Asymptotic distribution of least square estimators for linear models with dependent errors : regular designs

Emmanuel Caron
Directeurs: Bertrand Michel (LMJL), Jérôme Dedecker (MAP5)

Ecole Centrale Nantes, Laboratoire de Mathématiques Jean Leray

4 Décembre 2017

Sommaire

Sommaire

Linear Regression Model

Linear Regression model:

$$Y = X\beta + \epsilon,$$

- X is a fixed design, $[n \times p]$
- Y is a n random vector
- β is a p vector of unknown parameters
- ϵ are the errors, $\epsilon \in \mathbb{R}^n$.

Usual assumptions in the i.i.d. case:

- the errors are i.i.d. ($\epsilon \sim \mathcal{N}(0, \sigma^2 I_n)$)
- $\mathbb{E}(\epsilon) = 0$ and $Var(\epsilon) = \sigma^2 I_n$

Least Square Estimator

$$\hat{\beta} = argmin_{\beta \in \mathbb{R}^p} \|Y - X\beta\|_2^2 = (X^t X)^{-1} X^t Y.$$

$\hat{Y} = X\hat{\beta}$: Orthogonal Projection of Y on $\mathcal{M}_X = Vect\{X_{.,1}, ..., X_{.,p}\}$

- $\mathbb{E}(\hat{\beta}) = \beta$ and $Cov(\hat{\beta}) = \sigma^2 (X^t X)^{-1}$
- Residual vector: $\hat{\epsilon} = Y - \hat{Y} = Y - X\hat{\beta} \in \mathcal{M}_X^{\perp}$
- $\hat{\sigma}^2 = \frac{\|\hat{\epsilon}\|_2^2}{n-p}$.

Distribution of the LSE:

- Gaussian Case: $\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2 (X^t X)^{-1})$
- Non-Gaussian Case: $D(n)(\hat{\beta} - \beta) \xrightarrow[n \to \infty]{\mathcal{L}} \mathcal{N}(0, \sigma^2 Q^{-1})$.

Goals and Plan

Main Goal : Remove the independence hypothesis and find results similar to the i.i.d. case.

Plan :

1. Hannan's Theorem (1973) [4]: convergence of the LSE in the stationary case under very mild conditions
2. Show that for a large class of designs, the asymptotic covariance matrix is as simple as the i.i.d. case
3. Estimation of the covariance matrix
4. Applications with Fisher's tests.

Stationarity

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. $(\epsilon_i)_{i\in\mathbb{Z}}$ is an error process defined on $(\Omega, \mathcal{F}, \mathbb{P})$, supposed strictly stationary, with zero mean, and $\epsilon_0 \in \mathbb{L}^2$.

Definition : Strict Stationarity

A stochastic process $(\epsilon_i)_{i\in\mathbb{Z}}$ is said to be strictly stationary if the joint distributions of $(\epsilon_{t_1}, \ldots, \epsilon_{t_k})$ and $(\epsilon_{t_1+h}, \ldots, \epsilon_{t_k+h})$ are the same for all positive integers k and for all $t_1, \ldots, t_k, h \in \mathbb{Z}$.

Let $(\mathcal{F}_i)_{i\in\mathbb{Z}}$ be a non-decreasing filtration on $(\Omega, \mathcal{F}, \mathbb{P})$
(Example: $\mathcal{F}_i = \sigma(\epsilon_k, k \leq i)$).

We always suppose that $\mathcal{F}_{-\infty} = \bigcap_{i\in\mathbb{Z}} \mathcal{F}_i$ is trivial, and ϵ_0 $\mathcal{F}_\infty$-measurable.

Spectral density

Autocovariance function:

$$\gamma(k) = \mathrm{Cov}(\epsilon_m, \epsilon_{m+k}) = \mathbb{E}(\epsilon_m \epsilon_{m+k}).$$

Let f be the associated spectral density, $\lambda \in [-\pi, \pi]$:

$$\gamma(k) = \int_{-\pi}^{\pi} e^{ik\lambda} f(\lambda) d\lambda,$$

$$f(\lambda) = \frac{1}{2\pi} \sum_{k=-\infty}^{\infty} \gamma(k) e^{-ik\lambda}.$$

Sommaire

Hannan's condition on the error process

Stationary case: Hannan (1973) $\rightarrow$ Central Limit Theorem for the usual LSE $\hat{\beta}$, under very mild conditions.

