta_2 z_i^2 + \dots + \beta_K z_i^K + \text{other} + \varepsilon_i,$$

since a polynomial function can approximate any shape if K is high enough.

Main idea: $\{z_i^0, \dots, z_i^K\}$ can be collinear

Leaving aside the issue of choosing K , this is a possible solution (which may actually work) but is not very efficient due also to possible collinearity among the covariates z_i^k , $k = 1, \dots, K$.

Luckily, we can keep the idea and make things work smoothly (and efficiently) using smoothing splines instead.

Some theory

Semiparametric regression: a basis representation

In semiparametric regression, we use the splines obtained by appropriate specification within the model

$$y_i = \beta_0 + \sum_{k=1}^K b_k B_k(z_i) + \text{other variables} + \varepsilon;$$

of a fixed set of known **basis** functions B_1, \dots, B_K .

In other words, we seek a function describing the relationship between y and z , among those which can be written as

$$s(z) = \sum_{k=1}^K b_k B_k(z).$$

Known in advance

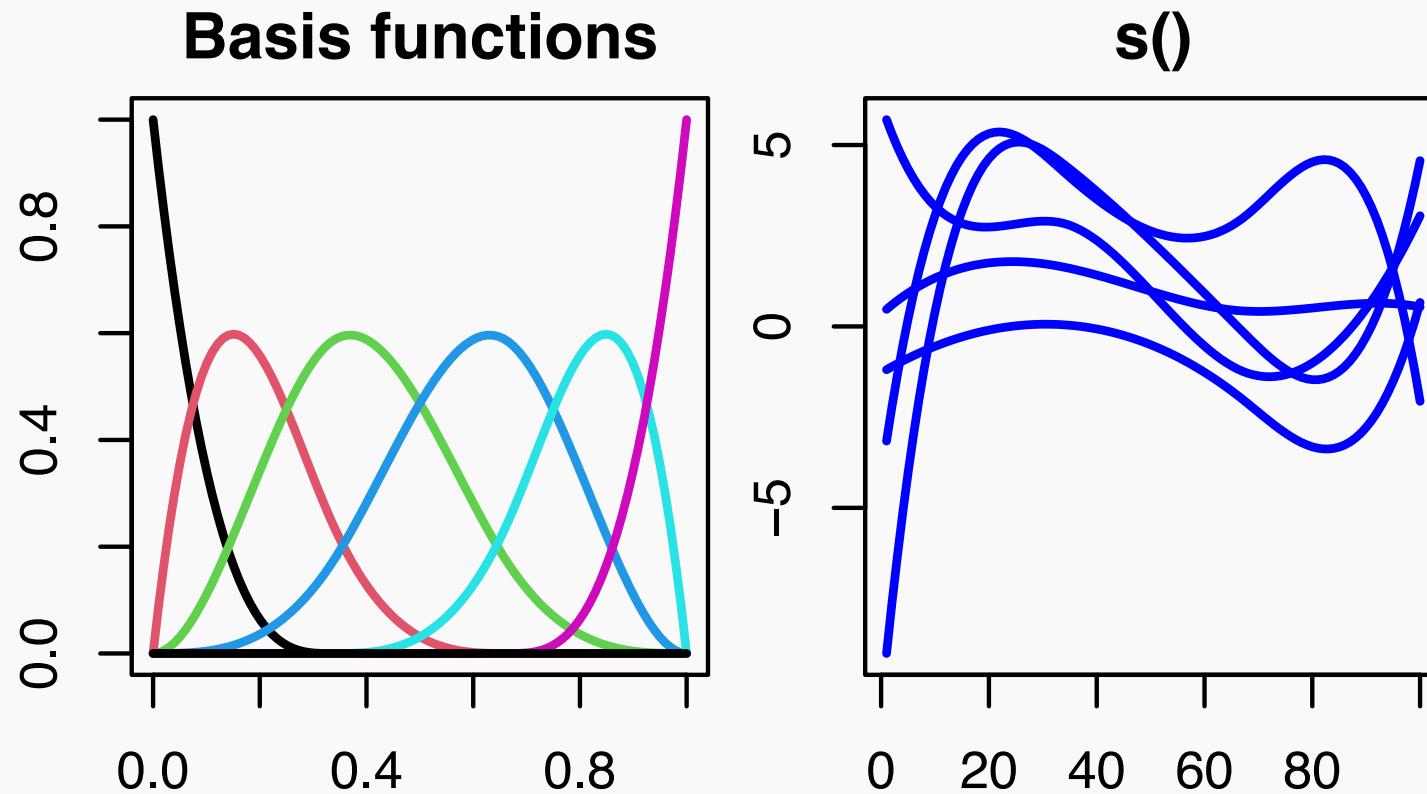
For a suitable choice of the basis function $B_k(z)$ we can obtain very different shapes using relatively few $B_k(z)$, i.e. possibly a low value of K , in an efficient manner.

For a high number of basis functions K the function can be more flexible and we can fit reducing the prediction error.

Semiparametric regression: basis representation

Various possibilities exist for the choice of the basis set.

Below we depict a simple example ($K = 6$) based on **B-splines**, which represent one of the most compelling choices and is more stable than truncated power basis. An example of B-splines is given below



Semiparametric regression: estimation

Given the specification

$$y_i = \beta_0 + \sum_{j=1}^p \beta_j x_{ij} + \sum_{k=1}^K b_k B_k(z_i) + \varepsilon_i$$

this contribution enters additively and don't interact w/ the previous term

and being the function $B_k(\cdot)$ known, estimation may proceed as usual for linear models.

Namely, we minimize with respect to both \mathbf{b} and $\boldsymbol{\beta}$ the sum of squares

$$\sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij} - \sum_{k=1}^K b_k B_k(z_i) \right)^2$$

Residual Sum of Squares \rightarrow MLE \rightarrow Linear Model

but max. K minimizes RSS though this is overfitting

Semiparametric regression: estimation

It may be convenient to define a new matrix

$$\tilde{X} = \begin{bmatrix} 1 & x_{11} & \dots & x_{1p} & B_1(z_1) & \dots & B_K(z_1) \\ \vdots & \vdots & & \vdots & & & \vdots \\ 1 & x_{n1} & \dots & x_{np} & B_1(z_n) & \dots & B_K(z_n) \end{bmatrix}$$

and a new coefficient vector

$$\tilde{\beta} = \begin{bmatrix} \beta \\ \mathbf{b} \end{bmatrix},$$

* Basis of white functions \rightarrow B-splines

* These are defined over the entire interval - in advance
and can be transformed into a piece-wise basis
w/ a suitable matrix transformation

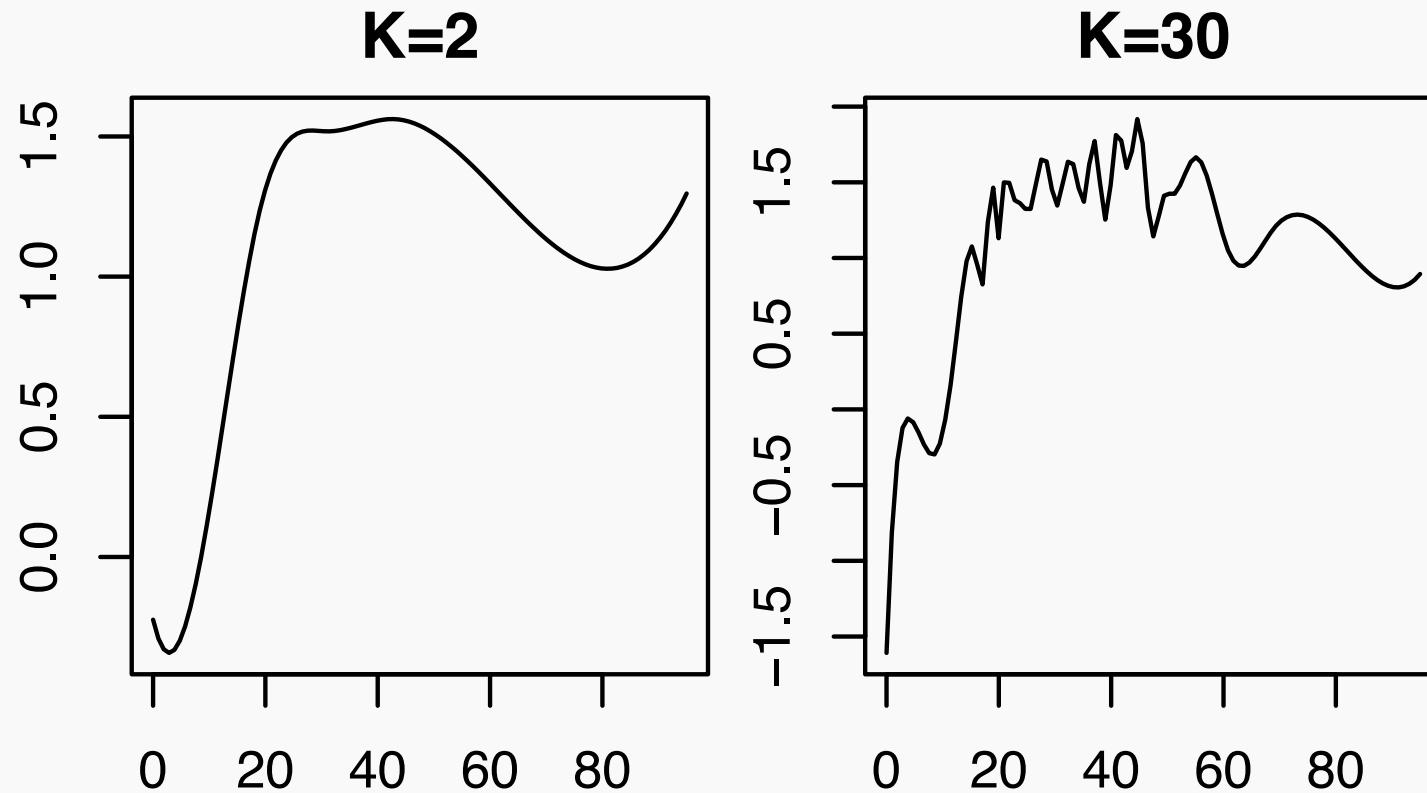
* B-splines "can be more robust" to collinearity

then the sum of squares becomes

$$(\mathbf{y} - \tilde{X}\tilde{\beta})^\top (\mathbf{y} - \tilde{X}\tilde{\beta}) \quad \hat{\tilde{\beta}} = \left(\tilde{X}^\top \tilde{X} \right)^{-1} \tilde{X}^\top \mathbf{y}$$

AutoBi data: estimation

For two models including ATTORNEY, SEATBELT and different sets of $B_k(\text{CLMAGE})$ functions, the estimated $s(\cdot)$ functions are



A larger basis set (higher K) leads to a less smooth estimated function.

Smoothness of regression curve and choice of K

By changing the number of basis functions we estimate curves with different *degrees of smoothness*.

The more basis functions are used

- the better the final curve fits the observations,
- the higher the uncertainty of the estimates (because there are more coefficients to be estimated).

This is the classical **bias-variance trade off**, where more basis functions means less bias but more variance and viceversa.

Choosing the *optimal balance* is part of the estimation procedure, as discussed in the following.

Quantifying the smoothness of the curve

Although changing the number of basis functions is a strategy to tune the smoothness of the curve, it is not the optimal strategy. Using a penalized version of the least squares fit can be considered.

We have defined a measure of the roughness of the spline function $s(\cdot)$ as follows

$$R(s) = \int s''(z)^2 dz,$$

it is null for a straight line (maximum smoothness) and increases as the curve gets less smooth.

Same as before and thus we introduce the penalization factor λ .

It can be shown that **roughness penalty** can be also expressed as a function of the coefficients \mathbf{b} . In particular, there exists a matrix \mathbf{G} such that

$$R(\mathbf{b}) = \mathbf{b}^\top \mathbf{G} \mathbf{b}$$

The specifics of \mathbf{G} depends on the chosen basis.

Penalized sum of squares

A convenient method to tune the smoothness of the estimated curve is then to fix the number of basis functions at a relatively large value, and then penalize the sum of squares according to the roughness penalty adopting the approach of **smoothing splines**.

We then define the penalized sum of squares as

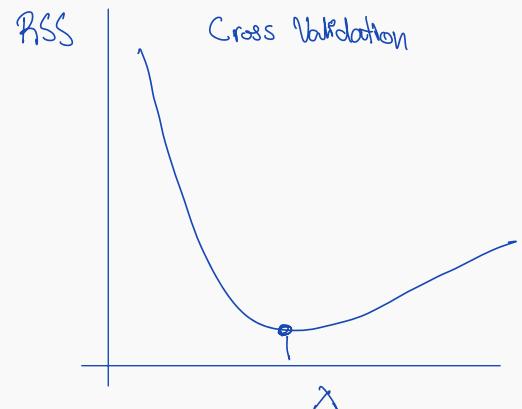
$$(\mathbf{y} - \tilde{\mathbf{X}} \tilde{\boldsymbol{\beta}})^\top (\mathbf{y} - \tilde{\mathbf{X}} \tilde{\boldsymbol{\beta}}) + \lambda \mathbf{b}^\top \mathbf{G} \mathbf{b}$$

where $\lambda > 0$ is a *tuning parameter*.

This gives an estimated $s(\cdot)$ function which is smoother with higher values of λ :

- if $\lambda \rightarrow \infty$ the fitted curve $s(\cdot)$ is a straight line.
- if $\lambda = 0$ no penalty is considered (the curve will be as wiggly as allowed by the number of basis functions).

- versatility
+ versatility



This procedure defines a regression curve for each value of λ .

Choice of tuning parameter

The last step is to choose an optimal degree of smoothness of the estimated curve, that is, an *optimal* λ .

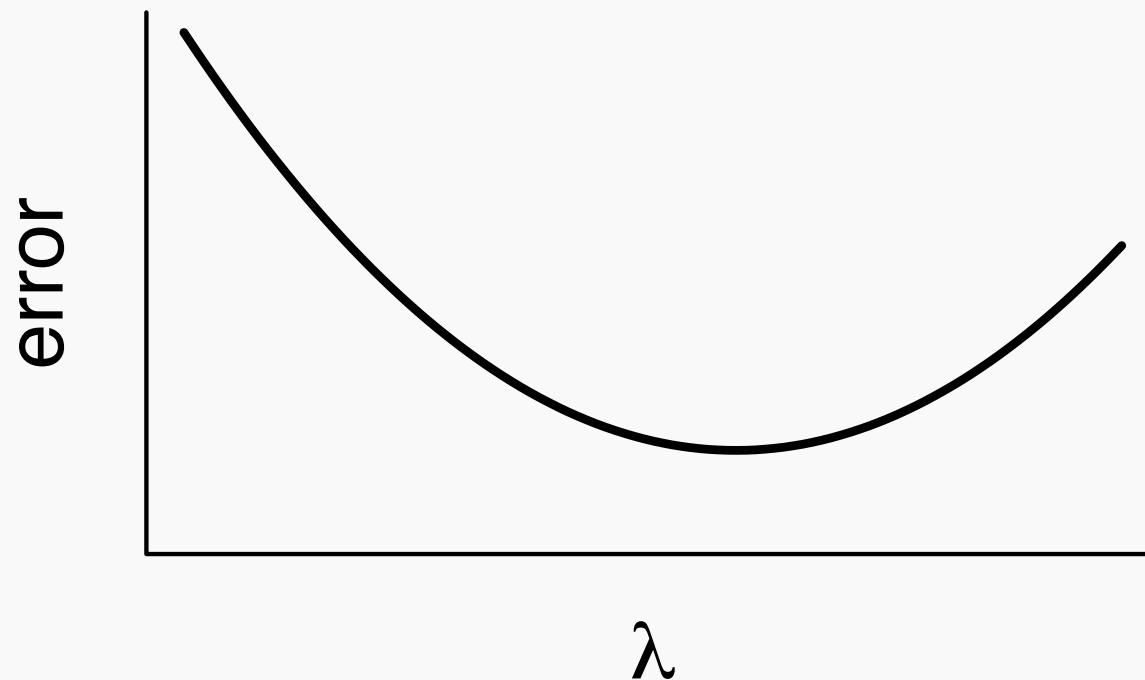
This entails choosing the optimal bias-variance balance as outlined above where a lower λ leads to

- a curve $s(\cdot)$ which better fits sample observations (**smaller bias**)
- a more uncertain estimate (**larger variance**)

The optimal balance can be found looking at the prediction error, that is the error we make when we use the model to perform **prediction on new units**.

Degree of smoothness and predictive accuracy

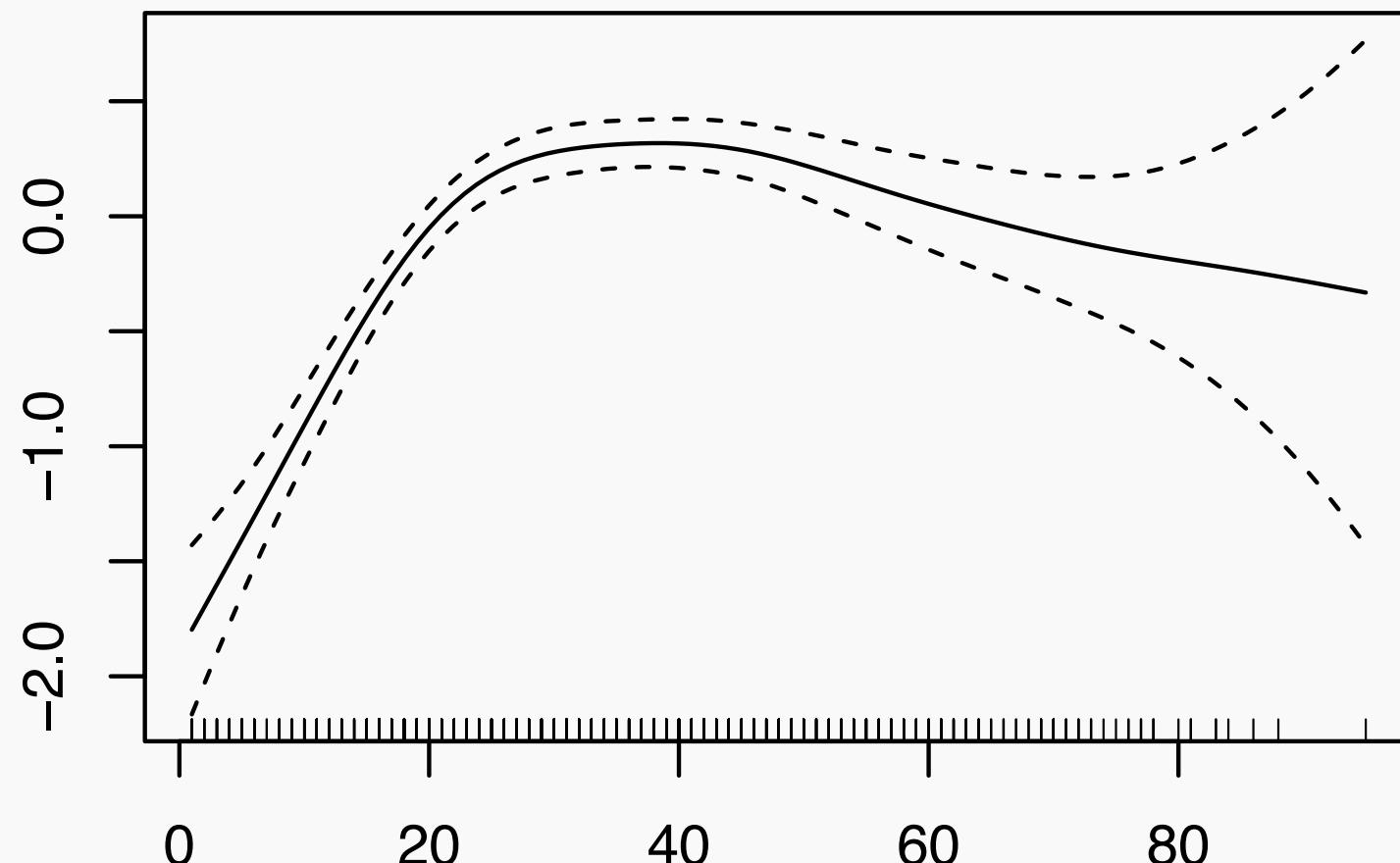
If we measure the prediction error for the different models (as λ varies), we generally get a picture such as



The optimal λ is then the one where the minimum prediction error is attained. As usual with likelihood methods, this can be obtained by **cross validation**, or by less costly alternatives such as the **Generalized Cross Validation (GCV) criterion**, which is similar to the AIC.

