

Data driven tests  
for homoscedastic linear regression model

Tadeusz Inglot and Teresa Ledwina

Model and null hypothesis

$Z = (X, Y)$  random vector in  $[0, 1] \times R$ ,  $X, \epsilon$  independent, distributions of  $X$  and  $\epsilon$  unknown,  $X \sim g$ ,  $\epsilon \sim f$ ,  $E_f \epsilon = 0$ ,  $\tau = E_f \epsilon^2 \in (0, \infty)$ ,

$$H_0 : Y = \beta[v(X)]^T + \epsilon, \quad \beta \in R^q,$$

$v(x) = (v_1(x), \dots, v_q(x))$  given vector of bounded functions.

Overfitting. Auxiliary models  $M(k)$ ,  $k = 1, 2, \dots$

$$Y = \theta[u(x)]^T + \beta[v(x)]^T + \epsilon, \quad \theta \in R^k,$$

$u(x) = (u_1(x), \dots, u_k(x))$  vector of bounded functions lineary independent on  $v(x)$ .

Auxiliary solution

Given  $k$  and  $M(k)$ ,  $Y = \theta[u(x)]^T + \beta[v(x)]^T + \epsilon$ ,  $\theta \in R^k$ , we construct efficient score statistic for

$$H_0(k) : \theta = 0, \quad \eta,$$

$\eta = (\beta, \sqrt{g}, \sqrt{f})$  - nuisance parameter.  $\ell^*$  -  $k$ -dimensional efficient score vector, i.e. the residuals from projections [derived under  $H_0(k)$ ] of scores for the parameters of interest  $[\theta_1, \dots, \theta_k]$  onto scores for the nuisance parameters  $[\eta]$ ; Neyman (1959).

$$W_k(\eta) = \left[ \frac{1}{\sqrt{n}} \sum_{i=1}^n \ell^*(Z_i; \eta) \right] \mathbf{L} \left[ \frac{1}{\sqrt{n}} \sum_{i=1}^n \ell^*(Z_i; \eta) \right]^T.$$

where

$$E_\eta \ell^*(Z; \eta) = 0, \quad \left\{ E_\eta [\ell^*(Z; \eta)]^T [\ell^*(Z; \eta)] \right\}^{-1} = \mathbf{L}, \quad W_k(\eta) \xrightarrow{D} \chi_k^2.$$

Define

$$W_k(\hat{\eta}) = \left[ \frac{1}{\sqrt{n}} \sum_{i=1}^n \hat{\ell}^*(Z_i; \hat{\eta}) \right] \hat{\mathbf{L}} \left[ \frac{1}{\sqrt{n}} \sum_{i=1}^n \hat{\ell}^*(Z_i; \hat{\eta}) \right]^T,$$

where  $\hat{\ell}^*(\bullet; \hat{\eta})$  is an estimator of  $\ell^*(\bullet; \eta)$ , while  $\hat{\mathbf{L}}$  is an estimator of  $\mathbf{L}$ .

**Theorem 1.** Assume the null hypothesis  $H_0(k) : \theta = 0$  is true and some mild extra assumptions are fulfilled. Suppose that  $\hat{\mathbf{L}}$  is a consistent estimator of  $\mathbf{L}$  and the estimator  $\hat{\ell}^*(\bullet; \hat{\eta})$  satisfies the following condition

$$P_{\eta}^n \left( \left\| \sum_{i=1}^n [\hat{\ell}^*(Z_i; \hat{\eta}) - \ell^*(Z_i; \eta)] \right\| \geq \delta \sqrt{n} \right) \rightarrow 0$$

for every  $\delta > 0$ , as  $n \rightarrow \infty$ .

Then for the test statistic  $W_k(\hat{\eta})$  it holds that

$$W_k(\hat{\eta}) \xrightarrow{\mathcal{D}} \chi_k^2, \text{ as } n \rightarrow \infty.$$

Some class of estimators is proposed for which Theorem 1 holds.

### Selecting $k$ in $W_k(\hat{\eta})$

**Score-based rule mimicing Schwarz's BIC**

$$S1 = \min\{1 \leq k \leq d : W_k(\hat{\eta}) - k \log n \geq W_s(\hat{\eta}) - s \log n, \quad s = 1, \dots, d\}.$$

**Score-based rule imitating Akaike's AIC**

$$A1 = \min\{1 \leq k \leq d : W_k(\hat{\eta}) - 2k \geq W_s(\hat{\eta}) - 2s, \quad s = 1, \dots, d\}.$$