- $\forall j \in \mathbb{Z}$ and $\forall Z \in \mathbb{L}^2(\Omega)$: $P_j(Z) = \mathbb{E}(Z|\mathcal{F}_j) - \mathbb{E}(Z|\mathcal{F}_{j-1})$.
- **Hannan's condition** on the error process:

$$\sum_{i \in \mathbb{Z}} \|P_0(\epsilon_i)\|_{\mathbb{L}^2} < +\infty.$$

  Notice that implies: $\sum_k |\gamma(k)| < +\infty$.

- Examples which verify Hannan's condition:
  - Linear Processes, functions of linear processes (Dedecker, Merlevède, Vólny (2007) [2])
  - Conditions à la Gordin (DMV (2007))
  - Framework of Wu (2005 [7])
  - Weakly dependent sequences

Hannan's conditions on the design

- Let $X_{.,j}$ be the column j of the matrix X, $j \in \{1, \dots, p\}$:

$$d_j(n) = \|X_{.,j}\|_2 = \sqrt{\sum_{i=1}^{n} x_{i,j}^2},$$

and let $D(n)$ be the diagonal matrix with diagonal term $d_j(n)$.

- **Conditions on the design**:

$$\forall j \in \{1, \dots, p\}, \qquad \lim_{n \to \infty} d_j(n) = \infty,$$

$$\forall j \in \{1, \dots, p\}, \qquad \lim_{n \to \infty} \frac{\sup_{1 \le i \le n} |x_{i,j}|}{d_j(n)} = 0,$$

and the following limits exist:

$$\forall j, l \in \{1, \dots, p\}, \qquad \rho_{j,l}(k) = \lim_{n \to \infty} \sum_{m=1}^{n-k} \frac{x_{m,j} x_{m+k,l}}{d_j(n) d_l(n)}.$$

We define the $p \times p$ matrix $R(k)$:

$$R(k) = [\rho_{j,l}(k)] = \int_{-\pi}^{\pi} e^{ik\lambda} F_X(d\lambda),$$

where F_X is the spectral measure associated with the matrix $R(k)$.

Moreover, we suppose:

$$R(0) > 0.$$

Theorem (Hannan [4])

Under the previous conditions:

$$D(n)(\hat{\beta} - \beta) \xrightarrow[n \to \infty]{\mathcal{L}} \mathcal{N}(0, F^{-1}GF^{-1}),$$

and we have the convergence of second order moment:

$$\mathbb{E}\left(D(n)(\hat{\beta} - \beta)(\hat{\beta} - \beta)^t D(n)^t\right) \xrightarrow[n \to \infty]{} F^{-1}GF^{-1},$$

with F and G the matrices:

$$F = \frac{1}{2\pi} \int_{-\pi}^{\pi} F_X(d\lambda),$$

$$G = \frac{1}{2\pi} \int_{-\pi}^{\pi} F_X(d\lambda) \otimes f(\lambda).$$

Sommaire

Regular design

Definition (Regular design)

A fixed design X is called regular if, for any j, l in $\{1, \ldots, p\}$, the coefficients $\rho_{j,l}(k)$ do not depend on k.

Interest:

- the asymptotic covariance matrix is easy to compute and similar to the i.i.d. case
- Not restrictive class.

Regularly varying sequence

A large class of regular designs is the one for which the columns are regularly varying sequences:

Definition (Regularly varying sequence [6])

A sequence $S(\cdot)$ is regularly varying if and only if it can be written as:

$$S(i) = i^{\alpha} L(i),$$

where $-\infty < \alpha < \infty$ and $L(\cdot)$ is a slowly varying sequence.

This includes the case of polynomial regression: $x_{i,j} = i^j$.

Proposition

Assume that each column $X_{.,j}$ is regularly varying with parameter $\alpha_j > -\frac{1}{2}$. Then the conditions on the design are satisfied. Moreover:

- $\forall j, l \in \{1, \dots, p\}$: $\rho_{j,l}(k) = \frac{\sqrt{2\alpha_j+1}\sqrt{2\alpha_l+1}}{\alpha_j+\alpha_l+1}$, $\Rightarrow$ *Regular design.*
- *the condition* $R(0) > 0$ *is satisfied provided* $\alpha_j \neq \alpha_l$ *for any distinct* j, l *in* $\{1, \dots, p\}$.

Other class of regular designs: ANOVA type designs.

