Statistical Analysis of Time Series and Forecasting

Pergamenshchikov Serguei, Pchelintsev Evgeny

Statistical Analysis of Time Series and Forecasting *

Pergamenshchikov Serguei† and Pchelintsev Evgeny, ‡

February 2, 2023

Abstract

In this course, we present the principal parts of the time series analysis. First, stationary processes and trends in times series are introduced. Then we consider the linear regression models for which we study the main problems such that point estimation, the construction of confidence intervals, hypothesis testing, and forecasting. In addition, big data models and the main methods for their analysis are presented. Finally, we introduce the autoregressive and moving average autoregressive processes (ARMA) and study their basic properties, including the problem of forecasting.

*This work was done under financial support of the Russian Federal Professor program (project no. 1.472.2016/1.4, Ministry of Education and Science)

†Laboratoire de Mathématiques Raphael Salem, UMR 6085 CNRS- Université de Rouen Normandie, France and International Laboratory of Statistics of Stochastic Processes and Quantitative Finance of National Research Tomsk State University, e-mail: Serge.Pergamenshchikov@univ-rouen.fr

‡International Laboratory of Statistics of Stochastic Processes and Quantitative Finance of National Research Tomsk State University

1 Times series and stochastic processes

A sequence of random variables $(y_j)_{j\geq 1}$ is called a stochastic process en discrete time. A stochastic process is called *strictly stationary* if for any $k \geq 1$ the joint distribution of the random variables $y_j, y_{j-1}, \ldots, y_{j-k+1}$ is the same for all $n > k$, i.e. for any bounded $\mathbb{R}^k \to \mathbb{R}$ functions h

$$\mathbf{E}\, h(y_j, \ldots, y_{j-k+1}) = \mathbf{E}\, h(y_k, \ldots, y_1)\,.$$

Moreover, sometime we will use a *weak stationary or covariance-stationary* process, i.e. process $(y_j)_{j\geq 1}$ for which $\mathbf{E}y_j$ and $\mathbf{E}y_j^2$ are constant and for some $\mathbb{R} \to \mathbb{R}$ function g the auto covariation

$$\mathbf{E}y_j\, y_l = g(j-l) \quad \text{for any} \quad j, l\,.$$

A weak stationary process $(\varepsilon_j)_{-\infty<j<\infty}$ is called a *white noise* if $\mathbf{E}\,\varepsilon_j = 0$, $\mathbf{E}\,\varepsilon_j^2 = \sigma^2$ and $\mathbf{E}\varepsilon_i\varepsilon_j = 0$ for any $i \neq j$. In the sequel we will use the well known Wold's decomposition or the Wold representation theorem (see, for example, in [1]).

Theorem 1.1. *Any weak stationary process $(y_j)_{j\geq 1}$ with $\mathbf{E}\, y_j = \mu$ can be represented as*

$$y_j = \mu + \sum_{j=0}^{\infty} \mathbf{b}_l \varepsilon_{j-l}\,, \tag{1.1}$$

where the coefficients $(\mathbf{b}_j)_{j\geq 0}$ are such that $\mathbf{b}_j = 1$ and $\sum_{j\geq 1} \mathbf{b}_j^2 < \infty$ and $(\varepsilon_l)_{-\infty<l<\infty}$

A stochastic process in discrete time $(y_j)_{j\geq 1}$ is called *time series* if it can be represented as

$$y_j = f(j) + \xi_j\,, \tag{1.2}$$

where $f(\cdot)$ is non random function and $(\xi_j)_{j\geq 1}$ is a stochastic weak stationary process with $\mathbf{E}\,\xi_j = 0$. The function f is called *trend* and the process $(\xi_j)_{j\geq 1}$ is called the *stochastic part* of the time series (1.2).
If $f(\cdot)$ has the polynomial form, then it is called *polynomial trend*, i.e.

$$f(x) = \sum_{i=1}^{p} \mathbf{a}_i\, x^{i-1}\,, \tag{1.3}$$

where $\mathbf{a}_1, \ldots, \mathbf{a}_p$ are the polynomial coefficients. Usually, we consider the time series (1.2) on the finite interval, i.e. $1 \leq j \leq n$ with $n > p$. The more comfortable form for the polynomial trend if we replace in (1.3) the power functions t^i with the orthogonal polynomials $\phi_1, \ldots, \Theta_p$, i.e. such that for any $l \neq i$

$$\sum_{j=1}^{n} \phi_l(j)\,\phi_i(j) = 0\,. \tag{1.4}$$

For example, for $p = 3$ we can take $\phi_1 \equiv 1$, $\phi_2(x) = x - (n+1)/2$ and

$$\phi_3(x) = x^2 - (n+1)x + \frac{(n+1)(n+2)}{6}\,.$$

Generally, we can represent the orthogonal polynomials as

$$\phi_j(x) = \sum_{i=1}^{j} \gamma_i x^{i-1}$$

and therefore, the trend (1.4) can be represented as

$$f(x) = \sum_{j=1}^{p} \theta_j\,\phi_j(x)\,, \tag{1.5}$$

where the coefficients $\mathbf{a}_i = \gamma_i \sum_{l=i}^{p} \theta_l$. Moreover, if the function $f(\cdot)$ has a trigonometric form, then it is *called trigonometric or cycle trend*, i.e.

$$f(x) = \mathbf{a}_0 + \sum_{i=1}^{m} (\mathbf{a}_i \cos(\omega_i x) + \mathbf{b}_i \sin(\omega_i x)\,, \tag{1.6}$$

where $\mathbf{a}_0, \ldots, \mathbf{a}_m$ and $\mathbf{b}_1, \ldots, \mathbf{b}_m$ are coefficients and ω_i are frequencies. Of course we consider the case when the number of the parameters $p = 2m+1 < n$. To obtain the property (1.2) one can take, for example,

$$\omega_i = \frac{\pi i}{n}\,.$$

To this end, first we chose the trigonometric basis $(\phi_j)_{j\geq 1}$ in $\mathbf{L}_2[0,1]$, i.e.

$$\phi_1 \equiv 1\ , \quad \phi_j(x) = \sqrt{2}\,\mathrm{Tr}_j\left(\frac{2\pi[j/2]}{n}x\right)\ , \quad j \geq 2\,, \tag{1.7}$$

where the function $\mathrm{Tr}_j(t) = \cos(t)$ for even j and $\mathrm{Tr}_j(t) = \sin(x)$ for odd j. Using these functions we can represent the trend (1.6) as

$$f(x) = \sum_{j=1}^{p} \theta_j\,\phi_j(x)\,, \tag{1.8}$$

where $\mathbf{a}_0 = \theta_1$, $\mathbf{a}_i = \theta_{2i}$ and $\mathbf{b}_i = \theta_{2i+1}$ for $1 \leq i \leq p$.

There exist non linear over parameters forms for the trend functions $f(x,\theta)$ (see, for example, [3] and the references therein). For example, hyperbolic regression

$$f(x,\theta) = \frac{1}{\theta_1 + \theta_2 x} \quad \text{and} \quad \theta = (\theta_1, \theta_2)\,,$$

exponential regression

$$f(x,\theta) = \theta_1\, e^{\theta_2 x} \quad \text{and} \quad \theta = (\theta_1, \theta_2)\,.$$

In clinical trials is used the logistics regression (see, e.g. [6])

$$f(x,\theta) = \theta_1 + (\theta_2 - \theta_1)\,\frac{x^{\theta_4}}{x^{\theta_4} + \theta_3} \quad \text{and} \quad \theta = (\theta_1, \theta_2, \theta_3, \theta_4)\,.$$

The main goal of the time series analysis is to develop statistical identification and forecasting methods for the different models of the time series (1.2).

Exercises

1. To show that anu sequence of independent identically distributed random variable, i.i.d. random variables $(y_j)_{j\geq 1}$ form a *strictly stationary* process.

2. Let $(\xi_j)_{j\geq 1}$ be i.i.d. sequence then for any fixed integer $m \geq 1$ and $\mathbb{R}^m \to \mathbb{R}$ function g. Now for any $n \geq m$ we set

$$y_j = g(\xi_j, \ldots, \xi_{j-m+1})\,. \tag{1.9}$$

   To show that the process $(y_j)_{j\geq m}$ is stationary.

3. To check that any strict square integrated stationary process is a weak stationary process.

2 Scalar regression analysis

First, we consider the scalar linear regression model, i.e.

$$y_j = \theta x_j + \varepsilon_j\,, \quad 1 \le j \le n\,, \tag{2.1}$$

where θ is unknown parameter, $(x_j)_{1\le j\le n}$ are non random regression variables and $(\varepsilon_j)_{1\le j\le n}$ is unobservable white noise, i.e. $\mathbf{E}\varepsilon_j = 0$ and $\mathbf{E}\,\varepsilon_j^2 = \sigma^2$ for any $1 \le j \le n$ and $\mathbf{E}\,\varepsilon_j\varepsilon_l = 0$ for $j \ne l$.

The identification problem for the model (2.1) is to estimate the parameter θ on, the basis of the observations $(y_j)_{1\le j\le n}$. To this end we will use the Least Square Estimator (LSE) method according to which one needs to minimize over unknown parameter the integral noise intensity, i.e.

$$\sum_{j=1}^{n} (y_j - \theta x_j)^2 \quad \to \quad \min_{\theta\in\mathbb{R}}\,. \tag{2.2}$$

Therefore, if

$$\sum_{j=1}^{n} x_j^2 > 0 \tag{2.3}$$

then we obtain immediately that least square estimator is

$$\widehat{\theta}_n = \frac{\sum_{j=1}^{n} y_j x_j}{\sum_{j=1}^{n} x_j^2}\,. \tag{2.4}$$

From the model (2.1) it is easy to deduce that

$$\widehat{\theta}_n = \theta + \frac{\sum_{j=1}^{n} x_j \varepsilon_j}{\sum_{j=1}^{n} x_j^2}\,.$$

Therefore,

$$\mathbf{E}\,\widehat{\theta}_n = \theta + \frac{\sum_{j=1}^{n} x_j \mathbf{E}\,\varepsilon_j}{\sum_{j=1}^{n} x_j^2} = 0$$

and, moreover, the mean square estimation accuracy in this case can be calculated as

$$\mathbf{V}(\widehat{\theta}_n) = \mathbf{E}\,(\widehat{\theta}_n - \theta)^2 = \frac{\mathbf{E}\left(\sum_{j=1}^{n} x_j\varepsilon_j\right)^2}{\left(\sum_{j=1}^{n} x_j^2\right)^2} = \frac{\sigma^2}{\sum_{j=1}^{n} x_j^2}\,. \tag{2.5}$$

From this we can obtain immediately the necessary and sufficient condition for the convergence in $\mathbf{L}_2$ as $n \to \infty$.

Proposition 2.1. *The least square estimator* (2.4) *tends to* θ *in in* $\mathbf{L}_2$ *if and only if*

$$\lim_{n\to\infty}\sum_{l=1}^{n} x_l^2 = +\infty\,. \tag{2.6}$$

For this estimator one can show the following theorem.

