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

Emmanuel Caron, PhD Student in Mathematics (with Jérôme Dedecker (MAP5) and Bertrand Michel (LMJL)), emmanuel.caron@ec-nantes.fr
Ecole Centrale Nantes, Laboratoire de Mathématiques Jean Leray, 1 Rue de la Noë, 44300 Nantes

Introduction

We consider the Linear Regression Model:

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

- X: $n \times p$ fixed design matrix,
- ϵ: strictly stationary process with zero mean.

The autocovariance function γ of the process ϵ and its spectral density f satisfy:

$$\gamma(k) = \mathrm{Cov}(\epsilon_m, \epsilon_{m+k}) = \mathbb{E}(\epsilon_m \epsilon_{m+k}) = \int_{-\pi}^{\pi} e^{ik\lambda} f(\lambda) d\lambda.$$

Hannan's Central Limit Theorem

Hannan's condition on the error process:

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

where $P_j(Z) = \mathbb{E}(Z|\mathcal{F}_j) - \mathbb{E}(Z|\mathcal{F}_{j-1})$. This implies that:

$$\sum_{k \in \mathbb{Z}} |\gamma(k)| < \infty.$$

Hannan's condition is satisfied for most short-range dependent stationary processes.

Let us define: $d_j(n) = \|X_{.,j}\|_2 = \sqrt{\sum_{i=1}^n x_{i,j}^2}$. Hannan's assumptions on the design, $\forall j \in \{1, \dots, p\}$:

- $\lim_{n \to \infty} d_j(n) = \infty$,
- $\lim_{n \to \infty} \frac{\sup_{1 \le i \le n} |x_{i,j}|}{d_j(n)} = 0$,
- the following limits exist: $\rho_{j,l}(k) = \lim_{n \to \infty} \sum_{m=1}^{n-k} \frac{x_{m,j} x_{m+k,l}}{d_j(n) d_l(n)}$.

Let $R(k)$ be the matrix:

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

with F_X the spectral measure associated with the matrix $R(k)$. Moreover $R(0)$ is supposed to be positive definite. Let then F and G be the matrices:

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

Theorem (Hannan's theorem [2])

Under the previous conditions, we have:

$$D(n)(\hat{\beta} - \beta) \xrightarrow[n \to \infty]{} \mathcal{N}(0, F^{-1} G F^{-1}),$$
$$\mathbb{E}\left(D(n)(\hat{\beta} - \beta)(\hat{\beta} - \beta)^t D(n)^t\right) \xrightarrow[n \to \infty]{} F^{-1} G F^{-1}.$$

Regular design

Hannan's theorem is very general because it includes a very large class of designs:

Definition (Regular design)

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

A large class of regular designs: the regularly varying sequences (i.e. of the form $S(i) = i^\alpha L(i)$, where $\alpha \in \mathbb{R}$ and $L(\cdot)$ a slowly varying sequence).
For regular design, the asymptotic covariance matrix is easy to compute.

Corollary (Hannan's theorem with regular design)

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:

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

In the case of regular design, the asymptotic covariance matrix is similar to the one in the case where the random variables (ϵ_i) are i.i.d.; the variance term σ^2 is replaced by the series of covariances.
Thus, to obtain confidence regions and tests for the parameter β, an estimator of: $\sum_{k=-\infty}^{\infty} \gamma(k)$ is needed.

Estimation of the covariance matrix

Consider the following estimator of the spectral density, for λ in $[-\pi, \pi]$:

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

where:

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

with $\hat{\epsilon}$ the residuals: $\hat{\epsilon} = Y - X\hat{\beta}$.
The kernel K is defined by:

$$K(x) = \mathbb{1}_{|x| \le 1} + (2 - |x|)\mathbb{1}_{1 \le |x| \le 2},$$

and the sequence of positive integers c_n is such that $c_n \xrightarrow[n \to \infty]{} \infty$ and $\frac{c_n}{n} \xrightarrow[n \to \infty]{} 0$.

Theorem (Consistence [1])

Let c_n be a sequence of positive integers such that $c_n \xrightarrow[n \to \infty]{} \infty$, and:

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

Then, under the assumptions of Hannan's theorem:

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

Combining Hannan's theorem and the previous result, we get:

Corollary

If $f(0) > 0$, then:

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

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

Tests

Thanks to these results, the usual Fischer tests on the linear model can be adapted to the case where the errors are short-range dependent. As usual, the null hypothesis H_0 means that the parameter β belongs to a vector space with dimension equal to p_0 (strictly smaller than p), and we denote by H_1 the alternative hypothesis.
Recall that if the errors are i.i.d. Gaussian random variables, the test statistic is:

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

where $RSS = \|\hat{\epsilon}\|_2^2$, $RSS_0 = \|\hat{\epsilon}_{H_0}\|_2^2$ and $\hat{\sigma}_\epsilon^2 = \frac{RSS}{n-p}$. Under H_0, $F \sim \mathcal{F}_{n-p}^{p-p_0}$.
If the error process $(\epsilon_i)_{i \in \mathbb{Z}}$ is stationary, the test statistic must be corrected as follows:

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

It converges to a χ^2-distribution with parameter $p - p_0$.

Simulations

Let us simulate the process $(\epsilon_i)_{1 \le i \le n}$ according to the AR(1) equation:

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

where ϵ_1 is uniformly distributed over $[-\frac{1}{2}, \frac{1}{2}]$, and $(\eta_i)_{i \ge 2}$ is a sequence of i.i.d. random variables, independent of ϵ_1, such that $\mathbb{P}(\eta_i = -\frac{1}{2}) = \mathbb{P}(\eta_i = \frac{1}{2}) = \frac{1}{2}$.
The model simulated with this error process is, $\forall i \in \{1, ..., n\}$:

$$Y_i = \beta_0 + \beta_1 \sqrt{i} + \beta_2 \log(i) + 10\epsilon_i.$$

We test H_0: $\beta_1 = \beta_2 = 0$ against H_1: $\beta_1 \ne 0$ or $\beta_2 \ne 0$, and we want an estimated level close to 5%.

- Case $\beta_1 = \beta_2 = 0$, no correction:

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

- Case $\beta_1 = \beta_2 = 0$, with correction:

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

If one increases the size of the samples n, we are getting closer to the estimated level 5%.

References

[1] E. Caron and S. Dede.
Asymptotic distribution of least square estimators for linear models with dependent errors : regular designs.
working paper or preprint, Oct. 2017.

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