Classical Linear Models and beyond

Master in Modelling for Science and Engineering

1. Linear models
2. Generalized Linear Models and Random Effects Models
3. Time Series Models

Linear Models

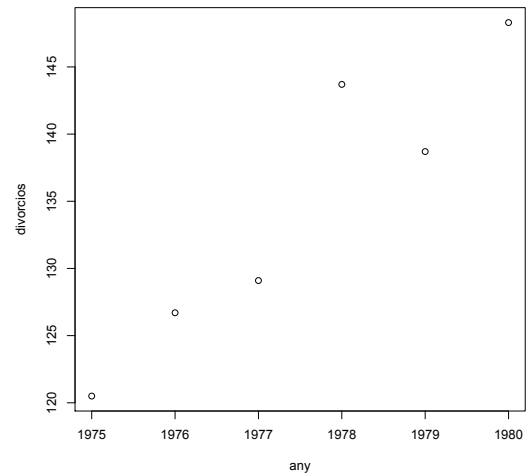
Regression is all about relationships between variables. In its simplest form, it constitutes a technique for modelling a relationship between two variables.

It was the pioneering work of Sir Francis Galton in the 1880s that gave rise to the technique, the original idea being the direct result of an experiment on sweet peas.

He noticed that the seeds of the progeny of parents with seeds heavier than average were also heavier than average, but the difference was not as pronounced; the same effect was true for the seeds of the progeny of parents with light seeds, where again the differences from the average were not as great. He called this phenomenon *reversion* and wrote that the mean weight "reverted, or regressed, toward mediocrity".

Annual number of divorces recorded in England and Wales between 1975 and 1980[1]

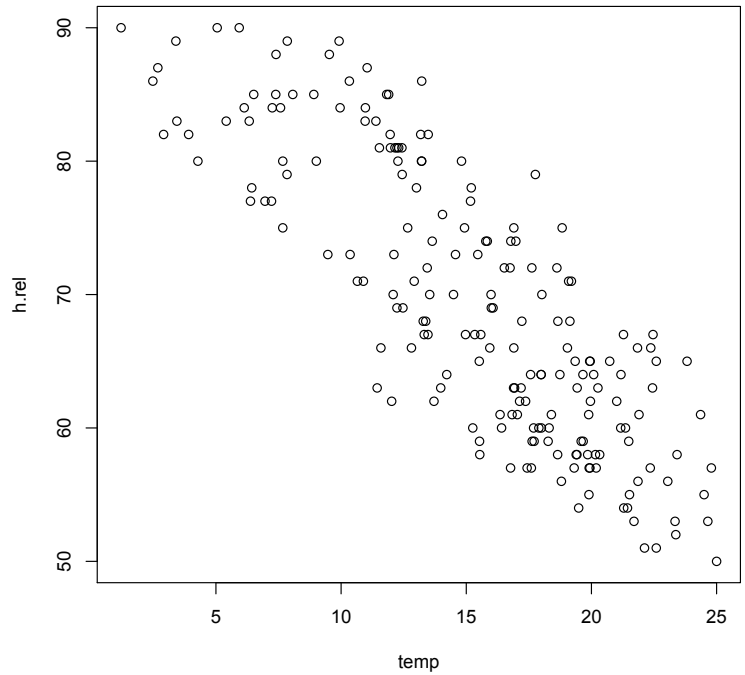
Year	1975	1976	1977	1978	1979	1980
Divorces (thousands)	120.5	126.7	129.1	143.7	138.7	148.3



The plot suggests a roughly linear trend with some random scatter:
$divorces_i = \beta_0 + \beta_1 year_i + error$

[1] *Marriage and Divorce Statistics*, Office of Population Censuses and Surveys, HMSO

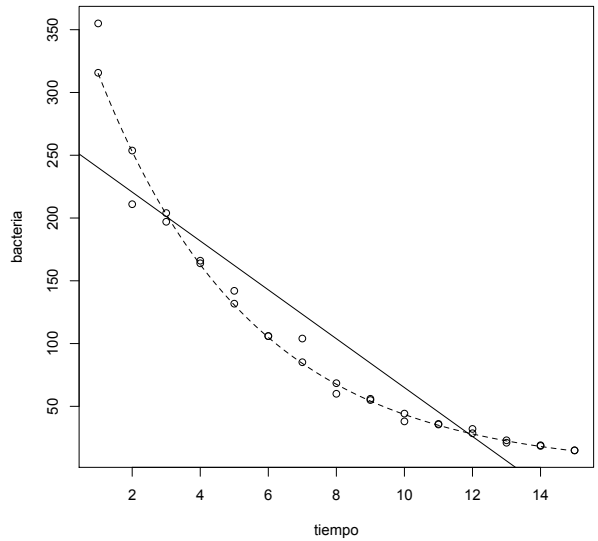
Next plot shows registers of temperature and relative humidity taken in a meteorological station in Castilla y León taken between May and November, 2000.



$$hrel_i = \beta_0 + \beta_1 temp_i + error$$

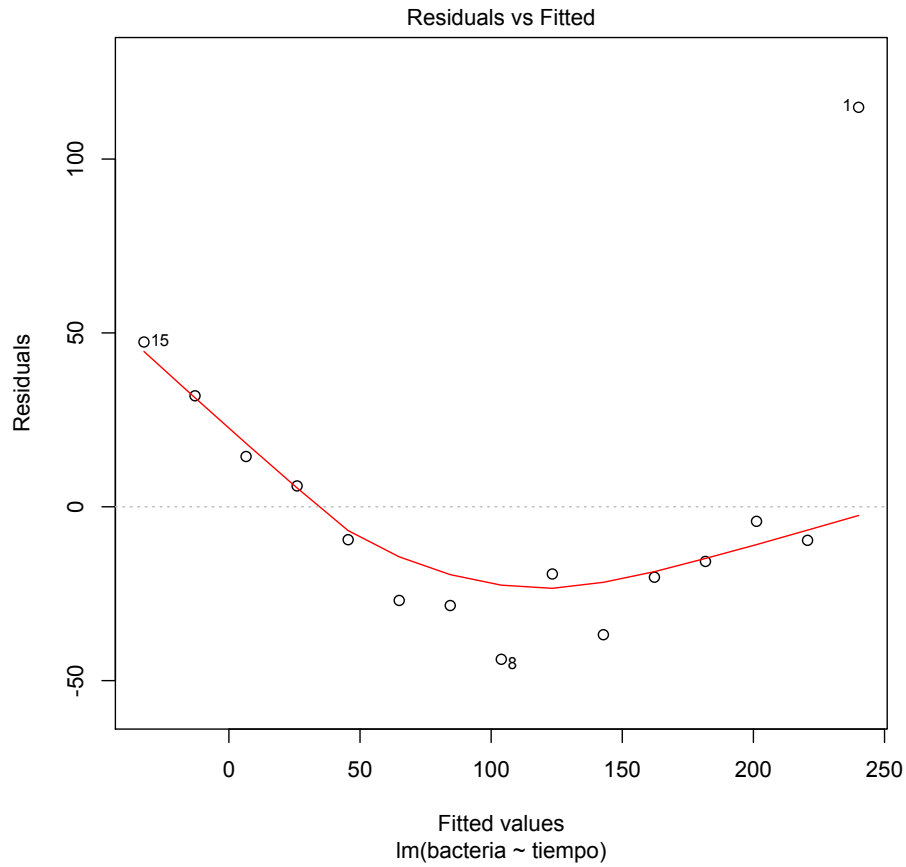
A study on the effects of X rays applied to marine bacteria[2]. According to the theory, each bacteria can be annihilated if an X-ray hits its vital centre. These kind ob bacteries are easy to count with a microscope. If the theory is correct, the logarithm number of surviving bacteria should be a linear function of radiation time.