N.B.: A design whose columns are either ANOVA or regularly varying is again a regular design.

Sommaire

For regular design, the asymptotic covariance matrix is very similar to the i.i.d. case: $\sigma^2 = \mathbb{E}(\epsilon_0^2)$ should be replaced by $\sum_k \gamma(k)$.

Since the coefficients $\rho_{j,l}(k)$ are constant:

- $R(k) = \int_{-\pi}^{\pi} e^{ik\lambda} F_X(d\lambda) = R(0) = \int_{-\pi}^{\pi} F_X(d\lambda)$,
- the spectral measure F_X is equal to $\delta_0 R(0)$, with δ_0 a Dirac mass at 0.

Thereby the matrix F and G can be computed explicitly:

$$F = \frac{1}{2\pi}\int_{-\pi}^{\pi} F_X(d\lambda) = \frac{1}{2\pi}\int_{-\pi}^{\pi} R(0)\delta_0(d\lambda) = \frac{1}{2\pi}R(0),$$

$$G = \frac{1}{2\pi}\int_{-\pi}^{\pi} F_X(d\lambda)\otimes f(\lambda) = \frac{1}{2\pi}\int_{-\pi}^{\pi} R(0)\otimes f(\lambda)\delta_0(d\lambda)$$
$$= \frac{1}{2\pi}R(0)\otimes f(0) = f(0)F.$$

Thus the covariance matrix can be written as:

$$F^{-1}GF^{-1} = f(0)F^{-1}.$$

But:

$$f(\lambda) = \frac{1}{2\pi}\sum_{k=-\infty}^{\infty}\gamma(k)e^{-ik\lambda}, \quad \text{and} \quad f(0) = \frac{1}{2\pi}\sum_{k=-\infty}^{\infty}\gamma(k).$$

Hence the covariance matrix:

$$f(0)F^{-1} = \left(\frac{1}{2\pi}\sum_{k=-\infty}^{\infty}\gamma(k)\right)F^{-1} = \left(\sum_{k=-\infty}^{\infty}\gamma(k)\right)R(0)^{-1},$$

since $F = \frac{R(0)}{2\pi}$ and $F^{-1} = 2\pi R(0)^{-1}$.

Corollary

Under the assumptions of Hannan's Theorem, if moreover the design X is regular, then:

$$D(n)(\hat{\beta}-\beta) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}\left(0, \left(\sum_{k=-\infty}^{\infty} \gamma(k)\right) R(0)^{-1}\right),$$

and we have the convergence of the second order moment:

$$\mathbb{E}\left(D(n)(\hat{\beta}-\beta)(\hat{\beta}-\beta)^t D(n)^t\right) \xrightarrow[n\to\infty]{} \left(\sum_{k=-\infty}^{\infty} \gamma(k)\right) R(0)^{-1}.$$

For the i.i.d. case: $D(n)(\hat{\beta}-\beta) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}\left(0, \sigma^2 Q^{-1}\right)$.

Thus, to obtain confidence regions and tests for β, we need an estimator of $\sum_{k=-\infty}^{\infty} \gamma(k)$.

Sommaire

- Since $f(0) = 2\pi \sum_{k=-\infty}^{\infty} \gamma(k)$, we need an estimator of the spectral density.
- Properties of spectral density estimates have been discussed in many classical textbooks on time series.
- Wu and Liu [5] have developed an asymptotic theory (Central Limit Theorem, deviation inequalities, ...) for the spectral density estimate. However they use a notion of dependence that is more restrictive than Hannan's.
- We propose an estimator of the spectral density under Hannan's dependence condition. Only the consistency of the estimator is proved, but under very mild conditions.

Spectral density estimate

Let us first consider a preliminary random function:

$$f_n(\lambda) = \frac{1}{2\pi} \sum_{|k| \leq n-1} K\left(\frac{|k|}{c_n}\right) \hat{\gamma}_k e^{ik\lambda}, \quad \lambda \in [-\pi, \pi],$$

with:

$$\hat{\gamma}_k = \frac{1}{n} \sum_{j=1}^{n-|k|} \epsilon_j \epsilon_{j+|k|}, \qquad 0 \leq |k| \leq (n-1).$$

K is the kernel:

$$\begin{cases} K(x) = 1 & if\ |x| \leq 1 \\ K(x) = 2 - |x| & if\ 1 \leq |x| \leq 2 \\ K(x) = 0 & if\ |x| > 2. \end{cases}$$

$c_n \xrightarrow[n \to \infty]{} \infty$ and $\frac{c_n}{n} \xrightarrow[n \to \infty]{} 0$.