**Refined score-based selection rule [Inglot and Ledwina (2006c)].**

Set  $(\mathcal{Y}_1, \dots, \mathcal{Y}_k) = \left[ n^{-1/2} \sum_{i=1}^n \hat{\ell}^*(Z_i; \hat{\eta}) \right] \hat{\mathbf{L}}^{1/2}$ . Then,  $W_k(\hat{\eta}) = \|(\mathcal{Y}_1, \dots, \mathcal{Y}_k)\|^2$ . Define new penalty

$$\pi(s, p) = \begin{cases} s \log n, & \text{if } \max_{1 \leq t \leq d} |\mathcal{Y}_t| \leq \sqrt{p \log n}, \\ 2s, & \text{if } \max_{1 \leq t \leq d} |\mathcal{Y}_t| > \sqrt{p \log n}, \end{cases}$$

where  $p$  is some fixed positive number. Then the refined selection rule is given by

$$T1 = \min\{1 \leq k \leq d : W_k(\hat{\eta}) - \pi(k, p) \geq W_s(\hat{\eta}) - \pi(s, p), \quad s = 1, \dots, d\}.$$

### Data driven score test statistics

$$W_{S1}(\hat{\eta}), \quad W_{T1}(\hat{\eta}).$$

### Asymptotic behaviour under $H_0$

For simplicity we assumed that  $d$ , the number of models on the list, does not depend on  $n$ .

**Theorem 2.** Under the null hypothesis  $H_0 : Y = \beta[v(X)]^T + \epsilon$ , the assumptions of Theorem 1 and  $n \rightarrow \infty$ , it holds that

$$P_\eta^n(S1 > 1) \rightarrow 0, \quad W_{S1}(\hat{\eta}) \xrightarrow{\mathcal{D}} \chi_1^2, \quad \text{and} \quad P_\eta^n(T1 > 1) \rightarrow 0, \quad W_{T1}(\hat{\eta}) \xrightarrow{\mathcal{D}} \chi_1^2.$$

### Example

$$H_0 : Y = \beta_1 + \beta_2 X + \epsilon.$$

Simulated critical values of  $W_{S1}$  and  $W_{T1}$  under the null model  $Y = 1 + 2X + \epsilon$  with  $X$  uniform on  $[0,1]$  and different errors. Sample size  $n = 300$ . 5% significance level,  $N = 10000$  MC runs.  $p = 2.4$ .

Error distribution	Parameter	Variance	Critical values		
			$W_{S1}$	$W_{T1}$	
Gaussian	$G(\sigma)$	0.25	0.063	5.91	6.11
		0.50	0.250	5.63	5.92
		0.75	0.563	5.83	6.04
		1.00	1.000	5.79	6.02
Laplace	$L(\varphi)$	4.00	0.125	5.29	5.57
		2.00	0.500	5.27	5.50
		1.00	2.000	5.75	5.93
		0.50	8.000	5.61	5.82
Normal Mixture	$NM(\mu)$	0.20	1.191	5.94	6.08
		0.40	1.762	5.67	6.00
		0.60	2.714	5.81	6.05
		0.80	4.048	5.66	5.85

## Empirical powers

**Alternatives:**

$$Y = 1 + 2X + r_j(X) + \epsilon, \quad j = 1, \dots, 4.$$

**Auxiliary models  $M(k)$ :**

$$Y = 1 + 2X + \sum_{j=1}^k \theta_j \cos([j + 1]\pi x), \quad k = 1, \dots, 10.$$

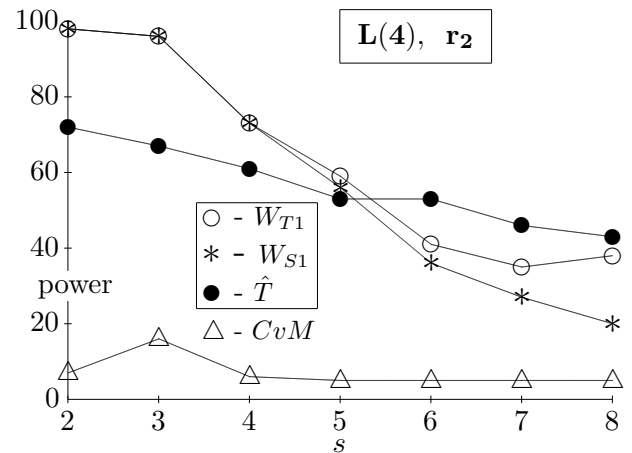
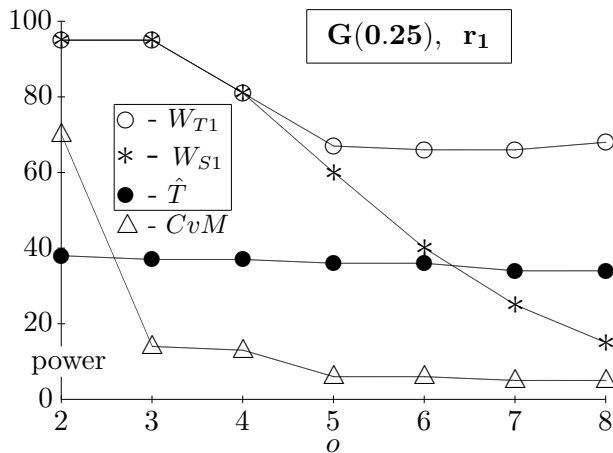
$$v(x) = (1, x), \quad u(x) = (\cos(2\pi x), \dots, \cos([k + 1]\pi x)).$$