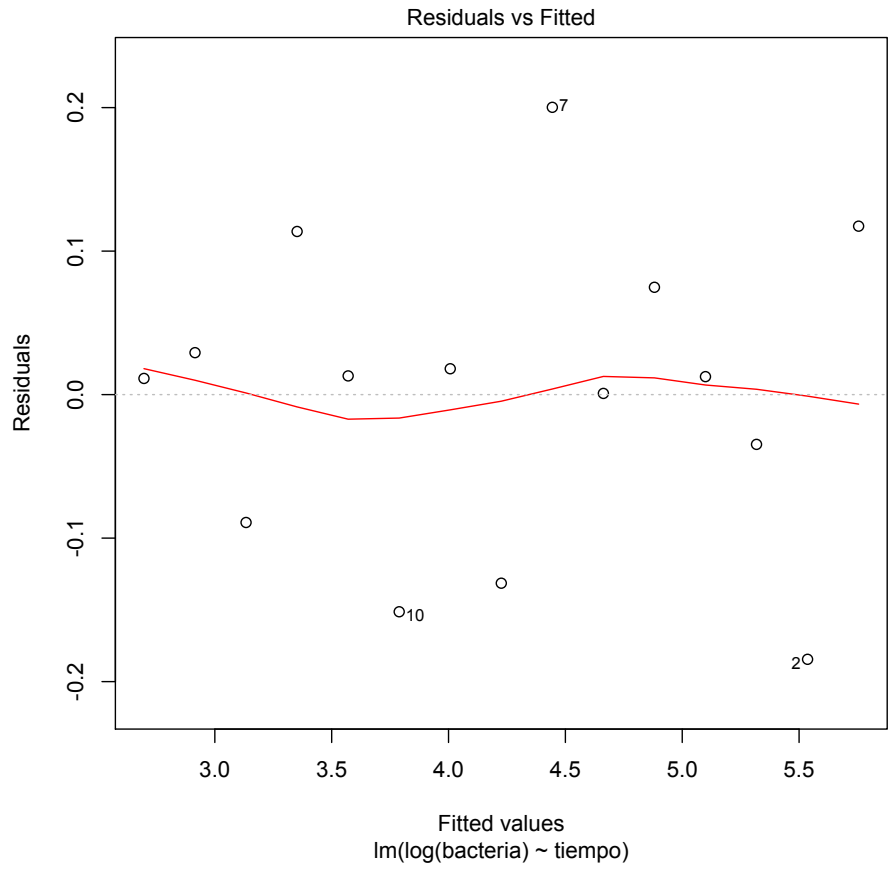
`http://mat.uab.cat/~acabana/data/rayosXbac.R`



$$\ln(bacteria) = \ln(a) + b * time + error$$

[2]Chatterjee, S. and Price, B. (1991). *Regression Analysis By Example*. Wiley-Interscience, Table 2.6, pag 36

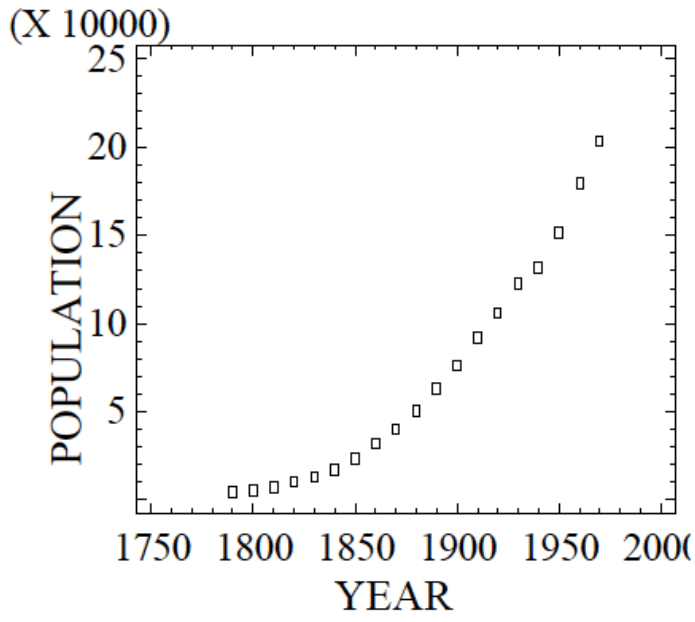

Residuals vs Fitted
Residuals
Fitted values
lm(bacteria ~ tiempo)
100
50
0
-50
0
50
100
150
200
250
1
15
8


Residuals vs Fitted
Residuals
0.2
0.1
0.0
-0.1
-0.2
7
10
2
3.0
3.5
4.0
4.5
5.0
5.5
Fitted values
lm(log(bacteria) ~ tiempo)

EVOLUCIÓN DE LA POBLACIÓN EN U.S.A.

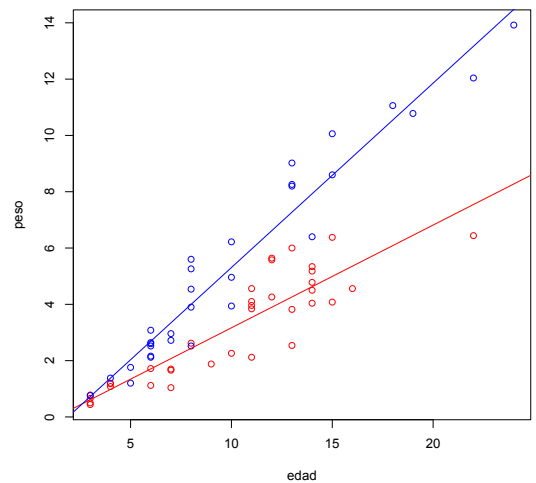
La tabla muestra la evolución de la población americana (POPUL.) en miles entre los años 1790 y 1970.

YEAR	POPUL.
1790	3929
1800	5308
1810	7239
1820	9638
1830	12866
1840	17069
1850	23191
1860	31443
1870	39818
1880	50155
1890	62947
1900	75994
1910	91972
1920	105710
1930	122775
1940	131669
1950	151325
1960	179323
1970	203211



$$pop_i = \beta_0 + \beta_1 * year_i + \beta_2 * year_i^2 + error$$

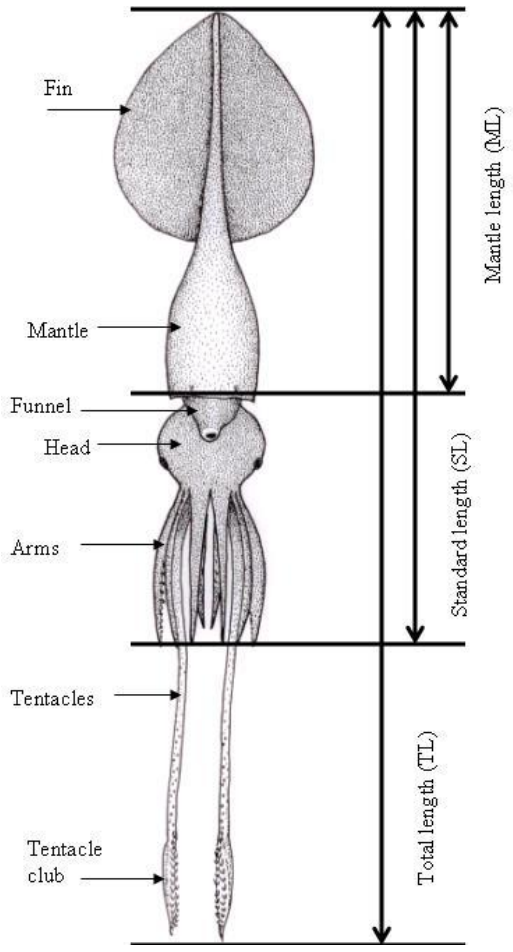
The data in
`http://mat.uab.cat/~acabana/data/mejillones.txt`
are part of a study on the age and growth characteristics of some species of mussels in two different locations in SW Virginia[3].
The idea is to find an adequate model to find out whether the location is an influential factor in growth.

