

Test of no-effect hypothesis by nonparametric regression

Test di indipendenza in media tramite regressione nonparametrica

Dembo Gadiaga¹
UFR/SEG
Université de Ouagadougou
Burkina Faso

Rosaria Ignaccolo²
Dipartimento di Statistica e Matematica Applicata
Università degli Studi di Torino
Italy

Riassunto: Nel presente lavoro vengono stabilite alcune proprietà di una classe di test per verificare l'indipendenza in media di Y da X , in un modello di regressione $Y = r(X) + \varepsilon$, sia nel caso di osservazioni indipendenti che nel caso di osservazioni correlate generate da processi stocastici α -mixing a tempo discreto. Tali test sono definiti a partire da stimatori per proiezione della funzione di regressione. L'analisi del comportamento asintotico della statistica test, con la determinazione del tasso di convergenza alla distribuzione limite, permette la costruzione effettiva del test. Vengono inoltre determinate le condizioni necessarie e sufficienti per la consistenza del test e fornite indicazioni per l'implementazione.

Keywords: regression, nonparametric test, strong mixing process.

1. Introduction

The nonparametric test for a regression function has been studied by several authors for processes both in discrete and continuous time. Several methods have been used, among others the kernel, the orthogonal series and the least squares method. Hart (1997) examines the relation between a variable Y and a deterministic variable x , expressed as $Y = r(x) + \varepsilon$, where r denotes the regression function and ε the error term. More recently Hart and Aerts *et al.* (2000) have developed tests under different assumptions. In Lee and Hart (2000), they use the trigonometric functions to estimate r and also supply a method to choose the order of the Fourier coefficients.

In our work we have assumed the existence of a relationship expressed by the model $Y = r(X) + \varepsilon$ where X is a random variable, to test the no-effect hypothesis $H_0 : r(x) = c$, where c is a constant. We have hence considered a projection estimator of $r(x)$ and defined the statistics of our test. Then we have analyzed the asymptotic behavior of the test for independent and correlated observations, obtaining in both cases the limit distribution and the necessary and sufficient conditions for the consistency of the test.

2. Definition of the test

We consider the sequences of r.v.s (X_i, Y_i) with $1 \leq i \leq n$, defined on $(\Omega, \mathcal{A}, \mathbb{P})$ with values in a measurable space $(E \times \mathbb{R}, \mathcal{B} \times \mathcal{B}_{\mathbb{R}}, \mu \otimes \lambda)$ where μ is a σ -finite measure on E

¹This research was mainly conducted while the authors visited the *Laboratoire de Statistique Théorique et Appliquée* (LSTA) of the University “Pierre et Marie Curie” in Paris. The authors thank all LSTA members for their kindness and, in particular, Prof. Denis Bosq for his helpful advices and valuable comments.

²Corresponding author: ignaccolo@econ.unito.it

and λ the Lebesgue measure. We suppose that the pairs (X_i, Y_i) are identically distributed and marginally $X_i \sim \mu$ with density $f(x)$.

We consider $r(x) = \mathbb{E}(Y|X = x)$ in the regression model $Y = r(X) + \varepsilon$ and we want to test the hypothesis $H_0 : r(x) = c$ where c is a constant. Without loss of generality, we pose $c = 0$ (as c is known) since it is possible to apply the test on the observations $Y_i - c$ considering the model $Y - c = r(X) + \varepsilon$. Moreover, if $r_0(x)$ is a specified function we can use the test for the hypothesis $H_0 : r(x) = c + r_0(x)$ running on the observations $Y_i - c - r_0(X_i)$.

We suppose that $r \in L^2(\mu)$, so that we can write $r(x) = \sum_{j=0}^{\infty} b_j e_j(x)$ with the Fourier coefficients $b_j = \langle r, e_j \rangle = \int r(x) e_j(x) d\mu(x)$.

For a fixed positive integer k , let $\{e_0, e_1, \dots, e_k\}$ be an orthonormal system with $e_0 \equiv 1$, in $L^2(\mu)$, which generates a subspace E_k , with $\dim(E_k) = k + 1$.

Denoting $\mathbf{X} = (X_1, \dots, X_n)$ and $\mathbf{Y} = (Y_1, \dots, Y_n)$, the estimator of the regression function by projection on E_k is defined by (see Bosq and Lecoutre (1987))

$$r_n(\mathbf{X}, \mathbf{Y}; x) = r_n(x) = \sum_{j=0}^k \hat{b}_{jn} e_j(x) \quad \text{for a fixed } k$$

where $\hat{b}_{jn} = \frac{1}{n} \sum_{i=1}^n Y_i e_j(X_i)$ is the unbiased estimator of the Fourier coefficient b_j . We observe that for $j = 0$ one has $\hat{b}_{0n} = \frac{1}{n} \sum_{i=1}^n Y_i = \bar{Y}_n$.