Theorem 2.1. *(Gauss - Markov) The least square estimator* (2.5) *is the best estimator in the class of all linear unbiased estimators of the non zero parameter* θ *in the model* (2.1) *with the condition* (2.3) *in the means square accuracy sense*

$$\mathbf{E}\,(\widetilde{\theta}_n - \theta)^2 \geq \mathbf{E}\,(\widehat{\theta}_n - \theta)^2\,, \tag{2.7}$$

where $\widetilde{\theta}_n$ *is an arbitrary linear estimator, i.e. an estimator of the form*

$$\widetilde{\theta}_n = \sum_{j=1}^{n} \mathbf{g}_j y_j$$

and $(\mathbf{g}_j)_{1\leq j\leq n}$ *are non random coefficients.*

Proof. Indeed, note that for unbiased estimators we have

$$\theta = \mathbf{E}\,\widetilde{\theta}_n = \sum_{j=1}^{n} \mathbf{g}_j \mathbf{E}\,y_j = \theta \sum_{j=1}^{n} \mathbf{g}_j \mathbf{x}_j\,,$$

i.e. $\sum_{j=1}^{n} \mathbf{g}_j \mathbf{x}_j = 1$. Using here the Cauchy Bunyakovsky Schwarz we get

$$1 = \left(\sum_{j=1}^{n} \mathbf{g}_j \mathbf{x}_j\right)^2 \leq \sum_{j=1}^{n} \mathbf{g}_j^2 \sum_{j=1}^{n} \mathbf{x}_j^2\,.$$

Therefore,

$$\mathbf{E}\,(\widetilde{\theta}_n - \theta)^2 = \mathbf{E}\,\left(\sum_{j=1}^{n} \mathbf{g}_j \varepsilon_j\right)^2 = \sigma^2 \sum_{j=1}^{n} \mathbf{g}_j^2 \geq \frac{\sigma^2}{\sum_{j=1}^{n} \mathbf{x}_j^2}\,.$$

Now, the property (2.5) implies directly (2.7). Hence Theorem 2.1. □
Note now, that the estimation accuracy (2.5) depend on the coefficient σ^2. Therefore, if it is unknown, then the estimation accuracy is unknown as well. To estimate it we use the model estimation defined as

$$\widehat{y}_j = \widehat{\theta}_n\, x_j\,. \tag{2.8}$$

Therefore, in this case the deviation is given as

$$\widehat{\varepsilon}_j = y_j - \widehat{y}_j = (\theta - \widehat{\theta}_n)x_j + \varepsilon_j$$

and, therefore,

$$\sum_{j=1}^{n} \widehat{\varepsilon}_j^2 = \sum_{j=1}^{n} \varepsilon_j^2 - \frac{\left(\sum_{j=1}^{n} x_j \varepsilon_j\right)^2}{\sum_{j=1}^{n} x_j^2}\,. \tag{2.9}$$

It is clear that

$$\mathbf{E} \sum_{j=1}^{n} \widehat{\varepsilon}_j^2 = (n-1)\sigma^2\,.$$

Therefore, for any $n > 1$

$$\widehat{\sigma}_n = \frac{1}{n-1} \sum_{j=1}^{n} \widehat{\varepsilon}_j^2 \tag{2.10}$$

is unbiased estimator for the variance σ^2.
Using the estimator (2.5), we estimate now estimation accuracy (2.5) as

$$\widehat{\mathbf{V}}(\widehat{\theta}_n) = \frac{\widehat{\sigma}_n}{\sum_{j=1}^{n} x_j^2}\,. \tag{2.11}$$

Now a natural question arises, what happens if we replace the unknown normalized coefficient in (2.13) with the known coefficient (2.11), i.e. the question now is the following: is it possible to calculate the distribution of the fraction

$$\Upsilon_n = \frac{\widehat{\theta}_n - \theta}{\sqrt{\widehat{\mathbf{V}}(\widehat{\theta}_n)}}\,. \tag{2.12}$$

To study this question one needs to add the condition that the noise variables $(\varepsilon_j)_{1\le j\le n}$ in the regression model (2.1) are i.i.d. Gaussian with the parameters $(0, \sigma^2)$. It is clear that in this case the estimator (2.4) is Gaussian

$$\widehat{\theta}_n \quad \sim \quad \mathcal{N}\left(\theta\,, \mathbf{V}(\widehat{\theta}_n)\right)\,,$$

i.e. for any $n > 1$

$$\frac{\widehat{\theta}_n - \theta}{\sqrt{\mathbf{V}(\widehat{\theta}_n)}} \quad \sim \quad \mathcal{N}(0\,,\,1)\,. \tag{2.13}$$

To study the fraction (2.12) we need the following definition.

Definition 2.1. *A positive random variable ζ is said to be a random variable distributed according to the $\mathcal{X}^2_{\mathbf{p}}$ law with $\mathbf{p}$ degree of liberty if for any measurable set $A \subseteq \mathbb{R}$*

$$\mathbf{P}(\zeta \in A) = \int_A \rho_p(u)\mathrm{d}u$$

where

$$\rho_{\mathbf{p}}(u) = \frac{u^{\mathbf{p}/2-1}\, e^{-u/2}}{2^{\mathbf{p}}\Gamma(\mathbf{p}/2)}\, \mathbf{1}_{\{u\geq 0\}}\,, \tag{2.14}$$

and $\Gamma(v) = \int_0^{+\infty} t^{v-1}e^{-t}\mathrm{d}t$ for $v > 0$

In the sequel we will use the following property.

Proposition 2.2. *If $\xi_1, \ldots, \xi_{\mathbf{p}}$ are i.i.d. $(0, 1)$ Gaussian random variables then the sum $\sum_{j=1}^{\mathbf{p}} \xi_j^2$ has the $\mathcal{X}_{\mathbf{p}}$ distribution.*

Using this definition we will show the following property.

Proposition 2.3. *For the Gaussian regression model* (2.1) *with the condition* (2.3) *and $n > 1$ the random variable*

$$\gamma_n = \frac{\sum_{j=1}^n \widehat{\varepsilon}_j^2}{\sigma^2} \quad \sim \quad \mathcal{X}^2_{n-1}\,.$$

Proof. First we set

$$\xi = (\xi_1, \ldots, \xi_n)' \quad \text{and} \quad \xi_i = \frac{\varepsilon}{\sigma}\,. \tag{2.15}$$

Here the prime $'$ denotes the transposition. It is clear that for the Gaussian model (2.1) the vector ξ is Gaussian in $\mathbb{R}^n$ with the parameters $(0, \mathbf{I}_n)$, where $\mathbf{I}_n$ is identity matrix of the order n. Now, using the vector (2.15) and the property (2.9), we can represent the random variable γ_n as the following quadratic form

$$\gamma_n = \xi^{'}\, (\mathbf{I}_n - \mathbf{A})\, \xi \quad \text{and} \quad \mathbf{A} = (\mathbf{a}_{i,j})_{1\leq i,j\leq n}\,, \tag{2.16}$$

where the elements

$$\mathbf{a}_{i,j} = \mathbf{g}_i\, \mathbf{g}_j \quad \text{and} \quad \mathbf{g}_i = \frac{x_i}{\sqrt{\sum_{\iota=1}^n x_\iota^2}}\,. \tag{2.17}$$

Taking into account that $\sum_{j=1}^n \mathbf{g}_j^2 = 1$, we can obtain that the matrix $\mathbf{A}$ is idempotent, i.e. $\mathbf{A}^2 = \mathbf{A}$. This means that the eigenvalues either 1 or 0. Moreover, note that

$$\mathrm{tr}\mathbf{A} = \sum_{j=1}^n \mathbf{a}_{j,j} = \sum_{j=1}^n \mathbf{g}_j^2 = 1\,,$$

i.e. the matrix $\mathbf{A}$ has $n-1$ eigenvalues 0 and one 1. Therefore, there exists an orthogonal $n \times n$ matrix $\mathbf{Q}$, i.e. $\mathbf{Q}'\mathbf{Q} = \mathbf{I}_n$, such that

$$\mathbf{I}_n - \mathbf{A} = \mathbf{Q}' J_n \mathbf{Q} \quad \text{and} \quad J_n = \operatorname{diag}(1, \dots, 1, 0) = \begin{bmatrix} 1 & \dots & 0 & 0 \\ 0 & \ddots & 0 & 0 \\ 0 & \dots & 1 & 0 \\ 0 & \dots & 0 & 0 \end{bmatrix}. \quad (2.18)$$

Thus, using this in (2.16) we can write that

$$\gamma_n = \eta' J_n \eta \quad \text{with} \quad \eta = (\eta_1, \dots, \eta_n)' = \mathbf{Q}\,\xi\,. \quad (2.19)$$

Note here, that η is a gaussian vector in $\mathbb{R}^n$ with the parameters $(0, \mathbf{I}_n)$, i.e. $\eta_1, \dots, \eta_n$ are i.i.d. random $(0,1)$ gaussian random variables, i.e. γ_n can be represented as

$$\gamma_n = \sum_{j=1}^{n-1} \eta_j^2\,,$$

i.e. in view of the exercise 2 the random variable γ_n has χ_{n-1} distribution. Hence Proposition 2.3. $\square$

Definition 2.2. *A random variable ζ is said to be a random variable distributed according to the Student law with $\mathbf{p} \geq 1$ freedom degrees, denoted as τ_p, if for any measurable set $A \subset \mathbb{R}$*

$$\mathbf{P}(\zeta \in A) = \int_A \rho_p^*(u)\mathrm{d}u\,,$$

where

$$\rho_{\mathbf{p}}^*(x) = \frac{\Gamma\left(\frac{\mathbf{p}-1}{2}\right)}{\Gamma\left(\frac{\mathbf{p}}{2}\right)} \frac{1}{(\mathbf{p}\pi)^{1/2}} \frac{1}{\left(1 + \frac{x^2}{\mathbf{p}}\right)^{\frac{\mathbf{p}+1}{2}}}\,. \quad (2.20)$$

To study the distribution of the fraction (2.12) we will use the following property.

Proposition 2.4. *Let U and γ be two independent random variables distributed as $\mathcal{N}(0,1)$ and $\mathcal{X}^2_{\mathbf{p}}$ respectively. Then the random variable*

$$\frac{U}{\sqrt{\gamma/\mathbf{p}}} \quad \sim \quad \tau_{\mathbf{p}}\,,$$

i.e. follows the Student law with $\mathbf{p}$ degree of liberty.

Now we can calculate the distribution for the random variable (2.12).

