3.6. INTRODUCTION TO NON LINEAR REGRESSION

MODEL OF NON LINEAR REGRESSION

Dependent variable: y
Explanatory variable: x
Dependence through a function h of a parameter θ,
Relation affected by a random error e:

$$y = h(x;\ \theta) + e\ .$$

For convenience we suppose that the parameter θ is scalar.
We suppose n observations (x_i, y_i), $i = 1, ..., n$.
Estimation of θ by the least squares method, so as to minimise the function of θ

$$S(\theta) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left[y_i - h(x_i ; \theta) \right]^2 \ .(1)$$

That function is not quadratic in θ, in general.
It is thus often impossible to solve the equation

$$S'(\theta) = 0\ .\ (2)$$

For obtaining the LSE (least square estimator), search the minimum of (1) numerically or to solve (2) numerically (by taking care to check the second order condition in order to obtain a minimum).

METHOD OF MAXIMUM LIKELIHOOD

General information

The likelihood function: density of the vector y of observations considered as a function of the parameters.
Its logarithm is written in the form

$$l(\theta; y) = \sum_{i=1}^{n} l_i(\theta; y).$$

If the observations are independent (the case of the simple random sample), one can write $l_i(\theta; y) = l_i(\theta; y_i)$.

In the case of time series, conditional densities $l_i(\theta; y) = l_i(\theta; y_i | y_{i-1}, \ldots, y_1)$ are used.

In the sequel

- y is omitted from the notation: $l(\theta)$ is the log-likelihood and $l_i(\theta)$, the i-th term.
- we consider a single parameter θ, to simplify the presentation.

The likelihood equation has the form $l'(\theta) = 0$, where $l'(\theta)$ is the score function.

Consistency: under rather general conditions, the MLE (maximum likelihood estimator) converges towards the true value of the parameter, θ_0.

Quasi-maximum likelihood estimator (QMLE)

Convergence remains valid even if the true law is not the one used in the MLE
LSE coincides with QMLE when the likelihood is computed as for a normal distribution

Example, the sample mean $\bar{y}$ is the MLE estimator of the mean *m*, in the case of a normal law, but converges to *m*, even if the true law is different from a normal.
Asymptotic normality: the asymptotic distribution of the estimator $\hat{\theta}$ is normal with mean θ_0 and variance $J^{-1} I J^{-1}$, (sandwich estimator, Huber 1967, White 1982) where

$$J = -E_0(l_i''(\theta)) \qquad I = E_0(l_i'(\theta)^2)$$

and $E_0(.)$ represents the mathematical expectation which respect to the true law, the law with θ_0 as parameter.

Remarks

1. If the true law is compatible with the likelihood function, then $I = J$ and hence, the variance of the asymptotic distribution equals $J^{-1}\ I\ J^{-1} = J^{-1} = I^{-1}$, where

$$I(\theta_0)^{-1} = \left[-E(l''(\theta))\right]^{-1}\Big|_{\theta=\theta_0} = \left[E(l'(\theta)^2)\right]^{-1}\Big|_{\theta=\theta_0},$$

involving minus the likelihood function, at value θ_0, and the second-order derivative of the log-likelihood $l''(\theta)$, evaluated at $\theta = \theta_0$, which is called the Fisher information.
In other words, the asymptotic variance is the inverse of the Fisher information.

2. In the case of a vector parameter, the Fisher information is a symmetric matrix,
Its inverse: the asymptotic covariance matrix,
Standard errors = square roots of the diagonal elements)
+ asymptotic correlations between estimates.

3. Since θ_0 is not known, we cannot determine $I(\theta_0)$.
Hence, we have to use

- the so-called estimated Fisher information $I(\hat{\theta})$ (which requires anyway evaluating a mathematical expectation), or
- the observed Fisher information $-l''(\hat{\theta})$.

Application to heteroscedastic non-linear regression

Context of non-linear regression but we simplify notation: $h(x_i;\theta) = h_i(\theta)$

We suppose the errors are independent realisations of a (non necessarily normal) law with mean 0 and variance $g_i(\theta)\,\sigma^2$

The logarithm of the quasi-likelihood function (with respect to the normal distribution) is written

$$l(\theta,\sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{n}\log\left[g_i(\theta)\sigma^2\right] - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\frac{\left[y_i - h_i(\theta)\right]^2}{g_i(\theta)}.$$

Quasi-likelihood equation with respect to $\sigma^2 \Rightarrow$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\frac{\left[y_i - h_i(\theta)\right]^2}{g_i(\theta)}.$$

Substitution in the likelihood equation $\Rightarrow$ the concentrated quasi-log-likelihood:

$$l(\theta,\hat{\sigma}^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\left(\left(\prod_{i=1}^{n} g_i(\theta)\right)^{1/n}\frac{1}{n}\sum_{i=1}^{n}\frac{\left[y_i - h_i(\theta)\right]^2}{g_i(\theta)}\right) - \frac{n}{2}.$$

Remark. If it is possible to standardise $g_i(\theta)$ such that

$$\prod_{i=1}^{n} g_i(\theta) = 1$$

then maximising the concentrated log-likelihood with respect to θ is equivalent to minimising the weighted sum of square

$$\sum_{i=1}^{n}\frac{\left[y_i - h_i(\theta)\right]^2}{g_i(\theta)}$$

Variance of the asymptotic distribution of the LSE.
If the true law is compatible with the likelihood, hence normal in this case, that variance equals:

$$\left[-E(l''(\theta))\right]^{-1}\Big|_{\theta=\theta_0} = I(\theta_0)^{-1}.$$

Remarks.
1. Simplest case of homoscedastic errors

$$-l''(\theta) = \frac{1}{\sigma^2}\sum_{i=1}^{n}\left[h_i'(\theta)h_i'(\theta) + \left(y_i - h_i(\theta)\right)h_i''(\theta)\right]$$

and the expectation of that random variable equals

$$E\left[-l''(\theta)\right] = \frac{1}{\sigma^2}\sum_{i=1}^{n}\left[h_i'(\theta)h_i'(\theta)\right]$$

since $E(e_i) = 0$ and the coefficient $h_i''(\theta)$ is a non-random.
2. A further expectation with respect to x should be applied if x is random).
3. Be careful: the information is computed that way in some non-linear least-squares program which is not valid in more general frameworks (e.g. with heteroscedastic errors or when a parametric transformation of the variable is used, such as the power or Box-Cox transformation y_i^{λ})

Case of two parameters: $\theta = (\alpha, \beta)$ such that $h_i(\theta) = h_i(\alpha)$ and $g_i(\theta) = g_i(\beta)$. Suppose that

$$E_0\left[e_i^3 / (\sigma^3 g_i^{3/2}(\beta))\right] = A_i \text{ and } E_0\left[e_i^4 / (\sigma^4 g_i^2(\beta))\right] = K_i$$

Considering term *i* of the quasi log-likelihood. The 1st-order derivatives are

$$l_i'(\alpha) = \frac{y_i - h_i(\alpha)}{\sigma^2 g_i(\beta)} h_i'(\alpha)$$

$$l_i'(\beta) = -\frac{1}{2}\frac{g_i'(\beta)}{g_i(\beta)} + \frac{[y_i - h_i(\alpha)]^2}{2\sigma^2}\frac{g_i'(\beta)}{g_i(\beta)}$$