To test the hypothesis $H_0 : r(x) = 0$, we consider the distance $d(r_n, 0) = \|r_n - 0\|$ in $L^2(\mu)$ between the estimated regression function and the true one under H_0 , where $\|\cdot\|$ denotes the $L^2(\mu)$ -norm. Now we consider the statistic $R_n = \sqrt{n} r_n$ and its norm in $L^2(\mu)$, $\|R_n\| = \sqrt{n} \|r_n\|$, obtaining

$$\|R_n\|^2 = n \sum_{j=0}^k \hat{b}_{jn}^2 \quad \text{for a fixed } k.$$

So the test rejects the null hypothesis $H_0 : r(x) = 0$ for every x versus $H_1 : r(x) \neq 0$ for large values of $\|R_n\|^2$. We observe that $\|R_n\|^2$ depends only on the estimated Fourier coefficients and that the previously specified hypothesis system is equivalent to the system $H_0 : b_j = 0 \forall j \geq 0$ versus $H_1 : \exists j \geq 0 : b_j \neq 0$. Indeed we shall consider the alternative hypothesis $H_1(k) : r(x) \neq 0$ with r such that it exists $j \in \{0, 1, \dots, k\}$ for which $b_j \neq 0$.

3. Large sample behaviour

In this section we give some results, without proofs. The relative theorems are proved in Chapter 5 of Ignaccolo (2002).

INDEPENDENT DATA. If we suppose that the observations $(x_1, y_1), \dots, (x_n, y_n)$ are independent identically distributed, under $H_0 : r = 0$ and assuming that the conditional second moment $\mathbb{E}(Y_i^2|X_i)$ is a positive finite constant $\gamma > 0$, we have proved that

$$\|R_n\|^2 \xrightarrow{d} \gamma Q$$

where $Q \sim \chi_{k+1}^2$.

Moreover, posing $D_{ij} = Y_i e_j(X_i) - \mathbb{E}(Y_i e_j(X_i))$, if $\mathbb{E}(|D_{ij}|^3) < \infty$ for every j we have

obtained the rate of the previous convergence:

$$\Delta_n = \sup_{t \in \mathbb{R}} |\mathbb{P}(\|R_n\|^2 \leq t) - \mathbb{P}(\gamma Q \leq t)| \leq c_0 \frac{(k+1)^3}{\gamma^{3/2}} \sum_{j=0}^k \mathbb{E}(|D_{1j}|^3) n^{-1/2},$$

where c_0 is a constant.

We have also proved that under $H_1(k)$ one has $\|R_n\|^2 \xrightarrow{\text{a.s.}} \infty$.

So we can say that with $\alpha \in]0, 1[$ such that

$$\mathbb{P}(\gamma \chi_{k+1}^2 > w) = \alpha$$

it is possible to construct a test with rejection region $\{(x_1, \dots, x_n) : \|R_n\|^2 > w\}$ with asymptotic size α and, under the null hypothesis, one has $\left| \mathbb{P}\left(\sum_{j=0}^k \hat{b}_{jn}^2 > \frac{w}{n}\right) - \alpha \right| = O(n^{-1/2})$.

Since γ is unknown, to apply the test we can estimate it by

$$\hat{\gamma} = \frac{1}{k+1} \sum_{j=0}^k \hat{\gamma}_j = \frac{1}{n(k+1)} \sum_{j=0}^k \sum_{i=1}^n Y_i^2 e_j^2(X_i).$$

Moreover the test with rejection region $\|R_n\|^2 > w_n$ is consistent with respect to $H_1(k)$ if and only if $w_n \xrightarrow{n \rightarrow \infty} \infty$ and $\frac{w_n}{n} \xrightarrow{n \rightarrow \infty} 0$.

CORRELATED DATA. To model a weak dependence, in the sequel we suppose to have observations generated by $(X_t, Y_t)_{t \in \mathbb{Z}}$ that is α -mixing with coefficients α_{XY} . We recall that a process $(X_t)_{t \in \mathbb{Z}}$ is said strong mixing (or α -mixing) if

$$\alpha(n) = \sup_{t \in \mathbb{Z}} \sup_{A \in \mathcal{F}_{-\infty}^t, B \in \mathcal{F}_{t+n}^\infty} |\mathbb{P}(A \cap B) - \mathbb{P}(A) \mathbb{P}(B)| \xrightarrow{n \rightarrow \infty} 0$$

where $\alpha(n)$ are the mixing coefficients and $\mathcal{F}_{-\infty}^t = \sigma(X_i, i \leq t)$ denotes the σ -algebra generated by $(X_i, i \leq t)$ and $\mathcal{F}_{t+n}^\infty = \sigma(X_i, i \geq t+n)$.