Proposition 2.5. *For the Gaussian regression model* (2.1) *with the condition* (2.3) *and* $n > 1$ *the random variable*

$$\Upsilon_n = \frac{\widehat{\theta}_n - \theta}{\sqrt{\widehat{\mathbf{V}}(\widehat{\theta}_n)}} \quad \sim \quad \tau_{n-1}\,,$$

i.e. has the Student distribution with $n-1$ *freedom degrees.*

Proof. First note that we can represent the random variable (2.12) as

$$\Upsilon_n = \frac{U}{\sqrt{\gamma_n/(n-1)}}\,, \tag{2.21}$$

where

$$U = \frac{\widehat{\theta}_n - \theta}{\sqrt{\mathbf{V}(\widehat{\theta}_n)}} \quad \text{and} \quad \gamma_n = \frac{\sum_{j=1}^n \widehat{\varepsilon}_j^2}{\sigma^2}\,.$$

Note that the covariances

$$\mathbf{E}\,(\widehat{\theta}_n - \theta)\widehat{\varepsilon}_j = -x_j \mathbf{E}\,(\widehat{\theta}_n - \theta)^2 + \mathbf{E}\,(\widehat{\theta}_n - \theta)\varepsilon_j = \frac{-\sigma^2 x_j + \sigma^2 x_j}{\sum_{i=1}^n x_i^2} = 0\,.$$

Therefore, taking into account that the random variables $\widehat{\theta}_n$ and $(\widehat{\varepsilon}_j)_{1\le j\le n}$ are Gaussian, the last property means that the estimator $\widehat{\theta}_n$ is independent of the sequence $(\widehat{\varepsilon}_j)_{1\le j\le n}$ and, therefore, in (2.21) the random variables U and γ_n are independent. So, Propositions 2.3 and 2.4 imply the desired result. Hence Proposition 2.5. $\square$

2.1 Coefficient testing

For the Gaussian model (2.1) we consider the following hypothesis testing problem

$$\mathbf{H}_0 \;:\; \theta = 0 \quad \text{and} \quad \mathbf{H}_1 \;:\; \theta \neq 0\,. \tag{2.22}$$

We set a threshold $0 < \alpha < 1$. This means, that if, for example, $\alpha = 0,05$ then this means that the risk of wrongly rejecting $\mathbf{H}_0$ is 5%. In the context of this problem we call *test function* any statistics

$$\varphi : \mathbb{R}^n \times \mathbb{R}^n \to \{0,1\}\,.$$

The observation in this case is $z = (y_1 \dots y_n, x_1 \dots x_n)$. The probability $\mathbf{P}\,(\varphi = 1 \,|\, \mathbf{H}_0)$ to reject the hypothesis $\mathbf{H}_0$ while it is true is called the *error of the first kind*. Moreover, the probability $\mathbf{P}\,(\varphi = 0 \,|\, \mathbf{H}_1)$ to accept the

hypothesis $\mathbf{H}_0$ while it is false is the *error of the second kind*. The *power* of the test is by definition $1 - \mathbf{P}\,(\varphi = 0\,|\,\mathbf{H}_1)$. We know that under the condition (2.3) for $n > 1$

$$\frac{\widehat{\theta}_n - \theta}{\sqrt{\widehat{\mathbf{V}}(\widehat{\theta}_n)}} \;\sim\; \tau_{n-1}\,.$$

So if $\theta = 0$ then

$$\frac{\widehat{\theta}_n}{\sqrt{\widehat{\mathbf{V}}(\widehat{\theta}_n)}} \;\sim\; \tau_{n-1}\,.$$

We construct the test as

$$\varphi \;=\; \begin{cases} 0\,, & \text{if} \quad |\widehat{\theta}_n| \;\le\; z_\alpha \sqrt{\widehat{\mathbf{V}}(\widehat{\theta}_n)}\,, \\ 1\,, & \text{if} \quad |\widehat{\theta}_n| \;>\; z_\alpha \sqrt{\widehat{\mathbf{V}}(\widehat{\theta}_n)} \end{cases} \tag{2.23}$$

(we accept the hypothesis $\mathbf{H}_0$ if $\varphi = 0$, we refuse it if $\varphi = 1$) where $z_\alpha > 0$ is the quantile of the threshold $1 - \alpha$ for the absolute value of the Student distribution, i.e.

$$\mathbf{P}\left(|\tau_{n-1}| \le z_\alpha\right) = 1 - \alpha\,.$$

It is clear, that for this test function

$$\begin{aligned} \mathbf{P}\,(\varphi = 1\,|\,\mathbf{H}_0\,) \;&=\; \mathbf{P}\left(\frac{|\widehat{\theta}_n|}{\sqrt{\widehat{\mathbf{V}}(\widehat{\theta}_n)}} \;>\; z_\alpha \,\middle|\, \mathbf{H}_0\right) \\ &=\; \mathbf{P}\left(|\tau_{n-1}| \;>\; z_\alpha\right) \;=\; \alpha\,. \end{aligned}$$

In addition, to study the second kind error.

Proposition 2.6. *Assume that the condition* (2.6) *holds. Then the second kind error for the test* (2.23) *goes to zero as* $n \to \infty$.

Proof. First, note, that

$$\begin{aligned} \mathbf{P}\,(\varphi = 0\,|\,\mathbf{H}_1) \;&=\; \mathbf{P}\left(\frac{|\widehat{\theta}_n|}{\sqrt{\widehat{\mathbf{V}}(\widehat{\theta}_n)}} \;\le\; z_\alpha \,|\, \theta \neq 0\right) \\ &\le\; \mathbf{P}\left(\frac{|\theta|}{\sqrt{\widehat{\mathbf{V}}(\widehat{\theta}_n)}} \;\le\; z_\alpha + \frac{|\widehat{\theta}_n - \theta|}{\sqrt{\widehat{\mathbf{V}}(\widehat{\theta}_n)}}\right). \end{aligned}$$

Moreover, for any fixed $N \geq 1$ we can estimate this error as

$$\mathbf{P}\left(\varphi = 0 \,|\, \mathbf{H}_1\right) \leq \mathbf{P}\left(|\theta| \,\leq\, (z_\alpha + N)\cdot\sqrt{\widehat{\mathbf{V}}(\widehat{\theta}_n}\right) \,+\, \mathbf{P}\left(\frac{|\widehat{\theta}_n - \theta|}{\sqrt{\widehat{\mathbf{V}}(\theta)}} \,>\, N\right)$$

$$= \mathbf{P}\left(\widehat{\mathbf{V}}(\widehat{\theta}_n) > \frac{(\theta)^2}{(z_\alpha + N)^2}\right) + \mathbf{P}(|\tau_{n-1}| \,>\, N)\,.$$

Taking here into account that under the condition (2.6) the term $\widehat{\mathbf{V}}(\widehat{\theta}_n) \to 0$ in probability as $n \to \infty$, we find that for any $N > 1$

$$\limsup_{n\to\infty} \mathbf{P}\left(\varphi = 0 | \mathbf{H}_1\right) \,\leq\, \limsup_{n\to\infty} \mathbf{P}\left(|\tau_{n-1}| > N\right)\,.$$

Let now $(\eta_j)_{j\geq 1}$ be i.i.d. $(0,1)$ gaussian random variables and U be also $(0,1)$ gaussian random variable independent of the sequence $(\eta_j)_{j\geq 1}$. Then in view of Propositions 2.2 and 2.4 for any $n > 1$ and $N > 1$

$$\mathbf{P}\left(|\tau_{n-1}| > N\right) = \mathbf{P}\left(|U| > N\sqrt{\frac{1}{n-1}\sum_{j=1}^{n-1}\eta_j^2}\right)\,.$$

Therefore, taking into account the large numbers law, we get that

$$\limsup_{n\to\infty} \mathbf{P}\left(|\tau_{n-1}| > N\right) \leq \mathbf{P}\left(|U| > N/2\right)\,.$$

tending here $N \to \infty$ we obtain that

$$\lim_{n\to\infty} \mathbf{P}(\varphi = 0 \,|\, \mathbf{H}_1) = 0\,.$$

Hence Proposition 2.6. $\square$

2.2 Confidence interval estimation

Now we remind that a confidence interval of a threshold $0 < \alpha < 1$ is a random interval $\widehat{J}_\theta = [\theta_*, \theta^*]$, where θ_* and θ^* are random variables such as

$$\mathbf{P}\left(\theta \in \widehat{J}_\theta\right) \,=\, \mathbf{P}(\theta_* \leq\, \theta \leq\, \theta^*) \,=\, 1 - \alpha\,.$$

We know that for any $n > 1$

$$\frac{\widehat{\theta}_n - \theta}{\sqrt{\widehat{\mathbf{V}}(\widehat{\theta}_n)}} \quad\sim\quad \tau_{n-1}\,.$$

We define the confidence interval for θ as

$$\widehat{J}_\theta = \left[\widehat{\theta}_n - z_\alpha \sqrt{\widehat{\mathbf{V}}(\widehat{\theta}_n)}\,,\, \widehat{\theta}_n + z_\alpha \sqrt{\widehat{\mathbf{V}}(\widehat{\theta}_n)}\right], \tag{2.24}$$

where $z_\alpha > 0$ is $1-\alpha$ quantile for the Student distribution, i.e.

$$\mathbf{P}\left(|\tau_{n-1}| \le z_\alpha\right) = 1-\alpha\,. \tag{2.25}$$

It should be noted, that asymptotically, if $\sum_{j=1}^n x_j^2 \to \infty$ as $n \to$ the length $2z_\alpha\sqrt{\widehat{\mathbf{V}}(\widehat{\theta}_n)}$ tend to zero.

2.3 Forescasting problem

Let us consider now the forecasting problem for the Gaussian model (2.1), i.e. we consider the estimation problem for the $y_{n+\mathbf{l}}$ for some fixed $\mathbf{l} \ge 1$ on the basis of the observations $y_1, \ldots, y_n$, i.e. on the basis of the σ - field $\mathcal{F}_n = \sigma\{y_1, \ldots, y_n\}$.

Proposition 2.7. *For any* $\mathbf{l} \ge 1$ *the conditional expectation* $\overline{y}_{n+\mathbf{l}} = \mathbf{E}(y_{n+\mathbf{l}}|\mathcal{F}_n)$ *is the best forecasting for* $y_{n+\mathbf{l}}$ *in* $\mathbf{L}_2(\Omega, \mathcal{F}_n, \mathbf{P})$*, i.e.*

$$\inf_{\varsigma \in \mathbf{L}_2(\Omega,\mathcal{F}_n,\mathbf{P})} \mathbf{E}(\varsigma - y_{n+\mathbf{l}})^2 = \mathbf{E}(\overline{y}_{n+\mathbf{l}} - y_{n+\mathbf{l}})^2\,. \tag{2.26}$$

Proof. Indeed, using the properties of the conditional expectations, we get that for any $\varsigma \in \mathbf{L}_2(\Omega, \mathcal{F}_n, \mathbf{P})$

$$\mathbf{E}\left(\varsigma - y_{n+\mathbf{l}}\right)^2 = \mathbf{E}\left(\overline{y}_{n+\mathbf{l}} - y_{n+\mathbf{l}}\right)^2 + \mathbf{E}\left(\varsigma - y_{n+\mathbf{l}}\right)^2.$$

This implies the property (2.26), i.e. hence Proposition 2.7. □

To this end we use the following estimation

$$\widehat{y}_{n+\mathbf{l}} = \widehat{\theta}_n\, x_{n+\mathbf{l}}\,. \tag{2.27}$$

From the model (2.1) we can obtain that the deviation from the value $y_{n+\mathbf{l}}$ is given as

$$\widehat{y}_{n+\mathbf{l}} - y_{n+\mathbf{l}} = (\widehat{\theta}_n - \theta)\, x_{n+\mathbf{l}} - \varepsilon_{n+\mathbf{l}}\,. \tag{2.28}$$

It is easy to obtain that $\mathbf{E}(\widehat{y}_{n+\mathbf{l}} - y_{n+\mathbf{l}}) = 0$ and the forecasting variance

$$\mathbf{V}(\widehat{y}_{n+\mathbf{l}}) = \mathbf{E}(\widehat{y}_{n+\mathbf{l}} - y_{n+\mathbf{l}})^2 = \sigma^2\left(1 + \frac{x_{n+\mathbf{l}}^2}{\sum_{\iota=1}^n x_\iota^2}\right). \tag{2.29}$$