In our context, $(\epsilon_i)_{i\in\{1,\ldots,n\}}$ is not observed. Only the residuals are available:

$$\hat{\epsilon}_i = Y_i - (x_i)^t\hat{\beta} = Y_i - \sum_{j=1}^{p} x_{i,j}\hat{\beta}_j,$$

because only the data Y and the design X are observed. Consequently, we consider the following estimator:

$$f_n^*(\lambda) = \frac{1}{2\pi}\sum_{|k|\leq n-1} K\left(\frac{|k|}{c_n}\right)\hat{\gamma}_k^* e^{ik\lambda}, \qquad \lambda\in[-\pi,\pi],$$

where:

$$\hat{\gamma}_k^* = \frac{1}{n}\sum_{j=1}^{n-|k|}\hat{\epsilon}_j\hat{\epsilon}_{j+|k|}, \qquad 0\leq|k|\leq(n-1).$$

Consistence

Theorem

Let c_n be a sequence of positive integers such that $c_n \to \infty$ as n tends to infinity, and:

$$c_n \mathbb{E}\left(\left|\epsilon_0\right|^2\left(1 \wedge \frac{c_n}{n}\left|\epsilon_0\right|^2\right)\right) \xrightarrow[n\to\infty]{} 0.$$

Then, under the assumptions of Hannan's Theorem:

$$\sup_{\lambda\in[-\pi,\pi]} \left\|f_n^*(\lambda) - f(\lambda)\right\|_{\mathbb{L}^1} \xrightarrow[n\to\infty]{} 0.$$

Remark

If $\epsilon_0 \in \mathbb{L}^2$, then there exists $c_n \to \infty$ such that $c_n \mathbb{E}\left(\left|\epsilon_0\right|^2\left(1 \wedge \frac{c_n}{n}\left|\epsilon_0\right|^2\right)\right) \xrightarrow[n\to\infty]{} 0$ holds.

Proof (ideas)

By the triangle inequality:

$$\|f_n^*(\lambda) - f(\lambda)\|_{\mathbb{L}^1} \leq \|f_n^*(\lambda) - f_n(\lambda)\|_{\mathbb{L}^1} + \|f_n(\lambda) - f(\lambda)\|_{\mathbb{L}^1}$$

Two parts:

1. $\sup_{\lambda \in [-\pi,\pi]} \|f_n(\lambda) - f(\lambda)\|_{\mathbb{L}^1} \xrightarrow[n\to\infty]{} 0$,
2. $\sup_{\lambda \in [-\pi,\pi]} \|f_n^*(\lambda) - f_n(\lambda)\|_{\mathbb{L}^1} \xrightarrow[n\to\infty]{} 0$.

1: $\sup_{\lambda\in[-\pi,\pi]} \|f_n(\lambda) - f(\lambda)\|_{\mathbb{L}^1}$

- $\tilde{\epsilon}_{i,m} = \mathbb{E}(\epsilon_i|\mathcal{F}_{i+m}) - \mathbb{E}(\epsilon_i|\mathcal{F}_{i-m})$,
- $\tilde{f}_n^m(\lambda) = \frac{1}{2\pi}\sum_{|k|\leq n-1} K\left(\frac{|k|}{c_n}\right)\hat{\tilde{\gamma}}_{k,m}e^{ik\lambda}, \quad \lambda\in[-\pi,\pi]$,
- $\hat{\tilde{\gamma}}_{k,m} = \frac{1}{n}\sum_{j=1}^{n-|k|}\tilde{\epsilon}_{j,m}\tilde{\epsilon}_{j+|k|,m}, \quad |k|\leq n-1$.