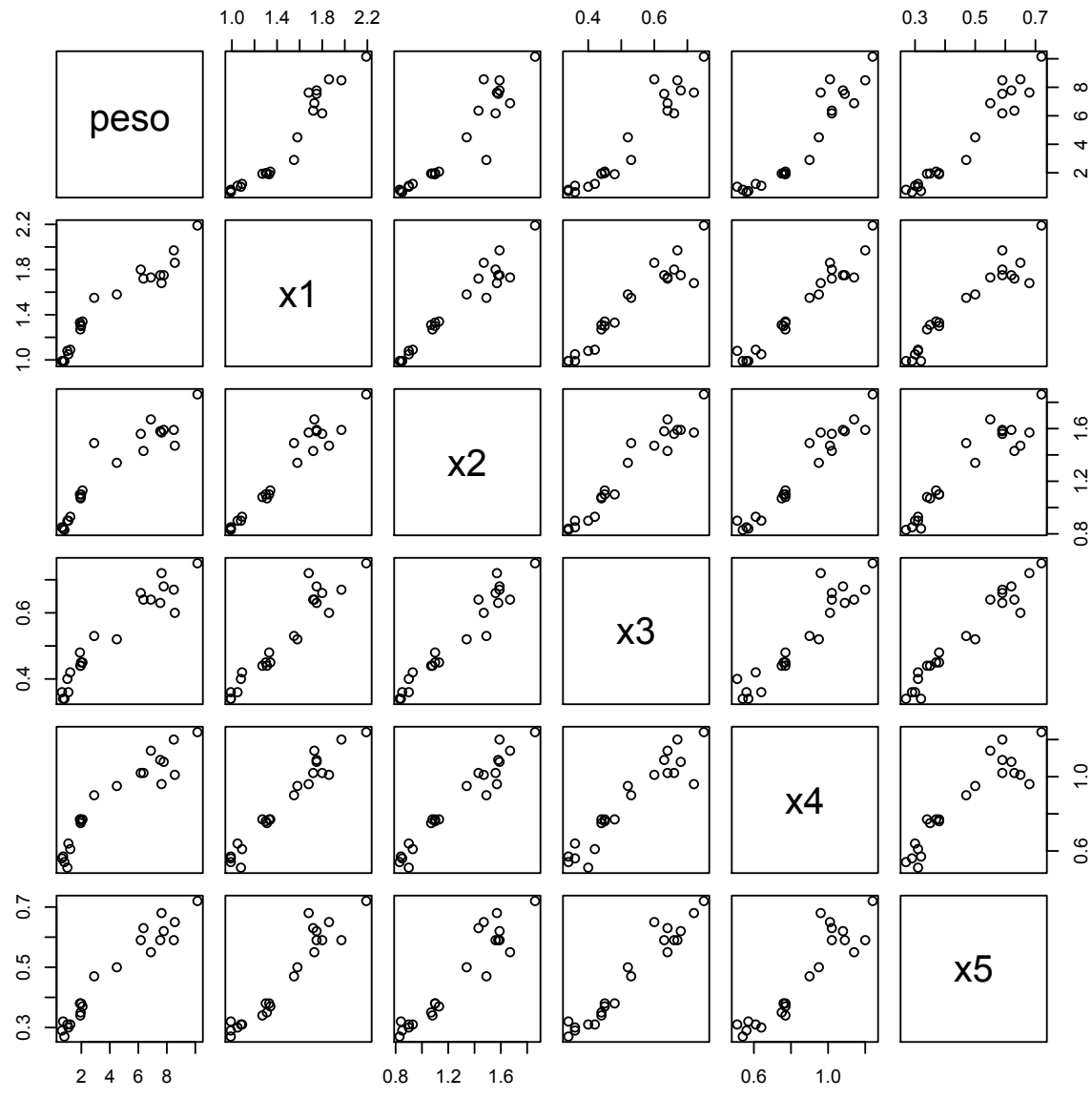


$$weight_i = \beta_0 + \beta_1 * 1_{L1} + \beta_2 * 1_{L1} * age_i + \beta_3 * 1_{L2} + \beta_4 * 1_{L2} * age_i + error$$

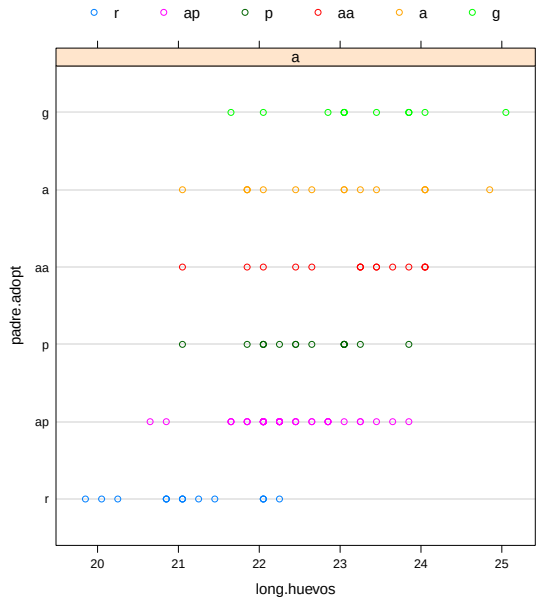
[3]Myers, R. H. (1990) Classical and Modern Regression with Applications, 2nd Ed

The data in
`http://mat.uab.cat/~acabana/data/calamar.txt`
are part of an experiment for studying squid that are eaten by sharks and tuna. We have measurements of 5 morphological variables of the squid to be related to weight in 22 specimens.