The 2nd-order derivatives are

$$l_i''(\alpha) = \frac{y_i - h_i(\alpha)}{\sigma^2 g_i(\beta)} h_i''(\alpha) - \frac{1}{\sigma^2 g_i(\beta)}\left(h_i'(\alpha)\right)^2$$

$$l_i''(\beta) = \frac{1}{2g_i(\beta)}\left(g_i'(\beta)\right)^2 - \frac{1}{2g_i(\beta)} g_i''(\beta) + \frac{[y_i - h_i(\alpha)]^2}{2\sigma^2}\left[-\frac{1}{2g_i^2(\beta)}\left(g_i'(\beta)\right)^2 + \frac{1}{2g_i^2(\beta)} g_i''(\beta)\right]$$

$$l_i''(\alpha, \beta) = -\frac{y_i - h_i(\alpha)}{\sigma^2 g_i(\beta)} h_i'(\alpha) g_i'(\beta)$$

Hence

$$I(\theta)=\sum_{i=1}^{n}E\begin{bmatrix} l'(\alpha)l'(\alpha) & l'(\alpha)l'(\beta) \\ l'(\beta)l'(\alpha) & l'(\beta)l'(\beta)\end{bmatrix}=\sum_{i=1}^{n}E\begin{bmatrix} \dfrac{(h_i'(\alpha))^2}{\sigma^2 g_i(\beta)} & \dfrac{h_i'(\alpha)g_i'(\beta)}{\sigma\, g_i^{3/2}(\beta)}A_i \\ \dfrac{h_i'(\alpha)g_i'(\beta)}{\sigma\, g_i^{3/2}(\beta)}A_i & \dfrac{1}{4}\left(\dfrac{g_i'(\beta)}{g_i(\beta)}\right)^2(K_i-2+1)\end{bmatrix}$$

Note that the assumptions on the first two moments of e_i implies that

$$J(\theta)=\sum_{i=1}^{n}E\begin{bmatrix} \dfrac{(h_i'(\alpha))^2}{\sigma^2 g_i(\beta)} & 0 \\ 0 & \dfrac{1}{2}\left(\dfrac{g_i'(\beta)}{g_i(\beta)}\right)^2\end{bmatrix}$$

Because of A_i and K_i, $J^{-1}\neq I^{-1}$ so the covariance matrix is obtained by $J^{-1}\ I\ J^{-1}$

Only the element (2,2) of the covariance matrix will be proportional to K_i-1.

The asymptotic independence between the two parameters will depend on A_i.

If the true distribution of the e_i is normal, then $A_i=0$ and $K_i=3$, so that

$$I(\theta)=\sum_{i=1}^{n}E\begin{bmatrix} \dfrac{(h_i'(\alpha))^2}{\sigma^2 g_i(\beta)} & 0 \\ 0 & \dfrac{1}{2}\left(\dfrac{g_i'(\beta)}{g_i(\beta)}\right)^2\end{bmatrix}$$

ITERATIVE METHODS

Algorithm

Stopping criterion

Case of a vector parameter

Gradient method

Gauss-Newton method

Compromise methods

Quasi-Newton methods

Algorithm

1° start from an initial value $\theta^{(0)}$ (well chosen) of parameter θ;
2° use a rule $R(\theta^{(j)})$ for going from $\theta^{(j)}$ to $\theta^{(j+1)}$;
3° stop when a stopping criterion $C_j(\theta^{(j+1)}, \theta^{(j)})$, generally based on the difference $\theta^{(j+1)} - \theta^{(j)}$, the difference $S(\theta^{(j+1)}) - S(\theta^{(j)})$ or on j itself.
The algorithm is as follows

```
j ← 0
Itérer
    θ(j + 1) ← R(θ(j))
sortir si Cj(θ(j + 1), θ(j))
    j ← j + 1
finitérer
θ̂ ← θ(j)
```

Stopping criterion

$C_j(\theta^{(j+1)}, \theta^{(j)})$ can be one of the following:

(1) $\dfrac{\left|\theta^{(j+1)} - \theta^{(j)}\right|}{\max\left\{\left|\theta^{(j)}\right|, \eta_1\right\}} < \varepsilon_1$

(2) $\dfrac{S(\theta^{(j)}) - S(\theta^{(j+1)})}{S(\theta^{(j)})} < \varepsilon_2$

(3) $j \geq N_3$,

where ε_1, η_1, ε_2, N_3 are parameters of the optimisation method, chosen appropriately.

Note that criteria (1) and (2) are relative

In (1), $\varepsilon_1 > 0$ determines the precision of the estimator

and $\eta_1 > 0$ aims at avoiding division by 0

In (2), $\varepsilon_2 > 0$ can detect a badly conditioned problem, where the sum of squares converges while the estimator itself doesn't converge.

In (3), $N_3 > 0$ is the maximum number of iterations

The stopping criterion $C_j(\theta^{(j+1)}, \theta^{(j)})$ can also be a complex criterion based on the simple criteria given above: the algorithm stops at the first occurrence of one of these simple conditions

With a complex criterion, it is not enough to observe that it is reached. We should also take care of the simple criterion which triggers it

Criterion (1) is the most favourable but it is important to note that, in general, *it does not guarantee that a global minimum has been reached*.

Case of a vector parameter

Criterion (2) leaves the impression of model convergence, but not of the parameters
In the case of *k* parameters θ_l, $l = 1, ..., k$, Criterion (1) takes the form

$$\max_{l=1,\ldots k} \frac{\left|\theta_l^{(j+1)} - \theta_l^{(j)}\right|}{\max\left\{\left|\theta_l^{(j)}\right|, \eta_1\right\}} < \varepsilon_1$$

We consider only optimisation procedures for a scalar parameter.
It is clear that for a vector parameter, a direction should first be selected and, in then a point in that direction.
Certain methods concentrate on the choice of the direction, others on the search in a given direction.

Gradient method

or the steepest descent method. It consists to use a Taylor series expansion of $S(\theta)$ to the 1st order:

$$S(\theta) = S(\theta^{(j)}) + (\theta - \theta^{(j)}) S'(\theta^{*(j)})$$

where $\theta^{*(j)}$ is located between θ and $\theta^{(j)}$.
The unknown $\theta^{*(j)}$ is replaced by $\theta^{(j)}$.
Since $S(\theta) < S(\theta^{(j)})$ if and only if $(\theta - \theta^{(j)}) S'(\theta^{(j)}) < 0$, it is proposed to select

$$\theta^{(j+1)} = \theta^{(j)} - k_1 S'(\theta^{(j)}),$$

where $k_1 > 0$ is chosen appropriately.
If $S(\theta^{(j+1)}) > S(\theta^{(j)})$, take a smaller k_1, for example by dividing it by two, and so on until the final $\theta^{(j+1)}$ is obtained.
The stopping criterion is then evaluated.