By the triangle inequality:

$$\begin{aligned}\|f_n(\lambda) - f(\lambda)\|_{\mathbb{L}^1} \quad &\leq \quad 2\left\|\tilde{f}_n^m(\lambda) - f_n(\lambda)\right\|_{\mathbb{L}^1} + \left\|\tilde{f}_n^m(\lambda) - \mathbb{E}(\tilde{f}_n^m(\lambda))\right\|_{\mathbb{L}^1} \\ &\quad + \|\mathbb{E}(f_n(\lambda)) - f(\lambda)\|_{\mathbb{L}^1}.\end{aligned}$$

2: $\sup_{\lambda \in [-\pi,\pi]} \|f_n^*(\lambda) - f_n(\lambda)\|_{\mathbb{L}^1}$

Recall:

$$f_n^*(\lambda) = \frac{1}{2\pi} \sum_{|k| \leq n-1} K\left(\frac{|k|}{c_n}\right) \hat{\gamma}_k^* e^{ik\lambda}.$$

After calculations:

$$\|f_n^*(\lambda) - f_n(\lambda)\|_{\mathbb{L}^1} \leq \frac{1}{2\pi} \sum_{|k| \leq 2c_n} \|\hat{\gamma}_k^* - \hat{\gamma}_k\|_{\mathbb{L}^1}.$$

Since $\frac{c_n}{n} \xrightarrow[n \to \infty]{} 0$, it remains to prove that:

$$\sup_{|k| \leq 2c_n} \|\hat{\gamma}_k^* - \hat{\gamma}_k\|_{\mathbb{L}^1} = \mathcal{O}\left(\frac{1}{n}\right).$$

Lemma

$$
\begin{aligned}
\left\|\hat{\gamma}_k^* - \hat{\gamma}_k\right\|_{\mathbb{L}^1} \leq \;& \frac{1}{2n}\sum_{l=1}^{p}\sum_{l'=1}^{p}\left\|\left(\beta_l - \hat{\beta}_l\right)^2 \sum_{j=1}^{n-|k|} x_{j,l}^2\right\|_{\mathbb{L}^1} \\
&+ \frac{1}{2n}\sum_{l=1}^{p}\sum_{l'=1}^{p}\left\|\left(\beta_{l'} - \hat{\beta}_{l'}\right)^2 \sum_{j=1}^{n-|k|} x_{j+|k|,l'}^2\right\|_{\mathbb{L}^1} \\
&+ \frac{1}{n}\sum_{l=1}^{p}\left\|\sum_{j=1}^{n-|k|} \epsilon_j x_{j+|k|,l}\left(\beta_l - \hat{\beta}_l\right)\right\|_{\mathbb{L}^1} \\
&+ \frac{1}{n}\sum_{l=1}^{p}\left\|\sum_{j=1}^{n-|k|} \epsilon_{j+|k|} x_{j,l}\left(\beta_l - \hat{\beta}_l\right)\right\|_{\mathbb{L}^1} .
\end{aligned}
$$

Corollary

Corollary

Under the assumptions of Hannan's Theorem, if the design X is regular and if $f(0) > 0$, then:

$$\frac{R(0)^{\frac{1}{2}}}{\sqrt{2\pi f_n^*(0)}} D(n)(\hat{\beta} - \beta) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, I_p), \tag{1}$$

where I_p is the $p \times p$ identity matrix.

Sommaire

Functions of Linear processes

- $\Omega = \mathbb{R}^{\mathbb{Z}}$, $\mathbb{P} = \mu^{\otimes\mathbb{Z}}$, where μ is a probability measure on $\mathbb{R}$.
- $(\eta_i, i \in \mathbb{Z})$: sequence of i.i.d. random variables with marginal distribution μ.
- $(a_i)_{i\in\mathbb{Z}}$ is a sequence of real numbers in l^1. Assume that $\sum_{i\in\mathbb{Z}} a_i\eta_i$ is defined almost surely.
- $\epsilon_0 \in \mathbb{L}^2$ and is regular with respect to the σ-algebras : $\mathcal{F}_i = \sigma(\eta_j, j \leq i)$.

A large class of stationary processes satisfying Hannan's condition is the class of functions of real-valued linear processes:

$$\epsilon_k = f\left(\sum_{i\in\mathbb{Z}} a_i\eta_{k-i}\right) - \mathbb{E}\left(f\left(\sum_{i\in\mathbb{Z}} a_i\eta_{k-i}\right)\right).$$

- Modulus of continuity of f on $[-M, M]$:

$$\omega_{\infty,f}(h, M) = \sup_{|t|\leq h, |x|\leq M, |x+t|\leq M} |f(x+t) - f(x)| .$$

- Let $(\eta'_i)_{i\in\mathbb{Z}}$ be an independent copy of $(\eta_i)_{i\in\mathbb{Z}}$, and let:

$$M_k = \max\left\{ \left|\sum_{i\in\mathbb{Z}} a_i \eta'_i\right|, \left|a_k\eta_0 + \sum_{i\neq k} a_i\eta'_i\right| \right\}.$$