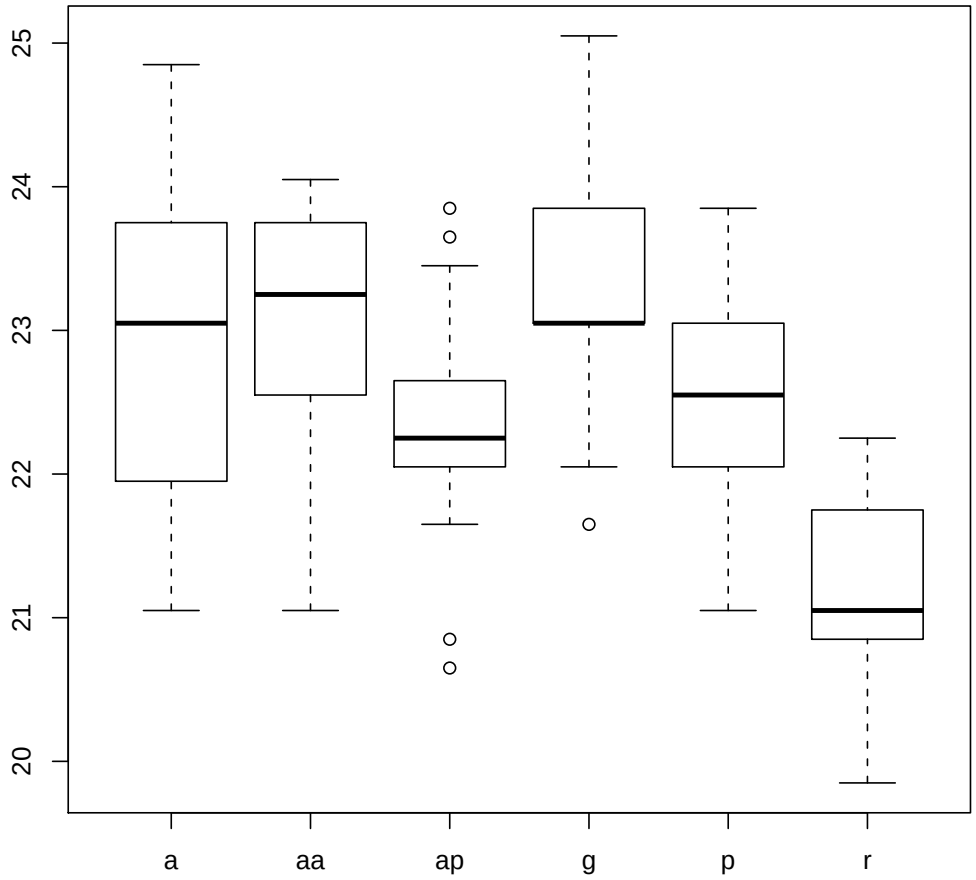



peso
x1
x2
x3
x4
x5

It is known at least since 1892 that the size of cuckoo eggs changes according to location. Moreover, in a study from1940 [4] they claim that cuckoos breed only in their particular location and that there are several geographical sub-species because they adopt different foster fathers.



[4] E.B. Chance, *The truth about the cuckoo* `http://lib.stat.cmu.edu/DASL/Stories/cuckoo.html`


25
24
23
22
21
20
a
aa
ap
g
p
r

All of the previous examples can be reduced to this situation:
Consider a variable Y in $\mathbf{R}^n$ obtained as a sum of an unknown parameter $\mu \in \mathbf{R}^n$ and a variable of errors e whose distribution is assumed to be isotropic Normal with variance σ^2 in $\mathbf{R}^n$:

$$Y = \mu + e.$$

In other words, Y is Gaussian in $\mathbf{R}^n$ with expectation μ and variance $\sigma^2 I$.
We will make inferences on μ and/or σ^2, from the information contained in a sample of Y.

If the former model is written expressing μ as $X\beta$, where X is a $p \times n$ matrix whose columns (that are assumed to be independent) span $\mathcal{R}$, and β is a vector in $\mathbf{R}^p$.

It's just a change of parameters, instead of using μ in a subspace $\mathcal{R}$ of dimension p, we use β in $\mathbb{R}^p$, linearly related to μ through the equation

$$\mu = X\beta$$

so that the model is written as

$$Y = X\beta + e$$

If we assume that the expectation of the error vector is zero, then the expectation of Y is

$$X\beta = \beta_1 X_1 + \ldots \beta_p X_p$$

where X_i is the i-th column of X. That is, $\mathbf{E}Y$ is in the subspace $\mathcal{R}$ spanned by the columns of X. So, we can think of approximating Y by an element of $\mathcal{R}$. We choose the orthogonal projection of Y in $\mathcal{R}$, denoted by $\hat{Y} = X\hat{\beta}$ that satisfies

$$X^t(Y - X\hat{\beta}) = 0 \iff (X^tX)\hat{\beta} = X^tY \quad \text{normal equations}$$

If $X_{n\times p}$ has maximal rank, X^tX has an inverse and $\hat{\beta} = (X^tX)^{-1}X^tY$. If not, $\hat{Y}$ is unique but there are infinitely many $\hat{\beta}$ that generate it.

Geometry of the linear model

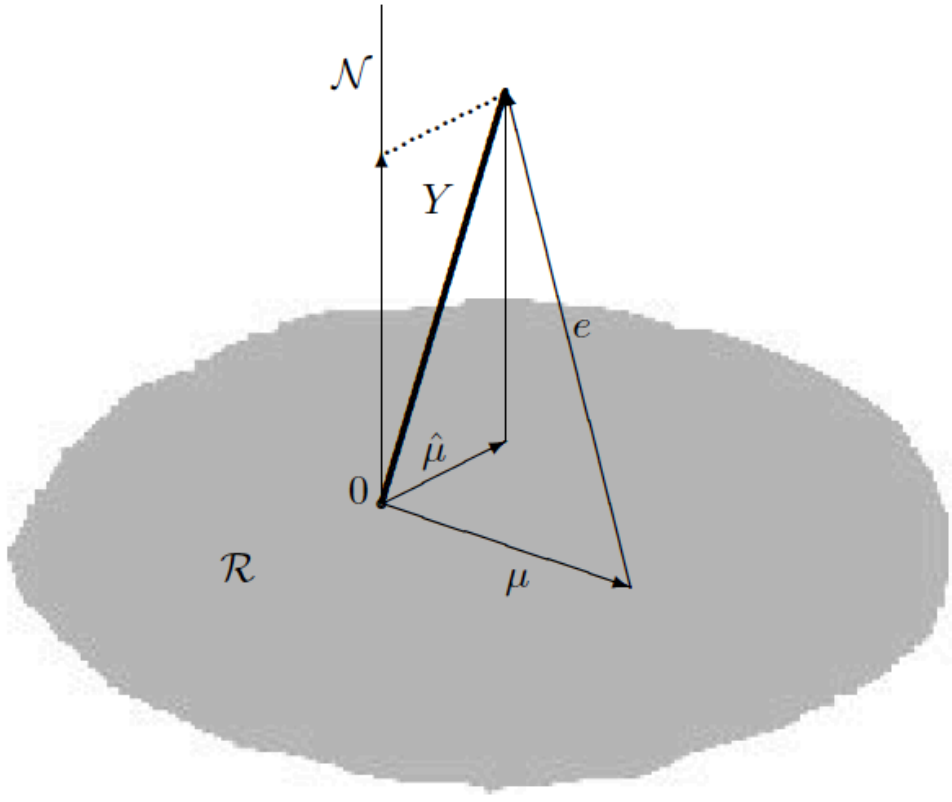


Figura 1: Proyección $\hat{\mu}$ de Y sobre el espacio $\mathcal{R}$ de estimación, sombreado, y sobre su complemento ortogonal $\mathcal{N}$. Esta última coincide con la proyección del vector de errores e.

Assuming that $e \sim \mathrm{N}(0, \sigma^2 I)$, where I is the identity matrix $(n \times n)$, then

$$Y \sim \mathrm{N}(X\beta, \sigma^2 I).$$

Since we are considering Y as a random variable, $\hat{\beta}$ is also a random variable.

$$\mathbf{Var}\hat{\beta} = \mathbf{E}\left((\hat{\beta} - \beta)(\hat{\beta} - \beta)^t\right)$$

Observe that

$$(\hat{\beta} - \beta) = (X^t X)^{-1} X^t Y - \beta = (X^t X)^{-1} X^t (X\beta + e) - \beta =$$

$$\beta + (X^t X)^{-1} X^t e - \beta = (X^t X)^{-1} X^t e$$

$$\mathbf{Var}\hat{\beta} = \mathbf{E}((X^tX)^{-1}X^te)((X^tX)^{-1}X^te)^t =$$

$$\mathbf{E}((X^tX)^{-1}X^te)(e^tX(X^tX)^{-1})$$

$$(X^tX)^{-1}X^t\mathbf{E}ee^tX(X^tX)^{-1} = (X^tX)^{-1}X^t\sigma^2IX(X^tX)^{-1}$$

so that

$$\mathbf{Var}\hat{\beta} = \sigma^2(X^tX)^{-1}$$

hence, each

$$\hat{\beta}_i \sim \mathbf{N}(\beta, \sigma^2 c_{ii})$$

where c_{ii} is in the diagonal of $C = (X^tX)^{-1}$ corresponding to β_i.