Note that the gradient method doesn't exploit the particular form of the objective function, a sum of squares.
Except in very special cases, convergence is slow after the first iterations.

Gauss-Newton method

It is a Newton-Raphson method applied on the Jacobian.
It does exploit the particular form of the objective function, a sum of squares.
We start from the model residual evaluated at point θ, $e_i(\theta) = y_i - h(x_i, \theta)$, and more precisely a Taylor series expansion of $e_i(\theta)$ restricted to the 1st order:

$$e_i(\theta) = e_i(\theta^{(j)}) + (\theta - \theta^{(j)})\, e_i'(\theta^{*(j)}),$$

$i = 1, ..., n$, where $\theta^{*(j)}$ is between θ and $\theta^{(j)}$.
Again the unknown $\theta^{*(j)}$ is replaced by $\theta^{(j)}$. We write

$$e_i(\theta^{(j)}) = - (\theta - \theta^{(j)})\, e_i'(\theta^{(j)}) + e_i(\theta)$$

$i = 1, ..., n$.
That equation can be interpreted as a simple linear regression without a constant

- $Y_i = e_i(\theta^{(j)})$ which is known is an observation of the dependent variable,
- where $X_i = e_i'(\theta^{(j)})$, also known, is an observation of the explanatory variable,
- $\beta = - (\theta - \theta^{(j)})$ is the unknown parameter
- $e_i(\theta)$ is the error term.

Minimising the sum of squares gives the LSE

$$-(\theta - \theta^{(j)}) = \beta = \frac{\sum_i X_i Y_i}{\sum_i X_i^2} = \frac{\sum_i e_i'(\theta^{(j)}) e_i(\theta^{(j)})}{\sum_i \left[e_i'(\theta^{(j)})\right]^2}$$

hence

$$\theta^{(j+1)} = \theta^{(j)} - k_2 \left\{ \sum_i \left[e_i'(\theta^{(j)})\right]^2 \right\}^{-1} \sum_i e_i'(\theta^{(j)}) e_i(\theta^{(j)})$$

where $k_2 > 0$ is selected.
If $S(\theta^{(j+1)}) > S(\theta^{(j)})$, a smaller k_2 should be tried, for example by dividing it by 2, and so on until the final $\theta^{(j+1)}$ is obtained.
Convergence of the Gauss-Newton procedure is slow at the beginning because the approximation is not valid but improves when the parameter is close enough from the true value so that linearisation of the objective function is justified.

Remark.

The derivative $e_i'(\theta^{(j)})$ can be evaluated numerically

An approximation can be provide d by the divided difference

$$\frac{e_i(\theta^{(j)} + \Delta\theta) - e_i(\theta^{(j)})}{\Delta\theta}$$

where $\Delta\theta$ is asmall increasing of the parameter.

Compromise methods

- The gradient method is quickly convergent at the beginning and more slowly thereafter
- The Gauss Newton method is slowly convergent at the beginning and more quickly thereafter

This is a justification for a compromise procedure quickly convergent at the beginning (like the gradient method) and also thereafter (like the Gauss-Newton method).
The methods of Marquardt and Lindeberg use a rule like

$$\theta^{(j+1)} = \theta^{(j)} - k_3\left\{\lambda + \sum_i \left[e_i'(\theta^{(j)})\right]^2\right\}^{-1} \sum_i e_i'(\theta^{(j)}) e_i(\theta^{(j)})$$

where the method additional parameter λ is chosen to be large at the beginning and is decreased progressively so as to tend toward 0
That way, by choosing $k_3 > 0$ in an appropriate way, it is the factor

$$\sum_i e_i'(\theta^{(j)}) e_i(\theta^{(j)}) = \frac{1}{2} S'(\theta^{(j)})$$

which determines propagation at the beginning (and direction in the case of a vector parameter θ), like in the gradient method, while during the next iterations, progressively the factor

$$\left\{\sum_i \left[e_i'(\theta^{(j)})\right]^2\right\}^{-1} \sum_i e_i'(\theta^{(j)}) e_i(\theta^{(j)})$$

determines propagation, like in the Gauss-Newton method.

Quasi-Newton methods

Some optimisation methods don't use derivatives at all (like the so-called “simplex” method of Nelder and Mead). Others use second derivatives. A good compromise consists in using first-order derivatives in a clever way. Let us mention one of these methods in the case of a vector parameter θ.

The problem amounts solving the equation $S'(\theta) = 0$.

We start from the Taylor series expansion of $S'(\theta)$ limited to the first order:

$$S'(\theta) = S'(\theta^{(j)}) + (\theta - \theta^{(j)})\, S''(\theta^{*(j)})$$

where $\theta^{*(j)}$, located between θ and $\theta^{(j)}$, is replaced by $\theta^{(j)}$.

We obtain

$$\theta - \theta^{(j)} = - [S''(\theta^{(j)})]^{-1}\, S'(\theta^{(j)})$$

Denote $[H^{(j)}]^{-1} = S''(\theta^{(j)})$, the Hessian matrix, and the gradient $g^{(j)} = S'(\theta^{(j)})$.

Suppose that $S(\theta)$ is a quadratic function of θ, e.g. $S(\theta) = \frac{1}{2}\theta^T A\theta + B\theta + C$.

We have:

$$g^{(j)} = S'(\theta^{(j)}) = A\theta^{(j)} + B, \qquad \left[H^{(j+1)}\right]^{-1} = S''(\theta^{(j+1)}) = A$$

Denote

$$\Delta\theta^{(j)} = \theta^{(j+1)} - \theta^{(j)}$$
$$\Delta g^{(j)} = g^{(j+1)} - g^{(j)}$$
$$\Delta H^{(j)} = H^{(j+1)} - H^{(j)}$$

In the quadratic case:

$$H^{(j+1)}\Delta g^{(j)} = A^{-1}(g^{(j+1)} - g^{(j)}) = A^{-1}(A\theta^{(j+1)} + B - A\theta^{(j)} - B) = \theta^{(j+1)} - \theta^{(j)} = \Delta\theta^{(j)}$$

Methods of type quasi-Newton impose a constraint $H^{(j+1)}\Delta g^{(j)} = \Delta\theta^{(j)}$.