AutoBi: optimal λ

Using the `gam` function in the `mgcv` package, The following curve for $\log(\text{LOSS})$ as a function of CLMAGE is obtained; note that the effect of age is now *conditional* on the other predictors



$$\log \text{loss} \sim \text{Circles} + s(\text{CLMAge}) + \dots$$

AutoBi: inference on the other coefficients

Inference for the coefficients is carried out similarly to the linear case, thus we have a coefficient table

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	1.3	0.05	25	3.8e-112
ATTORNEYno	-1.4	0.072	-19	6e-69
SEATBELTno	0.9	0.27	3.4	0.00073

The **b** coefficients are usually not included in the table; an overall significance test for the $s(\cdot)$ function may be reported instead

adaptive spline	edf	Ref.df	F	p-value
s(CLMAGE)	4.6	5.6	28	0

→ we need a line
→ a flexible curve

so just use linear models

The edf is the *estimated degrees of freedom*: the larger the number, the more wiggly the fitted model, with values around 1 close to a linear effect.

R Packages

GCMV simpler to use

GAM

Generalized Additive Models (GAMs)

GAMs: the basic ideas

Generalized Additive Models extend semiparametric regression in two directions:

1. *More than one nonlinear term*: for linear regression with a dependent variable Y and a set of predictor variables X_1, \dots, X_p , the model for the i -th observation is

$$y_i = \beta_0 + \sum_{j=1}^p s_j(x_{ij}) + \text{other variables} + \varepsilon_i,$$

with one smooth term for each predictor, plus possibly other standard (linear) terms. The specification is named **additive** since the various nonlinear terms enter the specification in an additive fashion, with no interaction effects.

(There are ways to introduce interactions, but they require some non-trivial extensions.)

GAMs: the basic ideas

2. *Generalized response*: like for GLMs, binary or count responses are handled by a link function.

The nonlinear terms are introduced in the linear predictor, which now becomes

$$\eta_i = \beta_0 + \sum_{j=1}^p s_j(x_{ij}) + \text{other variables}$$

The estimation proceeds by representing the smooth terms using a suitable basis, and then maximizing the **penalized log-likelihood** to jointly estimate the model coefficients and the basis coefficients

$$\ell(\boldsymbol{\beta}, \mathbf{b}) - \lambda R(\mathbf{b}), \quad \lambda \text{ is a tuning parameter}$$

where like before $R(\mathbf{b})$ is a measure of roughness. The estimation is often carried out using the **backfitting algorithm**, which updates one set of coefficients at a time, though other possibilities exist.

R lab: an example with binary data

As a simple example, we use the function gamSim from mgcv to generate a binary data set from a model with four additive terms

```
dat <- gamSim(1, n = 400, dist = "binary", scale = .33)
```

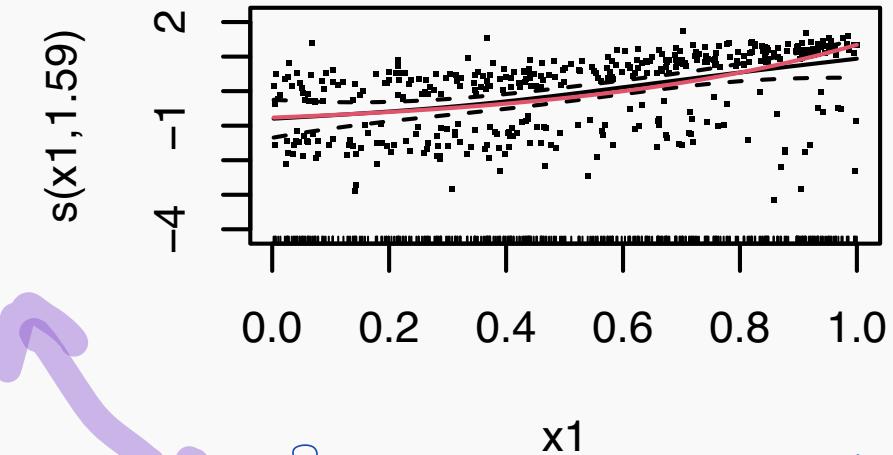
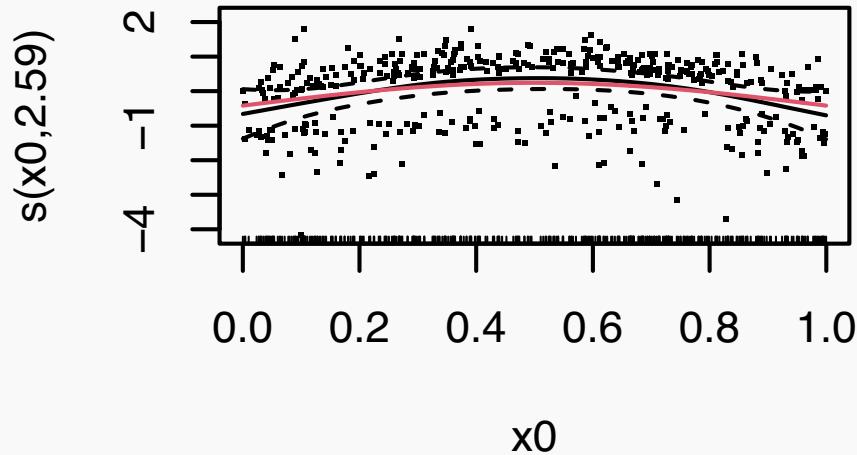
```
## Gu & Wahba 4 term additive model
```

It uses a link function

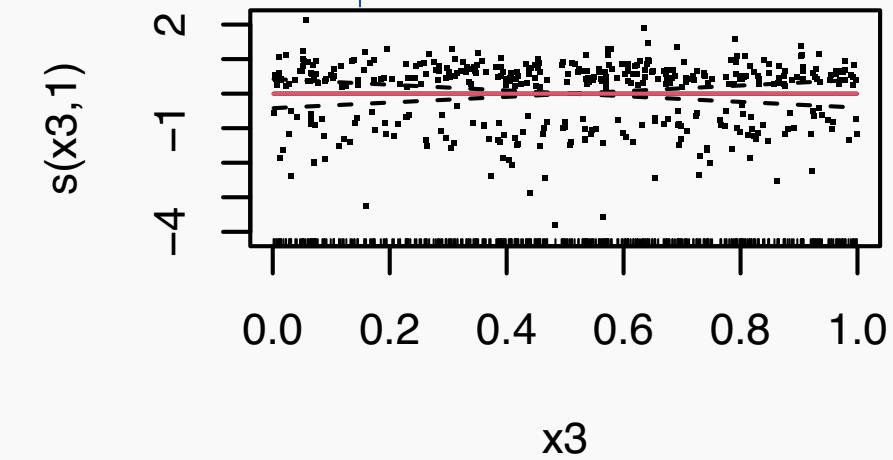
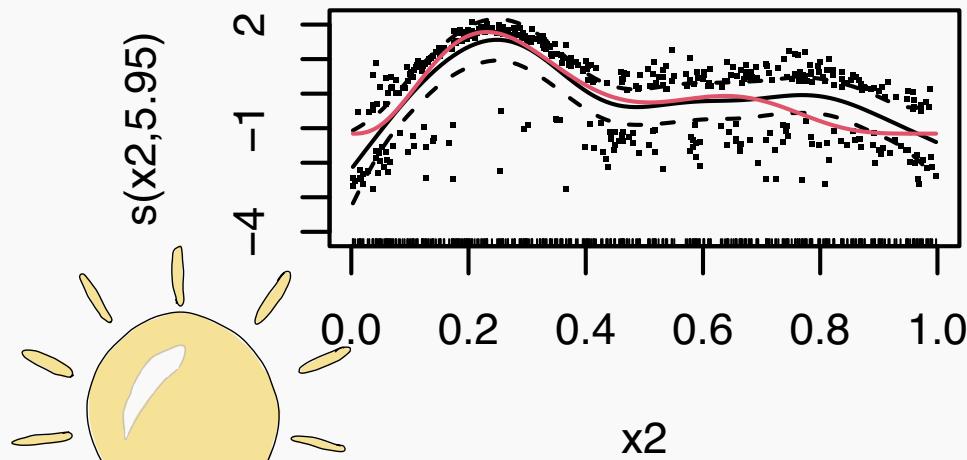
```
lr.fit <- gam(y ~ s(x0) + s(x1) + s(x2) + s(x3),  
               family = binomial,  
               data = dat, method = "REML")
```

Since we know the true model, we can compare it with the results.

Plot model components with truth overlaid in red



↑
x1
Seem to be linear and not spline.



Pay attention to the edf and not necessarily to the p-value.

Start by combining all variables using splines and then select

Don't use splines - flexible function - when not necessary

those who might be used in a linear fashion

R lab: model selection

```
lr.fit1 <- gam(y ~ s(x0) + s(x1) + s(x2), family = binomial,  
                 data = dat, method = "REML")  
  
lr.fit2 <- gam(y ~ s(x1) + s(x2), family = binomial,  
                 data = dat, method = "REML")  
  
AIC(lr.fit, lr.fit1, lr.fit2)  
  
##          df      AIC  
## lr.fit  14.283336 454.8213  
## lr.fit1 13.266372 452.8258  
## lr.fit2  8.473195 455.4590
```

less complex model.

Winding up

- What we have seen is just a glimpse of a very large body of methods.
- Indeed, the approach based on smoothing splines with a roughness penalty is just one of several available in statistics. It has the strong advantage of being extendable in several directions, to cover also more complex settings for non-independent data.

Classification and Regression trees

(Recursive partitioning)

N. Torelli, G. Di Credico, V. Gioia
2023

University of Trieste

Regression Trees

Classification Trees

MARS: Multivariate Adaptive Regression Splines

Regression Trees

Step functions as approximators

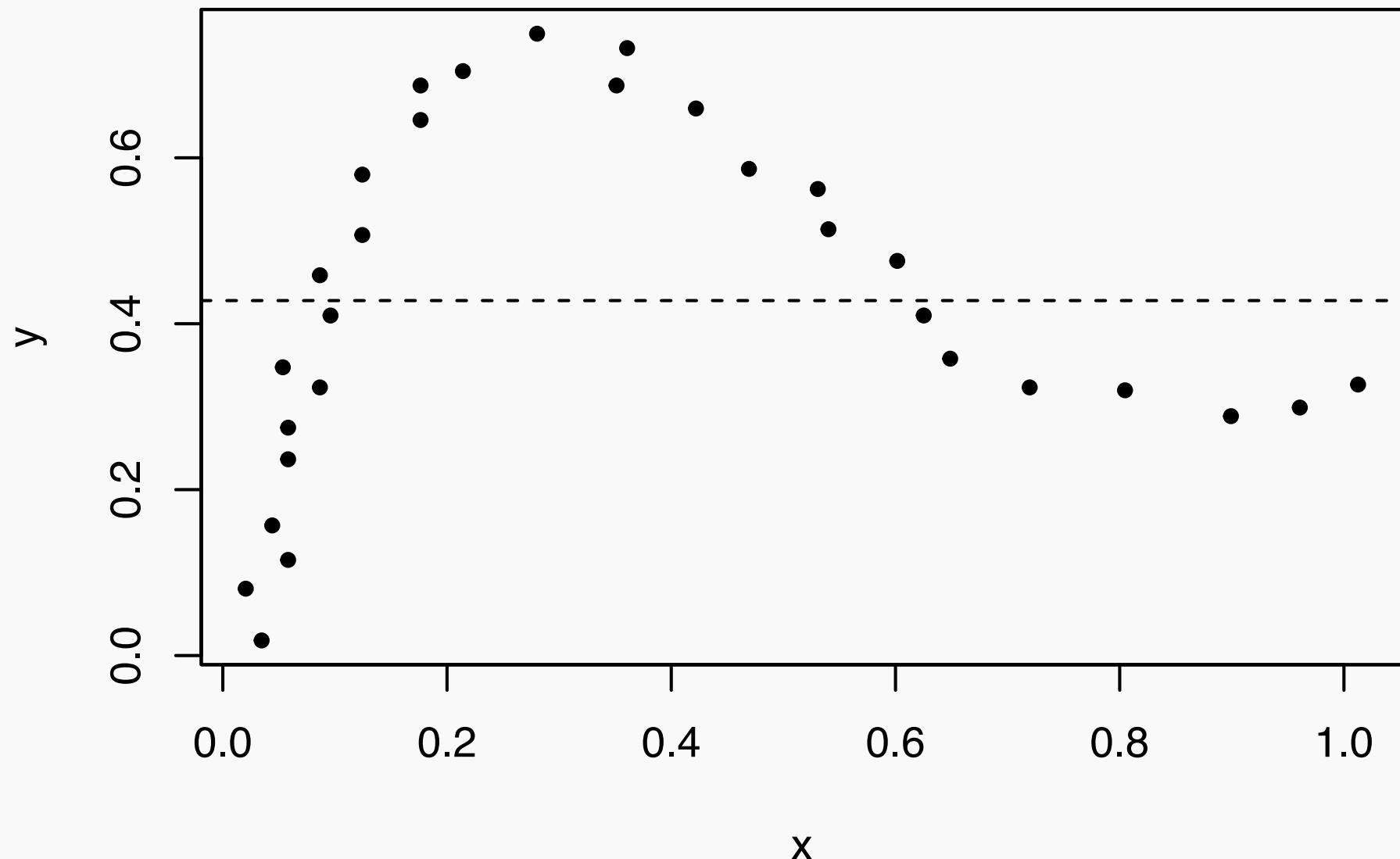
- A simple, yet effective, way to approximate a generic function $f(x)$ is to use a step function, that is, a piecewise constant function
- In such a case, there are various choices to be made:
 - where are the subdivision points to be placed?
 - which value of y must be assigned to each interval?
 - how many subdivisions of the x axis must be considered?
- The idea is to generalize the use of step functions to approximate (or predict) a response Y as function of some covariates.
- Note that Y could be of different nature: numeric, factor, counts

Step functions as a spline

- A step function actually is a spline of degree 0. Assume we want to fit such a function to a simple set of data.
- Subdivision points are now the knots and their position should be chosen to reflect changes of the function $f(x)$ (for instance more knots where the function is steeper)
- In a given interval the value of the constant can be chosen to be an average of the level of the function itself
- The choice of the number of subdivisions is critical: any increase in the number of steps increases the quality of the approximation, and therefore we are led to think of infinite subdivisions.
- However, this is counter to the requirement to use a approximate representation using few parameters and therefore to adopt a finite number of subdivisions.

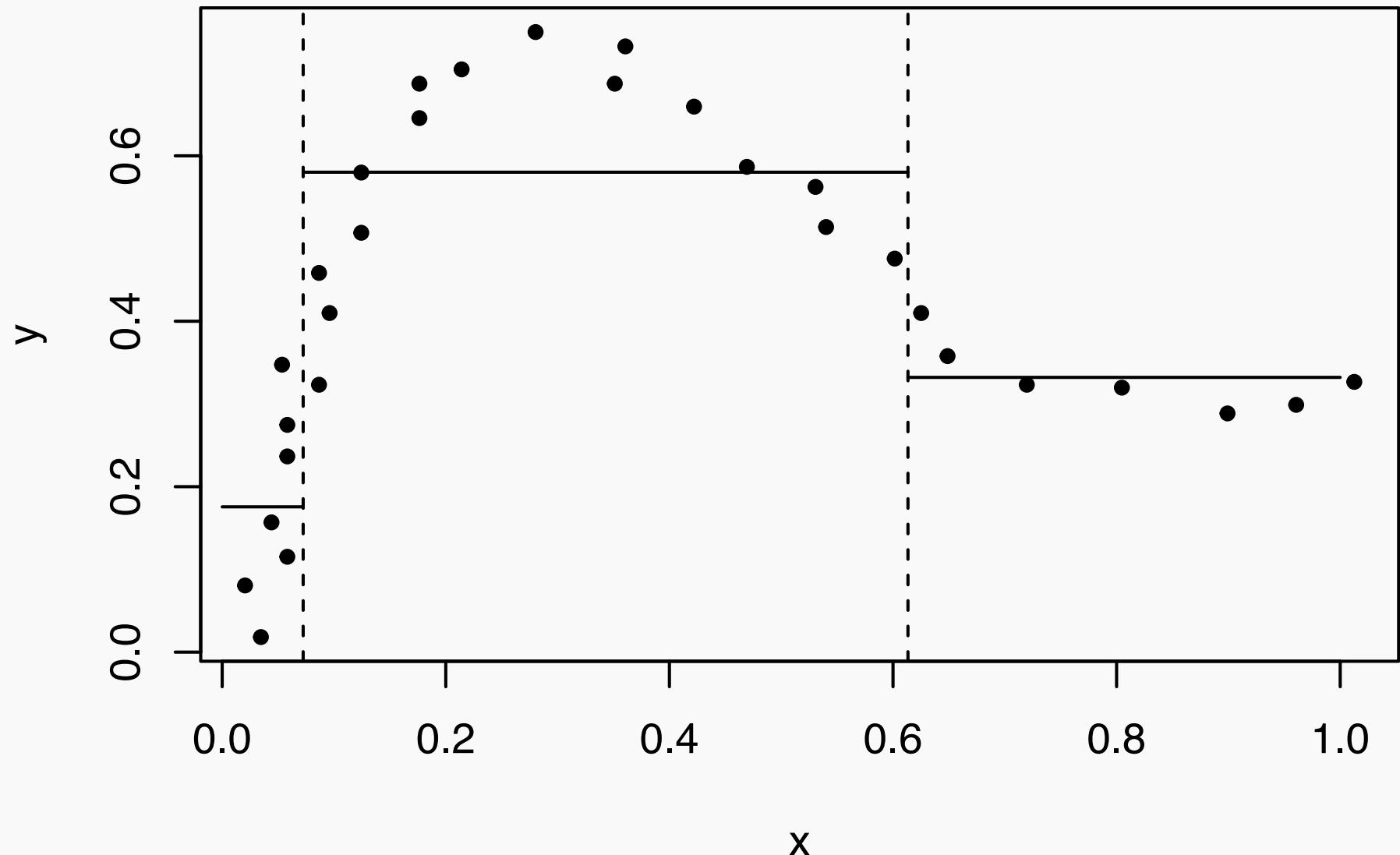
An introductory example

- If y is quantitative a global approximation of y could be its mean. Or we can use a (regression) function $g(\cdot)$



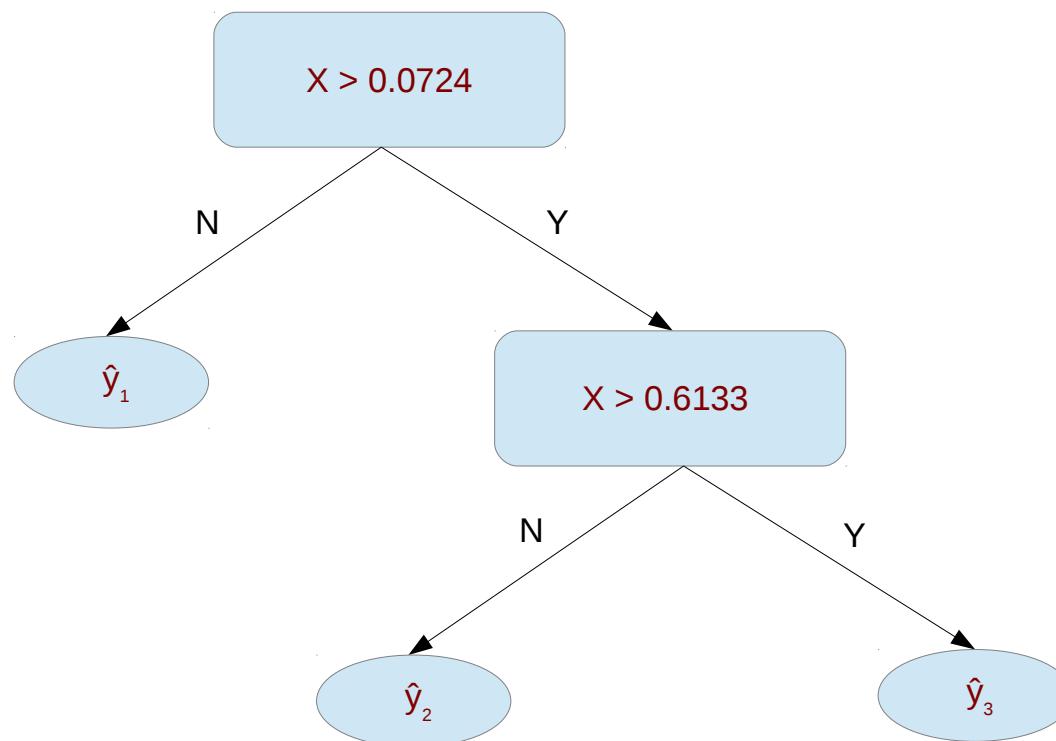
An introductory example

- Now consider a subdivision on X and approximate y with its local mean \hat{y}_i in the i -th interval and g is a piecewise constant function



