Data Driven Score Test of Fit for Semiparametric Homoscedastic Linear Regression Model

Tadeusz Inglot[1,2] **and Teresa Ledwina**[2]

[1] *Wrocław University of Technology, Wrocław, Poland*
[2] *Polish Academy of Sciences, Wrocław, Poland*

Abstract

We shall present new test for asserting validity of the following semiparametric linear regression model $\mathtt{M(0)}$

$$Y = \beta[v(X)]^T + \epsilon,$$

where X and ϵ are independent with unknown densities g and f, supported on [0,1] and R, respectively. We assume $E_f\epsilon = 0$ and $E_f\epsilon^2 < \infty$. $\beta \in R^q$ is a vector of unknown parameters while $v(x) = (v_1(x), ..., v_q(x))$ is a vector of known functions. The symbol ^{T} denotes transposition. Throughout we consider row vectors.

The test construction combines classical ideas with some modern smoothing methods.

The classical, mostly analytical, part relies, first of all, on following the idea of overfitting and replacing the basic problem by a series of auxiliary subproblems. More precisely, we embed $\mathtt{M(0)}$ into extended model $\mathtt{M_k}(\theta)$

$$Y = \theta[u(X)]^T + \beta[v(X)]^T + \epsilon,$$

where, for each given k, $\theta \in R^k$ is a vector of unknown parameters while $u(x) = (u_1(x), ..., u_k(x))$ is a vector of known functions. Note that the joint density of (X, Y), under $\mathtt{M_k}(\theta)$, has the form

$$p(z; \kappa) = g(x)f(y - (u, v)(\theta, \beta)^T) \quad \text{with} \quad \kappa = (\theta, \beta, f, g) \quad \text{and} \quad z = (x, y).$$

Next classical, in principle, idea is to construct efficient score test for testing $\theta = 0$ against $\theta \neq 0$ in $\mathtt{M_k}(\theta)$. This requires a derivation of suitable derivative of $p(z; \kappa)$ over κ from appropriate Banach space. The derivative is determined by a vector, which is called the score vector. Additionally, one has to calculate efficient score vector which results as residual from projections [derived under the null hypothesis] of scores for the parameters of interest on scores for nuisance parameters. Finally, an auxiliary statistic is defined as quadratic form of the efficient score vector and the inverse if its covariance matrix.

This is analytical part of the work, which is discussed in details in Inglot and Ledwina (2004). In that paper the above programme is carried out for heteroscedastic model as well.

Probabilistic part relies on exploiting some ideas of adaptive semiparametric estimation to propose suitable estimators of the involved parameters of the auxiliary statistic, defined above. The basic focus is on ensuring that the limiting null distribution of resulting object shall be independent of unknown nuisance parameters β, f, g. We call such statistic efficient score statistic.

Last step of our construction is to propose data driven selection rule to choose the right subproblem. So, the final result is the efficient score statistic with the dimension k fitted by the selection rule.

The construction shall be presented in some details. Also some simulations shall be shown, to demonstrate good performance of our test.

Keywords

Data driven test, Efficient score, Linear regression, Selection rule, Semiparametric model.

References:

Inglot, T., and T. Ledwina (2004). Semiparametric regression: Hadamard differentiability and efficient score functions for some testing problems. Shall be submitted to *Linear Algebra Appl.*, the workshop volume.

Inglot, T., and T. Ledwina (2003). Data driven score test of fit for semiparametric homoscedastic linear regression model. Submitted for publication.