Under $H_0 : r(x) = 0$ and under suitable assumptions on the existence of moments with order $2 + \delta$ (with $\delta > 0$) and on the strict positivity of a covariances' series related to the considered stochastic process, the test statistics $\|R_n\|^2$ converges to a linear combination of χ_1^2 r.v.s, more precisely

$$\|R_n\|^2 \xrightarrow{d} \sum_{j=0}^k \lambda_j^{*2} U_j^2$$

where λ_j^{*2} are the eigenvalues of the matrix Σ^* defined by

$$\sigma_{jl}^* = \mathbb{E}(Y_1^2 e_j(X_1) e_l(X_1)) + \sum_{i=2}^{\infty} \mathbb{E}(Y_1 e_j(X_1) Y_i e_l(X_i)) + \sum_{i=2}^{\infty} \mathbb{E}(Y_i e_j(X_i) Y_1 e_l(X_1))$$

and the r.v.s $U_j \sim \mathcal{N}(0, 1)$ are independent.

The rate of this convergence depends on the decreasing order of the mixing coefficients.

If $\alpha_{XY}(n) = O(n^{-\beta(2+\delta)(1+\delta)/\delta^2})$ for some $\beta > 1$ then the rate is of order $n^{-\frac{\delta}{2}} \log^{1+\delta} n$,

while if $\alpha_{XY}(n) = O(e^{-\beta n})$ for some $\beta > 1$ then the order becomes $n^{-\frac{\delta}{2}} \log^{1+\delta} n$. We have proved again that under $H_1(k)$ one has $\|R_n\|^2 \xrightarrow{\text{a. s.}} \infty$.

Hence if we get $\alpha \in]0, 1[$ such that

$$\mathbb{P} \left(\sum_{j=0}^k \lambda_j^{*2} U_j^2 > w \right) = \alpha$$

we can construct a test with rejection region $\{(x_1, \dots, x_n) : \|R_n\|^2 > w\}$ with asymptotic size α .

Really the estimation of the eigenvalues λ_j^{*2} needs. Then we have defined an estimator of σ_{jl}^* as $\hat{\sigma}_{jl} = \sum_{v=-\ell_n}^{\ell_n} w_n(v) \hat{\sigma}_{jl}(v)$, where $\hat{\sigma}_{jl}(v)$ is the classic estimator of the crossed covariance of the stationary bivariate process $(D_{tj}, D_{tl})_{t \in \mathbb{Z}}$ with $D_{tj} = Y_t e_j(X_t) - \mathbb{E}(Y_t e_j(X_t))$, $w_n(v)$ are weights satisfying some conditions and ℓ_n is a sequence of positive integers such that $\ell_n < n$ and $\lim_{n \rightarrow \infty} \ell_n = \lim_{n \rightarrow \infty} \frac{n}{\ell_n} = \infty$. Under suitable conditions $\hat{\sigma}_{jl}$ is consistent, so we can use the estimated eigenvalues to have an approximation of the limit distribution.

To obtain the consistency of the test we add a boundedness condition on the $|Y_i e_j(x)|$ for $x \in E$ and if $\sum_{i=1}^{\infty} \alpha^*(i) < \infty$, where $\alpha^*(i)$ are the mixing coefficients for the r.v.s $Y_i e_j(X_i)$, the test with rejection region $\|R_n\|^2 > w_n$ is consistent with respect to $H_1(k)$ if and only if $w_n \xrightarrow{n \rightarrow \infty} \infty$ and $\frac{w_n}{n} \xrightarrow{n \rightarrow \infty} 0$.

4. Conclusions

This work follows on the wake of the functional tests of fit associated with the projection estimators, introduced by Bosq (for general results see Bosq (2002)). The results of this work give us the limit distributions for the case of no effect hypothesis, for correlated and independent data. To perform the test for correlated data, the method requires an estimation of the eigenvalues and the determination of the quantiles of a linear combination of χ_1^2 . Simulations to evaluate the empirical power of the test are in progress.

References

- Aerts M., Claeskens G. and Hart J.D. (2000) Testing lack of fit in multiple regression, *Biometrika*, 87, 2, 405–424.
- Bosq D. (2002) Functional test of fit, in: *Goodness-of-Fit Tests and Model Validity*, Birkhäuser, Boston, chapter 25, 341–356, in: C. Huber-Carol et al.: (Eds.).
- Bosq D. and Lecoutre J.P. (1987) *Théorie de l'estimation fonctionnelle*, Economica, Paris.
- Hart J.D. (1997) *Nonparametric smoothing and lack-of-fit tests*, Springer-Verlag, New York.
- Ignaccolo R. (2002) *Tests d'ajustement fonctionnels pour des observations corrélées*, Ph.D. thesis, Università degli Studi di Milano and Université Paris VI.
- Ignaccolo R. (2004) Goodness-of-fit test for dependent data, *Journal of Nonparametric Statistics*, 16, 1-2, 19–38.
- Lee G. and Hart J.D. (2000) Model selection criteria with data dependent penalty, with applications to data-driven Neyman smooth tests, *Journal of Nonparametric Statistics*, 12, 5, 683–707.