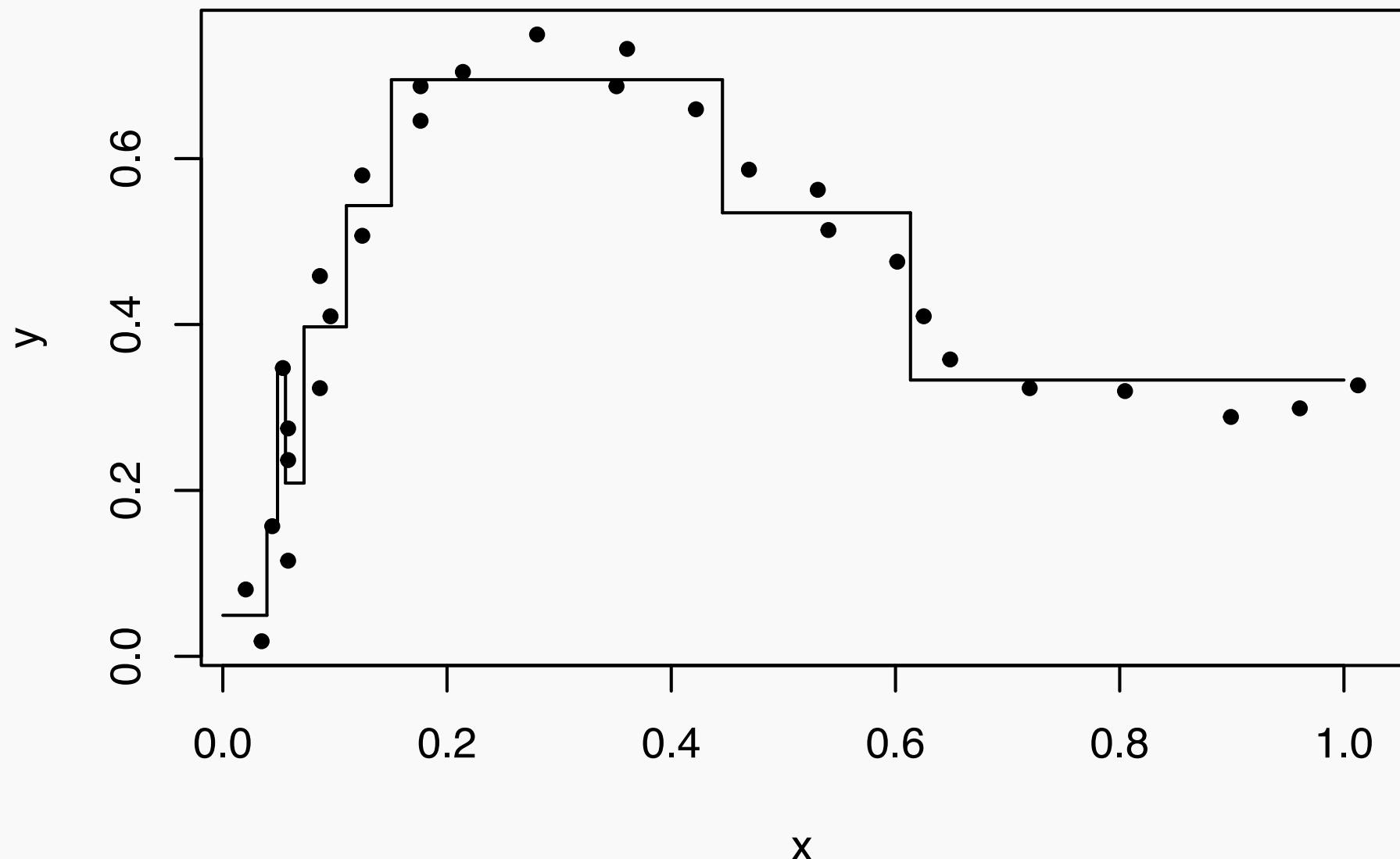
The tree

Note that the value \hat{y}_i of the function g can be also described by the following tree



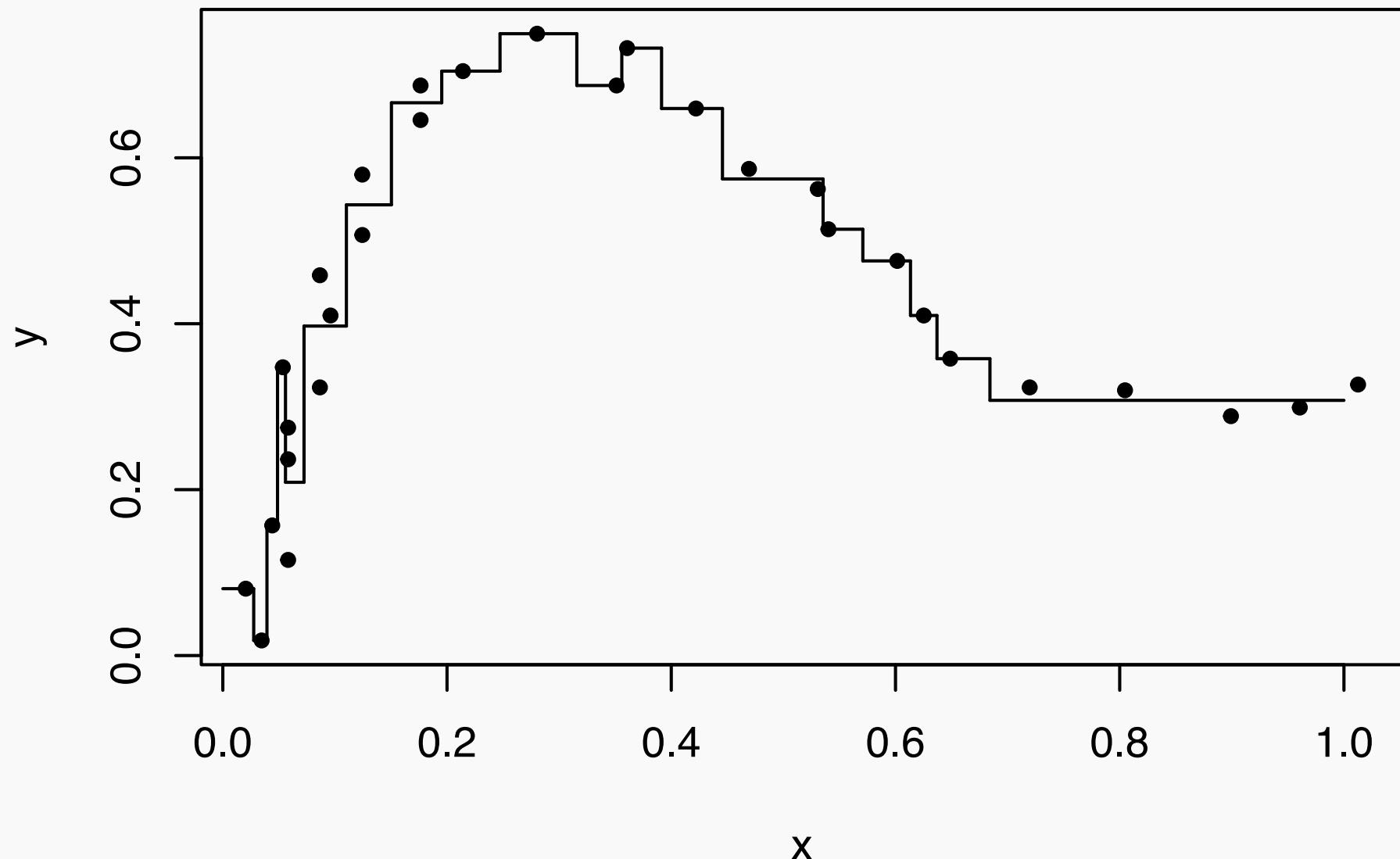
An introductory example

- As the number of intervals increase, we could achieve a very accurate description of the data



An introductory example

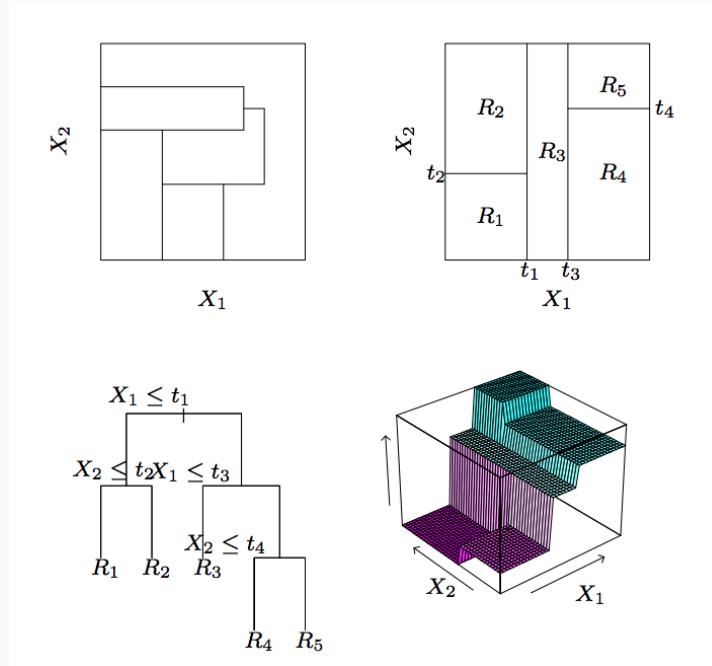
- As the number of intervals increase, we could achieve a very accurate description of the data (leading to overfitting)



Tree approximation

- Let's now consider a regression problem with continuous response Y and two covariates X_1 and X_2 . We want to estimate the generic regression curve $E(Y) = f(x_1, x_2)$. $= E[Y|X]$
- The idea is again to partition the space spanned by the covariates and to model Y with a different constant in each element of the partition
- we restrict attention to recursive binary partitions.
 - First split the space into two regions, and model the response by the mean of Y in each region.
 - variable and split-point are chosen in order to achieve the best fit.
 - one or both of these regions are split into two more regions,
 - the process is continued, until some stopping rule is applied.

A simple example of tree partitioning for two covariates



In the top right panel first split at $X_1 = t_1$. Then the region $X_1 \leq t_1$ is split at $X_2 = t_2$ and the region $X_1 > t_1$ is split at $X_1 = t_3$. Finally, the region $X_1 > t_3$ is split at $X_2 = t_4$. The result of such a recursive binary splitting is a partition into the five regions R_1, R_2, \dots, R_5 shown in the figure.

- The corresponding regression model predicts Y with a constant c_m in region R_m , that is, $\hat{f}(X_1, X_2) = \sum_{m=1}^5 c_m I\{(X_1, X_2) \in R_m\}$
- The sets R_m are rectangles, in the 2-dimensional space, with their edges parallel to the coordinate axes) and c_1, \dots, c_5 are constants. Note that the top left panel represents a partition that cannot be obtained by recursive binary splitting

A regression tree

- More generally:

- we want estimate a regression curve $f(x_1, x_2, \dots, x_p)$ underlying the data by $\hat{f}(x_1, x_2, \dots, x_p) = \sum_{m=1}^M c_m I\{(x_1, x_2, \dots, x_p) \in R_m\}$ where $I(x_1, x_2, \dots, x_p \in R_m)$ is the indicator function of the set R_m (R_m are rectangles, in the p -dimensional sense, with their edges parallel to the coordinate axes) and c_1, \dots, c_M are constants.
- Given an objective function such as the Deviance

$$D = \sum_{i=1}^n (y_i - \hat{f}(x_{1i}, x_{2i}, \dots, x_{pi}))^2$$

- the goal is to define a partition of the space of the covariates that minimizes D

Regression trees are used to consider interaction between variables.

Building the Regression tree

- this minimization, even if we fix the number of the elements of the partition, involves very complex computation
- a sub-optimal approach is considered using a step-by-step optimization: we construct a sequence of gradually more refined approximations and to each of these we minimize the deviance relative to the passage from the current approximation to the previous one
- It is not ensured that we get the global maximum. This procedure is called greedy-algorithm
- This operation is represented by a series of binary splits
- Each internal node represents a value query on one of the variables – e.g. “Is $x_3 > 0.4?$ ”. If the answer is ‘Yes’, go right, else go left.
- The terminal nodes are the decision nodes. Typically each terminal node is assigned a value, c_h , given by the arithmetic mean of the observed y_i having component x_{ji} falling in this node.

Growing the tree

- Trees are grown using a random subset of the available data (*the training data*), by recursive splitting
- A terminal node g is split into the left and right daughters (g_L and g_R) that increase the split criterion

$$D_g - D_{g_L} - D_{g_R}$$

the most, where D is the deviance associated to a given node.

- To avoid the overfitting, a large tree T_0 is grown and then pruned backward
- Indeed a tree with n leaves is equivalent to a polynomial regression of degree $n - 1$
- detection of the variable X_j that achieve the best split at each node and which is the split point can be done very quickly and hence by scanning through all of the inputs
- Deviance can be adapted for dealing with a response that is a count or a duration

Pruning the tree

- Pruning criterion: cost of a subtree $T \in T_0$, is defined by

$$C_\alpha(J) = \sum_{j=1}^J D_j + \alpha_j$$

- Here the sum is over the terminal nodes of T , J is the number of terminal nodes in T and α is a cost-complexity parameter
- The choice of an optimal size is evaluated by cross-validation, or on a validation set.
- For each α the best subtree T_α is found via weakest link pruning
- Larger α gives smaller trees
- A best value $\hat{\alpha}$ is estimated via cross-validation (or on a validation set)
- Final chosen tree is $T_{\hat{\alpha}}$
- New observations are classified by passing their x down to a terminal node of the tree, and then using the relative c_h .

An example

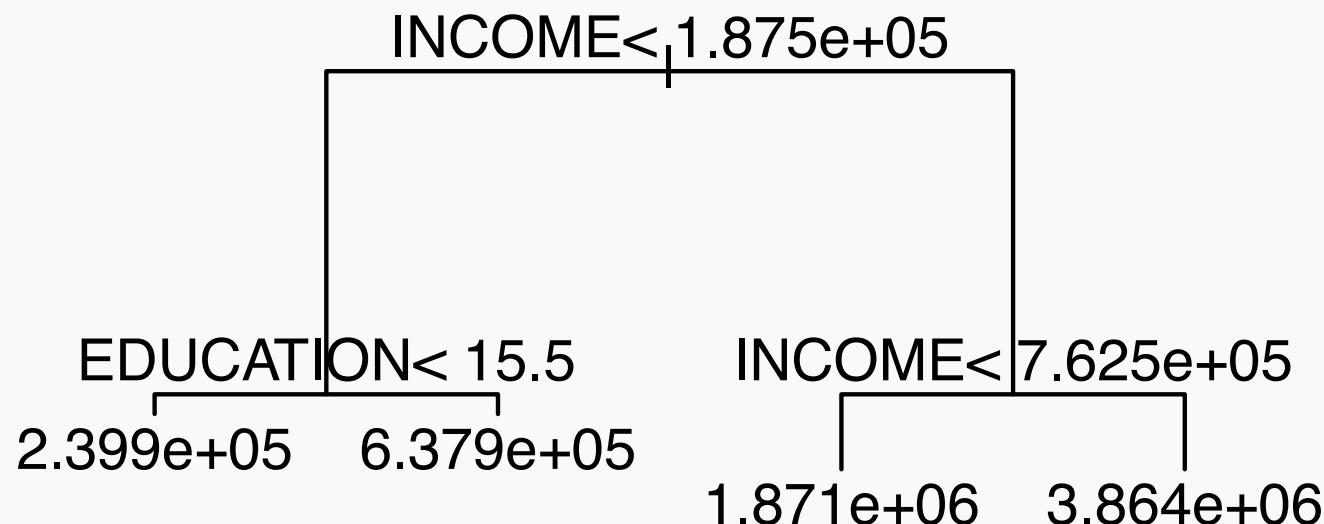
The variable FACE refer to the amount of life insurance bought by the head of a household. We want to predict it by using “INCOME”, number of household members, AGE, Education, etc. For illustration, a tree with maximum depth=2 is considered. Package `rpart` is used.

```
TL <- read.csv("TL.csv", header=TRUE, sep=",", row.names=1)
library(rpart)
attach(TL)

m2 <- rpart(FACE~INCOME+MARSTAT+NUMHH+EDUCATION+AGE,
             control=rpart.control(maxdepth=2))
m2

## n= 275
##
## node), split, n, deviance, yval
##      * denotes terminal node
##
## 1) root 275 7.681561e+14  747581.5
##    2) INCOME< 187500 227 2.629158e+14  413511.5
##      4) EDUCATION< 15.5 128 1.075360e+14  239930.5 *
##      5) EDUCATION>=15.5 99 1.465367e+14  637939.4 *
##    3) INCOME>=187500 48 3.600986e+14  2327454.0
##      6) INCOME< 762500 37 1.905974e+14  1870751.0 *
##      7) INCOME>=762500 11 1.358255e+14  3863636.0 *
```

The tree



Interactions between variables.

Classification Trees

Classification Trees

- If the target (response) variable is a categorical variable taking values $1, 2, \dots, K$, the only changes needed in the tree algorithm pertain to the criteria for splitting nodes and possibly pruning the tree.
- In these cases the tree will be used for predicting the categorical response and this is labeled as a **classification problem**. And the tree is then a **Classification tree**.
- Also in this case a tree is a hierarchical structure formed by:
 - root: the predictor space
 - nodes:
 1. internal: test an explanatory variable (and splits the predictor space)
 2. terminal (leaf): assign a label class
 - branches: corresponds to values of the explanatory variables
- A tree is constructed by repeated splits of the predictor space (root) into subregions (nodes). Each terminal region is associated with a prediction and their union form a partition of the predictor space.

Growing a classification tree

The following elements are needed

- A set of splits
- A goodness of split criterion
- A stop-splitting rule
- A rule for assigning every terminal node to a class
- Each split depends on the value of a single predictor x_j and depends on the nature of x_j :
 - qualitative, with values in $\mathcal{L} = \{l_1, \dots, l_K\}$: a split is any question as “is $x_j \in S_{\mathcal{L}}$?” with $S_{\mathcal{L}}$ a subset of \mathcal{L} ;
 - quantitative, with range (a, b) : a split is any question as “is $x_j \leq s$?” with $a \leq s < b$
- Examples
 - “Is the age of the subject not greater than 60?”
 - “Is the weather cloudy or rainy?”
- At each step of the tree growing procedure, the best split is identified for each predictor and, among these, the best of the best is selected.