The method of Davidon-Fletcher-Powell (DFP) consists in performing an update of $H^{(j)}$ as follows:

$$H^{(j+1)} = H^{(j)} + \frac{\Delta\theta^{(j)}(\Delta\theta^{(j)})^T}{(\Delta\theta^{(j)})^T\Delta g^{(j)}} - \frac{H^{(j)}\Delta g^{(j)}(\Delta g^{(j)})^T H^{(j)}}{(\Delta g^{(j)})^T H^{(j)}\Delta g^{(j)}}$$

That update adds two matrices which are each a product of a column vector by a row vector, hence rank 1 matrices.
The correction is thus a rank 2 matrix.
Let us check that that expression is compatible with the intention:

$$H^{(j+1)}\Delta g^{(j)} = H^{(j)}\Delta g^{(j)} + \frac{\Delta\theta^{(j)}(\Delta\theta^{(j)})^T\Delta g^{(j)}}{(\Delta\theta^{(j)})^T\Delta g^{(j)}} - \frac{H^{(j)}\Delta g^{(j)}(\Delta g^{(j)})^T H^{(j)}\Delta g^{(j)}}{(\Delta g^{(j)})^T H^{(j)}\Delta g^{(j)}}$$
$$= H^{(j)}\Delta g^{(j)} + \Delta\theta^{(j)} - H^{(j)}\Delta g^{(j)} = \Delta\theta^{(j)}$$

A variant of that method bears the name Broyden-Fletcher-Goldfarb-Shanno (BFGS) and consists in updating not $H^{(j)}$, but well the Hessian matrix itself:

$$\left[H^{(j+1)}\right]^{-1} = \left[H^{(j)}\right]^{-1} - \frac{\left[H^{(j)}\right]^{-1}\Delta\theta^{(j)}(\Delta\theta^{(j)})^T\left[H^{(j)}\right]^{-1}}{(\Delta\theta^{(j)})^T\left[H^{(j)}\right]^{-1}\Delta\theta^{(j)}} + \frac{\Delta g^{(j)}(\Delta g^{(j)})^T}{(\Delta\theta^{(j)})^T\Delta g^{(j)}}$$

Let us check again that that expression is compatible with the intention, by right-multiplying by $\Delta\theta^{(j)}$:

$$\begin{aligned}\left[H^{(j+1)}\right]^{-1}\Delta\theta^{(j)} &= \left[H^{(j)}\right]^{-1}\Delta\theta^{(j)} - \frac{\left[H^{(j)}\right]^{-1}\Delta\theta^{(j)}(\Delta\theta^{(j)})^T\left[H^{(j)}\right]^{-1}\Delta\theta^{(j)}}{(\Delta\theta^{(j)})^T\left[H^{(j)}\right]^{-1}\Delta\theta^{(j)}} + \frac{\Delta g^{(j)}(\Delta g^{(j)})^T\Delta\theta^{(j)}}{(\Delta\theta^{(j)})^T\Delta g^{(j)}} \\ &= \left[H^{(j)}\right]^{-1}\Delta\theta^{(j)} - \left[H^{(j)}\right]^{-1}\Delta\theta^{(j)} + \Delta g^{(j)},\end{aligned}$$

EVALUATION OF DERIVATIVES

We are often lead to replace the gradient S '(θ) by a numerical approximation. An approximation can be given by the divided difference

$$\frac{S(\theta+\Delta\theta)-S(\theta)}{\Delta\theta},$$

where $\Delta\theta$ is a small increment of the parameter, neither too large (to be close to the definition of a derivative where that increment tends to 0), nor too small (to avoid cancellation and round-off error).

For example, we can take $\Delta\theta = \max(\alpha|\theta|, \gamma)$, where α and γ are of order 10^{-6}, for example.

A better approximation of the first derivative is given by the symmetric divided difference

$$\frac{S(\theta+\Delta\theta)-S(\theta-\Delta\theta)}{2\Delta\theta} \approx \frac{S(\theta)+\Delta\theta S'(\theta)-S(\theta)+\Delta\theta S'(\theta)}{2\Delta\theta} = S'(\theta).$$

For the Hessian, a numerical approximation is also used

$$\frac{S(\theta+\Delta\theta)+S(\theta-\Delta\theta)-2S(\theta)}{(\Delta\theta)^2}$$

$$\approx \frac{S(\theta)+\Delta\theta S'(\theta)+(\Delta\theta)^2 S''(\theta)/2+S(\theta)-\Delta\theta S'(\theta)+(\Delta\theta)^2 S''(\theta)/2-2S(\theta)}{(\Delta\theta)^2}$$

$$= \frac{(\Delta\theta)^2 S''(\theta)}{(\Delta\theta)^2} = S''(\theta)$$

A Hessian matrix is symmetric.

Remark.

It can happen that a matrix should theoretically be symmetric but that slight loss of precision are introduced which can be propagated.

For example, don't count on the distributivity rule: $(a + b)c + (a + b)d = a(c + d) + b(c + d)$, because from the numerical point of view, this is not true. It is better to symmetrise a symmetric matrix, either by letting $A_{ji} \leftarrow A_{ij}$, or by letting A_{ij} , $A_{ji} \leftarrow (A_{ij} + A_{ji})/2$.

OTHER METHODS

Optimisation methods with constraints

Methods without derivatives

EM algorithm or "expectation-maximisation"

Optimisation methods with constraints

A simple way to take care of a unique constraint like $g(\theta) > 0$ is to minimise an expression like $S(\theta) + \mu\, I[g(\theta) \leq 0]$ where $\mu > 0$ and the second term, called a penalty, has a positive contribution when the constraint is not satisfied. It is recommended to check that the initial value $\theta^{(0)}$ satisfies the constraint.

Methods without derivatives

Let us mention

- the simplex method of Nelder et Mead (without a relation with the method of the same name in operations research) which is an iterative method
- simulated annealing (Goffe *et al.*, 1994) which allows to find more easily the global optimum, not a local optimum.

The latter algorithm (inspired from thermodynamics) builds initially a rough sketch of the surface while moving with large steps in a random way. When the temperature increases, the step decreases and the algorithm concentrates in the most promising area and finds a local maximum. It escapes from it thanks to random displacements downwards. The temperature is then reduced and the whole procedure is repeated and so on.

EM algorithm or "expectation-maximisation"

It is appropriate in the following context: variable y^* defined by the model is not observable but the result of a transformation $y = t(y^*)$ is observed, where t is a surjective application (for example $t(y^*) = 1$ if $y^* > 0$ and $t(y^*) = 0$ if $y^* \leq 0$).
To simplify the presentation, let us consider missing observations y^* , hence $y^* = (y, y_{\text{miss}})$.
The expression of the log-likelihood is

$$l(\theta; y^*) = \sum_{i=1}^{n} l_i(\theta; y^*).$$

which cannot be evaluated. The EM algorithm will provide a MLE on the basis of observed data. At each iteration j the following two steps using the estimate $\theta^{(j)}$ of θ:
step E: compute the conditional expectation

$$Q(\theta|\theta^{(j)}) = E\left[\sum_{i=1}^{n} l_i(\theta; y^*|y, \theta^{(j)})\right] = \int \sum_{i=1}^{n} l_i(\theta; y, y_{\text{miss}})\, f(y^*|y, \theta = \theta^{(j)}) dy_{\text{miss}}{}^*,$$

where evaluation of the likelihood is done as if the y^* were observed and uses θ and where the conditional distribution of y^* with respect to y depends on $\theta^{(j)}$;
step M: maximise $Q(\theta|\theta^{(j)})$ with respect to θ, giving $\theta^{(j+1)}$.
For an illustration of that algorithm in a simple case, see Navidi (1997).

LIBRARIES

Algorithms need to be selected with care, with a particular attention for initial values
It is often not possible to justify performances of a given method with respect to others

- Libraries NAG and IMSL contain many routines.
- Another well known library exclusively for optimisation is GQOPT (Goldfeld et Quandt, 1982). It includes simulated annealing.
- GAUSS has also a famous optimisation module.