Now, taking into account that the random variables $\widehat{\theta}_n$ and ε_{n+1} are independent, we can conclude that

$$\widehat{y}_{n+1} - y_{n+1} \quad \sim \quad \mathcal{N}(0, \mathbf{V}(\widehat{y}_{n+1}))\,.$$

Moreover, using the estimator (2.10), we set

$$\widehat{\mathbf{V}}(\widehat{y}_{n+1}) = \widehat{\sigma}_n \left(1 + \frac{x_{n+1}^2}{\sum_{\iota=1}^{n} x_\iota^2}\right)\,. \tag{2.30}$$

Proposition 2.8. *For the Gaussian regression model* (2.1) *with the condition* (2.3) *and* $n > 1$ *the normalized forecasting accuracy*

$$\frac{\widehat{y}_{n+1} - y_{n+1}}{\sqrt{\widehat{\mathbf{V}}(\widehat{y}_{n+1})}} \tag{2.31}$$

has Student's law with $n-1$ *degree of liberty.*

Proof. First note, that

$$\frac{\widehat{\mathbf{V}}(\widehat{y}_{n+1})}{\mathbf{V}(\widehat{y}_{n+1})} = \frac{\widehat{\sigma}_n}{\sigma^2} = \frac{\gamma_n}{n-1}\,,$$

where the random variable γ_n is given in (2.21). Therefore, the fraction (2.31) can be represented as

$$\frac{\widehat{y}_{n+1} - y_{n+1}}{\widehat{\mathbf{V}}(\widehat{y}_{n+1})} = \frac{U}{\sqrt{\gamma_n/(n-1)}}\,,$$

where

$$U = \frac{\widehat{y}_{n+1} - y_{n+1}}{\sqrt{\mathbf{V}(\widehat{y}_{n+1})}} \quad \sim \quad \mathcal{N}(0,1)\,.$$

Now the desired result follows directly from Propositions 2.3 and 2.4. □

It is clear that using this property we can define the α confidential interval for the forecasting as

$$\widehat{J}_{y,1} = [\widehat{y}_*\,,\, \widehat{y}^*]\,, \tag{2.32}$$

where

$$\widehat{y}_* = \widehat{y}_{n+1} - z_\alpha \sqrt{\widehat{\mathbf{V}}(\widehat{y}_{n+1})}\,, \qquad \widehat{y}^* = \widehat{y}_{n+1} + z_\alpha \sqrt{\widehat{\mathbf{V}}(\widehat{y}_{n+1})}\,.$$

and z_α is α - quantile defined in (2.25).

Exercises

1. To check if for the model (2.1) with $x_j = \sin(2\pi j/n)$ the least square estimator $\widehat{\theta}_n$ converges to θ or not in $\mathbf{L}_2$ as $n \to \infty$.

2. Show that if there exists n for which the condition (2.3) holds, then the estimator (2.10) converges to σ^2 in probability as $n \to \infty$.

3. Show Proposition 2.2.

4. Show Proposition 2.4.

3 Multivariate regression analysis

In this section we consider the multivariate regression model defined as

$$y_j = \theta_1 x_{j1} + \ldots + \theta_p x_{j,p} + \varepsilon_j\,, \quad 1 \le j \le n\,, \tag{3.1}$$

where $\theta = (\theta_1, \ldots, \theta_p)' \in \mathbb{R}^p$ are parameters of the model, $(\varepsilon_j)_{1\le j\le n}$ is unobservable white noise, i.e. $\mathbf{E}\varepsilon_j = 0$ and $\mathbf{E}\,\varepsilon_j^2 = \sigma^2$ for any $1 \le j \le n$ and $\mathbf{E}\,\varepsilon_j\varepsilon_l = 0$ for $j \ne l$.
We can represent this model in the matrix form

$$\begin{array}{lll} y_1 & = \theta_1 x_{1,1} + \ldots + \theta_p x_{1,p} & +\varepsilon_1 \\ \vdots & \ldots & \vdots \\ y_n & = \theta_1 x_{n,1} + \ldots + \theta_p x_{n,p} & +\varepsilon_n\,. \end{array}$$

Setting here

$$Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}\,, \quad \mathbf{X} = \begin{pmatrix} x_{1,1} & \ldots & x_{1,p} \\ \vdots & \vdots & \vdots \\ x_{n,1} & \ldots & x_{n,p} \end{pmatrix} \quad \text{and} \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}\,,$$

we get the equation in space $\mathbb{R}^n$

$$Y \,=\, \mathbf{X}\,\theta \,+\, \varepsilon\,. \tag{3.2}$$

The problem is to estimate the vector θ in $\mathbb{R}^p$ on the basis of the observations Y et X.We apply the method of least squares which consists in minimizing the sum of the squares of the errors:

$$L(\theta) \,=\, \|\varepsilon\|^2 \,=\, \|Y - \mathbf{X}\theta\|^2 \,\to\, \min_{\theta\in\mathbb{R}^p}\,. \tag{3.3}$$

To minimize this function over θ, we differentiate the function $L(\theta)$ on θ. To do this we represent the design matrix $\mathbf{X}$ as

$$\mathbf{X} = \begin{pmatrix} \widetilde{X}_1' \\ \vdots \\ \widetilde{X}_n' \end{pmatrix} \quad \text{and} \quad \widetilde{X}_j = \begin{pmatrix} x_{j,1} \\ \vdots \\ x_{j,p} \end{pmatrix} . \tag{3.4}$$

Using this representation we can obtain that

$$\begin{aligned} L(\theta) \; &= \; \|Y\|^2 - 2\, Y'\mathbf{X}\,\theta + \|\mathbf{X}\,\theta\|^2 \\ &= \sum_{j=1}^{n} y_j^2 - 2\sum_{j=1}^{n} y_j \widetilde{X}_j'\theta + \sum_{j=1}^{n} (\widetilde{X}_j'\theta)^2 \, . \end{aligned}$$

So, for any $1 \le r \le p$ the partial derivative

$$\frac{\partial L(\theta)}{\partial \theta_r} \; = \; -2\sum_{j=1}^{n} x_{j,r}\, y_j + 2\sum_{j=1}^{n} x_{j,r} \widetilde{X}_j'\theta$$

and to minimize the function $L(\cdot)$ one needs to resolve the system

$$\mathbf{X}'\mathbf{X}\,\theta = \mathbf{X}'\,Y \, . \tag{3.5}$$

To this end we assume that following condition

$\mathbf{C}_1$) *The matrix* $\mathbf{X}'\mathbf{X}$ *is positively defined.*

It clear that under this condition we obtain that the least square estimator

$$\widehat{\theta}_n = (\mathbf{X}'\,\mathbf{X})^{-1}\,\mathbf{X}'\,Y \, . \tag{3.6}$$

First note that

$$\widehat{\theta}_n = (\mathbf{X}'\,\mathbf{X})^{-1}\,\mathbf{X}'\,Y \; = \; (\mathbf{X}'\,\mathbf{X})^{-1}\,\mathbf{X}'\,(\mathbf{X}\theta + \varepsilon) \; = \theta + (\mathbf{X}'\,\mathbf{X})^{-1}\,\mathbf{X}'\,\varepsilon \, .$$

That implies that the expectation

$$\mathbf{E}\,\widehat{\theta}_n = \theta$$

and the variance

$$\begin{aligned} \mathbf{V}(\widehat{\theta}_n) \; &= \; \mathbf{E}\,(\widehat{\theta}_n - \theta)\,(\widehat{\theta}_n - \theta)' \\ &= \; \mathbf{E}\,(\mathbf{X}'\,\mathbf{X})^{-1}\,\mathbf{X}'\,\varepsilon\,\varepsilon'\,\mathbf{X}\,(\mathbf{X}\,\mathbf{X}')^{-1} \\ &= \; \sigma^2 (\mathbf{X}'\,\mathbf{X})^{-1} \, . \end{aligned} \tag{3.7}$$

From here we obtain immediately the criteria for the mean square convergence, i.e.

$$\lim_{n\to\infty} \mathbf{E}\,\|\widehat{\theta}_n - \theta\|^2 = 0$$

if and only if

$$\lim_{n\to\infty} \mathrm{tr}(\mathbf{X}'\,\mathbf{X})^{-1} = 0\,. \tag{3.8}$$

Now we show a multivariate version of the Gauss - Markov theorem 2.1

Theorem 3.1. *For any unbiased linear estimator $\widetilde{\theta}_n$ of non zero parameter θ we have*

$$\mathbf{E}(\widetilde{\theta}_n - \theta)(\widetilde{\theta} - \theta)' \succeq \mathbf{V}(\widehat{\theta}_n)\,, \tag{3.9}$$

where for two symmetric matrices A and B the notation $A \succeq B$ means that the difference $A - B$ is positively defined.

Proof. Note first that $\widetilde{\theta}_n$ is of the form $\widetilde{\theta}_n = \mathbf{G}\,Y + \mathbf{b}$, where $\mathbf{G}$ is $p \times n$ matrix and $\mathbf{b} \in \mathbb{R}^p$. Moreover, since $\widetilde{\theta}_n$ is unbiased, we obtain that $\mathbf{b} = 0$. Indeed, we have

$$\mathbf{E}\widetilde{\theta}_n = \mathbf{G}\,\mathbf{X}\theta + \mathbf{b} = \theta\,, \quad \text{i.e.} \quad (\mathbf{G}\,\mathbf{X} - \mathbf{I}_p)\theta + \mathbf{b} = 0\,.$$

As this last equality must be verified which whatever $\theta \in \mathbf{R}^p$, we can deduce that $\mathbf{b} = 0$, and at the same time

$$\mathbf{G}\,\mathbf{X} = \mathbf{I}_p. \tag{3.10}$$

Therefore,

$$\mathbf{E}(\widetilde{\theta}_n - \theta)(\widetilde{\theta}_n - \theta)' = \sigma^2\,\mathbf{G}\,\mathbf{G}'\,.$$

However, we have seen already that

$$\mathbf{V}(\widehat{\theta}_n) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\,.$$

Using here the property (3.10), the result can be deduced immediately from Theorem A.1. Hence Theorem 3.1. □

Now, similarly to the scalar case to estimate the noise variance σ^2 we will use the estimators for the noise variables in the model (3.1) for $1 \le j \le n$ defined as

$$\widehat{\varepsilon}_j = y_j - \widetilde{X}_j'\widehat{\theta}_n = \widetilde{X}_j'(\theta - \widehat{\theta}_n) + \varepsilon_j\,.$$

From here we can obtain that

$$\widehat{\varepsilon} = \begin{pmatrix} \widehat{\varepsilon}_1 \\ \vdots \\ \widehat{\varepsilon}_n \end{pmatrix} = (\mathbf{I}_n - \mathbf{A})\,\varepsilon \quad \text{and} \quad \mathbf{A} = \mathbf{X}(\mathbf{X}'\,\mathbf{X})^{-1}\,\mathbf{X}'\,. \tag{3.11}$$

Moreover, taking into account that $\mathbf{A}^2 = \mathbf{A}$ and $\mathrm{tr}A = p$ we obtain that for $n > p$

$$\mathbf{E}\,\|\widehat{\varepsilon}\|^2 = \sigma^2 \mathrm{tr}(\mathbf{I}_n - \mathbf{A})^2 = \sigma^2 \mathrm{tr}(\mathbf{I}_n - \mathbf{A}) = \sigma^2\,(n-p)\,.$$

Therefore,

$$\widehat{\sigma}_n \;=\; \frac{1}{n-p}\,\|\widehat{\varepsilon}\|^2 = \frac{1}{n-p}\,\sum_{j=1}^{n}\widehat{\varepsilon}_j^2 \tag{3.12}$$

is the unbiased estimator for the variance σ^2. Therefore, to estimate the estimation accuracy we set

$$\widehat{\mathbf{V}}(\widehat{\theta}_n) \;=\; \widehat{\sigma}_n\,(\mathbf{X}'\mathbf{X})^{-1}\,. \tag{3.13}$$

Now we need to study the distribution properties for the estimator (3.11) in the case when the model (3.1) is Gaussian, i.e. when the noise variables $(\varepsilon_j)_{1\le j\le n}$ are i.i.d. Gaussian with the parameters $(0,\sigma^2)$.