Estimating σ^2

The (random) error can be estimated by

$$\hat{e} = Y - \hat{Y} = Y - X\hat{\beta} = Y - (X^tX)^{-1}X^tY$$

Denote by $H = X(X^tX)^{-1}X^t$, the matrix of the projection onto $\mathcal{R}$spanned by the columns of X. Then,

$$\hat{e} = (I - H)Y.$$

$$\hat{e} = (I - H)Y = (I - H)(X\beta + e) = (I - H)e$$

Denote $SSE = ||\hat{e}||^2 = \hat{e}^t\hat{e} = e^t(I - H)e$.
Then,

$$S^2 = \frac{SSE}{n - p}$$

is an unbiased estimator of σ^2.
Observe that

$$SSE = X^tY - \hat{\beta}X^tY.$$

These estimators are used in order to make inferences about β and in order to compare models.

The multiple correlation coefficient

According to the model $Y = \mu + e = X\beta + e$, the observations describe μ adequately if the error e is relatively small, for instance, if $\frac{\sigma}{||Y||}$ o $\frac{\sigma}{\mu}$ but since we don't know σ, we are happy if $\dfrac{S^2}{||Y||^2}$ is small.

Since

$$||Y||^2 = ||\hat{\mu}||^2 + S^2(n-p)$$

asking $\dfrac{S^2}{||Y||^2}$ to be small is equivalent to asking that

$$\tilde{R}^2 = \frac{||\hat{\mu}||^2}{||Y||^2} = \frac{\hat{\beta}^t X^t Y}{Y^t Y}$$

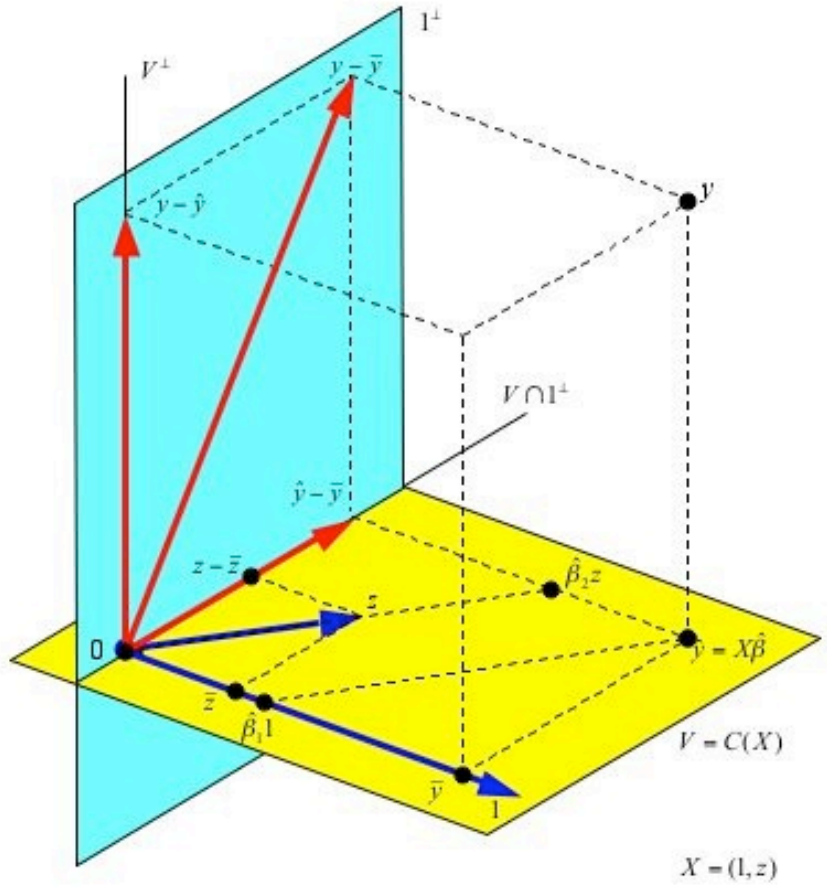
be big.

$\tilde{R}^2$ is the square of the cosine of the angle between Y and $\mathcal{R}$.
A change in the units of measure does not change the angle, but a change in the origin of the coordinate system does. In order that the measurement be invariant, we fix always the origin to be $\bar{Y}$ and define

$$R^2 = \frac{||\Pi_{\mathbf{1}^\perp} Y||^2}{||Y - \bar{Y}\mathbf{1}||^2}$$

where $\mathbf{1} = (1, 1, \ldots, 1)^t$, $\bar{Y} = \sum_{i=1}^n Y_i/n$, and $\Pi_{\mathbf{1}^\perp}$ is the matrix of the projection on the orthogonal complement of the subspace spanned by $\mathbf{1}$.

Geometry of the linear model II



Logistic Regression

It is a regression model where the dependent variable is categorical.

The simplest case is when we have binary dependent variables—that is, where it can take only two values, such as pass/fail, win/lose, alive/dead or healthy/sick, 0/1. The binary logistic model is used to estimate the probability of a binary response based on one or more predictor.

Probabilities are estimated using a logistic function:

$$f(x) = \frac{L}{1 + \mathrm{e}^{-k(x-x_0)}}$$

The Space Shuttle Challenger disaster occurred on January 28, 1986, when the NASA Space Shuttle Challenger broke apart 73 seconds into its flight, leading to the deaths of its seven crew members. Disintegration of the vehicle began after an O-ring seal in its right solid rocket booster failed at liftoff. The O-ring was not designed to fly under unusually cold conditions as in this launch (31 F). Data on temperatures and the results of previous launches.

temperature	66	70	69	68	67	72	73	70	57	63	70	78
result	0	1	0	0	0	0	0	0	1	1	1	0
temperature	67	53	67	75	70	81	76	79	75	76	58	
result	0	1	0	0	0	0	0	0	1	0	1	

The analysis with `R` is done with the instruction[5]