USE OF MATLAB

The following functions of MATLAB are related to this section:

fmin('function', x1, x2, options, p1, p2, ...)	minimum x of a 'function' of one variable, located between x1 and x2, with arguments p1, p2, ... N.B. options(1) ≠ 0: display of intermediate steps (default 0) options(2): stopping criterion on x (default: 1.E-4) options(14): maximum number of iterations (default: 500)
fmins('function', x0, options, p1, p2, ...)	minimum x of a 'function' of several variables, located in a neighbourhood of de x0, with arguments p1, p2, ... (algorithm of Nelder-Mead) N.B. options(1) ≠ 0: display of intermediate steps (default 0) options(2): stopping criterion on x (default: 1.E-4) options(3): stopping criterion on function (default: 1.E-4) options(14): maximum number of iterations (default: 500)

References

W. L. GOFFE, G. D. FERRIER and J. RODGERS, "Global optimization of statistical function with simulated annealing", Journal of Econometrics 12, 65-100, 1994.
S. M. GOLDFELD and R. E. QUANDT, Nonlinear Methods in Econometrics, North-Holland, Amsterdam, 1972.
P. GRIFFITHS and I. D. HILL (editors), «Applied Statistics Algorithms», Ellis Horwood, Chichester, 1985.
C. C. HEYDE, «Quasi-Likelihood and Its Application», Springer-Verlag, 1997.
P. J. HUBER, The behavior of maximum likelihood estimates under non-standard conditions, in Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1, pp. 221-233, 1967.
S. HUET, A. BOUVIER, M.-A. GRUET, E. JOLIVET, «Statistical Tools for Nonlinear Regression», Springer-Verlag, 1996.
W. J. KENNEDY, Jr. and J. E. GENTLE, «Statistical Computing»«», Marcel Dekker, New York, 1980.
W. NAVIDI, "A graphical illustration of the EM algorithm", American Statistician, 51, 29-31, 1997.
W. H. PRESS, B. P. FLANNERY, S. A. TEUKOLSKY, and W. T. VETTERLING, «Numerical Recipes: The Art of Scientific Computing», Cambridge University Press, Cambridge, 1986.
MARQUARDT, D.W., An algorithm for least-squares estimation of nonlinear parameters, 1963.
R. A. THISTED, «Elements of Statistical Computing: Numerical Computation», Chapman and Hall, New York, 1988.
H. WHITE, Maximum likelihood estimation of misspecified models, Econometrica 50, 1-25, 1982.

3.7. RESAMPLING AND THE BOOTSTRAP METHOD

Bootstrap: tirant de botte (Harrap's, New Shorter French and English Dictionary)

INTRODUCTION

- Monte Carlo method: generate artificial pseudo-random samples from an entirely specified distribution
- Bootstrap: similar but from the empirical distribution obtained from a given sample (equivalent to drawing randomly and with replacement a sample in the initial sample)

Example.

correlation coefficient *r* of a sample for 2 variables *X* and *Y*: $\{(x_i, y_i), i = 1, ..., n\}$

Classical theory is valid if the distribution of *X* and *Y* is bivariate normal (with correlation coefficient ρ)

It is shown that $r\sqrt{n-2} / \sqrt{1-r^2} \sim t_{n-2}$ if *X* and *Y* are independent (i.e. $\rho = 0$)

Also $r \sim N(\rho, ((1-\rho^2) / \sqrt{n-3})^2)$ as $n \to \infty$ (but the approximation for small *n* is very bad and depends on ρ and *n*)

Using a Monte Carlo study it is possible to obtain numerically the distribution of *r* but it is necessary to specify completely the joint distribution of *X* and *Y*, including ρ.

DEFINITION OF BOOTSTRAP

Nonparametric bootstrap

Consider a parameter of interest θ and an estimator $\hat{\theta}_n$ of θ obtained from a sample (x_i), i = 1, ..., n. Let us suppose it is obtained by the statistic $\hat{\theta}_n = T_n(x_1,...,x_n)$.
Let $(x_1^*,...,x_n^*)$ a random drawing with replacement in $\{x_1,...,x_n\}$
We compute the bootstrap statistic $\hat{\theta}_n^* = T_n\ (x_1^*,...,x_n^*)$
This is done B times, with bootstrap samples $(x_1^{*b},...,x_n^{*b})$, b = 1, ..., B.
Hence B bootstrap statistics $\hat{\theta}_n^{*b} = T_n\ (x_1^{*b},...,x_n^{*b})$, b = 1, ..., B are computed.
This is nonparametric bootstrap.

Parametric bootstrap

Here it is supposed that the distribution is known except for a vector of parameters θ.
The sample allows to obtain the estimator $\hat{\theta}_n$.
Parametric bootstrap consists in drawing random samples of the same size *n* in the completely specified distribution obtained by using the estimator $\hat{\theta}_n$ (taken from the initial sample) instead of the true parameter θ.

Example.

Consider again the correlation coefficient r of a sample from 2 variables X and Y : $\{(x_i, y_i), i = 1, ..., n\}$.

Suppose that the distribution of X and Y is bivariate normal

We estimate the means and the covariance matrix (hence the correlation coefficient) using the sample and perform a Monte Carlo simulation from that distribution: this is parametric bootstrap

On the contrary, for nonparametric bootstrap, we simply take simple random samples from the initial sample.

By using n = 15 and B = 3200, Efron and Tibshirani (1993, pp. 49-50) have obtained a standard error of 0,124 for parametric bootstrap, 0,131 for nonparametric bootstrap, while the asymptotic standard error is 0,115.

Taken from Efron et Tibshirani [1993], p.50.

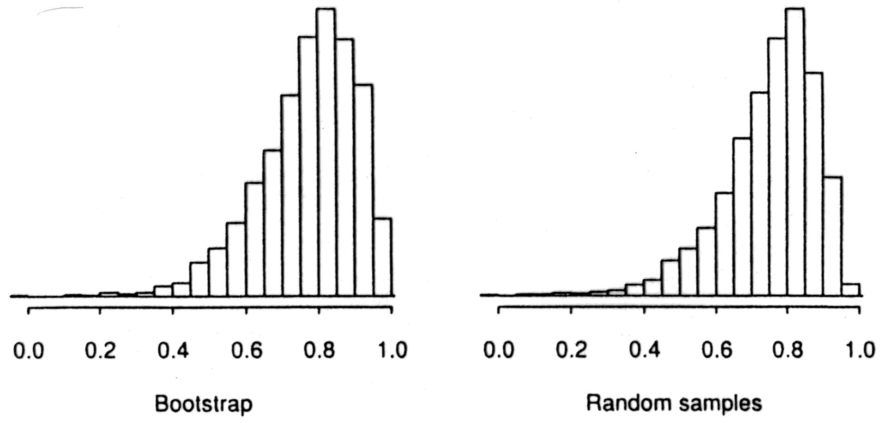

Figure 6.2. *Left panel: histogram of 3200 bootstrap replications of* $\widehat{corr}(\mathbf{x}^*)$, *from the law school data,* $n = 15$, *Table 3.1. Right panel: histogram of 3200 replications* $\widehat{corr}(\mathbf{x})$, *where* $\mathbf{x}$ *is a random sample of size* n *from the* $N = 82$ *points in the law school population, Table 3.2. The bootstrap histogram strongly resembles the population histogram. Both are notably non-normal.*

Nonparametric bootstrap estimation is less influenced by lack of validity of assumptions (here bivariate normality).
Of course the method used here is not very robust (a robust estimation or a nonparametric estimation, e.g. a rank statistic, can be used instead).