The goodness of split criterion

- The objective of classification tree construction is to finally obtain nodes that are as **pure** as possible, i.e., the split should send towards each branch observations of the same class
- It makes sense to consider good a split when it leads to a high **reduction of impurity** of the node (a high increase of the prediction/classification accuracy).
- Consider a node t for a two class classification problem, the two classes of y have frequency $p(t)$ and $1 - p(t)$. An **impurity measure** of a node t is a function of the proportion of units into the two classes. Let us consider the **misclassification error** defined as

$$i(t) = 1 - \max(p(t), (1 - p(t)))$$

as an impurity measure

- If the node is equipped with a split sending a proportion of p_L and p_R to the left and, respectively right, the gained reduction of impurity is:

$$\Delta i(t) = i(t) - p_L i(t_L) - p_R i(t_R)$$

- The best split is the split which maximizes the reduction of impurity

Impurity measures

More generally, for a multiclass problem, for a given node m that defines a region R_M with N_M observations, \hat{p}_{mk} is the observed proportion of cases in class k . The observation at the node will be classified in class $k(m)$ that is the class for which \hat{p}_{mk} is larger. The following impurity measures can be defined:

- Misclassification error:

$$\frac{1}{N_M} \sum_{i \in R_m} I(y_i \neq k(m)) = 1 - \hat{p}_{mk(m)}$$

- Gini index (heterogeneity index):

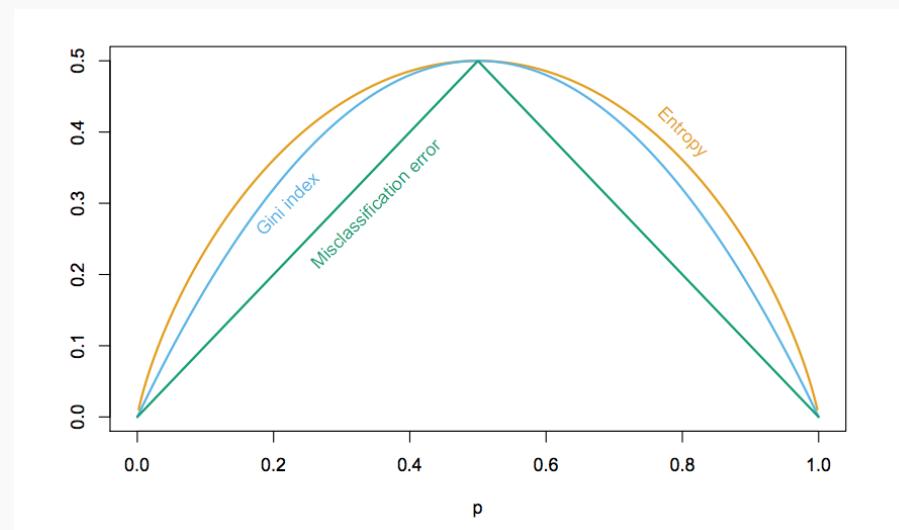
$$G = \sum_{k=1}^K \hat{p}_{mk}(1 - \hat{p}_{mk})$$

- Entropy:

$$H = - \sum_{k=1}^K \hat{p}_{mk} \log \hat{p}_{mk}$$

Measures of impurity in two-class problems

- for $K = 2$, with p the observed proportion in the second class, these three measures are respectively:
 - $1 - \max(p, 1 - p)$
 - $2p(1 - p) = 2(p - p^2)$
 - $-p \log p - (1 - p) \log(1 - p)$



Avoiding overfitting

- If the overall accuracy is too low we may always make the tree growing further
- The flexibility of the trees would in principle allow for building a perfect classification rule
- A tree that perfectly fits the sample data probably overfits the data: useless for predicting new data, not used for training the tree!
- A useful practice is to evaluate the accuracy of the estimated tree on a test set (out-of-sample).
- Often for Regression and Classification trees the available data are randomly subdivided into three sets:
 - the training set (to grow the tree)
 - the validation set (to prune it)
 - the test set (to evaluate it)
- Evaluation of the quality of the three can be achieved with usual tools for evaluating the prediction (classification) quality: Mean squared prediction errors, confusion matrices, ROC curves (see the R package ‘caret’)

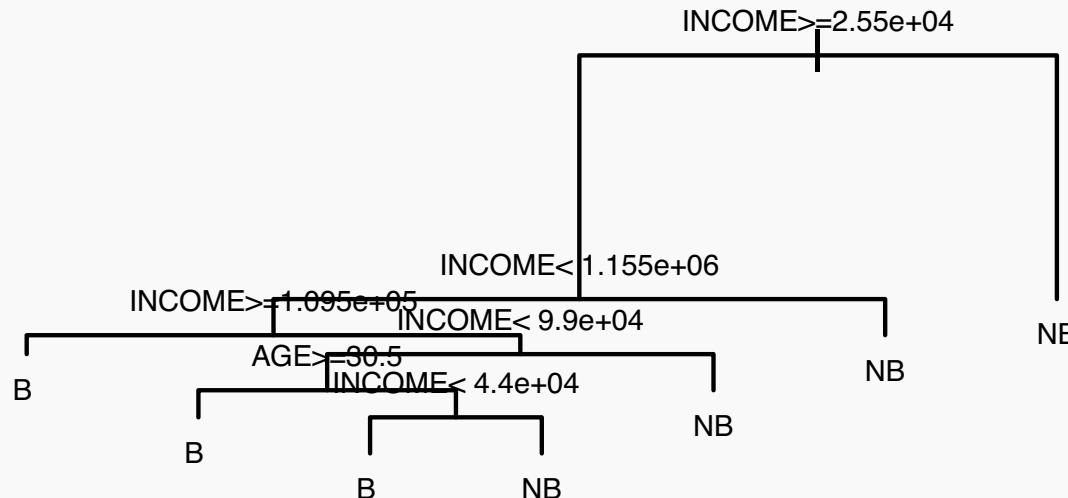
An example of two class tree

We want to predict now if a life insurance policy is bought using the same covariates

```
TL <- read.csv("TLbin.csv", header=TRUE, sep=",", row.names=1);
attach(TL); set.seed(4321); ind.train <- sample(1:500,300) ;
TL.train <- TL[ind.train,]; TL.test <- TL[-ind.train,]
tree <- rpart(FACEPOS~., data=TL.train); tree

## n= 300
##
## node), split, n, loss, yval, (yprob)
##      * denotes terminal node
##
## 1) root 300 136 B (0.5466667 0.4533333)
##    2) INCOME>=25500 235  90 B (0.6170213 0.3829787)
##      4) INCOME< 1155000 227   84 B (0.6299559 0.3700441)
##        8) INCOME>=109500 72   19 B (0.7361111 0.2638889) *
##        9) INCOME< 109500 155   65 B (0.5806452 0.4193548)
##          18) INCOME< 99000 145   58 B (0.6000000 0.4000000)
##            36) AGE>=30.5 122   45 B (0.6311475 0.3688525) *
##            37) AGE< 30.5 23   10 NB (0.4347826 0.5652174)
##              74) INCOME< 44000 14    5 B (0.6428571 0.3571429) *
##              75) INCOME>=44000 9     1 NB (0.1111111 0.8888889) *
##              19) INCOME>=99000 10    3 NB (0.3000000 0.7000000) *
##              5) INCOME>=1155000 8     2 NB (0.2500000 0.7500000) *
##              3) INCOME< 25500 65   19 NB (0.2923077 0.7076923) *
```

The tree



```
pred.test <- predict(tree, newdata=TL.test, type="class")
t <- table(TL.test$FACEPOS, pred.test)
t

##      pred.test
##      B NB
##      B 78 33
##      NB 49 40
sum(diag(t))/sum(t)

## [1] 0.59
```

Dealing with missing data

- It is quite common to have observations with missing values for one or more input features. The usual approach in statistics is to impute (fill-in) the missing values in some way.
- However, the first issue in dealing with missing data is whether the missing data introduce a sample selection that can bias results of analyses.
- It is important consider if missing data arise by a
 - Missing Completely at Random (MCAR) mechanism (*no bias*)
 - Missing at Random (MAR) mechanism (*possible bias if the dependence on missingness on some observed covariates are not recognized*)
 - Missing Not at Random (MNAR) mechanism (*huge problems, likely to have non negligible bias*)
- For the first, and possibly, the second case, in regression trees two approaches can be used when predictors have missing values:
 - if it is categorical, add a specific category for missing values
 - if it is continuous, use surrogate predictors to be used when observation is missing on the primary predictor.

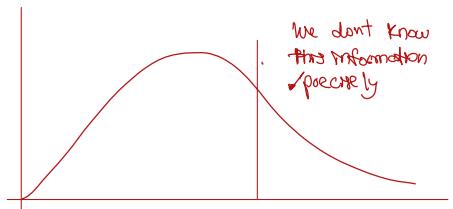
II Dropping variables w/ NA variables dependent if they are missing at random or not.

- Particularly Missing Completely at Random or
Missing at Random
Non missing at Random
- } We can use a third variable (z) to analyze the randomness behavior of y on x .
- ↳ Result from any model will be imprecise

II With trees we don't lose data!

- For categorical data \rightarrow add NA category
- For numerical data \rightarrow surrogate trees

Only for M&S



Regression and classification trees: Advantages

- Logical simplicity and ease of ‘communication’ (particularly those with a non-quantitative background)
- The step function has a simple, compact mathematical formulation in terms of information to be stored
- Speed of computation and can take advantage of parallel calculation
- Can handle huge datasets
- Can handle mixed predictors: quantitative and factors
- Easy ignore redundant variables and automatically detects interactions among variables
- Handle missing data elegantly
- Small trees are easy to interpret

Regression and classification trees: Disadvantages

- Instability of results: very sensitive to the insertion/changes in the sample
- Difficulty in upgrading: if more data arrive, they cannot be added to the already constructed tree; it is necessary to start again from the beginning.
- Difficulty of approximating some mathematically simple functions, particularly if they are steep,
- Statistical inference: formal procedures of statistical inference such as hypothesis testing, confidence intervals, and others are not available.
- (over?) emphasizes interactions
- large trees are hard to interpret
- prediction surface is not smooth

MARS: Multivariate Adaptive Regression Splines

trees + splines
knots suggested by data itself

trees good for interaction of covariates
splines interaction of covariates is complicated

MARS: Multivariate Adaptive Regression Splines

- MARS is an adaptive procedure for regression, and is well suited for high dimensional problems (i.e., a large number of inputs).
- It can be viewed as a generalization of stepwise linear regression or a modification of the CART. This latter approach for regression tree leads to smoother prediction surfaces
- A hybrid of MARS called PolyMARS specifically designed to handle classification problems has been also proposed
- MARS is a semi-parametric method that like CART uses a greedy algorithm and recursively adapt a curve to the regression surface
- At each step it is chosen a couple of basis functions recursively selecting the variable X that is most appropriate and the optimal position of the knot.

MARS

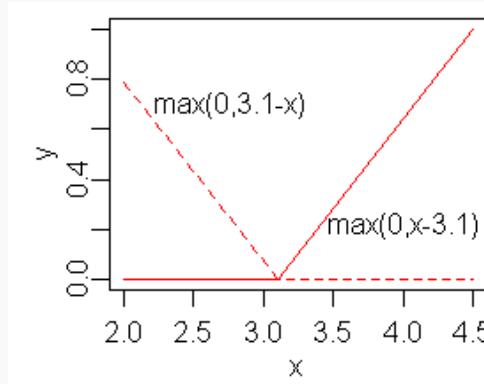
- MARS builds models of the form

$$\hat{f}(x) = \sum_{i=1}^k c_i B_i(x)$$

*f variable of a time
depending on the step.*

- The model is a weighted sum of basis functions $B_i(x)$. Each c_i is a constant coefficient.
- Each basis function $B_i(x)$ takes one of the following three forms:
 1. a constant
 2. a hinge function. A hinge function has the form $\max(0, x - const)$ or $\max(0, const - x)$.
MARS automatically selects variables and values of those variables for knots of the hinge functions.
 3. a product of two or more hinge functions. These basis functions can model interaction between two or more variables.

This is an example of a couple of Hinge functions

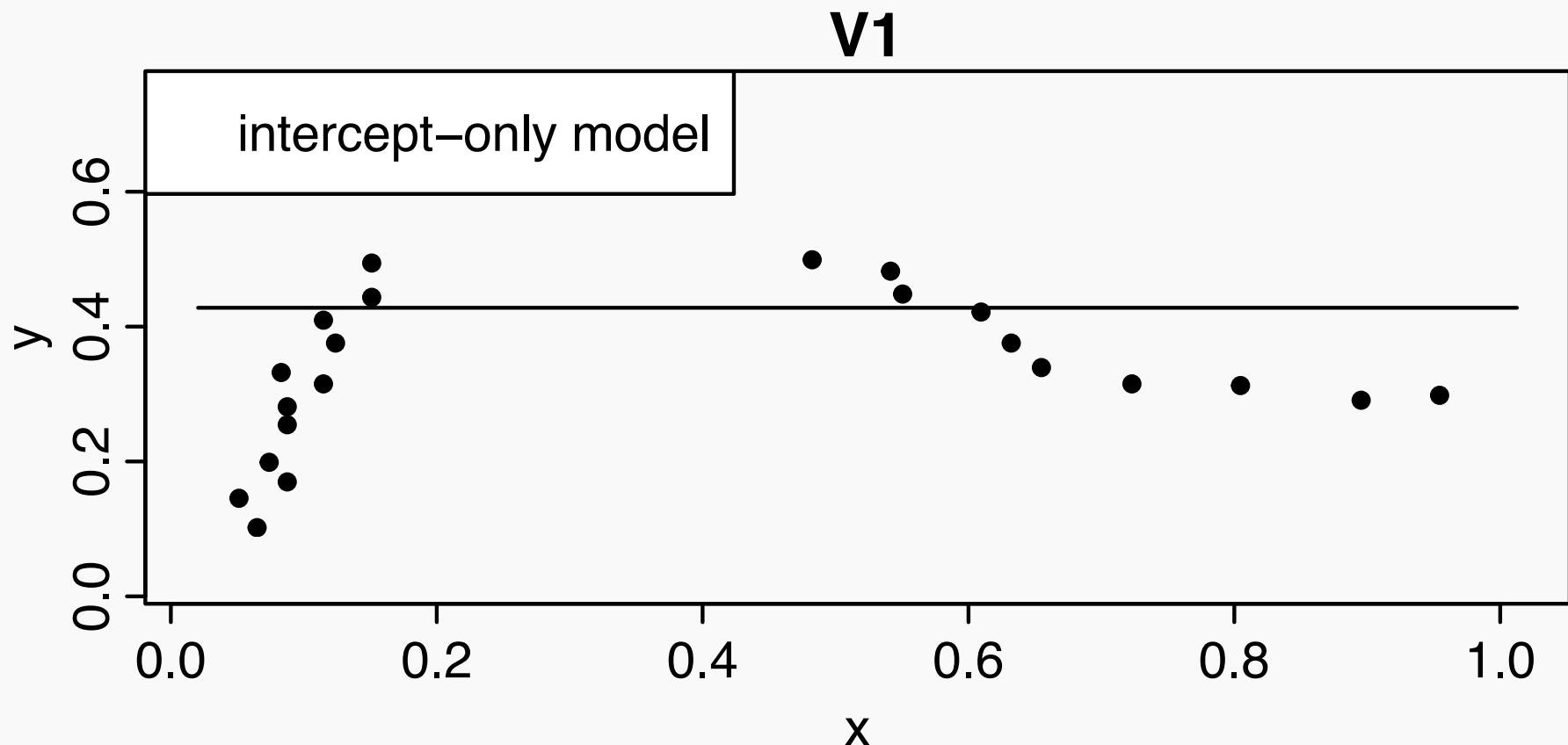


- Although they might seem quite different, the MARS and CART strategies actually have strong similarities.
- Suppose we take the CART procedure and make the following changes:
 - Replace step functions by the piecewise linear basis functions $I(x - t > 0)$ and $I(x - t \leq 0)$.
 - When a model term is involved in a multiplication by a candidate term, it gets replaced by the interaction, and hence is not available for further interactions.
 - With these changes, the MARS forward procedure is the same as the CART tree-growing algorithm.

An example

```
mod1=earth(V2~V1,data=x,nk=1)
plotmo(mod1,xlab="x",ylab="y")
points(x,pch=20)
```

V2 earth(V2~V1, data=x, nk=1)



An example

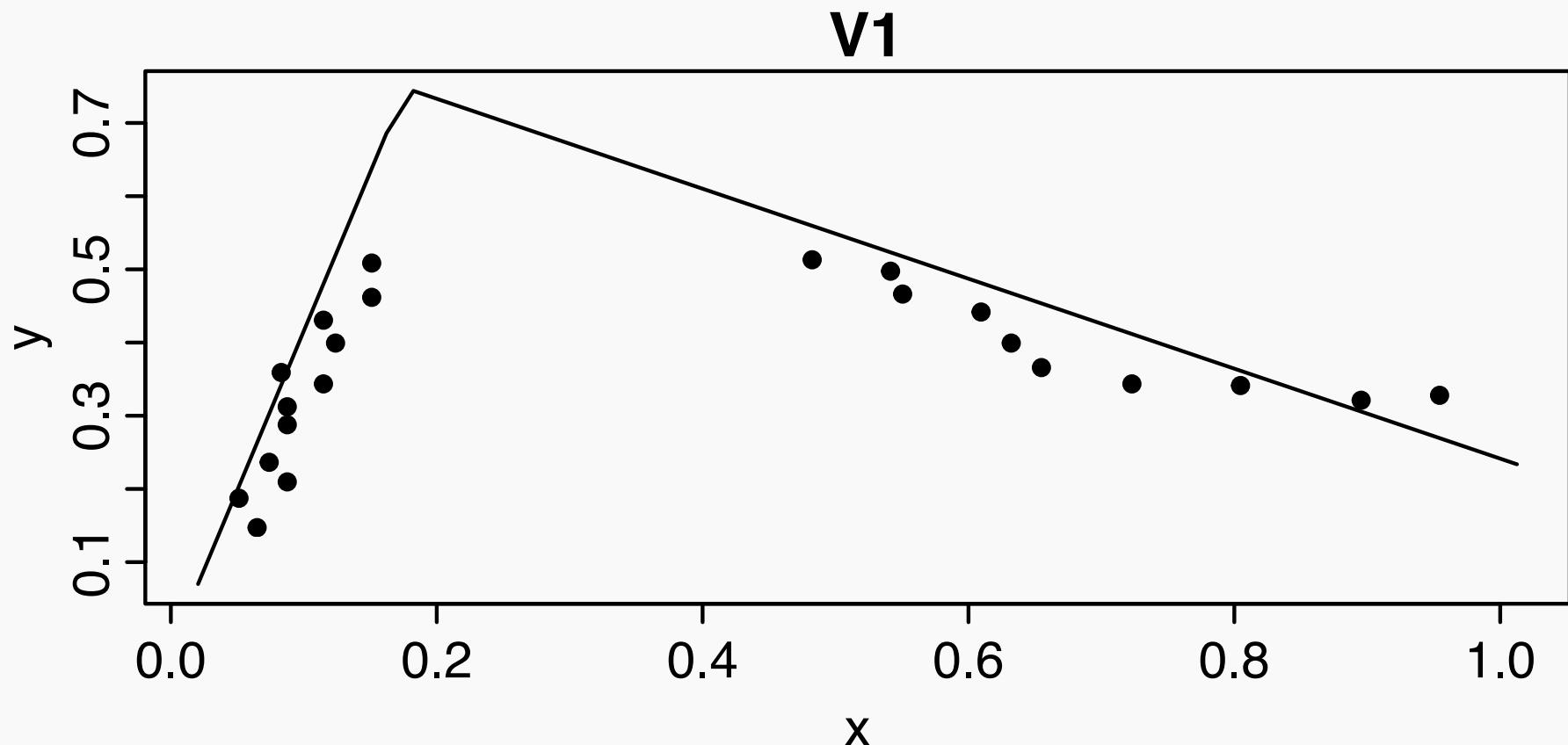
```
summary(mod1)

## Call: earth(formula=V2~V1, data=x, nk=1)
##
##           coefficients
## (Intercept)  0.4279076
##
## Selected 1 of 1 terms, and 0 of 1 predictors
## Termination condition: Reached nk 1
## Importance: V1-unused
## Number of terms at each degree of interaction: 1 (intercept only model)
## GCV 0.04290529    RSS 1.202778    GRSq 0    RSq 0
```

An example

```
mod2=earth(V2~V1,data=x,nk=4)  
plotmo(mod2,xlab="x",ylab="y")  
points(x,pch=20)
```

V2 earth(V2~V1, data=x, nk=4)



An example

```
summary(mod2)

## Call: earth(formula=V2~V1, data=x, nk=4)
##
##           coefficients
## (Intercept)      0.7476095
## h(0.176378-V1) -4.3458394 } 2 basis functions.
## h(V1-0.176378) -0.6146156
##
## Selected 3 of 3 terms, and 1 of 1 predictors
## Termination condition: Reached nk 4
## Importance: V1
## Number of terms at each degree of interaction: 1 2 (additive model)
## GCV 0.00632364    RSS 0.1317425    GRSq 0.852614    RSq 0.8904682
```

Ensemble methods

(Combining predictions and imbalanced learning)

N. Torelli, G. Di Credico, V. Gioia

2023

University of Trieste

Ensemble methods

Learning with imbalanced data

Ensemble methods

Combining multiple predictions: Model averaging

- Classification trees can be simple, but often produce noisy (bushy) and weak classifiers
 - Bagging (*Breiman, 1996*): Fit many large trees to bootstrap-resampled versions of the training data, and classify by majority vote
 - Boosting (*Freund & Shapire, 1996*): Fit many large or small trees to reweighted versions of the training data. Classify by weighted majority vote
 - Random Forests (*Breiman 1999*): Fancier version of bagging.
- Note however that the idea of combining multiple predictions or classifications can be used for any technique (i.e., logistic classification, NN, etc.) and it is not limited to trees
- This idea is closely related with model averaging: a strategy for model selection and evaluation of uncertainty in Bayesian analyses

Can be used
w/ many
classifiers

Combining predictions (classifications)

- The idea is to combine the output of different learners for each data point (y, \mathbf{x}) . This help when learners have complementary strengths.
- Suppose training data are available in the form of the p covariates $\mathbf{x} = (x_1, x_2, \dots, x_p)$ and the response is (target) is y . Let $\hat{y}_1 = f_1(\mathbf{x}), \dots, \hat{y}_M = f_1(\mathbf{x})$ be M different predictions (estimates, "experts evaluations") for the same data point.
- A simple combined vote takes their average

$$f_{comb}(\mathbf{x}) = \frac{1}{M} \sum_{m=1}^M f_m(\mathbf{x})$$

In the classification setting for each class k we have a prediction $f_m^k(\mathbf{x}, t)$ equal to 0 or 1. Then

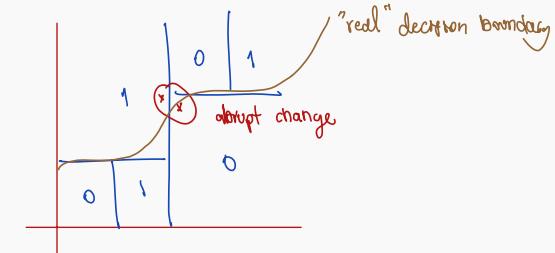
$$f_{comb}(\mathbf{x}) = \frac{1}{M} \sum_{m=1}^M f_m^k(\mathbf{x})$$

for each class k and $f_{comb}^k(\mathbf{x}, t)$ is the proportion of votes for class k . We predict the class with the most number of votes (**majority vote**). 