Proposition 3.1. *Assume that the condition* $\mathbf{C}_1$*) holds. Then for the Gaussian regression model* (3.1) *with* $n > p$ *the random variable*

$$\gamma_n = \frac{\sum_{j=1}^{n}\widehat{\varepsilon}_j^2}{\sigma^2} \quad\sim\quad \mathcal{X}^2_{n-p}\,. \tag{3.14}$$

Proof. First, taking into account that the matrix $\mathbf{I}_n - \mathbf{A}$ is idempotent, i.e. $(\mathbf{I}_n - \mathbf{A})^2 = \mathbf{I}_n - \mathbf{A}$, we can obtain from the representation (3.11) that

$$\gamma_n = \frac{\|\widehat{\varepsilon}\|^2}{\sigma^2} = \xi'(\mathbf{I}_n - \mathbf{A})\xi \quad\text{and}\quad \xi = \frac{1}{\sigma}\varepsilon\,. \tag{3.15}$$

We remind that the $\mathrm{tr}\mathbf{A} = p$, therefore this matrix has $n-p$ eigenvalues 0 and p eigenvalues 1. Therefore, there exists an orthogonal $n\times n$ matrix $\mathbf{Q}$, i.e. $\mathbf{Q}'\mathbf{Q} = \mathbf{I}_n$, such that

$$\mathbf{I}_n - \mathbf{A} = \mathbf{Q}' J_n \mathbf{Q} \quad\text{and}\quad J_n = \mathrm{diag}(\lambda_1,\ldots,\lambda_n) = \begin{bmatrix} \lambda_1 & \ldots & 0 \\ 0 & \ddots & 0 \\ 0 & \ldots & \lambda_n \end{bmatrix}, \tag{3.16}$$

where $\lambda_1 = \ldots = \lambda_{n-p} = 1$ and $\lambda_{n-p+1} = \ldots = \lambda_n = 0$. Using this form in (3.15), we obtain that

$$\gamma_n = \sum_{j=1}^{n-p}\eta_j^2 \quad\text{and}\quad \eta = \begin{pmatrix}\eta_1\\ \vdots \\ \eta_n\end{pmatrix} = \mathbf{Q}\xi \quad\sim\quad \mathcal{N}(0,\mathbf{I}_n)\,. \tag{3.17}$$

Now the desired result directly follows from Proposition 2.2. □

Now using this property we can show the estimation accuracy.

Proposition 3.2. *Assume, that the condition* $\mathbf{C}_1$*) holds true and* $n > p$ *and in the model* (3.1) *the noise variables* $(\varepsilon_j)_{1\le j\le n}$ *are i.i.d. Gaussian with the parameters* $(0, \sigma^2)$*. Then for any non random non zero vector* $\mathbf{u}$ *from* $\mathbb{R}^p$ *the normalized linear combination*

$$\frac{\mathbf{u}'(\widehat{\theta}_n - \theta)}{\sqrt{\mathbf{u}'\widehat{\mathbf{V}}(\widehat{\theta}_n)\mathbf{u}}} \quad \sim \quad \tau_{n-p}\,,$$

i.e. has the Student distribution with $n - p$ *degree of liberty.*

Proof. First, note that in view of (3.7) for the Gaussian model (3.1) for any non zero $\mathbf{u} \in \mathbb{R}^p$

$$\mathbf{u}'(\widehat{\theta}_n - \theta) \,= \sigma \mathbf{u}' \left(\mathbf{X}'\,\mathbf{X}\right)^{-1} \mathbf{X}'\,\xi \quad \sim \quad \mathcal{N}\left(0, \mathbf{u}'\mathbf{V}(\widehat{\theta}_n)\mathbf{u}\right)\,,$$

where the vector ξ is defined in (3.15). Therefore, using the definitions (3.12) - (3.14) we can write that

$$\frac{\mathbf{u}'(\widehat{\theta}_n - \theta)}{\sqrt{\mathbf{u}'\widehat{\mathbf{V}}(\widehat{\theta}_n)\mathbf{u}}} = \frac{U}{\sqrt{\gamma_n/(n-p)}} \tag{3.18}$$

and

$$U = \frac{\mathbf{u}'(\widehat{\theta}_n - \theta)}{\sqrt{\mathbf{u}'\mathbf{V}(\widehat{\theta}_n)\mathbf{u}}} \quad \sim \quad \mathcal{N}(0,1)\,.$$

Moreover, from (3.11) we can obtain the covariance matrix

$$\mathbf{E}\,\widehat{\varepsilon}\,(\widehat{\theta}_n - \theta)' = \sigma^2(\mathbf{I}_n - \mathbf{A})\,\mathbf{X}\,(\mathbf{X}'\,\mathbf{X})^{-1} = \sigma^2(\mathbf{X}\,(\mathbf{X}'\,\mathbf{X})^{-1} - \mathbf{X}\,(\mathbf{X}'\,\mathbf{X})^{-1}) = 0\,.$$

We note that for the Gaussian model (3.1) this property implies that in the representation (3.18) the random variables U and γ_n are independent. Therefore, using here Proposition 2.4 and Proposition 3.1, we come to the desired result. □

Definition 3.1. *Let* $p \ge 1$ *and* $q \ge 1$ *be two integers. The random variable* ζ *follows the Fisher-Snedecor law with the* p *and* q *freedom degrees denoted as* $\mathbf{f}_{p,q}$ *if it has density*

$$\varrho_{p,q}(u) = \frac{\Gamma(\frac{p+q}{2})}{\Gamma(\frac{p}{2})\Gamma(\frac{q}{2})} \left(\frac{p}{q}\right)^{\frac{p}{2}} \frac{u^{\frac{p}{2}-1}}{\left(1 + \frac{p}{2}\,u\right)^{\frac{p+q}{2}}}\,\mathbf{1}_{\{u\ge 0\}}\,, \tag{3.19}$$

i.e. for any measurable set $A \subset \mathbb{R}$

$$\mathbf{P}(\zeta \in A) = \int_A \varrho_{p,g}(u)\mathrm{d}u\,.$$

In the time series analysis such distributions are used to study the fraction of the two independent $\mathcal{X}^2$ random variables.

Proposition 3.3. *Let U and V be two independent random variables of the laws $\mathcal{X}^2_p$ and $\mathcal{X}^2_q$ respectevely. Then the fraction*

$$\frac{U/p}{V/q} \quad \sim \quad \mathbf{f}_{p,q}\,,$$

i;e. follows Fisher-Snedecor law with p and q freedom degrees.

Proposition 3.4. *Assume, that the condition $\mathbf{C}_1$) holds true and $n > p$ and in the model (3.1) the noise variables $(\varepsilon_j)_{1\le j\le n}$ are i.i.d. Gaussian with the parameters $(0, \sigma^2)$. Then the normalized quadratic form*

$$\frac{(\widehat{\theta}_n - \theta)'\widehat{\mathbf{V}}^{-1}(\widehat{\theta}_n)(\widehat{\theta}_n - \theta)}{p} \quad \sim \quad \mathbf{f}_{p,n-p} \tag{3.20}$$

i.e. has the Fischer - Snedecor distribution with p and $n-p$ freedom degrees.

Proof. First note that using the matrix $\mathbf{A}$ introduced in (3.11) and the Gaussian random variables defined in (3.17), we obtain that

$$(\widehat{\theta}_n - \theta)'\,\mathbf{V}^{-1}(\widehat{\theta}_n)\,(\widehat{\theta}_n - \theta) = \xi'\mathbf{A}\xi = \sum_{j=n-p+1}^{n} \eta_j^2 \quad \sim \quad \mathcal{X}^2_p\,.$$

Moreover, through the definitions (3.12) – (3.14), we can write that

$$\widehat{\mathbf{V}}(\widehat{\theta}_n) = \frac{\gamma_n}{n-p}\,\mathbf{V}(\widehat{\theta}_n)$$

and, therefore,

$$\frac{(\widehat{\theta}_n - \theta)'\widehat{\mathbf{V}}^{-1}(\widehat{\theta}_n)(\widehat{\theta}_n - \theta)}{p} = \frac{(\widehat{\theta}_n - \theta)'V^{-1}(\widehat{\theta}_n - \theta)/p}{\gamma_n/(n-p)}\,.$$

Taking into account here that $\widehat{\theta}_n$ and γ_n are independent, we obtain through Proposition 3.14 and Proposition 3.3 the distribution of the quadratic form in (3.20). Hence Proposition 3.4. $\square$

It should be noted that, in practice very often we need to estimate not the vector θ, but a linear transformation of this vector, i.e $\theta_T = T\theta$ for a fixed matrix T. In this case we have to use the estimator $\widehat{\theta}_{T,n} = T\widehat{\theta}_n$. One can obtain directly, that $\mathbf{E}\,\widehat{\theta}_{T,n} = \theta_T$ and

$$\mathbf{V}(\widehat{\theta}_{T,n}) = \mathbf{E}\,(\widehat{\theta}_{T,n} - \theta_T)\,(\widehat{\theta}_{T,n} - \theta_T)' = T\mathbf{V}(\widehat{\theta}_n)T'\,.$$

Similarly to (3.13) to estimate the mean square accuracy for the vector θ_T we set

$$\widehat{\mathbf{V}}(\widehat{\theta}_{T,n}) = T\widehat{\mathbf{V}}(\widehat{\theta}_n)T' \,. \tag{3.21}$$

Now we study the distribution of this vector for the Gaussian model (3.1).