Remark: Jackknife
A method related to and older than the bootstrap is the *jackknife*.
A jackknife sample is of size $n - 1$, not n, where one of the observations of the sample is randomly removed ("leave-one-out")
It can be shown that, in general, the jackknife make less use of the information and is therefore less efficient than the bootstrap.
In some cases it is even not consistent.

Example.

Consider the median of a sample. Because of a discontinuity in the definition of the median, the jackknife medians will take 3 values. Efron and Tibshirani (1993, pp. 148) show an example with $n = 16$ where the bootstrap standard error for $B = 100$ equals 9.58 while the jackknife standard error equals 6.68. This is a case where the jackknife standard error doesn't converge to the true standard error

JUSTIFICATION OF BOOTSTRAP

The nonparametric bootstrap is justified if and only if, conditionally to the initial sample, $\hat{\theta}_n^* - \hat{\theta}_n$ has the same asymptotic distribution than $\hat{\theta}_n - \theta$ when $n \to \infty$.

The parametric bootstrap is justified if and only if $\hat{\theta}_n^* - \theta$ has the same asymptotic distribution than $\hat{\theta}_n - \theta$.

Some results of that kind can be obtained.

For finite samples, it is necessary to perform Monte Carlo experiments.

The bootstrap will be less justified when information is available about the distribution $F(x)$ which is different from that contained in the empirical distribution function of $\hat{F}_n(x)$, e.g. when we can suppose that $F(x)$ belongs to a parametric family (normal law, for example) or when there is an explanatory variable (the case of regression, see later).

Example.

Consider a uniform law on (0, θ) and the MLE $\hat{\theta}_n = x_{(n)}$.

It appears that the nonparametric bootstrap method doesn't work very well (for example the probability that $\hat{\theta}_n^* = \hat{\theta}_n$ is close to 0,63). The reason is that the empirical distribution is not a good approximation of the true law in the queues.

OBJECTIVES OF BOOTSTRAP

The empirical distribution of the $\hat{\theta}_n^{*b}$, $b = 1, ..., B$ is called the bootstrap distribution.

We can be interested by the mean of that distribution $\overline{\hat{\theta}}_n^{*B}$.

The difference $\hat{\theta}_n - \overline{\hat{\theta}}_n^{*B}$ is called the bootstrap bias. There are other methods for estimating the bias.

The standard error of the bootstrap distribution is called the bootstrap standard error.

For determining the bias or the standard deviation, about 200 bootstrap samples are used, but sometimes as few as 20.
We can also be interested by a bootstrap confidence interval, which can be obtained using a normal approximation with the bootstrap standard error, or can be based on the quantiles of the bootstrap distribution.
Here the number of bootstrap samples should be much larger than 200.
There exist methods for building better confidence intervals: the BC_a method ("bias corrected and accelerated"), and the ABC ("approximate bootstrap confidence") method, see Efron and Tibshirani (1993, chapter 14).
The bootstrap method can also be used to perform a test of hypothesis (see later)
In a certain sense it is close to the permutation tests (Edgington, 1987; Good, 1994) although the interpretation is different and the latter cannot always be built (it is required to have something that can be permuted under the null hypothesis!).

Example.

Consider the test of hypothesis H_0 of equality of 2 distribution functions F_1 and F_2 from random samples, respectively, $(x_1^{(1)},\ldots,x_{n_1}^{(1)})$ and $(x_1^{(2)},\ldots,x_{n_2}^{(2)})$, with respective sizes n_1 and n_2, with the alternative hypothesis H_1 that $F_1(x) < F_2(x)$, for all x, i.e. that the variable has a tendency to be larger in F_1 than in F_2.

We select a test statistic $\hat{\theta}$, for example the difference of the means $\bar{x}_1 - \bar{x}_2$ (the Student statistic can be considered instead.

Under the alternative hypothesis H_1, we expect that the difference $\bar{x}_1 - \bar{x}_2$ is large.

In classical statistics, we rely on a parametric model (normal laws with the same standard deviations) or an approximation based on the central limit theorem in order to obtain the distribution of $\bar{x}_1 - \bar{x}_2$ which allows to determine the significance probability $P_{H_0}(\hat{\theta} \geq \bar{x}_1 - \bar{x}_2)$.

The approach of permutation tests is different: inference is conditional to the order statistics, i.e. the observed values, in the united sample; denote $n = n_1 + n_2$, $x_i = x_i^{(1)}, i = 1,\ldots,n_1$, $x_{i+n_1} = x_i^{(2)}, i = 1,\ldots,n_2$.

Under H_0 and conditionally to the order statistics, there are $n!/n_1!n_2!$ possible configurations, each one corresponding to a permutation of n_1 "1" and n_2 "2".

Each one has the same probability $1/(n!/n_1!n_2!)$. Each one gives a value for the statistic $\hat{\theta}$. Suppose we have chosen the probability level α.

The critical region of the test will be composed of the $\alpha/(n!/n_1!n_2!)$ values of $\hat{\theta}$ which are the largest ones.

The permutation test is however restricted to specific tests of hypothesis and is difficult to generalise for example to test simultaneously equality of the means and of the variances.

GENERALISATION OF BOOTSTRAP

Bootstrap with stabilisation of variance

In some cases, it is better to use that variant, particularly when the standard error depends on the parameter under study.

Example.

Consider again the correlation coefficient *r* of a sample for 2 variables *X* and *Y*, supposing a bivariate normal law.

Recall that *r* has asymptotically a $N(\rho, ((1-\rho^2)/\sqrt{n-3})^2)$ distribution when $n \to \infty$.

In classical statistics it is recommended to transform the correlation coefficient using the hyperbolic tangent transformation: $\zeta = \frac{1}{2}\log\frac{1+\rho}{1-\rho}$ such that, denoting $z = \frac{1}{2}\log\frac{1+r}{1-r}$, the distribution of *z* is asymptotically a $N(\zeta, 1/(n-3))$ distribution when $n \to \infty$, which no longer depends on ρ.

Application to problems of comparison

Consider for example the comparison of 2 means for *independent* sample (that condition is compulsory!). We take separately (hence independently) a random sample (with replacement) in each sample.

Example.

Like above, consider the test of hypothesis H_0 of equality of 2 distribution functions F_1 and F_2 using the respective random samples $(x_1^{(1)},\ldots,x_{n_1}^{(1)})$ and $(x_1^{(2)},\ldots,x_{n_2}^{(2)})$, with respective sizes n_1 and n_2, with the alternative hypothesis H_1 that $F_1(x) < F_2(x)$, for all x, again the variable has a tendency to be larger in F_1 than in F_2.

