Chování pořadových testů v lineárním modelu s chybami měření

RADIM NAVRÁTIL

navratil@karlin.mff.cuni.cz

Katedra pravděpodobnosti a matematické statistiky, MFF UK, Praha

V lineárním modelu s chybami měření budeme zkoumat chování pořadových testů, které byly původně navrženy pro testování hypotéz v modelech, které předpokládaly, že odezva i regresory byly měřeny přesně. Ukážeme, že pro některé hypotézy se použitím těchto testů v modelech s chybami měření zachovává hladina testu, přítomnost chyb však vede ke snížení síly použitého testu. Toto budeme ilustrovat numericky i pomocí simulací.

Úvod

O statistické inferenci v modelech s chybami měření (EV modely) již existuje bohatá literatura, reprezentovaná knihami Fuller (1987), Carroll et al. (2. vydání 2006), Cheng and van Ness (1999) a řadou článků. Některé články se již dotýkají robustních a neparametrických metod, avšak až donedávna neexistovaly žádné výsledky o chování a využití pořadových testů v EV modelech. Prakticky prvními pracemi o pořadových testech v EV modelech jsou práce [1]–[3] o pořadových testech v lineárních a částečně lineárních modelech s chybami měření. Ukazuje se, že právě pořadové testy jsou velmi výhodné pro EV modely, protože jsou invariantní vzhledem k mnoha transformacím a přinášejí rozumnou zprávu a závěry i v této situaci. S těmito modely se (někdy nevědomky) můžeme setkat například v lékařství, geologii, lesnictví, atd., vlastně všude, kde se výsledky získávají měřením.

Model bez chyb měření

Uvažujme lineární model

$$Y_i = \beta_0 + \boldsymbol{x}_i'\boldsymbol{\beta} + e_i, \quad i = 1, \ldots, n,$$

kde $\boldsymbol{x}_i$ je p-rozměrný vektor regresorů a e_i jsou i.i.d. náhodné veličiny s distribuční funkcí F, resp. absolutně spojitou hustotou f. Naším úkolem je testovat hypotézu

$$\mathrm{H}_0 : \boldsymbol{\beta} = 0 \quad \text{proti alternativě} \quad \mathrm{K} : \boldsymbol{\beta} \neq 0,$$

kde β_0 považujeme za rušivý parametr.

Budeme předpokládat, že při $n \to \infty$ platí:

$$\mathbf{Q}_n = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})' \to \mathbf{Q},$$

$$n^{-1}\max_{1\le i\le n}\{(\boldsymbol{x}_i - \overline{\boldsymbol{x}})'\mathbf{Q}_n^{-1}(\boldsymbol{x}_i - \overline{\boldsymbol{x}})\} \to 0,$$

kde $\mathbf{Q}$ je pozitivně definitní matice $p \times p$ a $\overline{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i$. Označme

$$\mathbf{S}_n = n^{-1/2}\sum_{i=1}^{n}(\boldsymbol{x}_i - \overline{\boldsymbol{x}})a_n(R_i),$$

kde $R_1, \ldots, R_n$ jsou pořadí $Y_1, \ldots, Y_n$ a skóry $a_n(i)$ jsou generované neklesající, s druhou mocninou integrovatelnou skórovou funkcí φ: $a_n(i) = \mathbb{E}\varphi(U_{n:i})$, nebo $a_n(i) = \varphi\left(\frac{i}{n+1}\right)$, kde $U_{n:1} \le \ldots \le U_{n:n}$ jsou pořádkové statistiky odpovídající výběru o rozsahu n z rovnoměrného rozdělení $R(0,1)$. K testování H_0 použijeme statistiku

$$T_n^2 = (A(\varphi))^{-2}\mathbf{S}_n'\mathbf{Q}_n^{-1}\mathbf{S}_n,$$

kde

$$A^2(\varphi) = \int_0^1(\varphi(t) - \overline{\varphi})^2 dt, \qquad \overline{\varphi} = \int_0^1 \varphi(t)dt.$$

Statistika T_n^2 má za hypotézy asymptoticky χ^2 rozdělení o p stupních volnosti, tedy H_0 zamítáme ve prospěch K, pokud $T_n^2 > \chi_p^2(1-\alpha)$, kde $\chi_p^2(1-\alpha)$ je $1-\alpha$ kvantil χ^2 rozdělení o p stupních volnosti.