Proposition 3.5. *Assume, that the condition* $\mathbf{C}_1$*) holds true and* $n > p$ *and in the model* (3.1) *the noise variables* $(\varepsilon_j)_{1\le j\le n}$ *are i.i.d. Gaussian with the parameters* $(0, \sigma^2)$*. Then for any non random* $m \times p$ *matrix* T *with* $m \le p$ *and* $Rang T = m$ *the normalized quadratic form*

$$\frac{(\widehat{\theta}_{T,n} - \theta_T)'\widehat{\mathbf{V}}^{-1}(\widehat{\theta}_{T,n})(\widehat{\theta}_{T,n} - \theta_T)}{m} \quad \sim \quad \mathbf{f}_{m,n-p} \tag{3.22}$$

i.e. has the Fischer - Snedecor distribution with m *and* $n-p$ *freedom degrees.*

Proof. First, note that for the gaussian model (3.1)

$$\widehat{\theta}_{T,n} - \theta_T \quad \sim \quad \mathcal{N}(0, \mathbf{V}(\widehat{\theta}_{T,n}) \quad \text{and} \quad \mathbf{V}(\widehat{\theta}_{T,n}) = T\mathbf{V}(\widehat{\theta}_n)T' \,.$$

It is clear, that if $\operatorname{Rang} T = m$, then $\mathbf{V}(\widehat{\theta}_{T,n})$ is invertible. In addition, we have

$$\begin{aligned}(\widehat{\theta}_{T,n} - \theta_T)'\,\mathbf{V}^{-1}(\widehat{\theta}_{T,n})(\widehat{\theta}_{T,n} - \theta_T) &= (\widehat{\theta}_n - \theta)'T'\mathbf{V}^{-1}(\widehat{\theta}_{T,n})T(\widehat{\theta}_n - \theta) \\ &= \eta\eta' \,,\end{aligned}$$

where $\eta = \mathbf{S}^{-1}T(\mathbf{X}'\mathbf{X})^{-1}\,\mathbf{X}\xi$ and $\mathbf{S}$ is a symmetric matrix, i.e. $\mathbf{S}' = \mathbf{S}$, such that $\mathbf{S}^2 = T(\mathbf{X}\mathbf{X}')^{-1}T'$. Note that, for the Gaussian model (3.1) the vector $\xi = \varepsilon/\sigma$ is a Gaussian vector with the parameters $\mathcal{N}(0, \mathbf{I}_n)$ and, therefore, the vector η is gaussian also with $\mathbf{E}\,\eta = 0$ and

$$\begin{aligned}\mathbf{E}\,\eta\,\eta' &= \mathbf{E}\,\mathbf{S}^{-1}\,T(\mathbf{X}'\,\mathbf{X})^{-1}\,\mathbf{X}'\,\xi\,\xi'\,\mathbf{X}(\mathbf{X}'\,\mathbf{X})^{-1}T'\,\mathbf{S}^{-1} \\ &= \mathbf{S}^{-1}\,T(\mathbf{X}'\,\mathbf{X})^{-1}\,T\,\mathbf{S}^{-1} = \mathbf{I}_m \,.\end{aligned}$$

This means that teh quadratic form

$$\mathbf{D}_T = (\widehat{\theta}_{T,n} - \theta_T)'\mathbf{V}^{-1}(\widehat{\theta}_{T,n})\,(\widehat{\theta}_{T,n} - \theta_T) \quad \sim \quad \mathcal{X}^2_m \,.$$

It is clear that from (3.21) it follows that

$$\widehat{\mathbf{V}}(\widehat{\theta}_{T,n}) = \frac{\gamma_n}{n-p}\,\mathbf{V}(\widehat{\theta}_{T,n})$$

and, therefore,

$$\frac{(\widehat{\theta}_{T,n} - a_T)'\widehat{\mathbf{V}}^{-1}(\widehat{\theta}_{T,n})(\widehat{\theta}_{T,n} - a_T)}{m} = \frac{\mathbf{D}_T/m}{\gamma_n/(n-p)} \,. \tag{3.23}$$

Then, taking into account that the random variable γ_n and the vector $\widehat{\theta}_{T,n}$ are independent, we can conclude through Proposition 3.14 and Proposition 3.3 that the fraction (3.23) follows Fisher-Snedecor law with m and $n-p$ freedom degrees. Hence Proposition 3.5. $\square$

3.1 Coefficients testing

- **Comparison of linear combinations of parameters**

  First for the gaussian model (3.1) we consider the hypothesis testing problem for a linear combination of the parameters, i.e.

$$\mathbf{H}_0 : \quad \mathbf{u}'\theta = \mathbf{u}'\overline{\theta} \quad \text{and} \quad \mathbf{H}_1 \quad \mathbf{u}'\theta \neq \mathbf{u}'\overline{\theta}\,, \tag{3.24}$$

  where $\mathbf{u} \neq 0$ and θ are some fixed vectors from $\mathbb{R}^p$. It should be noted if the combination vector

$$\mathbf{u} = (\underbrace{0, \ldots, 1}_{i}, 0 \ldots, 0)'\,,$$

  then we obtain the the testing problem for the ith parameter θ_i, i.e.

$$\mathbf{H}_0 : \quad \theta_i = \overline{\theta}_i \quad \text{and} \quad \mathbf{H}_1 \quad \theta_i \neq \overline{\theta}_i\,.$$

  Note that, under the hypothesis $\mathbf{H}_0$ in view of Proposition 3.2 we have

$$\frac{\mathbf{u}'(\widehat{\theta}_n - \overline{\theta})}{\sqrt{\mathbf{u}'\widehat{\mathbf{V}}(\widehat{\theta}_n)\mathbf{u}}} \quad \sim \quad \tau_{n-p}\,.$$

  Now, we denote by $z_\alpha > 0$ the quantile of threshold $1 - \alpha$ for $|\tau_{n-p}|$, i.e.

$$\mathbf{P}\left(|\tau_{n-p}| \leq z_\alpha\right) = 1 - \alpha\,.$$

  Therefore, if

$$\left| \frac{\mathbf{u}'(\widehat{\theta}_n - \overline{\theta})}{\sqrt{\mathbf{u}'\widehat{\mathbf{V}}(\widehat{\theta}_n)\mathbf{u}}} \right| > z_\alpha$$

  then we reject the hypothesis $\mathbf{H}_0$, i.e. the linear combination $\mathbf{u}'\theta$ is significantly different from $\mathbf{u}'\overline{\theta}$ (at the threshold α), if not, i.e.

$$\left| \frac{\mathbf{u}'(\widehat{\theta}_n - \overline{\theta})}{\sqrt{\mathbf{u}'\widehat{\mathbf{V}}(\widehat{\theta}_n)\mathbf{u}}} \right| \leq z_\alpha$$

  then we accept the hypothesis $\mathbf{H}_0$, the linear combination $\mathbf{u}'\theta$ is not significantly different from $\mathbf{u}'\overline{\theta}$ (at the threshold α).

- **Comparison of set of parameters with fixed values.**

  Now we consider the hypothesis testing problem for the equality of a subset of regression coefficients to some fixed values:

$$\mathbf{H}_0 : \theta_1 = \overline{\theta}_1 \,, \ldots \,, \theta_m = \overline{\theta}_m \quad \text{and} \quad \mathbf{H}_1 : \exists\, 1 \le i \le m : \quad \theta_i \ne \overline{\theta}_i \,. \tag{3.25}$$

  To this end we set the $m \times p$ - matrix T as

$$T = \begin{bmatrix} 1 & 0 & \ldots & \ldots & \ldots & 0 \\ 0 & 1 & \ldots & \ldots & \ldots & 0 \\ \vdots & & \ddots & & & \vdots \\ 0 & \ldots & 0 & 1 & \ldots & 0 \end{bmatrix} . \tag{3.26}$$

  So in this case we can rewrite the problem (3.25) as

$$\mathbf{H}_0 : \theta_T = \overline{\theta} \quad \text{and} \quad \mathbf{H}_1 : \theta_T \ne \overline{\theta} \,, \tag{3.27}$$

  where $a_T = Ta$ and $\overline{\theta} = (\overline{\theta}_1, \ldots, \overline{\theta}_m)$. So, in view of Proposition 3.5 under the hypothesis $\mathbf{H}_0$) the quadratic form of the inversed matrix (3.13)

$$\frac{1}{m}(\widehat{\theta}_T - \overline{\theta})' \widehat{\mathbf{V}}^{-1}(\widehat{\theta}_T)(\widehat{\theta}_T - \overline{\theta}) \quad \sim \quad \mathbf{f}_{m,n-p} \,,$$

  i.e. has the Fisher-Snedecor distribution with the m and $n-p$ liberty degrees. Let $z_\alpha > 0$ be the quantile of threshold $1-\alpha$ for the Fisher-Snedecor random variable $\mathbf{f}_{m,n-p}$, i.e.

$$\mathbf{P}(\mathbf{f}_{m,n-p} \le z_\alpha) = 1 - \alpha \,.$$

  Therefore we accept the hypothesis $\mathbf{H}_0$ if

$$\frac{1}{m}(\widehat{\theta}_T - \overline{\theta})' \widehat{\mathbf{V}}^{-1}(\widehat{\theta}_T)(\widehat{\theta}_T - \overline{\theta}) \le z_\alpha$$

  and we reject it otherwise.

3.2 Confidence interval estimation

Now for some fixed vector $\mathbf{u} \ne 0$ in $\mathbb{R}^p$ we estimate a liner combination $\mathbf{u}'\theta$ via the confidence interval of a threshold $0 < \alpha < 1$. Similarly to (2.24) we set

$$\widehat{J}_{\mathbf{u}'\theta} = \left[\mathbf{u}'\widehat{\theta}_n - z_\alpha \sqrt{\mathbf{u}'\widehat{\mathbf{V}}(\widehat{\theta}_n)\mathbf{u}} \,,\, \mathbf{u}'\widehat{\theta}_n + z_\alpha \sqrt{\mathbf{u}'\widehat{\mathbf{V}}(\widehat{\theta}_n)\mathbf{u}} \right] ,$$

where the matrix $\widehat{\mathbf{V}}(\widehat{\theta}_n)$ is defined in (3.13) and $z_\alpha > 0$ is $1-\alpha$ quantile for the Student distribution, i.e.

$$\mathbf{P}\left(|\tau_{n-p}| \le z_\alpha\right) = 1 - \alpha \,. \tag{3.28}$$

Note that in view of Proposition 3.2 for $n > p$ the normalized deviation

$$\frac{\mathbf{u}'(\widehat{\theta}_n - \theta)}{\sqrt{\mathbf{u}'\widehat{\mathbf{V}}(\widehat{\theta}_n)\mathbf{u}}} \quad \sim \quad \tau_{n-p}\,,$$

i.e. has the Student distribution with $n-p$ liberty degrees. Therefore, through the definition of z_α in (3.28), we obtain that

$$\mathbf{P}\left(\mathbf{u}'\theta \in \widehat{J}_{\mathbf{u}'\theta}\right) = \mathbf{P}\left(|\tau_{n-p}| \leq z_\alpha\right) \,=\, 1-\alpha\,.$$

It should be noted, that asymptotically, under the condition (3.8) the length $2z_\alpha\sqrt{\mathbf{u}'\widehat{\mathbf{V}}(\widehat{\theta}_n)\mathbf{u}}$ tend to zero.