We can take as a test statistic $\hat{\theta}$, for example the difference of the means $\bar{x}_1 - \bar{x}_2$ (or the Student statistics but with possibly different results).

We estimate the common distribution function from the two united samples.

The bootstrap differs from permutation tests in the sense that we take samples with replacement (whereas permutations are drawings without replacement).

Nevertheless, it is not surprising that B bootstrap samples will give results close to those of a random subset of B permutations.

We obtain therefore an approximation of the significance probability of the permutation test.

In the example, the test is one-sided and a large value of the statistic has a low probability under the null hypothesis. The significance probability equals $\#\{\hat{\theta}_n^{*b} > \hat{\theta}_n\}/B$.

For a two-sided test of the hypothesis $\theta = 0$, we should evaluate $\#\{|\hat{\theta}_n^{*b}| > |\hat{\theta}_n|\}/B$.

The bootstrap method can be adapted to different hypotheses, like the test of equality of the means when the variances are not equal.

Application in regression

To simplify, let us consider simple linear regression of a variable *Y* in function of a variable *X*. We have already shown the treatment of the correlation (sampling pairs $\{(x_i, y_i), i = 1, ..., n\}$). A second approach consists in keeping the x_i fixed and sampling in the distribution of résiduals:

- after having determined an estimation of the regression line, we compute the *n* residuals, let $\{e_i, i = 1, ..., n\}$;
- we draw a sample with replacement $(e_1^*, ..., e_n^*)$;
- we determine the corresponding $(y_1^*, ..., y_n^*)$, by using again the regression line estimated on basis of the initial sample;
- for each bootstrap sample, we compute the statistic.

That approach works well when, as is often the case, inference is done conditionally to the values taken by the explanatory variable(s). The first approach is however less sensitive to a bad model specification (for example, if linearity of the regression can be challenged).

Application in time series analysis

It is no longer possible to draw a random sample because it would break the chronological sequence of the data. Two approaches are possible according to a parametric model is considered or not.

If a parametric model can be considered, for example an autoregressive model of order 1, we can proceed as in the second approach above (regression):

- determine the residuals for the fitted model estimated over the initial sample;
- perform resampling on the residuals;
- generate the corresponding data using these resampled residuals
- estimate the parameters of the model for each bootstrap sample

If a parametric model cannot be considered, and provided that the process be *m*-dependent, i.e. variables distant from each others from more than *m* unit time intervals are independent, we can consider all the possible blocks of length *m* and perform resampling in these blocks.

For example, if $n = m.k$, we can take k blocks with replacement in order to constitute each bootstrap sample. It is the "moving blocks") methods (e.g. Li and Maddala (1996)).

USE OF MATLAB

The functions of MATLAB necessary for the bootstrap in the univariate case have already been seen:

```
xboot = x(floor(rand(n,1)*n) + 1)
```

References

E. S. EDGINGTON, "Randomization Tests", Dekker, New York, 2nd Ed., 1987.
B. EFRON, "The Jackknife, the Bootstrap and other Resampling Plans ", CBMS-NSF Regional Conference Series in Applied Mathematics, SIAM, 1982.
B. EFRON and R. J. TIBSHIRANI, "An Introduction to the Bootstrap", Chapman & Hall, New York, 1993.
P. GOOD, "Permutation Tests", Springer-Verlag, New York, 1994.
P. HALL, «The Bootstrap and Edgeworth Expansion», Springer-Verlag, 1997.
W. HÄRDLE, «Smoothing Techniques», Springer-Verlag, 1991.
H. LI and G. S. MADDALA, "Bootstrapping time series models", Econometric Reviews, 15, 115-158, 1996.
R. A. THISTED, «Elements of Statistical Computing: Numerical Computation», Chapman and Hall, New York, 1988.

3.8. COMPLEMENTS

Two real studies are treated:

one with Fortran:

8.1 RANK TESTS FOR TIME SERIES

the other one with MATLAB:

8.2 A MONTE CARLO ANALYSIS FOR LINEAR REGRESSION

3.8.1 RANK TESTS FOR TIME SERIES

Presentation of the problem

The paper (Hallin and Mélard [1988]) is concerned with the properties of several tests of randomness of a time series. These tests are based on ranks and take the form of rank autocorrelations:
VdW: Van der Waerden, which is optimal if the true law is normal
W: Wilcoxon, which is optimal if the true law is logistic
L: Laplace, which is optimal if the true law is double exponential
We can add the Spearman rank autocorrelation which has no optimality property but which is simpler:

$$r_S = \frac{\frac{1}{n-1}\sum_{t=2}^{n} R_t R_{t-1} - m_S}{\sigma_S},$$

where R_t is the rank of observation X_t in the sequence of n observations (X_1, ..., X_t).
A similar study has been conducted by Hallin *et al.* [1990] concerning signed rank tests.
The purpose of the paper is to study the four tests and compare them to the test based on the ordinary autocorrelation

$$r_1 = \frac{\frac{1}{n-1}\sum_{t=2}^{n} X_t X_{t-1} - m_1}{\sigma_1},$$

(and three more complex variants called Ljung-Box, Moran et Dufour-Roy).

Short description of the paper

1. Introduction

2. Definitions.

Note that the variance is given by an expression (which is not detailed in the paper):

$$(\sigma_f^{(n)})^2 = [n(n-1)]^{-1} \sum\sum_{1 \le i_1 \ne i_2 \le n} \left[\phi\left(F^{-1}\left(\frac{i_1}{n+1}\right)\right) F^{-1}\left(\frac{i_2}{n+1}\right)\right]^2$$

$$+2\left[n(n-1)^2\right]^{-1} \sum\sum\sum_{1 \le i_1 \ne i_2 \ne i_3 \le n} \phi\left(F^{-1}\left(\frac{i_1}{n+1}\right)\right) \phi\left(F^{-1}\left(\frac{i_2}{n+1}\right)\right) F^{-1}\left(\frac{i_2}{n+1}\right) F^{-1}\left(\frac{i_3}{n+1}\right)$$

$$+\left[n(n-1)^2\right]^{-1} \sum\sum\sum\sum_{1 \le i_1 \ne i_2 \ne i_3 \ne i_4 \le n} \phi\left(F^{-1}\left(\frac{i_1}{n+1}\right)\right) \phi\left(F^{-1}\left(\frac{i_2}{n+1}\right)\right) F^{-1}\left(\frac{i_3}{n+1}\right) F^{-1}\left(\frac{i_4}{n+1}\right)$$

$$-(n-1)(m_f^{(n)})^2.$$

where F denotes the distribution function and $\phi = -f'/f$, where f is the density.
Although terms with triple and quadruple sums are asymptotically negligible, we need to keep them for finite time series. We can reduce double, triple and quadruple sums with restrictions by simple sums. For example:

$$\sum\sum_{1 \le i_1 \ne i_2 \le n} \left[\phi\left(F^{-1}\left(\frac{i_1}{n+1}\right)\right) F^{-1}\left(\frac{i_2}{n+1}\right)\right]^2$$