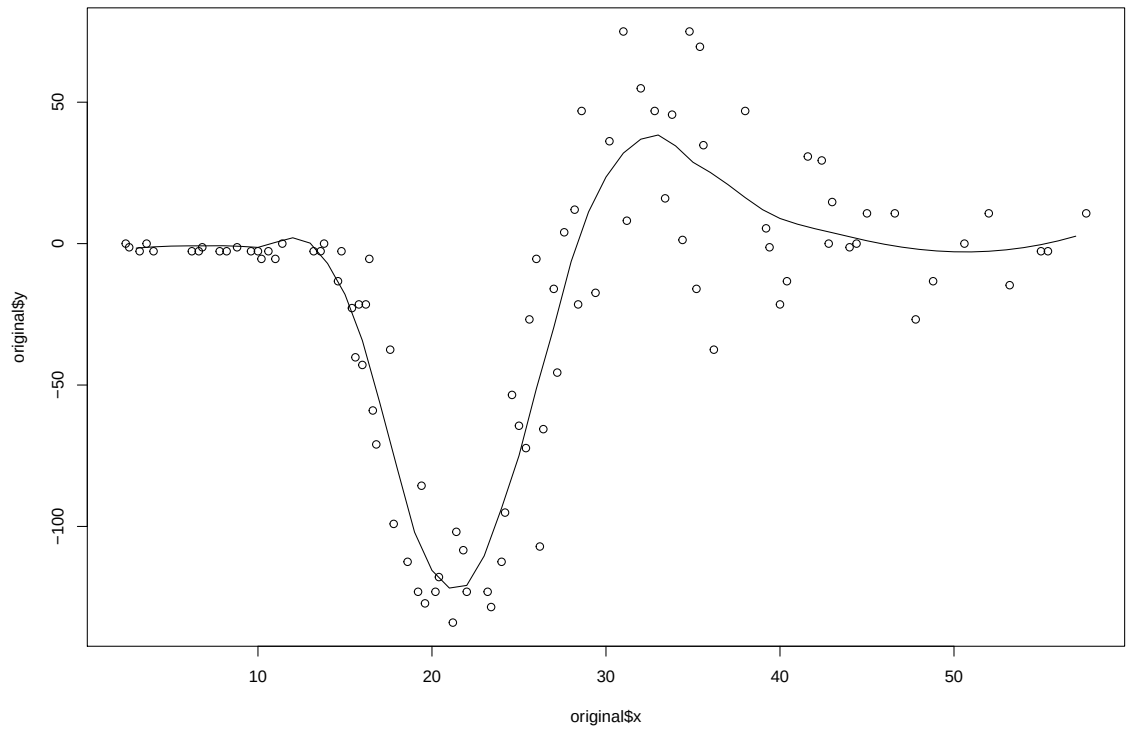
```
logit<-glm(y temp,family=binomial)
resu=summary(logit)
```

We get that the *odds* (how much more likely was a failure than a success?) $\frac{p}{1-p} = \mathrm{e}^{((b0+b1*31))} = 2555.125$ hence $p = 0.9996088$.

[5] see `modelos.R`

LOESS: local polynomial regression fitting

The plot corresponds to a series of measurements of head acceleration in a simulated motorcycle accident, used to test crash helmets[6].



[6] Silverman, B.W. (1985) Some aspects of the spline smoothing approach to non-parametric curve fitting, JRSS-B. See `modelos.R`

LOESS builds on classical methods, such as linear and nonlinear least squares regression. It addresses situations in which the classical procedures do not perform well or cannot be effectively applied without undue labor.

It does this by fitting simple models to **localised subsets of the data** to build up a function that describes the deterministic part of the variation in the data, point by point.

One of the chief attractions of this method is that the data analyst is not required to specify a global function of any form to fit a model to the data, only to fit segments of the data.

The command `loess` in R fits these kind of models.

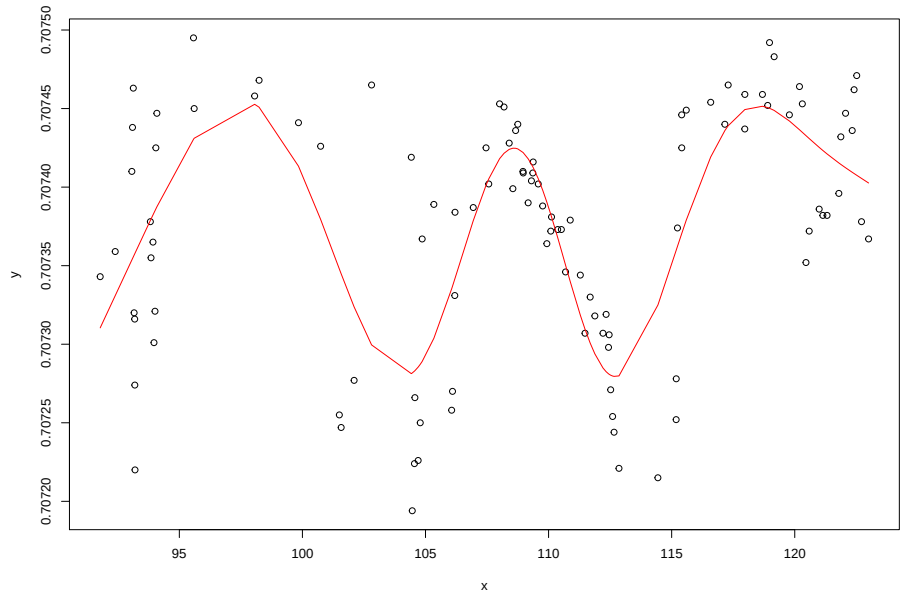
Generalized additive models (GAM)

A GAM is a generalized linear model in which the linear predictor depends linearly on unknown smooth functions of some predictor variables, and interest focuses on inference about these smooth functions.

$$g(\mathrm{E}(Y)) = \beta_0 + f_1(x_1) + f_2(x_2) + \cdots + f_m(x_m).$$

An exponential family distribution is specified for Y (for example normal, binomial or Poisson distributions) along with a link function g (for example the identity or log functions), and the functions f_i may be functions with a specified parametric form (for example a polynomial, or a spline) or may be specified non-parametrically, or semi-parametrically, simply as 'smooth functions', to be estimated by non-parametric means.

The data in `http://mat.uab.cat/ acabana/data/fossil.txt`[7] contains 106 observations of fossil shells from the Atlantic Ocean. For each shell the age (in millions of years) and the ratio of strontium isotopes are registered.



The model $strontium = m(age) + \varepsilon$ for m an unknown smooth function has been adjusted with the function `gam` in the library `mgcv`

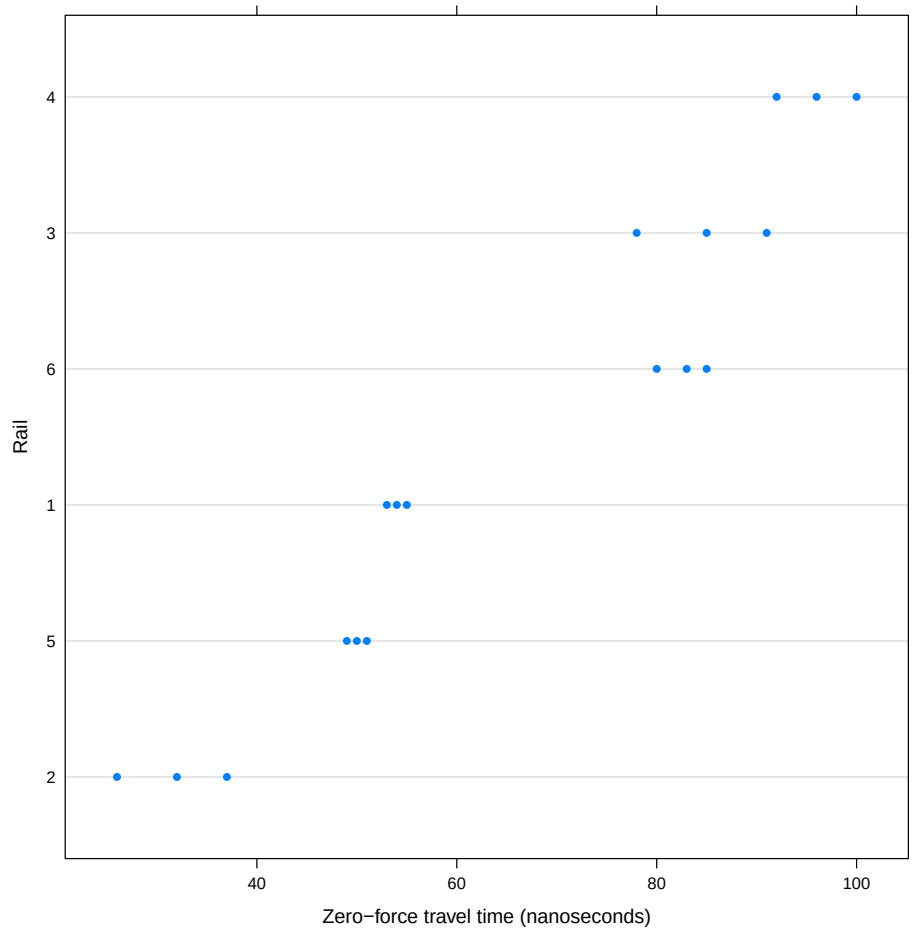
[7] Ruppert, D., Wand, M.P. and Carroll, R.J. (2003) Semiparametric Regression. Cambridge University Press.