3.3 Forescasting problem

Let us consider now the forecasting problem for the Gaussian model (2.1), i.e. we consider the estimation problem for the $y_{n+\mathbf{l}}$ for some fixed $\mathbf{l} \geq 1$ on the basis of the observations $y_1, \ldots, y_n$, i.e. on the basis of the σ - field $\mathcal{F}_n = \sigma\{y_1, \ldots, y_n\}$ and $n > p$. To this end, using the vector (3.4) we set

$$\widehat{y}_{n+\mathbf{l}} = \widetilde{X}'_{n+\mathbf{l}}\widehat{\theta}_n\,. \tag{3.29}$$

From the model (2.1) we can obtain that the deviation from the value $y_{n+\mathbf{l}}$ is given as

$$\widehat{y}_{n+\mathbf{l}} - y_{n+\mathbf{l}} = \widetilde{X}'_{n+\mathbf{l}}(\widehat{\theta}_n - \theta) - \varepsilon_{n+\mathbf{l}}\,. \tag{3.30}$$

It is easy to obtain that $\mathbf{E}(\widehat{y}_{n+\mathbf{l}} - y_{n+\mathbf{l}}) = 0$ and the forecasting variance

$$\mathbf{V}(\widehat{y}_{n+\mathbf{l}}) = \mathbf{E}(\widehat{y}_{n+\mathbf{l}} - y_{n+\mathbf{l}})^2 = \sigma^2\,\left(1 + \widetilde{X}'_{n+\mathbf{l}}(\mathbf{X}'\,\mathbf{X})^{-1}\widetilde{X}_{n+\mathbf{l}}\right)\,. \tag{3.31}$$

Now, taking into account that the random variables $\widehat{\theta}_n$ and $\varepsilon_{n+\mathbf{l}}$ are independent, we can conclude that

$$\widehat{y}_{n+\mathbf{l}} - y_{n+\mathbf{l}} \quad \sim \quad \mathcal{N}(0, \mathbf{V}(\widehat{y}_{n+\mathbf{l}}))\,.$$

Moreover, using the estimator (3.12), we set

$$\widehat{\mathbf{V}}(\widehat{y}_{n+\mathbf{l}}) = \widehat{\sigma}_n\,\left(1 + \widetilde{X}'_{n+\mathbf{l}}(\mathbf{X}'\,\mathbf{X})^{-1}\widetilde{X}_{n+\mathbf{l}}\right)\,. \tag{3.32}$$

Proposition 3.6. *For the Gaussian regression model* (2.1) *with the condition* (2.3) *and* $n > 1$ *the normalized forecasting accuracy*

$$\frac{\widehat{y}_{n+\mathbf{l}} - y_{n+\mathbf{l}}}{\sqrt{\widehat{\mathbf{V}}(\widehat{y}_{n+\mathbf{l}})}} \tag{3.33}$$

has Student's law with $n-p$ *degree of liberty.*

Proof. First note, that

$$\frac{\widehat{\mathbf{V}}(\widehat{y}_{n+1})}{\mathbf{V}(\widehat{y}_{n+1})} = \frac{\widehat{\sigma}_n}{\sigma^2} = \frac{\gamma_n}{n-p}\,,$$

where the random variable γ_n is given in (3.14). Therefore, the fraction (2.31) can be represented as

$$\frac{\widehat{y}_{n+1} - y_{n+1}}{\widehat{\mathbf{V}}(\widehat{y}_{n+1})} = \frac{U}{\sqrt{\gamma_n/(n-p)}}\,,$$

where

$$U = \frac{\widehat{y}_{n+1} - y_{n+1}}{\sqrt{\mathbf{V}(\widehat{y}_{n+1})}} \quad \sim \quad \mathcal{N}(0,1)\,.$$

Now the desired result follows directly from Propositions 2.3 and 2.4. □
It is clear that using this property we can define the α confidential interval for the forecasting as

$$\widehat{J}_{y,1} = [\widehat{y}_*\,,\,\widehat{y}^*]\,, \tag{3.34}$$

where

$$\widehat{y}_* = \widehat{y}_{n+1} - z_\alpha\sqrt{\widehat{\mathbf{V}}(\widehat{y}_{n+1})}\,, \qquad \widehat{y}^* = \widehat{y}_{n+1} + z_\alpha\sqrt{\widehat{\mathbf{V}}(\widehat{y}_{n+1})}\,.$$

and z_α is α - quantile defined in (3.28).

Exercises

1. Let ζ be a random variable distributed as $\mathbf{f}_{p,q}$. Show that

$$\mathbf{E}\zeta = \frac{q}{q-2} \quad \text{for} \quad q > 2$$

and

$$\mathbf{E}(\zeta - \mathbf{E}\,\zeta)^2 = \frac{2q^2(p+q-2)}{p(q-2^2)(q-4)} \quad \text{for} \quad q > 4\,.$$

2. Show Proposition 3.3.

4 Big Data models

Now we consider the parameter estimation problem for the model (3.2) in the case, when the Big Data setting, i.e. the problem is to estimate the vector $\theta = (\theta_1, \dots, \theta_p)' \in \mathbb{R}^p$ under the condition that $p > n$. We recall, that to find the least square estimators one needs to resolve the system (3.5). Note, that in the case when $p > n$ the matrix $\mathbf{X}'\mathbf{X}$ is degenerated, i.e. it is not invertible. The linear subspace $\{\theta \in \mathbb{R}^p : \mathbf{X}\theta = y\}$ of the dimension $m - n \geq 1$ gives the solution for the minimization problem (3.3).

4.1 LASSO criteria

LASSO (Least Absolute Shrinkage and Selection Operator) (see, for example, in [7]). The idea is to modify the cost function (3.3).

$$\sum_{j=1}^{n} \left(y_j - \sum_{l=1}^{p} x_{j,l}\theta_l \right)^2 + \sum_{j=1}^{p} |\theta_j| \quad \to \quad \min_{\theta \in \mathbb{R}^p}$$

In the sequel it was appeared different modifications

$$\sum_{j=1}^{n} \mathbf{V}_1 \left(y_j - \sum_{l=1}^{p} x_{j,l}\theta_l \right) + \sum_{j=1}^{p} \mathbf{V}_2(\theta_j) \quad \to \quad \min_{\theta \in \mathbb{R}^p},$$

where $\mathbf{V}_1(\cdot)$ and $\mathbf{V}_2(\cdot)$ are functions such that, for example, $\mathbf{V}_i(x) = |x|^{\gamma_i}$ for some $\gamma_i > 0$. Note, that if $\mathbf{V}_1(x) = x^2$ and $\mathbf{V}_2(x) = \delta x^2$ for some $\delta > 0$, we obtain the Tikhonov regularisation procedure, i.e.

$$\widehat{\theta}_n = (\mathbf{X}'\,\mathbf{X} + \delta I_p)^{-1}\,\mathbf{X}'\,Y\,, \tag{4.1}$$

where I_p is the identity matrix of order $p \geq 1$, the matrix $\mathbf{X}$ and the vector Y are defined in (3.2).

4.2 Dantzig selector

The another modification was proposed by Candes and Tao (see, for example, in [7])

$$\min_{\theta \in \mathbb{R}^p} \sum_{j=1}^{p} |\theta_j| \quad \text{subject to} \quad |y - \mathbf{X}\theta|_\infty \leq \epsilon$$

where for the vector $z = (z_1, \dots, z_n)'$ and the norm is defined as

$$|z|_\infty = \max_{1 \leq j \leq n} |z_j|\,.$$

The main difficulty is that usually these estimators can not be found in explicit form. Moreover, to calculate these estimators one needs to know the parameter dimension p.

5 Autoregressive and Moving Averaging processes (ARMA)

We start by the definitions.

Definition 5.1. *The process $(y_t)_{t\geq 0}$ is called the moving averaging process of the order $q \geq 1$ denoted as $MA(q)$ if it can be represented as*

$$y_t = \varepsilon_t + \phi_1 \varepsilon_{t-1} + \ldots + \phi_q \varepsilon_{t-q}\,, \tag{5.1}$$

where $\phi_1, \ldots, \phi_q$ are some fixed non random parameters, $(\varepsilon_t)_{-\infty<t<+\infty}$ are i.i.d. random variables with $\mathbf{E}\,\varepsilon_t = 0$ and $\mathbf{E}\,\varepsilon_t^2 < \infty$. The process $(y_t)_{t\geq 0}$ is called the autoregressive process of order $p \geq 1$ denoted as $AR(p)$, if it can be represented as

$$y_t = \theta_1 y_{t-1} + \ldots + \theta_p y_{t-p} + \varepsilon_t\,, \tag{5.2}$$

where $\theta_1, \ldots, \theta_p$ are some fixed non random parameters. The process $(y_t)_{t\geq 0}$ is called the autoregressive process of orders $p \geq 1$ and $q \geq 1$ denoted as $ARMA(p,q)$, if it can be represented as

$$y_t = \theta_1 y_{t-1} + \ldots + \theta_p y_{t-p} + \varepsilon_t + \phi_1 \varepsilon_{t-1} + \ldots + \phi_q \varepsilon_{t-q}\,. \tag{5.3}$$

Note also, that any time series $(y_t)_{t\geq 1}$ for which there exists a sequence $(\psi_k)_{k\geq 1}$ with $\sum_{k\geq 1} \psi_k^2 < \infty$ such that this series can be represented as

$$y_t = \varepsilon_t + \sum_{k=1}^{+\infty} \psi_k\, \varepsilon_{t-k} \tag{5.4}$$

is called *causal time series* and denoted as $MA(+\infty)$. And this series is called *invertible* if there exists a sequence $(\pi_k)_{k\geq 0}$

$$\varepsilon_t = \sum_{k=0}^{+\infty} \pi_k\, y_{t-k} \quad \text{and} \quad \sum_{k\geq 0} |\pi_k| < \infty\,, \tag{5.5}$$

where $(\varepsilon_t)_{-\infty<t<+\infty}$ are i.i.d. random variables with $\mathbf{E}\,\varepsilon_t = 0$ and $\mathbf{E}\,\varepsilon_t^2 < \infty$. Let now we represent the (5.2) in the vector form. To this end we set

$$X_t = \begin{pmatrix} y_t \\ \vdots \\ y_{t-p+1} \end{pmatrix}, \quad A = \begin{pmatrix} \theta_1 & \dots & \theta_p \\ 1 & 0 \dots & 0 \\ \vdots & \vdots & \vdots \\ 0 & \dots 1 & 0 \end{pmatrix} \quad \text{and} \quad \xi_t = \begin{pmatrix} \varepsilon_t \\ 0 \\ \vdots \\ 0 \end{pmatrix}. \tag{5.6}$$

Using these vectors, we can represent the process (5.2) as the autoregressive process of the first order in $\mathbb{R}^p$, i.e.

$$X_t = AX_{t-1} + \xi_t \,. \tag{5.7}$$

It clear, that if this process is stationary, then it can be represented in $\mathbb{R}^p$ as

$$X_t = \sum_{j=0}^{+\infty} A^j \xi_{t-j} \,. \tag{5.8}$$

This is possible if and only if all eigenvalues are less than one in the modules. Since if all the modules of the eigenvalues are less than one, then there exist $\mathbf{c} > 0$ and $0 < \varrho < 1$ such that for all $n \geq 1$

$$|A^n| \leq \mathbf{c}\varrho^n \,, \tag{5.9}$$

where $|\cdot|$ is the Euclidean norm of the matrix, i.e. $|A|^2 = \mathrm{tr} A\, A'$. Let us calculate now the eigenvalues of the matrix A defined in (5.6). To this end one needs to calculate the determinant of the matrix $A - \lambda\, I_p$ for any scalar $\lambda \in \mathbb{C}$, where I_p is the identity matrix of the order $p \geq 1$. To this end we set

$$\Delta_p = \det(A - \lambda\, I_p) = \det \begin{pmatrix} \theta_1 - \lambda & \dots & \theta_p \\ 1 & -\lambda \dots & 0 \\ \vdots & \vdots & \vdots \\ 0 & \dots 1 & -\lambda \end{pmatrix}.$$

This determinant can be represented as

$$\Delta_p = (-1)^{p+1}\theta_p \det \begin{pmatrix} 1; & -\lambda \dots & 0 \\ 0 & 1; -\lambda \dots & 0 \\ \vdots & \vdots & \vdots \\ 0 & \dots & 1; -\lambda \\ 0 & \dots 0 & 1 \end{pmatrix} - \lambda\Delta_{p-1} = -\lambda\Delta_{p-1} + (-1)^{p+1}\theta_p \,.$$

From this and taking into account, that $\Delta_1 = \theta_1 - \lambda$ we get, that

$$\Delta_p = (-\lambda)^{p-1}\Delta_1 + \sum_{j=2}^{p}(-\lambda)^{p-j}(-1)^{j+1}\theta_j$$
$$= (-\lambda)^{p-1}(\theta_1 - \lambda) + (-1)^{p+1}\sum_{j=2}^{p}\lambda^{p-j}\theta_j = (-1)^p\Theta_p(\lambda)\,,$$

where

$$\Theta_p(\lambda) = \lambda^p - \sum_{j=1}^{p}\lambda^{p-j}\theta_j\,. \tag{5.10}$$

The function $\Theta_p(\lambda)$ is called the *characteristic polynomial* of the process (5.2).