Bagging

- Bagging or bootstrap aggregation averages a given procedure over many samples, to reduce its variance
- A natural way to reduce the variance and increase the prediction accuracy of a statistical learning method is to take many training sets from the population, build a separate prediction model using each training set, and average the resulting predictions.
- Instead, we can
 - bootstrap, by taking repeated samples from the same training data set
 - use the b -th bootstrapped training set to get the prediction $\hat{f}^b(x)$ and
 - average all the predictions, to obtain

$$\hat{f}_{bag}(x) = \frac{1}{B} \sum_{b=1}^B \hat{f}^b(x)$$



- This is called bagging. In classification problems it uses majority votes.
- Bagging can dramatically reduce the variance of unstable procedure (like trees), leading to improved prediction.
- Bagging averages many trees, produces smoother decision boundaries, reduces the variance, but can slightly increase bias

Out-of-bag

- Using random samples of observations allows the use of the out-of-bag tool, for easy estimation of prediction errors.
- In each bootstrap sample, some of the data of the original training set are excluded.
- On average, each bagged sample makes use of around two-thirds of the observations $\frac{1}{3}$ don't appear in D_{train} \rightarrow can be used on D_{test}
- The remaining one-third of the observations not used to fit a given bagged tree are referred to as the out-of-bag (OOB) observations
- For each classifier $\hat{f}^b(x)$ the data of training set that are not in sample can be used as a test set. This will give $B/3$ predictions for the i -th observation.
- Estimate the misclassification error on these data outside the sample used for the fit (out-of-bag), so avoiding cross-validation for large data sets

Random Forest



- A quite popular refinement of bagging. Particularly when bagging trees for which was originally developed. We will describe this version.
- At each tree split, a random sample of m features is drawn, and only those m features are considered for splitting.
- Typically $m = \sqrt{p}$ or $\log_2 p$, where p is the number of features (covariates)
- For each tree grown on a bootstrap sample, the error rate for observations left out of the bootstrap sample is monitored (out-of-bag)
- Random forests tries to improve on bagging by “de-correlating” the trees, and reduce the variance.
- Each tree has the same expectation, so increasing the number of trees does not alter the bias of bagging or random forests.

Variable importance

Random forests can be used to rank the importance of variables in a regression or classification problem.

- For each tree grown in a random forest, calculate number of votes for the correct class in out-of-bag data.
- Now perform random permutation (shuffling) of a predictor's values (let's say variable-k) in the OOB data and then check the number of votes for correct class.
- Subtract the number of votes for the correct class in the variable-k-permuted data from the number of votes for the correct class in the original OOB data.
- The average of this number over all trees in the forest is the raw importance score for variable k. The score is normalized by taking the standard deviation.
- Variables having large values for this score are ranked as more important. It is because if building a current model without original values of a variable gives worse prediction, it means the variable is important.

Boosting

-  All data is used all the time but weights for data change \rightarrow changing probability of data to appear
- Designed, initially, exclusively for classification problems
 - Idea: Like bagging, but take unequal probability bootstrap samples. Put more weight on observations that are misclassified, to make the classifier work harder on those points. Invention of Freund e Schapire (1997)
 - Details
 - Start with equal observation weights $p_i = 1/n$
 - At iteration t , draw a bootstrap sample with the current probabilities p_1, p_2, \dots, p_n , compute the classifier and e_t , the error rate of the classifier on the original sample. Let $\beta_t = e_t / (1 - e_t) = \frac{1}{e_t}$
 - For those points that are classified correctly, decrease their probabilities $p_i = p_i \beta_t$ and normalize them
 - Do this for many (say 1000) iterations.

$$\lim_{e_t \rightarrow 0} \beta_t(e_t) \rightarrow \infty$$
$$e_t \nearrow 1 \rightarrow 0$$
$$e_t \searrow \frac{1}{2} \rightarrow \infty$$

Boosting

- At the end, take a weighted vote of the classifications, with weights $\alpha_t = \log(1/\beta_t)$ (more weight on classifiers with lower error).
- Boosting can improve bagging in many instances

Weighting decorrelates the trees, and focuses on regions missed by past trees.

In the classification setting for each class k we have a prediction $f_m^k(x, t)$ equal to 0 or 1. Then

$$f_{comb}(x) = \frac{1}{M} \sum_{m=1}^M f_m^k(\mathbf{x})$$

for each class k and $f_{comb}^k(x, t)$ is the proportion of votes for class k . We predict the class with the most number of votes (**majority vote**).

Learning with imbalanced data

Classification with imbalanced datasets

- The problem of data imbalance emerges in supervised classification problems and here we will only mention the (most relevant) case of binary classification
- It is an issue that occurs when one of the two classes which represents the target variable (usually also the class of main interest) is rare and then it is much less represented than the other class in the dataset
- It is a situation encountered in many real word applications:
 - in many fraud detection problems the number of observed frauds is (luckily) a rare event
 - when predicting the insolvency of a firm the event of failing in a given year is not very frequent
 - in medicine, many specific diseases in the population have usually a very low frequency
 - there are many examples where customers are very loyal and observing customer churn is rare

Degree of imbalance

Imbalance ratio = majority class / minority class



- The degree of imbalance for the response variable can be measured by the imbalance ratio IR which is defined as the ratio between the cases of the prevalent class divided by the number of data points in the rare class. Saying that IR is 100 mean that for 1 data point in the rare (positive) class there are 100 cases of the prevalent (negative) class.
- Actually any dataset presents a certain degree of imbalance, but the severity of the problem for a two-class classification problem emerges usually when the IR is at least larger than 10
 - IR larger than 100 defines a strong imbalance
 - IR larger than 1000 is an extremely skewed dataset (very strong imbalance)
- In the two last cases dealing with the imbalance before bulding any classification rule or machine learning algorithm cannot be avoided
- The size of the dataset also matters when defining the severity of the problem

Why standard ML algorithms can fail with imbalanced datasets?

- Standard classifiers such as logistic regression, Support Vector Machine (SVM), classification tree or other ML algorithms are suitable for balanced training sets.
- When facing imbalanced scenarios, these models often provide suboptimal classification results, i.e., a good coverage of the majority examples, whereas the minority examples are distorted
 - The learning process often guided by global performance metrics such as prediction accuracy induces a bias towards the majority class
 - Rare minority examples may possibly be treated as noise by the learning model.
- Many machine learning and statistical approaches have been developed in the past two decades to cope with imbalanced data classification, most of which have been based on three strategies: (i) sample techniques, (ii) cost sensitive learning and (iii) possible modification of the learning algorithm

Standard classifiers fail for imbalanced data
Ensemble methods deal approximately w/ this problem

Summary

Performance metrics in a two-class problems

- The quality of a learning algorithm is evaluated by looking at its performance on test data.
- The simplest performance measures are based on comparison between the predictions of the classifier and the true values (confusion matrix).

		predicted		total
		positive	negative	
Actual	positive	TP	FN	POS
	negative	FP	TN	NEG
total		PredPOS	PredNEG	N

- The simplest measure, accuracy (ACC) is defined as

$$ACC = \frac{TP + TN}{N} \underset{\text{if } TP \rightarrow 0}{\approx} \frac{TN}{N}$$

- note that ACC **cannot** be used as a performance measure in imbalanced dataset: a trivial classifier which always classify new cases into the majority class will have an accuracy equal to the proportion of the majority class in the sample
- If we have 0.1% of the sample of cases with a rare type of cancer, a (dull) strategy which always predicts that you are free from the disease will anyway obtain ACC equal to 99.9%

Other metrics and their use for imbalanced datasets

- True positive rate (**recall** or sensitivity) = $\frac{TP}{POS}$
- True negative rate (specificity) = $\frac{TN}{NEG}$
- Positive predictive value (**precision**) = $\frac{TP}{predPOS}$
- $F1 = (1 + \beta^2) \frac{\text{precision} \cdot \text{recall}}{(\beta^2 \text{precision}) + \text{recall}}$

β is often taken to be 1 (that is precision and recall have the same weight)

- Classification is obtained by setting a threshold for those methods (logistic regression, trees) which estimate a score (probability) and this threshold can be somewhat arbitrary and should be changed appropriately when using some of the remedies for imbalance (such as oversampling the rare class).

AUC (area under the ROC curve)

- AUC which does not depend on a given threshold is a most appropriate measure for comparing performances with imbalanced dataset.
- ROC (Receiver Operating Characteristics) measures the accuracy of a classification prediction when prediction comes in form of a numerical scoring (a probability).
- ROC Curve is obtained by plotting sensitivity versus specificity for different thresholds. AUC measures the area under this curve and the larger the better is the performance (for a perfect classifier AUC is 1).
- Note that ROC curve, and as a consequence AUC, can be very unstable when the test dataset is small and imbalanced

Approaches to imbalanced data classification

- Preprocessing techniques (resampling and synthetic data generation)

Preprocessing is often performed before building learning model in order to obtain balanced input data in building the classifier

- Cost-sensitive learning
- Specific modifications of classification algorithms for imbalanced learning

(re)Sampling techniques: undersampling

- This strategy belongs to the first category of remedies: preprocessing techniques
- The aim is to obtain a balanced sample (let's say one of the classes has no less than 30% of the cases, or even better the two classes are made equivalent) for both the training and the test set.
- Undersampling:

this method reduces majority class. It consists in randomly selecting a subset of observations (without replacement) from majority class to make the data set balanced. This method can be very successfully used when the data set is really huge. It can be improved on by adding strategy for selecting the data most relevant for classification.

(re)Sampling techniques: oversampling

TEST

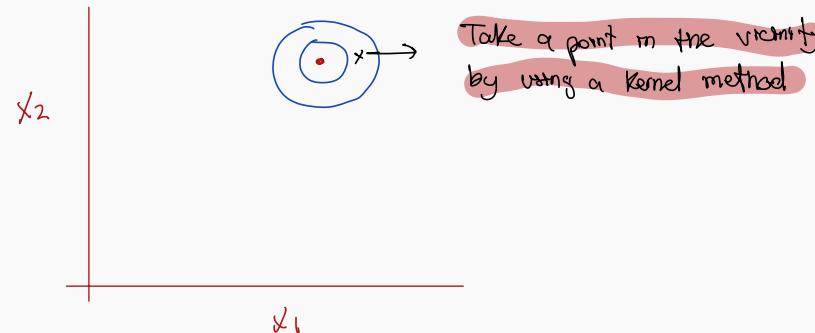
- It eliminates the harms of skewed distribution of the target by multiplying new minority class samples. Data from the rare class are re-sampled (with replacement) in order to balance the sample (note that it can induce overfitting)
- Oversampling can also be achieved by generating new instances from existing one.
"invent" data of the minority class similar to the actual data and then resampling
- There are many methods specifically designed for generating new synthetic data (data cloning) for the minority class. We will briefly describe two of them:
 - SMOTE (Chawla et al, 2002)
 - ROSE (Menardi & Torelli, 2014)
- Hybrid methods: are a combination of the over-sampling method and the under-sampling method.

ROSE: Random OverSampling Examples

- ROSE is aimed at oversampling the rare class by creating new (synthetic) data points that are as similar as possible to the existing (real) ones
- ROSE also suggests to under-sample the majority class (possibly by cloning data with the same strategy) **LOOK GRAPH BELOW**
- ROSE chooses a random point from the rare set and then a new point is generated in the neighborhood according to a kernel function (imagine to put a multivariate Gaussian distribution centered on the point and select a new point from that Gaussian)
- The cloning method is formally based on a kernel density estimation of the distribution of the predictor variables within the rare (or sometimes also the prevalent) class. This turns out to be equivalent to obtaining a **smoothed bootstrap** resampling scheme

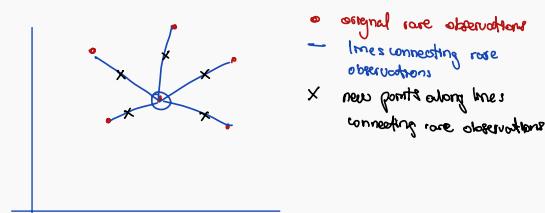
ROSE: Random OverSampling Examples

- A package exists in R named *ROSE*. It can be also called directly within the *caret* package. Recently a version of Rose has been made also available within *Scikit-learn* in Python
- *ROSE* can also be used to undersample the prevalent class.
- A combination of undersampling and oversampling (possibly cloning also data from the prevalent class) is sometimes useful
- Usually the test set is left unchanged. But the authors of ROSE suggest that for a more stable estimate of some of the metrics (such as AUC) also for the test set some new data can be cloned



SMOTE

- SMOTE is a popular method for data cloning.
- It generates new data in the rare class in order to obtain as many cases as in the prevalent class.
- The generation of new cases happens by first selecting randomly a case from the rare class and considering the K points which are closer to this point.
- New points are generated selecting a random position along the line connecting the selected point to one of the K neighbouring points
- There are many variants of SMOTE
- In R SMOTE is a function available within the R package *DMwR*. It can be also called directly within the *caret* package. SMOTE and its variants are available in Python Scikit-learn



Some practical issues

In real applications one has to choose:

1. when imbalance is a problem. Sometimes one can consider balancing the training set even when IR is not larger than 4-5. With small data sets it could be beneficial anyway. While in case of large (or huge datasets) simple/naive strategies such as undersampling the prevalent class may give good results
2. which method is most appropriate. This is a matter of tastes. Usually, considering alternative strategies and comparing the performances is suggested. Any method has its merits and which is the more appropriate depends on specific characteristics of the data
3. the nature of the predictors matters.
 - SMOTE and ROSE work better when most of the predictors are quantitative. With categorical predictors simple undersampling or oversampling can be more appropriate.
 - when applying ROSE some extra care is needed in the case of mixed variables (for instances zero inflated variables) or for limited variables.

Other remedies

1. Cost sensitive learning

In many cases the two errors have not the same importance. A different cost can be associated to the two errors FP and FN and a higher value is assigned to the most relevant error for the specific problem. The loss function to be minimized for the algorithm will change accordingly and it could help concentrating on predicting accurately the minority class

2. Modification of the standard algorithms

One of the most notable example is modification of boosting/bagging procedures to account for data imbalance. Note that actually the use of ensemble methods itself can alleviate the problem of data imbalance

Bayesian Inference

(An essential introduction)

N. Torelli, G. Di Credico, V. Gioia
2023

University of Trieste

Introduction

Introducing Bayesian inference

Classical and Bayesian Inference

Bayesian models

Bayesian interval estimation and testing

Selecting the prior

Bayes computation

Introduction

Bayes Theorem (basic)

- Bayes' theorem is a rule to compute **conditional probabilities**.
- In other words, it links probability measures on different spaces of events: given two events E and H , the **probability of H conditional on E** is the probability given to H *knowing that E is true* (i.e. E is the new sample space (Ω)).
- More precisely,
 - I have given a probability measure on E and H ,
 - I am told that E has occurred,
 - how do I change (if I change) my opinion on H :

$$P(H) \rightarrow P(H|E) = ?$$

Theorem of Bayes (for events) Let E and H be two events, assume $P(E) \neq 0$, then

$$P(H|E) = \frac{P(H \cap E)}{P(E)} = \frac{P(H)P(E|H)}{P(E)}$$

Bayes Theorem (an example: a crime case)

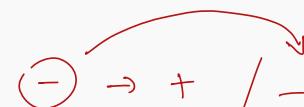
In an island a murder is committed

- there is no clue on who is the murderer;
- **but for** the DNA found on the victim (who had a fight with the murderer);
- within the population of the island 1000 persons may have committed the crime, it is a certain thing that the murderer is among them, with equal probability.

The police compares the DNA of all the 1000 suspects with that of the murderer.

The DNA test used by the island police force is not perfect, there is

- probability of a false positive 1%;
- probability of a false negative 2%.



Bayes Theorem (an example: a crime case)

Formally, if

- T is the ‘event “the person is positive at the test”’,
- C is the event “the person is guilty” ’,

$P(T| \cdot)$

↓

test

		T	\bar{T}	
		0.98	0.02	→ 1
guilty	C	0.98	0.02	→ 1
	\bar{C}	0.01	0.99	→ 2

↓

the two assumptions are then

- $\rightarrow + / -$
- probability of a false positive 1% : $P(T|\bar{C}) = 0.01$,
 - probability of a false negative 2% : $P(\bar{T}|C) = 0.02$.
- $\rightarrow - / +$

+ : guilty

The experimental observation is: the police starts testing the 1000 suspects and the 130-th is positive at testing.

Crime case: the investigation (likelihood inference)

The sheriff says that the experimental evidence, represented by the ratio between the likelihoods

$$\frac{P(T|C)}{P(T|\bar{C})} = \frac{0.98}{0.01} = 98 \quad \begin{matrix} \text{prob. of positive at test, gives the person is guilty} \\ \checkmark \end{matrix} \quad = \frac{\text{positive being guilty}}{\text{positive given not guilty}}$$

is overwhelmingly in favour of that person being guilty, thus constituting decisive evidence, so he asks the judge to arrest and condemn the guy.

Being more formal, who are model and likelihood?

- **model:** set of probability distributions which may have generated the sample T , there are two alternatives, represented by $\{C, \bar{C}\}$:

$$P(T|C) = 0.98 \quad P(T|\bar{C}) = 0.01$$

- **likelihood** the parameter space is $\{C, \bar{C}\}$, the likelihood takes two values

$$L_C = P(T|C) = 0.98 \quad L_{\bar{C}} = P(T|\bar{C}) = 0.01$$

- the maximum likelihood estimate is then C .

Crime case: the (Bayesian) defence

A Bayesian lawyer argues against the sheriff and notes that *despite the fact that the experimental evidence is much more compatible with the man being guilty, the verdict should be based on the probability of the man being guilty: the sheriff ignores prior probabilities*

- **a priori:** before the test, the suspect was only one among 1000 suspects, the probability of him being guilty is $P(C) = 0.001$.
- **data and likelihood:** having observed T and knowing that $P(T|C) = 0.98$

we obtain that

$$P(C|T) = \frac{P(C)P(T|C)}{P(\bar{T})} = \frac{0.001 \times 0.98}{0.001 \times 0.98 + 0.999 \times 0.02} = 0.0893$$

$$P(T) = \underbrace{P(T|C)P(C)}_{1/1000} + \underbrace{P(T|\bar{C})P(\bar{C})}_{1 - 1/1000}$$

Crime case: the (Bayesian) defence

To understand the result, consider what would happen, on average, if all 1000 suspects were tested:

- 999 are innocent, among them
 - $999P(T|\bar{C}) = 9.99$ test positive; *not being culprit.*
 - the other 989.01 test negative;
- 1 is guilty,
 1. he tests positive with probability 0.98
 2. he tests negative with probability 0.02

then, on average, the 1000 suspects partition as follows

	Pos	Neg
1 guilty	0.98	0.02
999 innocent	9.99	989.01
Tot	10.97	989.03

and the probability of being guilty given you teste positive is simply
 $\frac{0.98}{10.97} = 0.893$

Bayes' theorem (more than two hypotheses)

We now consider a more general version of Bayes' theorem where more than two events are involved,

Bayes (with n hypotheses)

$\{H_i | i = 1, \dots, n\}$ is a partition of the sample space Ω such that
 $\cup_{i=1}^n H_i = \Omega$; $H_i \cap H_j = \emptyset$ if $i \neq j$ then

$$P(H_j|E) = \frac{P(H_j)P(E|H_j)}{\sum_{j=1}^n P(H_j)P(E|H_j)} = P(E) \quad \begin{matrix} \text{posterior} & & \\ & \text{prior} & \text{likelihood} \\ & \text{information} & \end{matrix}$$

A second example: which box of seeds?