Random Effects Models

Six rails were chosen at random and tested three times each by measuring the time it took for a certain type of ultrasonic wave to travel the length of the rail.

The only experimental setting that changes between the observations is the rail.

The engineers were interested in estimating the average travel time for a "typical" rail (the expected travel time), the variation in average travel times among rails (the **between-rail variability**), and the variation in the observed travel times for a single rail (the **within-rail variability**).


Rail
4
3
6
1
5
2
40
60
80
100
Zero−force travel time (nanoseconds)

Data from a one-way classification like these can be analysed either with a

- **fixed-effects model**: inferences about those particular levels of the classification factor that were used in the experiment

  $$y_{i,j} = \beta + \epsilon_{i,j} \quad \text{or} \quad y_{i,j} = \beta_i + \epsilon_{i,j} \quad i = 1, \ldots, M, \quad j = 1, \ldots, n_i$$

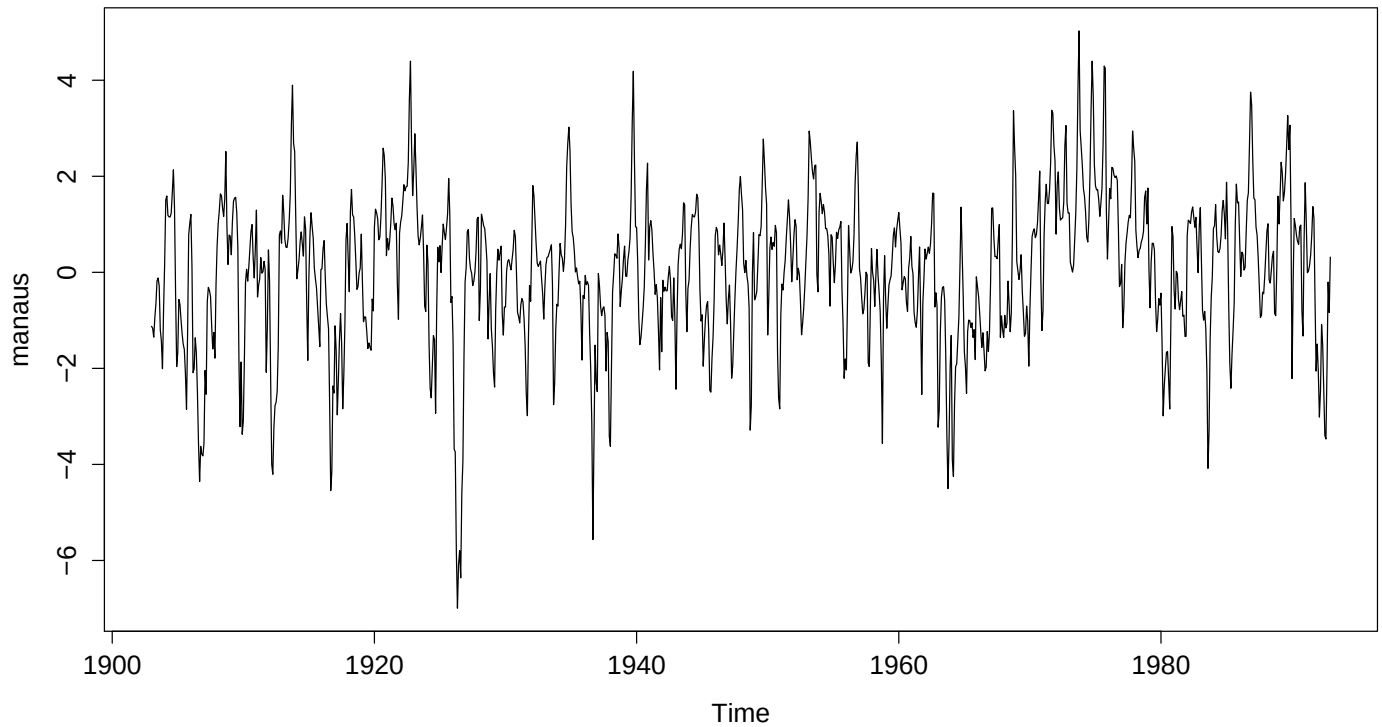
- **random-effects model**: inferences about the population from which these levels were drawn, that usually correspond to different subjects or different plots or different experimental units of some sort.

  $$y_{i,j} = \beta + b_i - + \epsilon_{i,j} \quad i = 1, \ldots, M, \quad j = 1, \ldots, n_i$$

  where b_i are random. If we suppose they are Gaussian, the parameters are β, σ_b^2, σ_ϵ^2.

Time series models

Monthly averages of the daily heights of Rio Negro at Manaus. The data here cover 90 years from January 1903 until December 1992.



Time series models

General paradigm for modelling a time series $\{X_t\}$:

$$X_t = H_t + Y_t$$

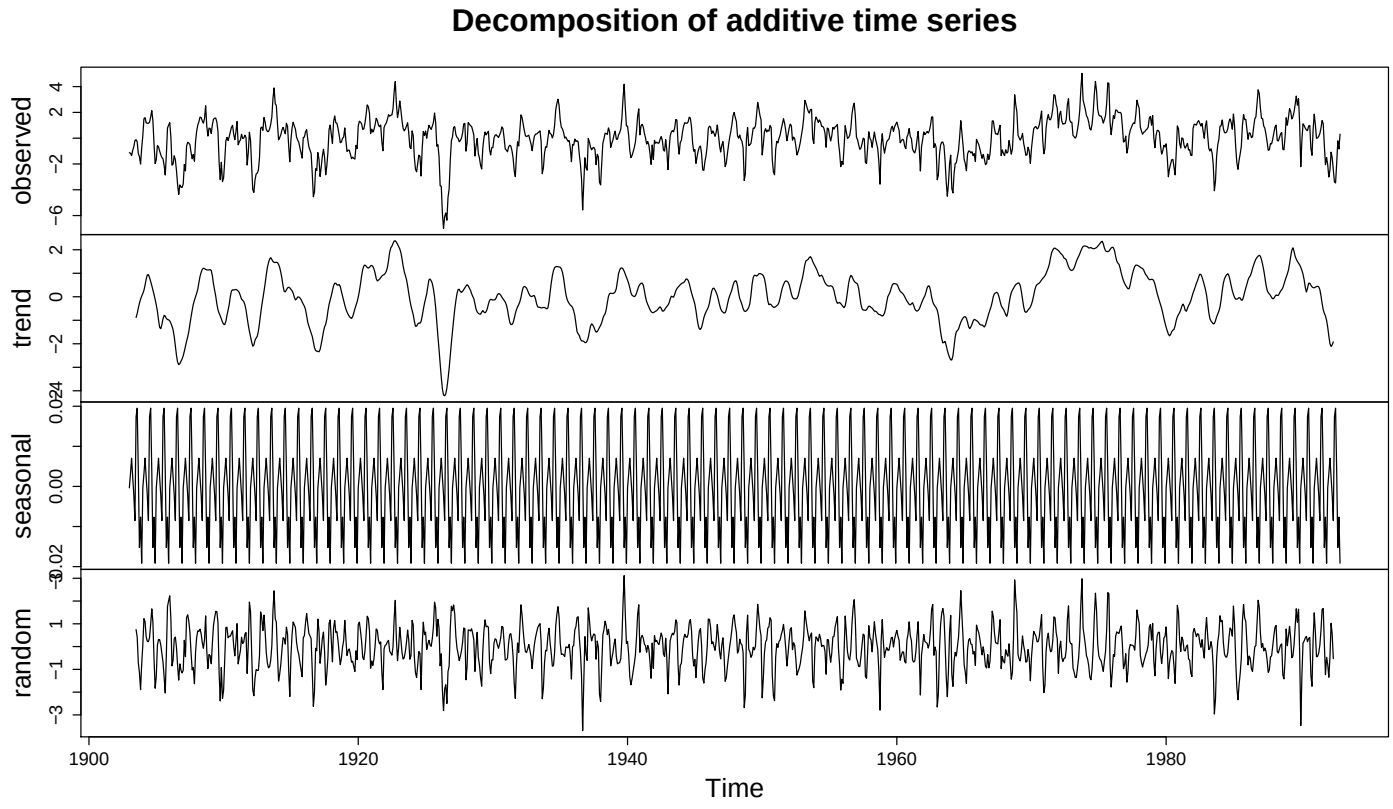
where H_t = deterministic component; Y_t = random noise