  If the following condition holds then Hannan's condition holds (cf. Section 5 of Dedecker, Merlevède, Volný [2]):

$$\sum_{k\in\mathbb{Z}} \left\| \omega_{\infty,f}(|a_k||\eta_0|, M_k) \wedge \|\epsilon_0\|_\infty \right\|_{\mathbb{L}^2} < \infty.$$

- **Application:** If f is γ-Hölder on any compact set; if $\omega_{\infty,f}(h, M) \leq C h^\gamma M^\alpha$ for some $C > 0$, $\gamma \in]0, 1]$ and $\alpha \geq 0$, then the condition above holds as soon as $\sum |a_k|^\gamma < \infty$ and $\mathbb{E}(|\eta_0|^{2(\alpha+\gamma)}) < \infty$.

Sommaire

Tests

- Equation of linear regression model: $Y = X\beta + \epsilon$, with $(\epsilon_i)_{i\in\mathbb{Z}}$ a stationary process
- Hannan's conditions are assumed true.
- Assume that ϵ_0 is $\mathcal{F}_\infty$-measurable and $\mathcal{F}_{-\infty}$ is trivial.
- The design X is regular.

With these conditions, the usual Fischer tests can be corrected by replacing the estimator of $\sigma^2 = \mathbb{E}(\epsilon_0^2)$ by an estimator of: $\sum_k \gamma(k)$.

If the errors are i.i.d. Gaussian, the test statistic is:

$$F = \frac{1}{p - p_0} \times \frac{RSS_0 - RSS}{\hat{\sigma}_\epsilon^2}.$$

- p_0 is the dimension of the model under the H_0-hypothesis
- $RSS = \|\hat{\epsilon}\|_2^2$ (for the complete model)
- RSS_0 is the corresponding quantity under H_0
- $\hat{\sigma}_\epsilon^2 = \frac{RSS}{n-p}$

Under H_0:

$$F \overset{\mathcal{L}}{\sim} \mathcal{F}_{n-p}^{p-p_0}.$$

- In the case where the design satisfies Hannan's conditions, if the random variables (ϵ_i) are i.i.d. but not Gaussian, the test statistic is the same as above and under H_0:

$$F \xrightarrow[n\to\infty]{\mathcal{L}} \frac{\chi^2(p-p_0)}{p-p_0}.$$

- If the error process $(\epsilon_i)_{i\in\mathbb{Z}}$ is stationary, the test statistic must be corrected:

$$\tilde{F}_c = \frac{1}{p-p_0} \times \frac{RSS_0 - RSS}{2\pi f_n^*(0)},$$

where $f_n^*(\lambda) = \frac{1}{2\pi}\sum_{|k|\leq n-1} K\left(\frac{|k|}{c_n}\right)\hat{\gamma}_k^* e^{ik\lambda}$. Thanks the previous results:

$$\tilde{F}_c \xrightarrow[n\to\infty]{\mathcal{L}} \frac{\chi^2(p-p_0)}{p-p_0}.$$

In practice, only a finite number of $\gamma(k)$ is estimated.

For the simulations, to choose this number (called a_n) we shall use the graph of the empirical autocovariance of the residuals.

Hence:

$$F_c = \frac{1}{p-p_0} \times \frac{RSS_0 - RSS}{\hat{\gamma}_0 + \sum_{k=1}^{a_n} \hat{\gamma}_k}.$$

Example: An autoregressive process

The process $(\epsilon_1, \dots, \epsilon_n)$ is simulated, according to the AR(1) equation:

$$\epsilon_{k+1} = \frac{1}{2}(\epsilon_k + \eta_{k+1}).$$

- ϵ_1 is uniformly distributed over $[-\frac{1}{2}, \frac{1}{2}]$
- $(\eta_i)_{i\geq 2}$ is a sequence of i.i.d. random variables, independent of ϵ_1, such that $\mathbb{P}(\eta_i = -\frac{1}{2}) = \mathbb{P}(\eta_i = \frac{1}{2}) = \frac{1}{2}$
- $\mathcal{F}_i = \sigma(\eta_k, k \leq i)$, and $\mathcal{F}_{-\infty}$ is trivial.
- Hannan's conditions are satisfied and the Fisher tests can be corrected.