The problem

A factory sells boxes of seeds. It produces 4 types of boxes and each box mixes 2 type of seeds: High Quality seeds and Normal seeds.

The boxes are labelled as Standard (S), Extra (E) and Platinum (P).

Platinum has 90% of High Quality seeds, Gold 80%, Extra 70%, Premium 50%.

Assume you have an unlabelled box and you want to decide which kind of box it is by selecting (with replacement) a sample of 30 seeds. Now assume that in your sample the number of High quality seeds is 23.

Which kind of box is it?

Plat. 90%

$$P(23 | \text{Plat.}) = \binom{30}{23} 0.9^{23} 0.1^7$$

$$\underset{b \in \mathcal{B}}{\operatorname{argmax}} P(HQ_S = 23 | b) = \underset{P \in \mathcal{P}}{\operatorname{argmax}} \binom{30}{23} P^{23} (1-P)^7$$

Maximum likelihood estimation

We can rephrase the problem as follows:

- let p be the proportion of High quality seeds in a box.
- we observe a sample x_1, x_2, \dots, x_{30} from a rv $X_i \sim Be(p)$ and let $x = \sum_i^{30} x_i$
- we want to use these data to estimate the parameter p where $p = \{0.5, 0.7, 0.8, 0.9\}$

We can write the likelihood function, *i.e.*, the probability of observing x (23 in our case) high quality seeds when the box is P, G, E or S and p can take on one of the 4 values $p_1 = 0.5, p_2 = 0.7, p_3 = 0.8, p_4 = 0.9$

Since we want to maximize the likelihood \Rightarrow the best option (?) We have different values of p . \rightarrow Bayesian Statistics

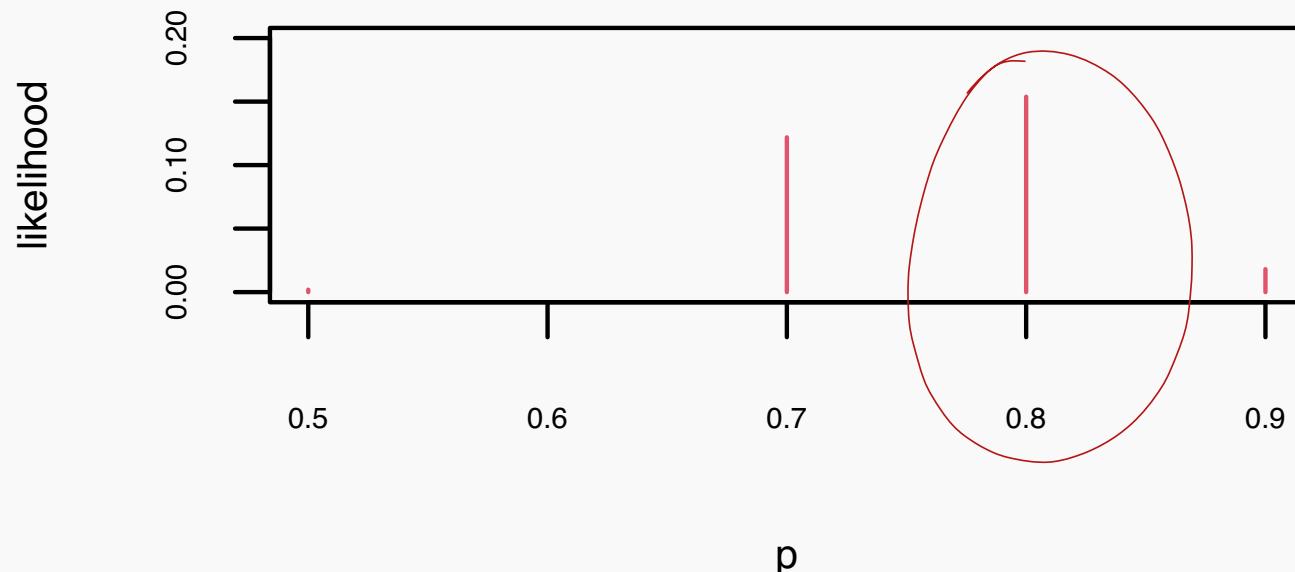
$$L(p_i) = \binom{30}{x} p_i^x (1 - p_i)^{30-x}$$

and calculate it. Note that p_i is our parameter and the parameter space contains only four elements.

Maximum likelihood estimation

```
p <- c(.5, .7, .8, .9); n <- 30; x=23;  
L <- choose(30,x)*p^x*(1-p)^(30-x)  
L  
## [1] 0.001895986 0.121853726 0.153820699 0.018043169  
plot(p,L,type="h",main="likelihood function", cex.lab=0.7,  
cex.axis=0.5, ylab="likelihood", ylim=c(0,0.2), col=2)
```

likelihood function



A different perspective: toward bayesian inference

Assume that we know that in the factory the proportions of Platinum, Gold, Extra and Standard boxes are as follows

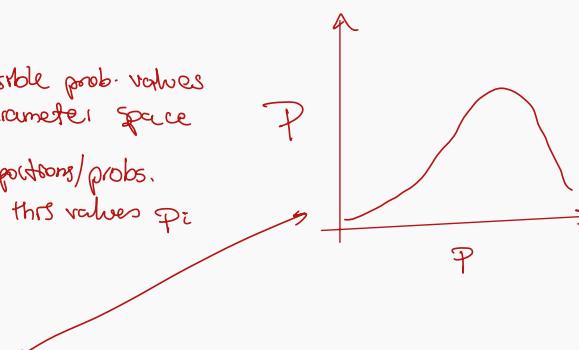
prop(Standard)	prop(Extra)	prop(Gold)	prop(Premium)
0.4	0.3	0.2	0.1

One can then assume, even before seeing the sample of 30 seeds, that the probability of getting one specific type of boxes, i.e., of getting a specific value for p_i is:

p_i	0.5	0.7	0.8	0.9
$P(p_i)$	0.4	0.3	0.2	0.1

parameter treated as a rv \rightarrow distribution

possible prob. values
parameter space
proportions/probs.
of this values p_i



Bayesian solution

In Bayesian inference we want to express our uncertainty about the parameter p by giving a probability distribution on it. The quantity p is now random.

Note that we have:

- a probability distribution on the possible values of p before observing the sample $P(p_i)$. This is called the **prior distribution**
- the **likelihood function** $L(p_i)$, but since now p is a rv, then we can rewrite it as the conditional probability $P(x|p_i)$, where x is evidence from the sample.
- Bayes theorem can then be applied to get the so called **posterior distribution**

$$P(p_i|x) = \frac{P(p_i)L(p_i)}{\sum_i^4 P(p_i)L(p_i)} = \frac{P(p_i)P(x|p_i)}{\sum_i^4 P(p_i)P(x|p_i)} \propto P(p_i)P(x|p_i)$$

not a probability distribution

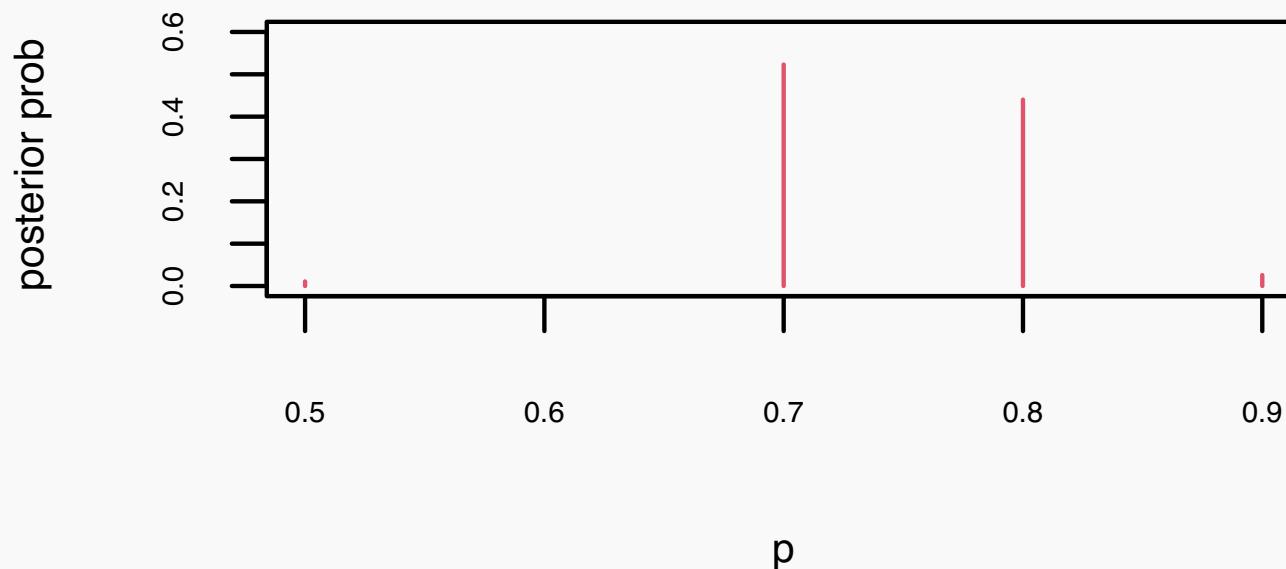
Bayesian solution

```
prior <- c(.4, .3, .2, .1)
like <- choose(30,x)*p^x*(1-p)^(30-x)
posterior <- prior*like/sum(prior*like); posterior

## [1] 0.01085235 0.52310482 0.44022371 0.02581912

plot(p,posterior,type="h", main="posterior", cex.lab=0.7,
cex.axis=0.5, ylab="posterior prob ", ylim=c(0,0.6), col=2)
```

posterior



Likelihood vs Bayesian

In the example:

- The likelihood estimate was $p = 0.8$, Gold, since this is value of the parameter with the highest value of the likelihood.
- In Bayesian inference we have a probability distribution over the parameter space. We can say that the value is $p = 0.7$ with probability ≈ 0.52 , Extra. This is a probability statement.
- Bayesian approach allows us to update our prior information with experimental data

information post experiment \propto information from experiment \times
information prior to experiment

posterior \propto prior \times likelihood

Introducing Bayesian inference

Bayes theorem: continuos variables

Bayes Theorem

If

- (i) $\pi(\theta)$ density function
- (ii) $f(y|\theta)$ density function of y given θ

then

$$\pi(\theta|y) = \frac{\pi(\theta)f(y|\theta)}{\int_{\Theta} f(y|\theta)\pi(\theta)d\theta} \propto \pi(\theta)f(y|\theta)$$

Note that $\int_{\Theta} \pi(\theta|y)d\theta = 1$ and that the quantity $\int_{\Theta} f(y|\theta)\pi(\theta)d\theta$ is called the normalization constant.

Bayesian paradigm: model and likelihood

Consider a **model**, a family of probability distributions indexed by a parameter θ among which we assume there is the distribution of y :

$$f(y|\theta), \quad \theta \in \Theta.$$

This is no different than the classical paradigm, but for the fact that the distributions are defined conditional on the value of the parameter (which is not a r.v. in the classical setting)

One defines then the likelihood

$$L(\theta; y) \propto f(y|\theta),$$

as in the classic paradigm.

Bayesian paradigm: prior distribution

A prior distribution is set on the parameter θ

$$\pi(\theta)$$

which is independent of observations (it is called prior since it comes before observation).

This is the new thing

Prior information and likelihood are combined in Bayes' theorem to give the posterior distribution

$$\pi(\theta|y) = \frac{\pi(\theta)f(y|\theta)}{\int_{\Theta} f(y|\theta)\pi(\theta)d\theta} \propto \pi(\theta)f(y|\theta) \propto \pi(\theta)L(\theta; y)$$

which sums up all the information we have on the parameter θ .

Inference on the quality of seeds

We are again interested in estimating the proportion of high quality seeds in the boxes. But now we assume that the proportion p can be any value in the interval $[0, 1]$. We still have a sample of n seeds drawn from a box (with replacement), and we count the number of high quality seeds x .

We want to infer on the value p . Data are i.i.d realizations from a $Be(p)$.

We can never know the real value of p , unless we can rely on a sample where $n \rightarrow \infty$.

We can design a procedure that selects values of p that are more supported by the data. We can then judge how uncertain is our procedure by looking at its behaviour in possible (not actual) replication of the sample under the same condition. **Classical inference**

We can try to give a probability distribution over possible values of the parameter p . And this probability distribution will summarize all the information we have about it: before and after observing the data

Bayesian inference

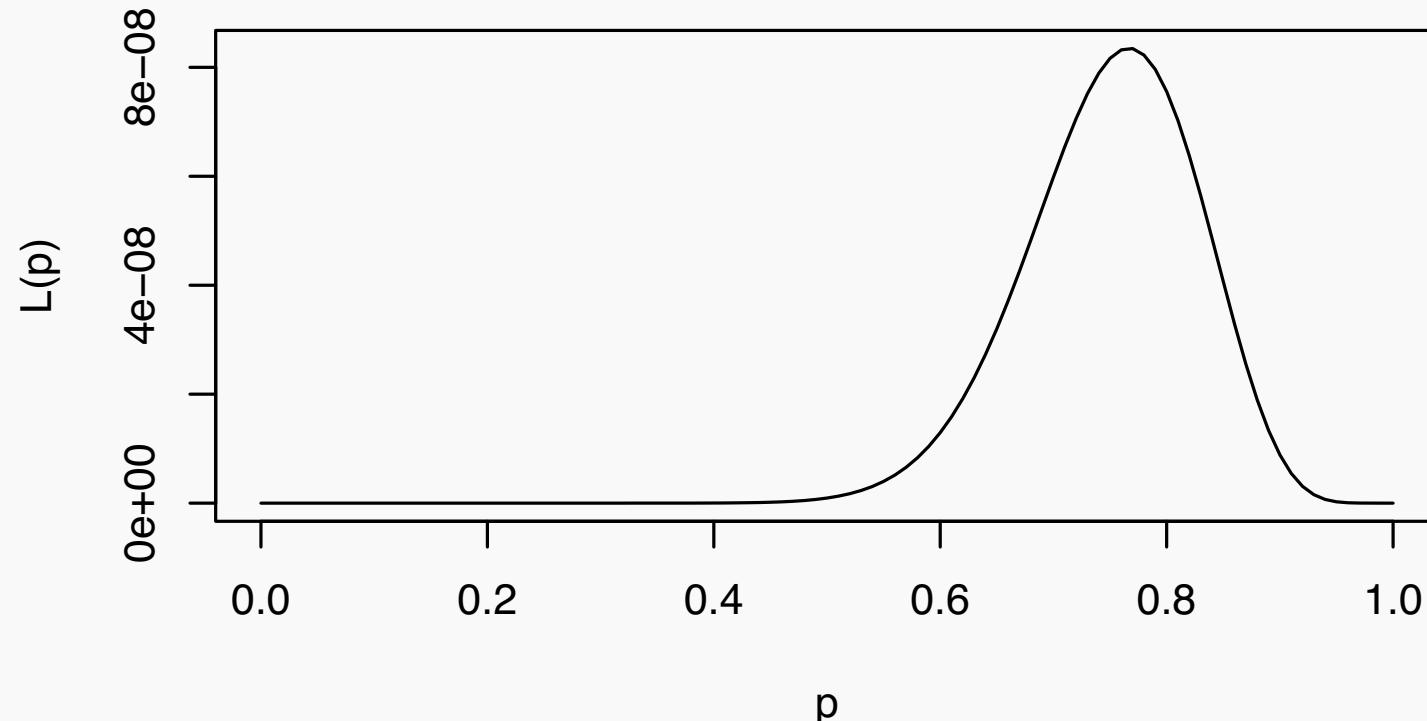
Inference on p

1. Likelihood estimation is straightforward:

- $L(p) \propto p^x(1 - p)^{n-x}$
 - it is easy to show that ML estimate is $\hat{p} = x/n$. The observed proportion of high quality seeds in the sample.
2. Bayesian solution requires specification of the probability distribution $\pi(p)$.
- Since $p \in [0, 1]$ candidates are probability models whose support is the interval $[0, 1]$.
 - Random variables belonging to the Beta family could be appropriate

The likelihood function

```
n <- 30; z=23;  
curve(x^z*(1-x)^(30-z), xlim=c(0,1), xlab="p",ylab="L(p)")
```



The Beta distributions

assumption

$$\pi(\theta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1} (1-\theta)^{\beta-1}$$

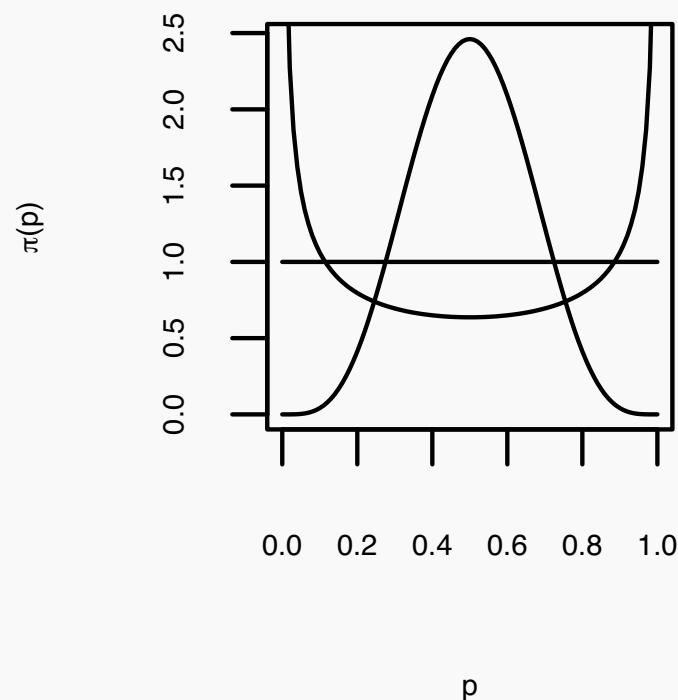
dove $0 < \theta < 1$ e $\alpha, \beta > 0$,

$$E(\theta) = \frac{\alpha}{\alpha + \beta} \quad V(\theta) = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}$$

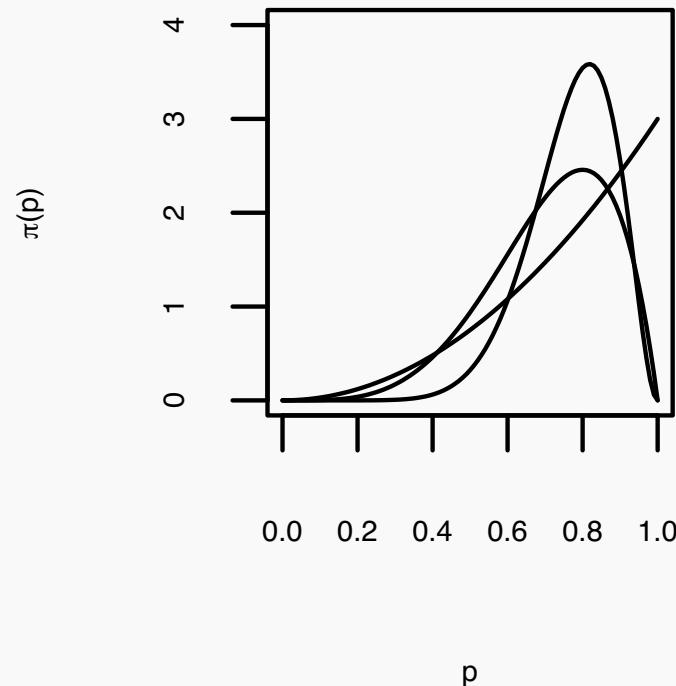
remind that $\Gamma(x) = \int_0^{+\infty} t^{x-1} e^t dt$ and if x is integer $\Gamma(x) = (x - 1)!$

The Beta distributions

$\alpha=\beta$



$\alpha>\beta$



The posterior distribution

Since

$$\begin{aligned}\pi(p|x) &\propto L(p)\pi(p) \propto p^x(1-p)^{n-x} p^{\alpha-1}(1-p)^{\beta-1} \\ &\propto p^{\alpha+x-1}(1-p)^{\beta+n-x-1}\end{aligned}$$

likelihood comes from data

then

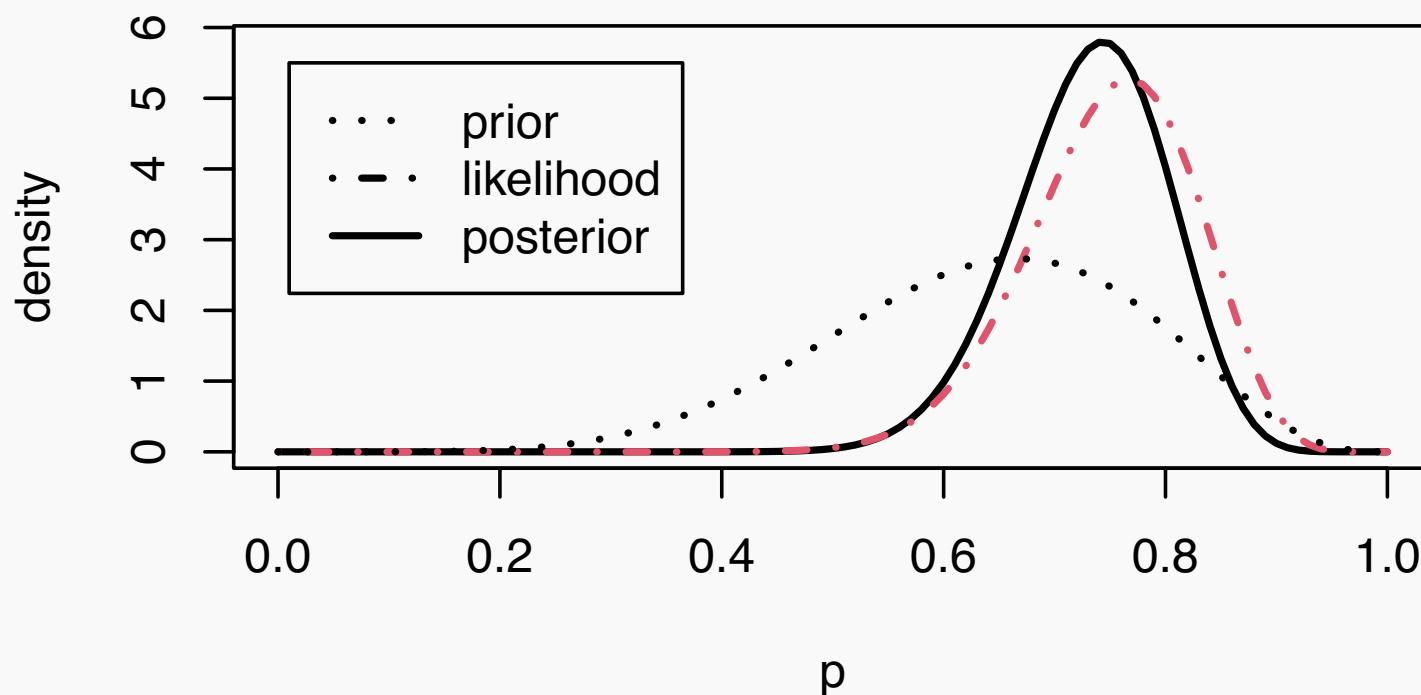
$$\pi(p|x) = \frac{\Gamma(\alpha+\beta+n)}{\Gamma(\alpha+x)\Gamma(\beta+n-x)} p^{\alpha+x-1}(1-p)^{\beta+n-x-1}$$

the posterior for θ is then a Beta with parameters $\alpha + x$ and $\beta + n - x$.