$$= \sum\sum_{1 \le i_1, i_2 \le n} \left[\phi\left(F^{-1}\left(\frac{i_1}{n+1}\right)\right) F^{-1}\left(\frac{i_2}{n+1}\right)\right]^2 - \sum\sum_{1 \le i_1 = i_2 \le n} \left[\phi\left(F^{-1}\left(\frac{i_1}{n+1}\right)\right) F^{-1}\left(\frac{i_2}{n+1}\right)\right]^2$$

$$= \sum_{1 \le i_1 \le n} \left[\phi\left(F^{-1}\left(\frac{i_1}{n+1}\right)\right)\right]^2 \sum_{1 \le i_2 \le n} \left[F^{-1}\left(\frac{i_2}{n+1}\right)\right]^2 - \sum_{1 \le i_1 \le n} \left[\phi\left(F^{-1}\left(\frac{i_1}{n+1}\right)\right) F^{-1}\left(\frac{i_1}{n+1}\right)\right]^2.$$

and similarly for the triple and quadruple sums.
We are lead to computing one and for all the n couples:

$$Y_i = F^{-1}\left(\frac{i}{n+1}\right)$$

$$Z_i = \phi\left(F^{-1}\left(\frac{i}{n+1}\right)\right) = \phi(Y_i)$$

and evaluate sums and sums of products.
The reduction in terms of number of operations is tremendous.

3. Distribution under the null hypothesis in the case of small samples

Are evaluated as a function of n:

- the mean and the variance
- the full (discrete) distribution (for very short series)
- the discrete critical values (up to $n = 10$)
- the approximate critical values, using a random sample of size 100000 among the $n!$ permutations up to $n = 25$
- approximate critical values obtained using a beta distribution
- approximate critical values obtained using a normal distribution

4. Study of the power of the tests

The alternative hypothesis is the autoregressive process: $X_t = \theta X_{t-1} + \varepsilon_t$, where the ε_t are independent random variables with density distribution *f*
The study is done by Monte Carlo simulation (using ANSECH now in TSE, see Mélard and Pasteels [1994] to generate the series) over 5000 replications
The same sequences, one for *n* = 20 and one for *n* = 100, of pseudo-random numbers are used in all cases (as a reduction of variance device)
Three kinds of distribution are used: normal, logistic, double exponential
Several values of θ are used: 0.5, 0.25, 0.125, 0.0625
The rank autocorrelations are computed for the 5000 replications and the results are output in SPSS format
The ordinary autocorrelation is used also (3 different approximations)
The output is handled by SPSS
According to the distribution, the size *n* and the value of θ, the frequency of rejections are different
Questions:

- what is the behaviour of the rank tests with respect to the classical tests
- does the theoretically optimal rank test (e.g. van der Waerden in the normal case) beat the other rank tests (Wilcoxon, Laplace)

5. Examples

5.1 An artificial series is produced by a 1st-order autoregressive process (in principle positively autocorrelated) which is perturbed at two points in order to reduce autocorrelation
5.2 A series from the literature (Bartels)
5.3 Another series from the literature (Anderson)
5.4 A series with seasonality and outliers
5.5 A series which is not a time series (at the request of a referee).

Programs

Four Fortran 77 program files are enclosed:
1° GNCORR.FOR: computes rank and ordinary autocorrelations; an example of output on a data set is given (GNCORR.OUT); this is the program used for handling the examples and studying the power. It is now included in Time Series Expert (Mélard and Pasteels [1994]).
GNCORR.FOR
DRGNC.OUT
2° TABEX.FOR: main program for studying the exact distribution of the rank statistics under the null hypothesis by considering ***all*** the permutations of numbers 1 to *n*; an example of output is given (untitled)
TABEX.FOR
TABEX.OUT
3° RNKMV.FOR: prepares the treatments for TABEX.FOR (called once for each *n*) and GNCORR.FOR;
RNKMV.FOR
4° TABQTL.FOR: computes the quantiles of the exact distribution exacte and of normal and beta approximations for TABEX.FOR.
TABQTL.FOR
The paper has used another main program (not given), TABSIM.FOR, for studying the distribution under the null hypothesis by taking random samples of permutations. TABSIM makes use of RNKMV.FOR and TABQTL.FOR.

Some delicate points

- Reduction of multiples sums (all the permutations, random samples of permutations)
- Exact distribution: large number of different values => condensation
- Attention: rank autocorrelations go outside of the [-1; 1] interval
- Realisation of the *n*! permutations (the counter exceeded the range of integer variables)
- Verification of means and variances
- Determination of quantiles without sorting
- Generation of random permutations
- Number of replications of simulations
- Implementation of the design of experiments
- Generation of series of random errors (normal, logistic, double exponential) with variance reduction
- Generation of autoregressive series
- Critical values and number of rejections
- Analysis of results by several methods
- For signed rank tests: take signs into account

Remark. The study has been done on the Control Data Cyber computer of the Centre de Calcul ULB-VUB, in Fortran 66 and with the aid of NAG library. Most computations have been done during the night. Later, the programs have been converted to Fortran 77 and then ported to PC's by replacing the modules of the NAG library.

References

M. HALLIN and G. MELARD, Rank-based tests for randomness against first-order serial dependence, Journal of the American Statistical Association 83, 1988, 1117-1128.
M. HALLIN, A. LAFORET and G. MELARD, Distribution-free tests against serial dependence: signed or unsigned ranks ?", Journal of Statistical Planning and Inference, 24, 1990, 151-165.
G. MELARD et J.-M. PASTEELS, Manuel d'utilisateur de Time Series Expert (TSE version 2.2), Institut de Statistique, Université Libre de Bruxelles, Bruxelles.

3.8.2 A MONTE CARLO ANALYSIS FOR LINEAR REGRESSION

Object

Method

Object

What is the real influence of the distribution of the errors on the distribution of the Student statistic for the regression coefficient of a simple regression fitted by the least squares method where the explanatory variable (X) represents time.
9 cases are considered with each time at most one of the conditions which is not satisfied:

1. none (assumptions of the general linear model)
2. explanatory variable measured with errors
3. multicolinearity
4. nonlinear relation
5. non normality (4 other distributions than the normal, with the same scatter (measure by the interquartile range equal to 1,348): uniform, Laplace, Student with à 2 degrees of freedom, Cauchy)
6. outliers
7. heteroscedasticity
8. positive autocorrelation of errors
9. negative autocorrelation of errors

Method

5000 (for example) simulations of a simple linear regression on $T = 100$ (for example) observations
Each time, 6 graphs are plotted:
(1) the data of the first simulation with the true relation and the fitted straight line;
(2) the residuals of the first simulation as a function of the fitted values, using the model fitted by the LS method, with 95% confidence intervals based on a normal distribution ;
(3) the residuals of the first simulation as a function of time, with 95% confidence intervals based on a normal distribution ;
(4) the empirical distribution function for date n° 110 (for example), compared with the normal law
(5) the empirical distribution function of the Student statistic, compared to the Student distribution with $T - 2$ degrees of freedom (*on order to examine the validity of the regression test*);
(6) the empirical distribution function of the residual at time 110, compared o the normal distribution (*in order to examine the validity of forecast intervals*).

MATLAB Program: CONDREG.M

Condreg.htm

Results