First model

First model simulated:

$$Y_i = \beta_0 + \beta_1 i + 10\epsilon_i, \quad \forall i \in \{1, ..., n\}$$

- $\beta_0 = 3$.
- H_0: $\beta_1 = 0$, against H_1: $\beta_1 \neq 0$.

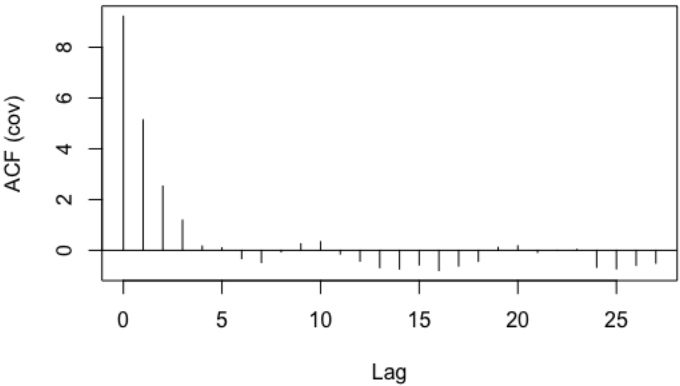
Under H_0, the same Fisher test is carried out 2000 times. Then we look at the estimated level of the test for different choices of n and a_n. (we want an estimated level close to 5%).

- Case $\beta_1 = 0$ and $a_n = 0$ (no correction):

n	200	400	600	800	1000
Estimated level	0.2745	0.2655	0.2615	0.2845	0.2445

The estimated levels are too large. The test reject the null hypothesis too often.

We choose a_n by analyzing the graph of the empirical autocovariances. For this example, this graph suggests $a_n = 2$ or 3.



- Case $\beta_1 = 0$, $a_n = 2$:

n	200	400	600	800	1000
Estimated level	0.0805	0.086	0.0745	0.0675	0.077

$a_n = 2$ gives a better estimated level than $a_n = 0$.

- Case $\beta_1 = 0$, $a_n = 3$:

n	200	400	600	800	1000
Estimated level	0.078	0.0725	0.074	0.059	0.0625

  - $a_n = 3$ seems slightly better than $a_n = 2$
  - the estimated level tends to 0.05
  - If $n = 5000$ and $a_n = 4$, the estimated level is around 0.05.

- Case $\beta_1 = 0.005$, $a_n = 3$:

In this example, H_0 is not satisfied. We perform the same tests as above to estimate the power of the test.

n	200	400	600	800	1000
Estimated power	0.2255	0.728	0.9945	1	1

The estimated power increases with the size of the samples and as soon as $n = 800$, the power is around 1 (the test always rejects the H_0-hypothesis).

Second model

Second model considered:

$$Y_i = \beta_0 + \beta_1 i + \beta_2 i^2 + 10\epsilon_i, \quad \forall i \in \{1, ..., n\}$$

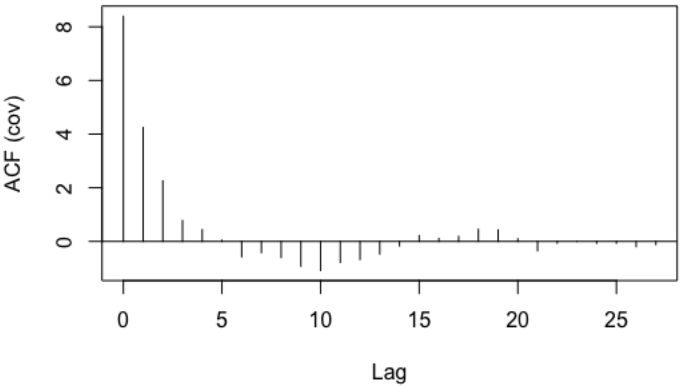
- H_0: $\beta_1 = \beta_2 = 0$ against H_1: $\beta_1 \neq 0$ or $\beta_2 \neq 0$
- $\beta_0 = 3$
- we use the same simulation scheme as above.

- Case $\beta_1 = \beta_2 = 0$ and $a_n = 0$ (no correction):

n	200	400	600	800	1000
Estimated level	0.402	0.378	0.385	0.393	0.376

As for the first simulation, if $a_n = 0$ the test will reject the null hypothesis too often.

As suggested by the graph of the estimated autocovariances, the choice $a_n = 4$ should give a better result for the estimated level.