Likelihood, prior and posterior

Assume that in our case we believe, before getting the sample, that values above 0.5 are more likely. I could use as a prior a Beta(7,4) whose mean is $7/11=0.636$. Note that the likelihood (normalized so that it integrate to 1) is also a Beta with parameters (24,8).

Then the posterior is still a Beta(30,11)



Classical and Bayesian Inference

Likelihood and Bayesian inference

A statistical problem is faced when, given observations, we want to assess what random mechanism generated them

- In other words,
 - there are two or more probability distributions which may have generated the observations;
 - analyzing the data we want to infer on the actual distribution which generated the data (or on some property of it).
- How? Let us discriminate between the two approaches.
 - Based on the likelihood we compare $P(\text{Data}|\text{Model})$ for the different models.
 - In Bayesian statistics comparing $P(\text{Model}|\text{Data})$.
$$P(\text{proportion } \in \text{Interval})$$

Exam Question

In the likelihood approach quality of the procedure is evaluated relying upon fictitious repetitions of the experiment.

Classical and Bayesian statistical inference, differences

In CLASSICAL INFERENCE

- the conclusion is not derived within probability calculus rules (these are used in fact, but the conclusion is not a direct consequence)
- the **likelihood** and the probability distribution of the sample are used;
- the parameter is a constant.

In BAYESIAN INFERENCE

- the reasoning and the conclusion is an immediate consequence of probability calculus rules (more specifically of Bayes' theorem);
- the **likelihood** and the **prior distribution** are used;
- the parameter is a random variable.

Bayesian vs classical inference

In the Bayesian approach the parameter is random: this is a fundamental difference between the two approaches, how can this be interpreted?

- In classical statistics, on the contrary, the parameter is a fixed quantity.
- the random character of θ represents our ignorance on it.
- random means, in this context, not known for lack of information.
We measure our uncertainty about the model $\rightarrow P(\theta|X)$
- The randomness and the probability distribution on θ are subjective.
- The probability in Bayesian approach is a subjective probability, i.e., the probability of a given event is defined as the “degree of belief of the subject on the event”.

The role of subjective probability

- Consider events such as *tail is observed when a coin is thrown*,
 - everyone (presumably) would agree on the value of the probability;
 - the frequentist definition is intuitively applied;
 - → this is an ‘objective’ probability.
- For events such as *Juventus will be Italian champion next year* or *Right wing parties will win next elections*,
 - it is still possible to state a probability;
 - everyone would assign a different probability;
 - the probability given by someone will change in time depending on available information.

One then accepts that the probability is not an objective property of a phenomenon but rather the opinion of a person and one defines

Subjective probability: definition (de Finetti)

The probability of an event is, for an individual, his degree of belief on the event.

Bayesian statistics and subjective probability

If the probability is a subjective degree of belief, it depends on the information which is subjectively available, and that by **random we mean not known for lack of information.**

The subjective definition of probability is most compatible with the Bayesian paradigm, in which:

- the parameter to be estimated is a well specified quantity but is not known for lack of information
- a probability distribution is (subjectively) specified for the parameter to be estimated, this is called **a priori**
- after seeing experimental results the probability distribution on the parameter is updated using Bayes' theorem to combine experimental results (likelihood) and the prior to obtain the posterior distribution.

Note that, starting in 1763 (the year Bayes' theorem was published), Bayesian statistics comes first, before the so-called classical statistics, initially developed by Galton and Pearson at the end of XIX century and then by Fisher in the twenties.

Bayesian models

Bayesian models

- A **prior distribution** is defined on the parameter θ

$$\pi(\theta)$$

- we assume an i.i.d. sample $y = (y_1; y_2, \dots, y_n)$ is obtained from a distribution belonging to a family indexed by θ , and

$$f(y|\theta), \quad \theta \in \Theta$$

is then proportional to the likelihood function.

- Bayes theorem allows us to combine prior information and likelihood to give the **posterior distribution**

$$\pi(\theta|y) = \frac{\pi(\theta)f(y|\theta)}{\int_{\Theta} f(y|\theta)\pi(\theta)d\theta} \propto \pi(\theta)f(y|\theta) \propto \pi(\theta)L(\theta; y)$$


The posterior distribution then sums up all the information (and the uncertainty) about the parameter θ . Inference on the parameter amount at studying and appropriately summarizing the posterior.

Again on the proportion of seeds

Consider again the example of the estimation the proportion p of High quality seeds in a box. Recall that for inference on p we assumed that:

- $\pi(p)$ is $Beta(\alpha, \beta)$, the prior
- $f(x_i|p)$ are $Be(p)$ and $f(x|p) \propto Bi(n, p)$ is the likelihood

(where $x = \sum_i^n x_i$ and x_i are i.i.d.).

- $\pi(p|x)$, the posterior, is $Beta(x + \alpha, \beta + n - x)$.

In our example, $n = 30, x = 23$, the prior is $Beta(7, 4)$

Then the posterior is a $Beta(30, 11)$

As stated the posterior distribution summarizes what we know about the parameter combining prior knowledge and experimental data.

So inference on p derives from the analysis of this distribution. And we will use it to illustrate the procedures for point and interval estimation

Point estimation of the proportion of seeds

If we want to select a single value as a point estimate of p , let say \hat{p} , we are back to a classical problem: how to select a single number to summarize a distribution $\pi(p)$.

Classical solutions are:

- the expected value $E(P)$ of the posterior distribution of the rv P ,
$$E(p) = \int_0^1 p\pi(p|x)dp$$
- the median Me of the posterior distribution, $Me : \int_0^{Me} \pi(p|x)dp = 0.5$
- the mode Mo of the posterior distribution, i.e., the value of p for which $\pi(p|x)$ is maximum.

One can choose one of these as point estimate and, provided that the posterior is unimodal, they provide an appropriate synthesis.

Obviously the three values are equivalent if the posterior is symmetric and unimodal.

Bayes risk

More formally, Bayes estimators can be defined as the quantity that minimizes the posterior expected value of a loss function $L(\theta, \hat{\theta})$

The quantity $E_{\pi|x}(L(\theta, \hat{\theta}))$ is called **Bayes risk**.

1. When $L(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$,

i.e. a quadratic loss is used, the quantity that minimizes the Bayes risk is the posterior mean.

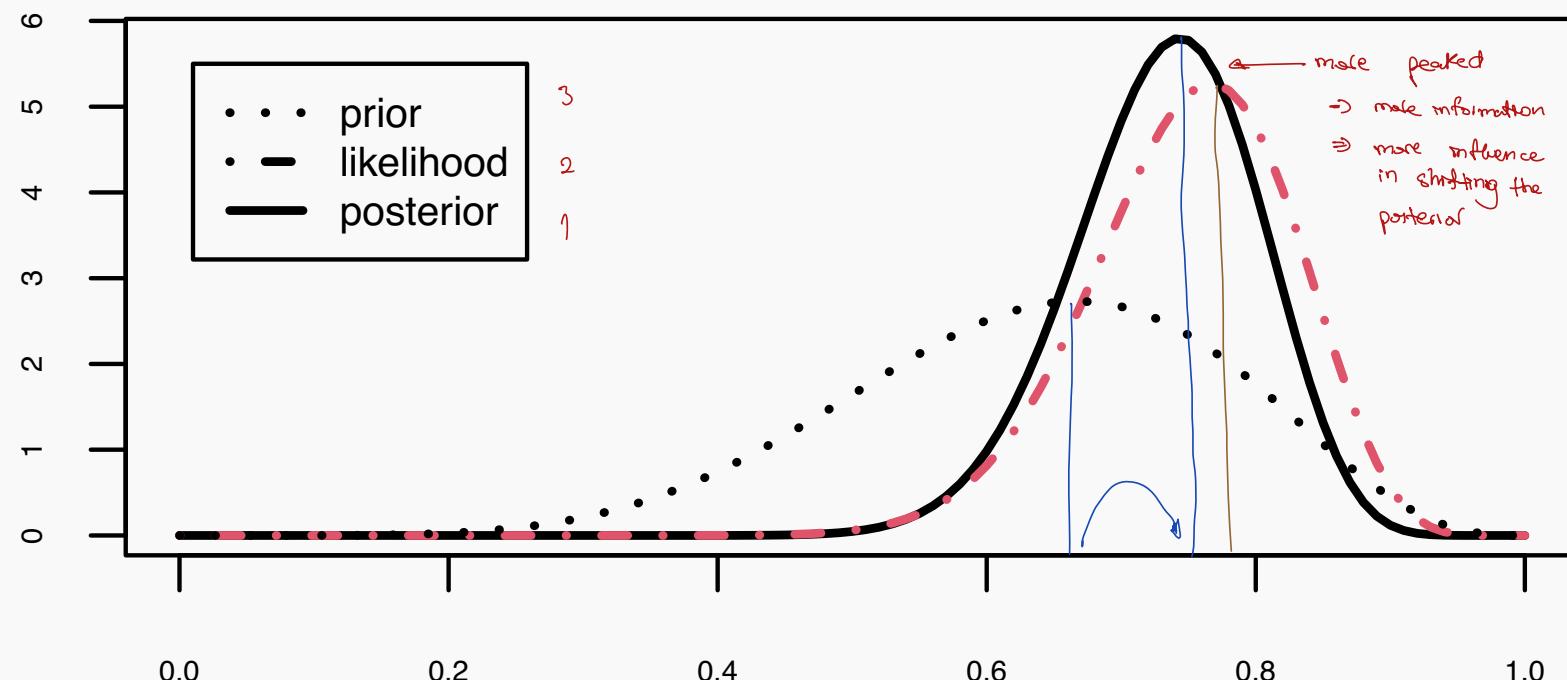
2. Posterior median can be also justified as the quantity that minimizes a “linear” loss function, with $a > 0$, defined as $L(\theta, \hat{\theta}) = a|\theta - \hat{\theta}|$.
3. Posterior mode (MAP: Maximum A-posteriori Probability) can be justified as the minimizer of the trickier loss function of the form

$$L(\theta, \hat{\theta}) = \begin{cases} 0 & \text{if } |\hat{\theta} - \theta| < c, \\ 1 & \text{otherwise,} \end{cases}$$

as c goes to 0

Estimating the quantity of seeds by the posterior mean

```
curve(dbeta(x,7+23,4+7),xlab="p", ylab="density",lty=1,lwd=2,  
cex.axis=.5, cex.lab=.6, ann=F)  
curve(dbeta(x,23+1,7+1),add=TRUE,lty=4,lwd=2, col=2)  
curve(dbeta(x,7,4),add=TRUE,lty=3,lwd=2)  
legend(.01,5.5,c("prior","likelihood","posterior"), lty=c(3,4,1))
```



The value of the parameter θ st. the posterior is maximum
is in between the values of θ that maximize the prior and the
likelihood

$$\theta_{\text{posterior, max}} \in (\theta_{\text{prior, max}}, \theta_{\text{likelihood, max}})$$

Using Posterior mean as an estimator of p

Recall that for if $X \sim \text{Beta}(\alpha, \beta)$ then $E(X) = \frac{\alpha}{\alpha+\beta}$

being $\pi(\theta|y)$ a Beta distribution, the posterior mean is

$$\begin{aligned} &= \frac{\alpha + x}{\alpha + \beta + n} \\ &= \frac{\alpha + \beta}{\alpha + \beta + n} \frac{\alpha}{\alpha + \beta} + \frac{n}{\alpha + \beta + n} \frac{x}{n} \\ &= \frac{\alpha + \beta}{\alpha + \beta + n} E_\pi(p) + \frac{n}{\alpha + \beta + n} \hat{p}_{ML} \end{aligned}$$

$$E(P|x) = \underbrace{\frac{\alpha + \beta}{\alpha + \beta + n}}_{\text{prior expectation}} + \underbrace{\frac{n}{\alpha + \beta + n} \hat{p}}_{MLE}$$

Seems to be
generalized

Obs: α, β are fixed
 $n \rightarrow \infty \Rightarrow E(P|x) \rightarrow \hat{p}_{MLE}$ and the prior expectation doesn't influence too much the posterior expectation

A closer look to posterior distribution

We have seen that the posterior mean is a weighted average of the prior expectation and the ML estimate, where

- ML estimate prevails if n is large;
- ML estimate prevails if α and β are small (the variance of the prior distribution is large). It is worth noting that $\alpha + \beta$ can be interpreted as the equivalent number of observation of the prior distribution.

The posterior distribution as a whole is a compromise between the prior and the likelihood, and the likelihood prevails if

- n is large;
- α and β are both close to 1 (the prior is diffuse)

To appreciate the quality of the posterior mean as an estimate we can look at the posterior variance (or at standard deviation)

$$V(\theta) = \frac{(\alpha+x)(\beta+n-x)}{(\alpha+\beta+n)^2(\alpha+\beta+n+1)} \text{ that for large } n \text{ is } \approx \frac{1}{n} \frac{x}{n} \left(1 - \frac{x}{n}\right)$$

A model for gaussian data

Assume that observations come from a gaussian distribution (variance known)

- $Y_1, \dots, Y_n \sim iid N(\mu, \sigma^2)$ conditional to parameter(s) value(s)

μ is the parameter, σ^2 is known;

the likelihood $L (= L(\mu))$ is

$$L \propto \prod_{i=1}^n \left(\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(y_i - \mu)^2} \right)$$

Gaussian model; σ^2 known

Likelihood:

$$\begin{aligned} L(\mu) &\propto \prod_{i=1}^n \left(\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(y_i - \mu)^2} \right) \\ &\propto e^{-\frac{n}{2\sigma^2}(\bar{y} - \mu)^2} \end{aligned}$$

Assume a gaussian prior on μ ,

$$\mu \sim N(\mu_0, \sigma_0^2)$$

The posterior distribution is then

$$\pi(\mu|y) \propto L(\mu)\pi(\mu)$$

Gaussian model; σ^2 known

$$\begin{aligned}
 \pi(\mu|y) &\propto e^{-\frac{n}{2\sigma^2}(\bar{y}-\mu)^2} e^{-\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2} \\
 &\propto e^{-\frac{n}{2\sigma^2}\mu^2 - \frac{1}{2\sigma_0^2}\mu^2 + \frac{\mu\bar{y}n}{\sigma^2} + \frac{\mu\mu_0}{\sigma_0^2}} \\
 &\propto e^{-\frac{1}{2}\left(\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}\right)\mu^2 + \mu\left(\frac{n}{\sigma^2}\bar{y} + \frac{1}{\sigma_0^2}\mu_0\right)} \\
 &\propto e^{-\frac{1}{2\left(\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}\right)}\left(\mu^2 - 2\mu\frac{\frac{n}{\sigma^2}\bar{y} + \frac{1}{\sigma_0^2}\mu_0}{\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}}\right)}
 \end{aligned}$$

get rid of constant
 that depends on \bar{y}

$$\begin{aligned}
 \pi(\mu|y) &\propto L(\mu)\pi(\mu) \\
 &\propto e^{-\frac{1}{2\left(\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}\right)}\left(\mu - \frac{\frac{n}{\sigma^2}\bar{y} + \frac{1}{\sigma_0^2}\mu_0}{\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}}\right)^2}
 \end{aligned}$$

Gaussian model; σ^2 known

$$\begin{aligned}\pi(\mu|y) &\propto L(\mu)\pi(\mu) \\ &\propto e^{-\frac{1}{2(\sigma^*)^2}(\mu-\mu^*)^2} \quad N(\mu^*, (\sigma^*)^2)\end{aligned}$$

that is, we obtain a gaussian posterior distribution with parameters μ^* and σ^* which are a function of prior distribution's parameters and of the data:

$$\mu^* = \frac{\frac{n}{\sigma^2}\bar{y} + \frac{1}{\sigma_0^2}\mu_0}{\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}} = \frac{\mu_0\sigma^2 + \bar{y}n\sigma_0^2}{\sigma^2 + n\sigma_0^2}$$

$$(\sigma^*)^2 = \left(\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2} \right)^{-1} = \frac{\sigma^2\sigma_0^2}{\sigma^2 + n\sigma_0^2}$$

Gaussian model; σ^2 known

The **posterior mean** is a weighted average of the prior mean and of the ML estimate, where the weights are the reciprocal of the respective variances

In this case the posterior distribution provided the mean directly

$$\mu^* = \mu_{n,\sigma_0}^* = \frac{\frac{n}{\sigma^2} \bar{y} + \frac{1}{\sigma_0^2} \mu_0}{\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}} = \frac{\frac{1}{V(\bar{y})} \bar{y} + \frac{1}{V(\mu)} \mu_0}{\frac{1}{V(\bar{y})} + \frac{1}{V(\mu)}}$$

- $\mu_{n,\sigma_0}^* \xrightarrow[n \rightarrow \infty]{} \bar{y}$ as n grows, the ML estimates weights more
- $\mu_{n,\sigma_0}^* \xrightarrow[\sigma_0 \rightarrow 0]{} \mu_0$ the more concentrated is the prior distribution, the more the prior mean weights.