Theorem 5.1. *The linear difference equation* (5.4) *has a stationary solution if and only if all roots of the characteristic polynomial* (5.10) *in the module are less than one.*

Now it should be noted, that for the stationary process (5.2) the covariation matrix is

$$\mathbf{E}X_t\,X_t^{'} = F = \sum_{j\geq 0}A^jB(A^{'})^j \quad \text{and} \quad B = \begin{pmatrix} 1 & 0\ldots & 0 \\ \vdots & \vdots & \vdots \\ 0 & \ldots 0 & 0 \end{pmatrix}. \tag{5.11}$$

One can check directly, that (see, for example, in [1]) that the matrix F is positive defined. Moreover, one can show, that through the large number law for stationary processes

$$\lim_{n\to\infty}\frac{1}{n}\sum_{j=1}^{n}X_j\,X_j^{'} = F \quad \text{a.s..} \tag{5.12}$$

From (5.8) it follows directly, that

$$y_t = \sum_{j=0}^{+\infty} < A^j >_{11}\varepsilon_{t-j}\,, \tag{5.13}$$

where $< A >_{ij}$ denotes the (i,j) element of the matrix A. Therefore, the process (5.2) is causal if the root of the polynomial (5.10) are less than one in the modules. It is clear, that the process (5.3) is inversible if the roots of the polynomial $\Phi_q(\lambda) = \lambda^q + \phi_1\lambda^{q-1} + \ldots + \phi_q$ are less than one in the modules. Now we recall the definition of the partial autocorrelation coefficient. For the weak stationary process $(y_t)_{t\geq 1}$ we set for $l \geq 2$

$$\mathbf{r}(l) = \frac{\mathbf{E}(\widetilde{y}_{l+1} - \mathrm{Pr}_l(\widetilde{y}_{l+1}))(\widetilde{y}_1 - \mathrm{Pr}_l(\widetilde{y}_1))}{\mathbf{E}(\widetilde{y}_{l+1} - \mathrm{Pr}_l(\widetilde{y}_{l+1}))^2}\,, \tag{5.14}$$

where $\widetilde{y}_j = \widetilde{y}_j - \mathbf{E}\, y_j$ and $\mathrm{Pr}_l(\xi)$ is the projection in $\mathcal{L}_2(\Omega, \mathcal{F}, \mathbf{P})$ of the random variable ξ into the linear subspace $\mathrm{Vect}(\widetilde{y}_1, \ldots, \widetilde{y}_l)$, i.e. $\mathrm{Pr}_l(\xi) = \sum_{j=2}^{l} \lambda_j^* \widetilde{y}_j$, where the coefficients $\lambda_2^*, \ldots, \lambda_l^*$ are such that

$$\mathbf{E}\,(\xi - \sum_{j=2}^{l} \lambda_j^* \widetilde{y}_j)^2 = \min_{\lambda_2, \ldots, \lambda_l} \mathbf{E}\,(\xi - \sum_{j=2}^{l} \lambda_j \widetilde{y}_j)^2 \,.$$

Moreover, for $l = 1$ we set $\mathbf{r}(1) = \mathbf{E}\widetilde{y}_2\widetilde{y}_1 / \mathbf{E}\widetilde{y}_2^2$. It should be noted also that for the autoregressive model the partial correlation coefficient $\mathbf{r}(l) = 0$ for $l \geq p$. Moreover, as to the forecasting problem, note, that for any $l \geq 1$ from (5.8) we get, that

$$y_{t+l} = \sum_{j=1}^{p} < A^l >_{1,j} y_{t+1-j} + \sum_{j=0}^{l-1} < A^j >_{1,1} \varepsilon_{t+l-j} \,. \tag{5.15}$$

Therefore, for any $t \geq p$

$$\mathbf{E}\left(y_{t+l} \,|\, y_1, \ldots, y_t\right) = \sum_{j=1}^{p} < A^l >_{1,j} y_{t+1-j} \,, \tag{5.16}$$

i.e. if we know the parameters $\theta_1, \ldots, \theta_p$, then the optimal forecasting is the conditional expectation (5.16).

A Appendix

A.1 Matrix calculus

Lemma A.1. *Let* $\mathbf{X}$ *and* $\mathbf{U}$ *two matrices of orders* $n \times p$ *and* $p \times n$ *respectively such that* $\mathbf{UX} = \mathbf{X}'\mathbf{X}$. *Then* $\mathbf{UU}' \succeq \mathbf{X}'\mathbf{X}$.

Proof. First, note that the $p \times n$ matrix $\mathbf{V} = \mathbf{U} - \mathbf{X}'$ is orthogonal to $\mathbf{X}$, i.e. $\mathbf{V}\,\mathbf{X} = 0$. Therefore,

$$\begin{aligned} \mathbf{UU}' &= (\mathbf{X}' + \mathbf{V})(\mathbf{X} + \mathbf{V}') \\ &= \mathbf{X}'\mathbf{X} + \mathbf{VX} + \mathbf{X}'\mathbf{V}' + \mathbf{VV}' \\ &= \mathbf{X}'\mathbf{X} + \mathbf{V}\,\mathbf{V}' \succeq \mathbf{X}'\mathbf{X} \,. \end{aligned}$$

Hence Lemma A.1. □

The theorem below is used in the proof of Theorem 3.1.

Theorem A.1. *Let* $\mathbf{G}$ *and* $\mathbf{X}$ *be two matrices of orders* $p \times n$ *and* $n \times p$ *respectively such that* $\mathbf{G}\mathbf{X} = \mathbf{I}_p$. *Assume that* $\mathbf{X}'\mathbf{X}$ *is invertible. Then*

$$\mathbf{G}\,\mathbf{G}' \succeq (\mathbf{X}'\mathbf{X})^{-1}\,.$$

Proof. Let $\mathbf{U} = (\mathbf{X}'\mathbf{X})\mathbf{G}$. Then we get

$$\mathbf{U}\mathbf{X} = \mathbf{X}'\mathbf{X}\mathbf{G}\mathbf{X} = \mathbf{X}'\mathbf{X}\,.$$

Thus, in view of Lemma A.1, $\mathbf{U}\mathbf{U}' \succeq \mathbf{X}'\mathbf{X}$, i.e.

$$(\mathbf{X}'\mathbf{X})(\mathbf{G}\,\mathbf{G}')(\mathbf{X}'\mathbf{X}) \succeq \mathbf{X}'\mathbf{X}\,.$$

Now for any $\mathbf{z} \in \mathbf{R}^p$, setting $\widetilde{\mathbf{z}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{z}$, we get that

$$\mathbf{z}'(\mathbf{G}\mathbf{G}')\mathbf{z} = \widetilde{\mathbf{z}}'(\mathbf{X}'\mathbf{X})(\mathbf{G}\,\mathbf{G}')(\mathbf{X}'\mathbf{X})\widetilde{\mathbf{z}} \geq \widetilde{\mathbf{z}}'(\mathbf{X}'\mathbf{X})\widetilde{\mathbf{z}} = \mathbf{z}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{z}\,,$$

which proves, that

$$\mathbf{G}\,\mathbf{G}' \succeq (\mathbf{X}'\mathbf{X})^{-1}\,.$$

Hence Theorem A.1. □

A.2 Invariant Donsker Principle

Now we explain how we can calculate the limit distribution in the Skorokhod space $\mathbf{D}[0,1]$. To this end for any $0 \leq t \leq 1$ we set

$$W_t^{(n)} = \frac{1}{\sqrt{n}} \sum_{j=1}^{[nt]} \xi_j\,, \tag{A.1}$$

where $(\xi_j)_{j\geq 1}$ are i.i.d. random variables with $\mathbf{E}\,\xi_j = 0$ and $\mathbf{E}\,\xi_j^2 = \sigma^2$.

Theorem A.2. *([5]) In the space* $\mathbf{D}[0,1]$ *the functional sequence* $W^{(n)} = (W_t^{(n)})_{0\leq t\leq 1}$ *defined in* (A.1) *weakly converges to Brownian motion* $W = (W_t)_{0\leq t\leq 1}$ *with the variance* $\mathbf{E}W_1^2 = \sigma^2$, *i.e. for any bounded* $\mathbf{D}[0,1] \to \mathbb{R}$ *functional* g

$$\lim_{n\to\infty} \mathbf{E}\,g(W^{(n)}) = \mathbf{E}\,g(W)\,.$$

References

[1] Anderson, T. W. The Statistical Analysis of Time Series. Wiley, 1971

[2] Ayvazyan S.A. Econometrics methods. Textbook: Moscow School of Economics, Moscow State University M. V. Lomonosov, Moscow: Master, 2010, 512 p.

[3] Ayvazyan S.A., Fantazzini D. Econometrics-2: advanced course with applications in finance. Textbook: Moscow School of Economics, Moscow State University M. V. Lomonosov, Moscow: Master, 2015, 942 p.

[4] Afanasyev, V.N., Yuzbashev, M.M. Time Series Analysis and Forecasting, Moscow: Finance and Statistics, 2012, 317 p.

[5] Billingsley, P. Convergence of Probability measures. Second Edition. - John Wiley & Sons, Inc, Wiley series in probability and statistics, Probability and Statistics Section, 1999.

[6] Dragalin, V. Sequential methods in multi-arm clinical trials. *Preprint Wyeth Research, Collegeville, PA, USA*, 2009.

[7] Hastie,T., Friedman, J. and Tibshirani, R. *The Elements of Statistical Leaning. Data Mining, Inference and Prediction.* Second Edition, Springer, Springer series in Statistics, 2008.

[8] Magnus, Ya.R. Katyshev, P.K., Peresetsky, A.A. Econometrics. - Moscow, "Delo", 2004 (in Russian).

[9] Mills T. The Econometric Modeling of Financial Time Series. - Cambridge Univ. Press, 1993.

[10] Podkorytova O.A., Sokolov M.V. Time series analysis: study guide for undergraduate and graduate programs: St. Petersburg. state university, European university in St. Petersburg, Moscow: Yurayt, 2016, 265 P

[11] Soft. *R - Project. - https://www.r-project.org/help.html*

[12] Shiryaev, A.N. *Probability.* Springer, Berlin, 1996.

[13] Tsay Ruey S. Analysis of Financial Time Series, 3rd Edition, Wiley, 2010.

[14] Tsay Ruey S. An Introduction to Analysis of Financial Data with R, John Wiley, 2013.