- Case $\beta_1 = \beta_2 = 0$, $a_n = 4$:

n	200	400	600	800	1000
Estimated level	0.103	0.076	0.069	0.056	0.063

- $a_n = 4$ works well. For $n = 1000$, the estimated level is around 0.06
- If $n = 2000$ and $a_n = 4$, the estimated level is around 0.05.

- Case $\beta_1 = 0.005$, $\beta_2 = 0$, $a_n = 4$:

Under H_1, we study the estimated power of the test:

n	200	400	600	800	1000
Estimated power	0.2145	0.634	0.9855	1	1

As expected, the estimated power increases with the size of the samples, and it is around 1 as soon as $n = 800$.

Third model

Third model considered :

$$Y_i = \beta_0 + \beta_1 \sqrt{i} + \beta_2 \log(i) + 10\epsilon_i, \quad \forall i \in \{1, ..., n\}$$

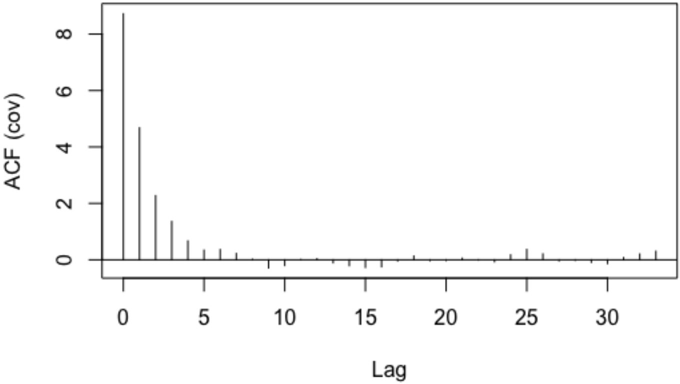
- H_0: $\beta_1 = \beta_2 = 0$ against H_1: $\beta_1 \neq 0$ or $\beta_2 \neq 0$
- $\beta_0 = 3$
- The conditions of the simulation are the same as above except for the size of the samples. For this model, n must be greater than previously to have an estimated level close to 5% with the correction.

- Case $\beta_1 = \beta_2 = 0$ and $a_n = 0$ (no correction):

n	500	1000	2000	3000	4000	5000
Estimated level	0.4435	0.4415	0.427	0.3925	0.397	0.4075

If $a_n = 0$ the test will reject the null hypothesis too often.

As suggested by the graph of the estimated autocovariances, $a_n = 4$ should give a better result for the estimated level.



- Case $\beta_1 = \beta_2 = 0$, $a_n = 4$:

n	500	1000	2000	3000	4000	5000
Estimated level	0.106	0.1	0.078	0.072	0.077	0.068

  - For $a_n = 4$ and $n = 5000$, the estimated level is around 0.07
  - If $n = 10000$, it is around 5%. Asymptotically, it converges to 0.05.

Then, the estimated power of the test:

- Case $\beta_1 = 0$, $\beta_2 = 0.2$, $a_n = 4$:

n	500	1000	2000	3000	4000	5000
Estimated power	0.2505	0.317	0.4965	0.6005	0.725	0.801

The estimated power increases with the size of the samples, and it is around 0.8 as soon as $n = 5000$.

R. C. Bradley.
Basic properties of strong mixing conditions.
In *Dependence in probability and statistics (Oberwolfach, 1985)*, volume 11 of *Progr. Probab. Statist.*, pages 165–192. Birkhäuser Boston, Boston, MA, 1986.

J. Dedecker, F. Merlevède, and D. Volnỳ.
On the weak invariance principle for non-adapted sequences under projective criteria.
Journal of Theoretical Probability, 20(4):971–1004, 2007.

J. Dedecker and C. Prieur.
New dependence coefficients. examples and applications to statistics.

Probability Theory and Related Fields, 132(2):203–236, 2005.

E. J. Hannan.
Central limit theorems for time series regression.
Probability theory and related fields, 26(2):157–170, 1973.

W. Liu and W. B. Wu.
Asymptotics of spectral density estimates.
Econometric Theory, pages 1218–1245, 2010.

E. Seneta.
Regularly varying functions, volume 508.
Springer, 2006.

W. B. Wu.
Nonlinear system theory: Another look at dependence.
Proceedings of the National Academy of Sciences of the United States of America, 102(40):14150–14154, 2005.

W. B. Wu.
Asymptotic theory for stationary processes.
Stat. Interface, 4(2):207–226, 2011.