**Errors :** as described above.

**Tests for comparison:**

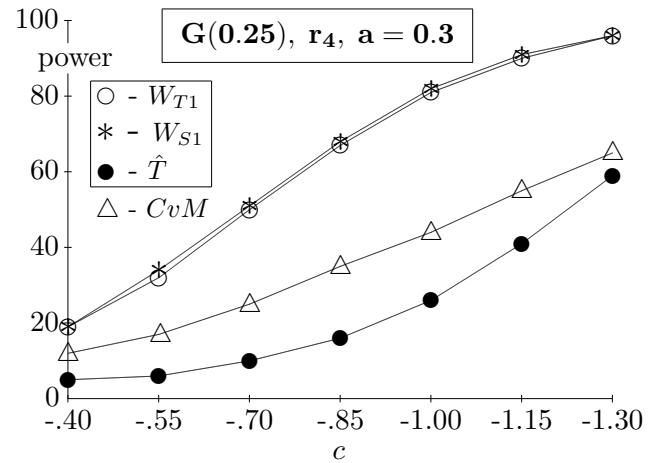
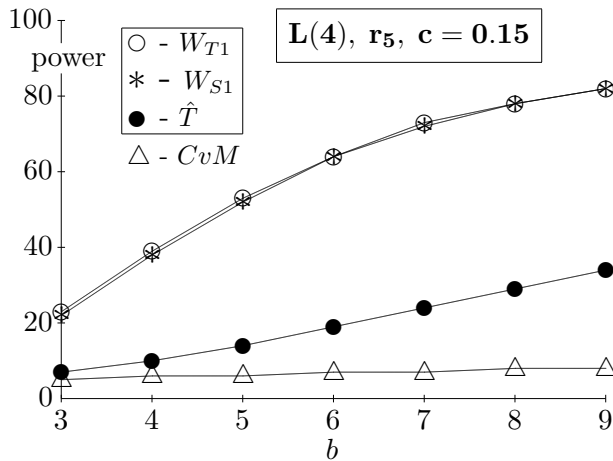
**CvM - Cramér-von Mises test,**

$\hat{T}$  - statistic of Guerre and Lavergne (2005).



$r_1(x) = c \times \cos(\pi \alpha x)$ ,  $r_2(x) = c \times L_s(x)$ ,  $L_s$ -sth normalized Legendre's polynomial on  $[0,1]$ .

**Simulated powers of tests based on  $W_{T1}$ ,  $W_{S1}$ ,  $\hat{T}$  and CvM under the alternatives  $Y = 1 + 2X + r_j(X) + \epsilon$ ,  $j = 1, 2$ ,  $X$  uniform on  $[0,1]$  and different errors. Signal/noise 0.25. 5% nominal level,  $n = 300$ ,  $N = 10000$  MC runs.  $p = 2.4$ .**



$$r_5(x) = c \times \arctan[b(2x - 1)], \quad r_4(x) = c \times (x - a)1_{[a,1]}(x).$$

**Simulated powers of tests based on  $W_{T1}$ ,  $W_{S1}$ ,  $\hat{T}$  and CvM under the alternatives  $Y = 1 + 2X + r_j(X) + \epsilon$ ,  $j = 4, 5$ ,  $X$  uniform on  $[0,1]$  and different errors. 5% nominal level,  $n = 300$ ,  $N = 10000$  MC runs.  $p = 2.4$**

## References

Guerre, E., Lavergne, P. (2005). Data-driven rate-optimal specification testing in regression models. *Ann. Statist.* **33**, 840-870.

Inglot, T., Ledwina, T. (2006a). Data driven score tests for a homoscedastic linear regression model: the construction and simulations. *Proc. Prague Stochastics 2006*, 124-137.

Inglot, T., Ledwina, T. (2006b). Data driven score tests for a homoscedastic linear regression model: asymptotic results. *Probab. Math. Statist.*, 41-61.

Inglot, T., Ledwina, T. (2006c). Towards data driven selection of a penalty function for data driven Neyman tests. *Linear Algebra and its Appl.*, 579-590.

Neyman, J. (1959). Optimal asymptotic tests of composite statistical hypotheses. In *Probability and Statistics: The Harald Cramér Volume* (U. Grenander, ed.) 213-234. Wiley, New York.