H_t decomposes in trend and seasonality

$$H_t = m_t + s_t$$

both estimated by basic methods from Box and Jenkins (1976)
H_t is removed (e.g. by taking successive differences) and we are left with the random component Y_t.
Y_t is assumed to be stationary or weakly stationary for some statistical modelling to be possible.

Time series models



We are interested in modelling the random component.

The basic structure of a model for $\{Y_t\}$

$$Y_t = E(Y_t|F_{t-1}) + a_t$$

where

- F_{t-1} = information set available at time $t-1$
- $E(Y_t|F_{t-1}) = \mu_t =: G(F_{t-1})$ (conditional mean)
- a_t is stochastic shock or innovation, assumed to have zero conditional mean, and hence
- (Conditional variance)
  $\sigma_t^2 = Var(Y_t|F_{t-1}) = E(a_t^2|F_{t-1}) =: H(F_{t-1})$

where G and H are well-defined functions with $H(\cdot) > 0$.

According to the form of $E(Y_t|F_{t-1})$ and $Var(Y_t|F_{t-1})$ as functions of F_{t-1}, i.e., $G(F_{t-1})$ and $H(F_{t-1})$, we have models of different nature for Y_t.

Weak stationarity (at least)

Some statistical property must be assumed or else there is no possibility of modelling at all.

- Stationarity
- Weak stationarity (more likely)

(linear models) Autoregressive models ARMA(p, q)

Consider $\{Y_t\} = \{r_t : t = 1, \ldots, T\}$ data from an observed time series

- $F_{t-1} = \{r_{t-1}, r_{t-2}, \ldots, r_{t-p}\}$ (p lags)
- G a linear function of F_{t-1}, so

$$\mu_t = \phi_1 r_{t-1} + \ldots + \phi_p r_{t-p}$$

- H is constant ($H(F_{t-1}) = \sigma^2$)

We get AR(p) model: $r_t = \phi_1 r_{t-1} + \ldots + \phi_p r_{t-p} + a_t$

This generalises to Auto Regressive and Moving Average of order p, q, ARMA(p, q):

$$r_t = \phi_1 r_{t-1} + \ldots + \phi_p r_{t-p} + a_t + \theta_1 a_{t-1} + \ldots + \theta_q a_{t-q}$$

where $\phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q$ are real numbers,

(linear models) ARIMA(p,d,q)

Given a series of obs. $\{X_t\}$. If:
1. it looks stationary
2. the ACF decreases quickly
Then we try to fit an ARMA to the centred data.

WARNING: Check for unit roots (Test of Dickey-Fuller) since their existence implies the AR model is not stationary

In which case ...
Do some transformation to data to make 1. and 2. hold. This is often attain by differencing $\mapsto$ ARIMA.

$\{X_t\}$ is ARIMA(p,d,q) $\iff$ $(1-B)^d X_t$ is ARMA(p,q)

An ARIMA model is stationary iff $d = 0$.
This is a good model for data with trend.

RLab 2.1 Fitting an ARIMA to EUR/USD.

Nonlinear models

We are concerned with nonlinear models, where G or H are non-linear functions
We look at

- ARCH and GARCH
- Neural Networks (NNet)
- Support Vector Machines (SVM)

A model typology

The quality of being semiparametric or parametric, and the existence of closed form solutions provides two important dimensions to characterize the forecasting models
(McNelis, 2005)

Closed-form solutions	Parametric	Semiparametric
YES	Linear regression	Taylor polynomial
NO	ARCH / GARCH	NNet / SVM

Heteroscedastic volatility (ARCH and GARCH)

These models have been thought for modelling financial returns, hence all the language associated to them contains economic jargon.ñ

Assume that today's asset return variance σ_t^2 depends on the past shocks,

$$a_t = \epsilon_t \sigma_t$$

σ_t^2 variance at time t, $\epsilon_t \sim N(0,1)$ iid random variables. and further, that the shock a_t of asset return is serially uncorrelated, but dependent in the following way: large shocks tend to be followed by another large shocks.
This has been observe in practice , in the form of volatility clusters.

This is modelled by Robert Engle, ARCH(m) model (1982)

Autoregressive Conditional Heteroscedasticity (ARCH)

$$\begin{aligned} r_t &= \mu_t + a_t \\ \sigma_t^2 &= \omega + \alpha_1 a_{t-1}^2 + \ldots + \alpha_p a_{t-p}^2 \end{aligned}$$

and further $a_t = \epsilon_t \sigma_t$, where $\epsilon_t \sim N(0,1)$ iid random vars., $\omega > 0$, $\alpha_i \geq 0$.

To build an ARCH(p) model, we must determine the order p, and estimate ω, $\alpha_1, \ldots, \alpha_p$ using Maximum Likelihood methods.
But DO NOT WORRY! The software R does it for us.

GARCH Model

The *Generalized Autoregressive Conditional Heteroscedasticity* model (Bollerslev, 1986) goes a step further and assumes that current volatility depends on:

- the q recent values for the volatility, and
- the magnitude of the p recent residuals in the data series; that is, $a_{t-i} = r_{t-i} - \mu_{t-i}$; which are also considered to exhibit autocorrelation.

Thus, given a log–return series r_t, let $a_t = r_t - \mu_t$ be the innovation at time t. Then a_t follows a $GARCH(p,q)$ model if:

$$\begin{aligned} a_t &= \sigma_t \epsilon_t \\ \sigma_t^2 &= \omega + \sum_{i=1}^{p} \alpha_i \cdot a_{t-i}^2 + \sum_{j=1}^{q} \beta_j \cdot \sigma_{t-j}^2 \end{aligned}$$

where $\epsilon_t \sim N(0,1)$ iid random variables, $\omega > 0$, $\alpha_i \geq 0$, $\beta_j \geq 0$ and $\sum_{i=1}^{\max(p,q)} (\alpha_i + \beta_i) < 1$

Note that $GARCH(p,0) = ARCH(p)$