It is interesting to write the posterior mean as

$$\mu^* = \mu_{n,\sigma_0}^* = \mu_0 + (\bar{y} - \mu_0) \frac{n\sigma_0^2}{\sigma^2 + n\sigma_0^2}$$

the posterior mean is the prior mean plus an adjustment toward the sample mean.

$$\mu^* = \mu_{n,\sigma_0}^* = \bar{y} - (\bar{y} - \mu_0) \frac{\sigma^2}{\sigma^2 + n\sigma_0^2}$$

Gaussian model; σ^2 known

The reciprocal of the **posterior variance** is the sum of the reciprocals of the prior variance and the variance of ML estimator

$$(\sigma^*)^2 = (\sigma_{n,\sigma_0}^*)^2 = \left(\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2} \right)^{-1} = \left(\frac{1}{V(\bar{y})} + \frac{1}{V(\mu)} \right)^{-1}$$

- $\sigma_{n,\sigma_0}^* \xrightarrow[n \rightarrow \infty]{} 0$ as n grows the variance of the posterior diminish
- $\sigma_{n,\sigma_0}^* \xrightarrow[\sigma_0 \rightarrow 0]{} 0$ also if the variance of the prior is reduced the posterior is more concentrated

Application of Bayesian mean principle: estimating the score for Tripadvisor ratings

Tripdavisor uses a formula for calculating and comparing the ratings of restaurants by its users. It derives from using a weighted mean that relies upon the Bayesian idea

The following formula was the base to calculate $W = \frac{Rv + Cm}{v + m}$

where:

$W \rightarrow C$ for few data

$W \rightarrow R$, $n \gg t$

- W = weighted rating
- R = average rating (stars) for the restaurant (1 to 5) - *the likelihood*
- v = number of votes/ratings for the restaurant = (votes)
- m = weight given to the prior estimate (in this case, the number of votes for a stable average rating)
- C = the mean vote across the whole pool - *the prior*

W is just the weighted arithmetic mean of R and C with weights v and m .

As the number of ratings surpasses m , the confidence of the average rating surpasses the confidence of the prior knowledge, and the weighted

Bayesian interval estimation and testing

Bayesian interval estimation

The posterior distribution can be summarized by posterior expectation and variance;

- these roughly correspond to point estimate and its standard error in classical inference (although the interpretation is a bit different).
- Given that θ is a random variable, it is natural to think at an analogue of confidence intervals;
- this analogue is called **credibility interval**.
- there is a big difference in interpretation where credibility interval are much more natural and close to common sense.
- most non statisticians actually interpret confidence intervals as if they were credibility intervals.

Classical confidence interval vs credibility interval

Classical interval estimate (confidence interval)

An interval is associated to the sample y such that with a confidence level $1 - \alpha$, contains the true value of the parameter.

$$P(L < \theta < U) = 1 - \alpha$$

$I = [L, U]$ is a r.v

Interpretation: if N samples were observed and for each of them a $1 - \alpha$ confidence interval were obtained, on average $100(1 - \alpha)$ of them would contain the true value of the parameter.

An interval is associated to the sample y such that it **contains the true value of the parameter with probability $1 - \alpha$** .

Bayesian interval estimate (credibility interval)

A credibility interval for θ is a pair of statistics $L(Y), U(Y) \in \Theta$ such that

$$P(L(Y) \leq \theta \leq U(Y)) \geq 1 - \alpha$$

where the probability is with respect to the distribution of θ ,

$$P(L(Y) \leq \theta \leq U(Y)) = \int_{L(Y)}^{U(Y)} \pi(\theta|y) d\theta$$

$\theta \in I$ w/
 $1 - \alpha$ prob.

Credibility intervals

Given a distribution for θ , $\pi(\theta|y)$ there is not a unique interval satisfying the condition

$$P(L(Y) \leq \theta \leq U(Y)) = \int_L^U \pi(\theta|y) d\theta = 1 - \alpha$$

the easiest choice is to set L and U equal to the quantiles $\alpha/2$ and $1 - \alpha/2$ of $\pi(\theta|y)$, that is, such that

$$\int_{-\infty}^L \pi(\theta|y) d\theta = \int_U^{+\infty} \pi(\theta|y) d\theta = \alpha/2$$

this interval satisfies the condition but is not, generally, the smallest one.

This interval is not the smallest one and the tails include values with higher density in one side than the other.

HPD (High Posterior Density) region

A better (smaller) interval is defined as

High posterior density (HPD)

The high posterior density credibility region is a set $C \subset \Theta$ such that

$$P(\theta \in C) = 1 - \alpha$$

and

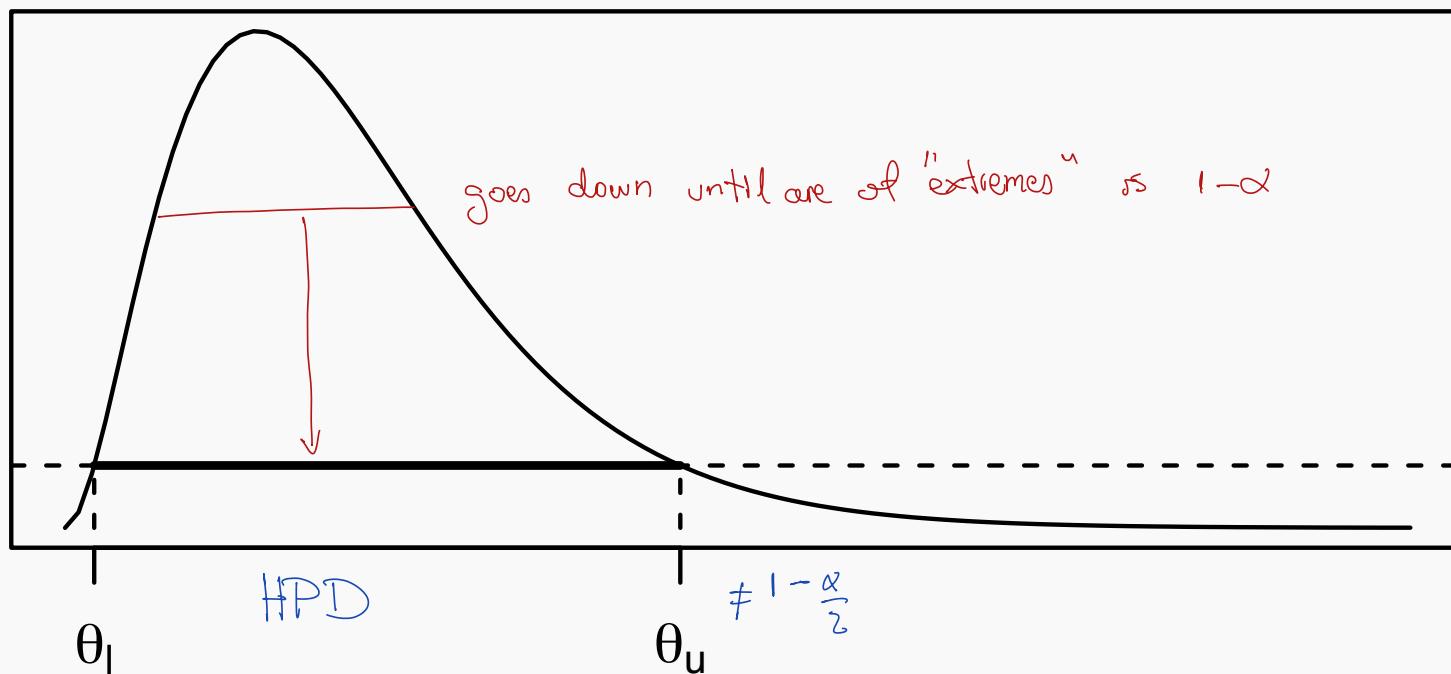
$$\pi(\theta_1|y) > \pi(\theta_2|y)$$

if $\theta_1 \in C$ and $\theta_2 \notin C$

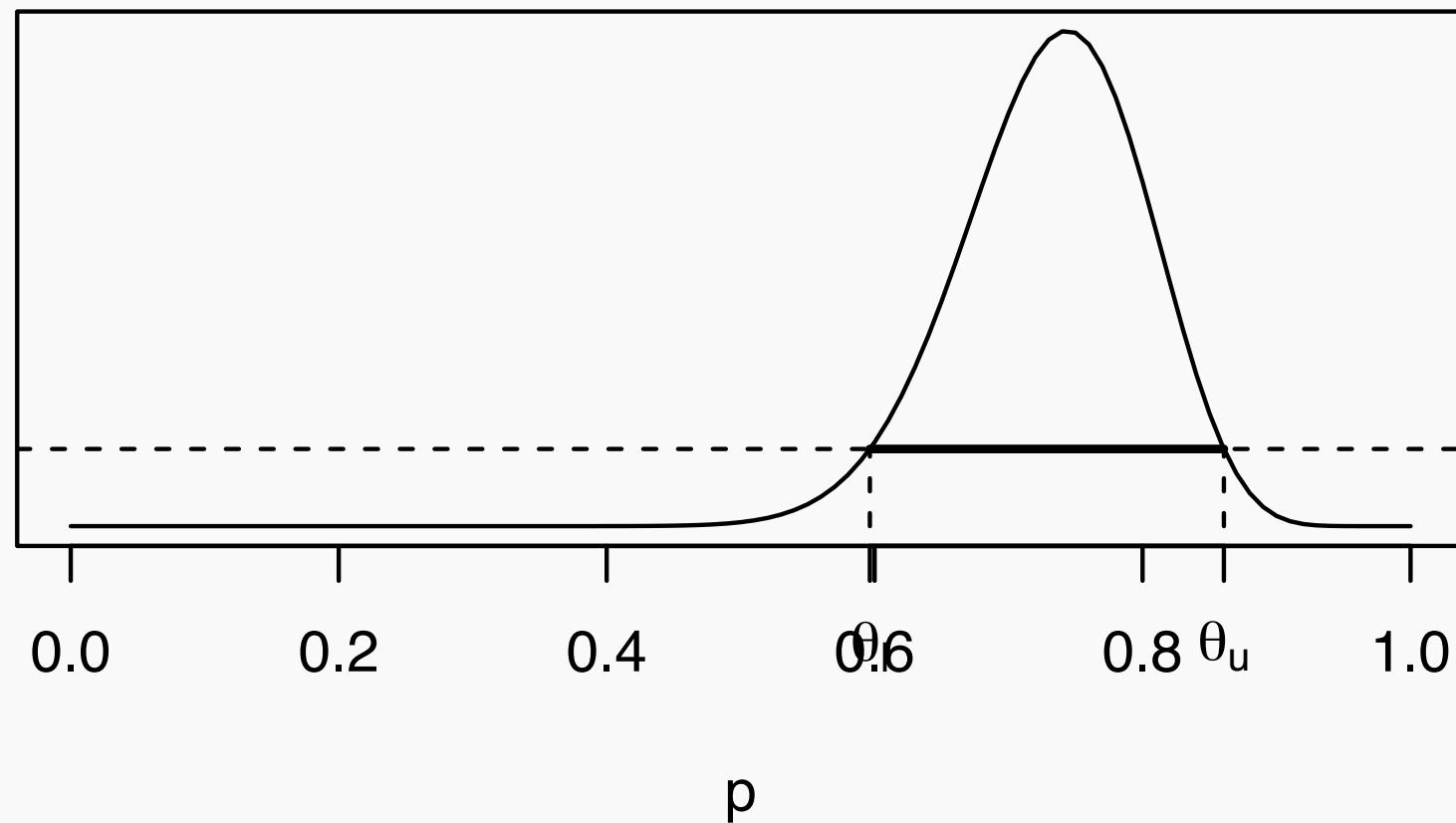
Given $\pi(\theta|y)$ the HPD interval C is obtained including the values of θ corresponding to a higher density

HPD region

HPD region

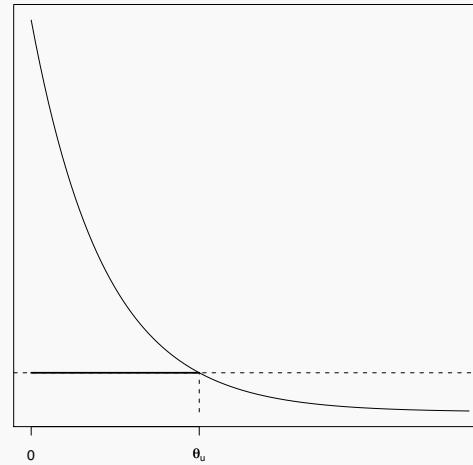


.95 HPD region for p in the seeds example

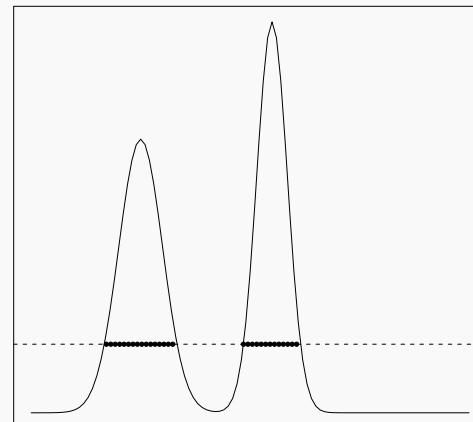


Special cases

monotone posterior



multimodal posterior the HPD region is not necessarily an interval but can be the union of disjoint intervals



Finding the HPD region

For a unimodal posterior (not necessarily symmetric) we may use an algorithm to find the interval:

start from $k_m = 0$, $k_M = \max_{\theta} \pi(\theta|y)$ then at step i

1. $k_i = (k_m + k_M)/2$
 2. determine $C = \{\theta | \pi(\theta|y) > k_i\}$
 3. compute $I = \int_C \pi(\theta|y) d\theta$
- if $I < 1 - \alpha$ $k_m \leftarrow k_i$ (shorter interval) return to 1
 - if $I > 1 - \alpha$ $k_M \leftarrow k_i$ (longer interval), return to 1
 - if $I = 1 - \alpha$ STOP C is the solution

Hypotheses testing

Suppose you want to test the Hypothesis

$$H_0 : \theta \in \Theta_0; \quad H_1 : \theta \in \Theta_1$$

Θ_0 and Θ_1 form a partition of the parameter space

The beliefs about the two hypotheses are summarized by the posterior odds ratio

$$\frac{p_0}{p_1} = \frac{P(\theta \in \Theta_0 | y)}{P(\theta \in \Theta_1 | y)} = \frac{\int_{\Theta_0} \pi(\theta | y) d\theta}{\int_{\Theta_1} \pi(\theta | y) d\theta}$$

A measure of the evidence provided by the data in support of H_0 is the
Bayes factor

$$BF = \frac{\text{posterior odds}}{\text{prior odds}} = \frac{p_0/p_1}{\pi_0/\pi_1}$$

Where π_0 and π_1 are respectively $\int_{\Theta_0} \pi(\theta) d\theta$ and $\int_{\Theta_1} \pi(\theta) d\theta$ the probability of the two hypotheses prior observing the data. Note that you can then evaluate the posterior probability that the null hypothesis is true

$$p_o = \frac{\pi_0 BF}{\pi_0 BF + 1 - \pi_0}$$

Selecting the prior

The prior distribution

The idea of using prior knowledge in Bayesian statistics is a critical issue and it is a new element in comparison with classical statistics.

Formally this knowledge is introduced by specifying a prior distribution that includes information other than what is directly observed in the process of inference.

The most common concerns are about the use of subjective probability in deriving the prior (for instance, by using experts' opinions for eliciting the prior).

For this reasons some relevant topics in Bayesian statistics refer to:

1. the choice of prior distributions that are diffuse (non informative) in
 $\theta \in U[a,b]$
 $\rightarrow \theta$ informative
order to give more (or exclusively) weight to experimental data or to obtain results that are consistent with results from likelihood based inference
2. the analysis of the sensitivity of the inference to the alternative choice of the prior (Bayesian robustness)

Objections on the use of prior distributions

One (non-Bayesian statistician) could argue that if I specify a subjective prior distribution, since I can chose any distribution, I can also modify the result and obtain whatever conclusion I want. The result could then be manipulated it is subjective and hence not scientific.

Counter-objections include

- classical procedures are also subjective, for example in the specification of the model;
- the relevance of the prior distribution is limited and tends to vanish if the sample size increases;
- actually, the information conveyed by the data would outweigh the information in the prior for any reasonable specification;
- a possible compromise is to use standard priors which do not involve personal (subjective) opinions.

Conjugacy

Note that in the two examples considered above prior and posterior distribution have the same functional form

For the seeds example the posterior distribution is a Beta like the prior as well as for the Normal mean example

Likelihood	Prior	Posterior
$L(\theta; y)$	$\pi(\theta)$	$\pi(\theta y)$
Binomial	$Beta(\alpha, \beta)$	$Beta(\alpha + \sum_i y_i, \beta + n - \sum_i y_i)$
Normal	$N(\mu, \sigma^2)$	$N(\mu^*, (\sigma^*)^2)$

This property relates the family of the prior distribution with the likelihood and is called **conjugacy**. Example of conjugate families are:

- Beta prior and Binomial likelihood
- Normal and Normal
- Gamma prior and Poisson likelihood

Use of conjugacy would lead to select the prior in order to make computation of the posterior easy and straightforward. None the less, conjugacy could not be the best solution to reflect real world knowledge about the parameter. Many other factors

Non informative priors



The prior distribution is meant to reflect the opinion of the researcher prior to observing any data. What if there is no opinion? (Whether this is realistic is disputable.)

This is a relevant issue and a possible answer to the objection that the results of inference should not depend on subjective opinions.

It has then been proposed to use ‘standard’ distributions which, in some sense, bring no (or very limited) information on the parameter.

An intuitive solution is to assume $\pi(\theta) \propto k$ so that no values of θ are privileged (principle of insufficient reason).

- Strictly speaking, this is admissible only if the parameter space is limited.
- If the parameter space is not limited a constant has an infinite integral and so is not a probability distribution.
- It is possible however, that a proper posterior distribution is obtained even starting from an improper prior. If this is the case, the inference is valid.

Jeffreys' prior

The non informative nature of the uniform distribution is disputable

- Let

$$\pi(\theta) \propto k$$

- consider the reparametrization $\psi = \psi(\theta)$, then

$$\pi(\psi) = \pi(\theta^{-1}(\psi)) \left| \frac{d\theta}{d\psi} \right|$$

which is not uniform in general.

- that is, assuming that uniform means non informative, by specifying a uniform distribution for the parameter θ , we are specifying instead an informative prior on its transform $\psi = \psi(\theta)$.

Jeffreys' prior

The above issue may be overcome by posing

$$\pi(\theta) = \sqrt{\det H(\theta)}$$

where H is the information matrix, that is, the matrix with (i, j) element

$$[H(\theta)]_{ij} = -E\left(\frac{\partial^2}{\partial\theta_i\partial\theta_j}I(\theta; y)\right)$$

then, for any parametrization $\psi = \psi(\theta)$

$$\pi(\psi) = \sqrt{\det H(\psi)} = \sqrt{\det H(\theta)} \left| \det \left(\frac{d\theta}{d\psi} \right) \right|$$

Consider, for instance, a Binomial experiment, so the log-likelihood is

$$I(\theta) = y \log \theta + (n - y) \log(1 - \theta)$$

then

$$[H(\theta)] = -E\left(\frac{d^2}{d\theta^2}I(\theta; y)\right) = \frac{n}{\theta(1 - \theta)}$$

Jeffrey's prior
is an informative prior.

the Jeffreys' prior is then a Beta($1/2, 1/2$)

$$\pi(\theta) \propto \theta^{-1/2}(1 - \theta)^{-1/2}$$

Bayes computation

Bayes computation

To answer the basic questions of statistical inference we need to know the posterior

$$\pi(\theta|y) = \frac{f(y|\theta)\pi(\theta)}{\int_{\Theta} f(y|\theta)\pi(\theta)d\theta}$$

The principal practical challenge is that $\int_{\Theta} f(y|\theta)\pi(\theta)d\theta$ is usually intractable for many interesting models and also quantity related to $\pi(\theta|y)$ of direct interest for summarizing inference (such as mean, median, percentiles, probabilities) cannot be evaluated.

There are then two main strategies to overcome the problem:

- approximate the integrals
- find a way to get a (simulated) sample from $\pi(\theta|y)$ without requiring evaluation of the integrals.

The latter strategy is based on the fact that if we simulating from a density is as good as being able to evaluate the density, and sometimes better. This is achieved mainly by Monte Carlo Markov Chain methods.

Monte Carlo Markov Chain

Monte Carlo Markov Chain methods simulate values from a Markov chain whose stationary distribution is exactly the posterior distribution of interest.

Once a sample of simulate values is given this can be used to evaluate all the quantities of interest for Bayesian inference

Two are the main algorithm to obtain this sample of simulate values

- Metropolis-Hastings algorithm
- Gibbs sampling

Variational inference

- **Variational inference** is a method from machine learning that approximates probability densities through optimization. The idea is to use this approach to approximate the posterior distribution
- It has been used in many applications with a complex parameters space (the most notable is topic modelling) and tends to be faster than methods based on sampling fromm the posterior distribution, such as Markov chain Monte Carlo.
- The idea behind variational inference is to first posit a family of densities over the parameter space and then to find the member of that family which is close to the target. Closeness can be measured, for instance, by Kullback-Leibler divergence.