Dále se dá ukázat, že za alternativy $\mathrm{K}^* : \boldsymbol{\beta} = \boldsymbol{\beta}_n = n^{-1/2}\boldsymbol{\beta}_*$, pro pevné $0 \neq \boldsymbol{\beta}_* \in \mathbb{R}^p$ je asymptotické rozdělení T_n^2 necentrální χ^2 rozdělení o p stupních volnosti s parametrem necentrality $\eta^2 = \boldsymbol{\beta}_*'\mathbf{Q}\boldsymbol{\beta}_*\frac{\gamma^2(\varphi,f)}{A^2(\varphi)}$, přičemž $\gamma(\varphi, f) = \int_0^1 \varphi(t)\varphi(t,f)dt$, $\varphi(t,f) = -\frac{f'(F^{-1}(t))}{f(F^{-1}(t))}$. Tedy síla odvozeného testu za této alternativy je $1 - \Psi_p(\chi_p^2(1-\alpha), \eta^2)$, kde $\Psi_p(x, \eta^2)$ značí distribuční fukci necentrálního χ^2 rozdělení o p stupních volnosti s parametrem necentrality η^2.

Model s chybami v regresorech

Nyní budeme předpokládat, že místo regresorů $\boldsymbol{x}_i$ pozorujeme $\boldsymbol{w}_i = \boldsymbol{x}_i + \boldsymbol{v}_i$, kde $\boldsymbol{v}_i$ jsou i.i.d. p-rozměrné náhodné vektory nezávislé na e_i. Předpokládejme, že pro jejich rozdělení při $n \to \infty$ platí

$$\mathbf{V}_n = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{v}_i - \overline{\boldsymbol{v}})(\boldsymbol{v}_i - \overline{\boldsymbol{v}})' \xrightarrow{p} \mathbf{V},$$

kde $\mathbf{V}$ je pozitivně definitní matice $p \times p$ a navíc předpokládejme, že

$$\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{v}_i - \overline{\boldsymbol{v}})'(\boldsymbol{x}_i - \overline{\boldsymbol{x}}) \xrightarrow{p} 0.$$

Analogicky jako v předchozím označme

$$\widetilde{\mathbf{S}}_n = n^{-1/2}\sum_{i=1}^{n}(\boldsymbol{w}_i - \overline{\boldsymbol{w}})a_n(R_i),$$

$$\widetilde{T}_n^2 = (A(\varphi))^{-2}\widetilde{\mathbf{S}}_n'(\mathbf{Q}_n + \mathbf{V}_n)^{-1}\widetilde{\mathbf{S}}_n.$$

Statistika $\widetilde{T}_n^2$ má za hypotézy asymptoticky χ^2 rozdělení o p stupních volnosti, zatímco za alternativy asymptoticky necentrální χ^2 rozdělení o p stupních volnosti s parametrem necentrality $\eta^2 = \boldsymbol{\beta}_*'\mathbf{Q}(\mathbf{Q}+\mathbf{V})^{-1}\mathbf{Q}\boldsymbol{\beta}_*\frac{\gamma^2(\varphi,f)}{A^2(\varphi)}$. Asymptotická relativní eficience (ARE) testu H_0 založeném na $\widetilde{T}_n^2$ vzhledem ke stejnému testu bez přítomnosti chyb měření je $\frac{\boldsymbol{\beta}_*'\mathbf{Q}(\mathbf{Q}+\mathbf{V})^{-1}\mathbf{Q}\boldsymbol{\beta}_*}{\boldsymbol{\beta}_*'\mathbf{Q}\boldsymbol{\beta}_*}$. Číslo $n \cdot \mathrm{ARE}$ může být interpretováno jako počet pozorování, které bychom potřebovali k dosažení asymptoticky stejné síly jako při použití stejného testu v modelu bez přítomnosti chyb měření.

Model s chybami v odezvě

Předpokládejme, že regresory $\boldsymbol{x}_i$ pozorujeme přesně, zatímco místo Y_i pozorujeme $Z_i = Y_i + W_i$, kde W_i jsou i.i.d. náhodné veličiny nezávislé na $\boldsymbol{x}_i$ i Y_i s distribuční funkcí G a absolutně spojitou hustotou g. Označme

$$\hat{\mathbf{S}}_n = n^{-1/2}\sum_{i=1}^{n}(\boldsymbol{x}_i - \overline{\boldsymbol{x}})a_n(\hat{R}_i),$$

$$\hat{T}_n^2 = (A(\varphi))^{-2}\hat{\mathbf{S}}_n'\mathbf{Q}_n^{-1}\hat{\mathbf{S}}_n,$$

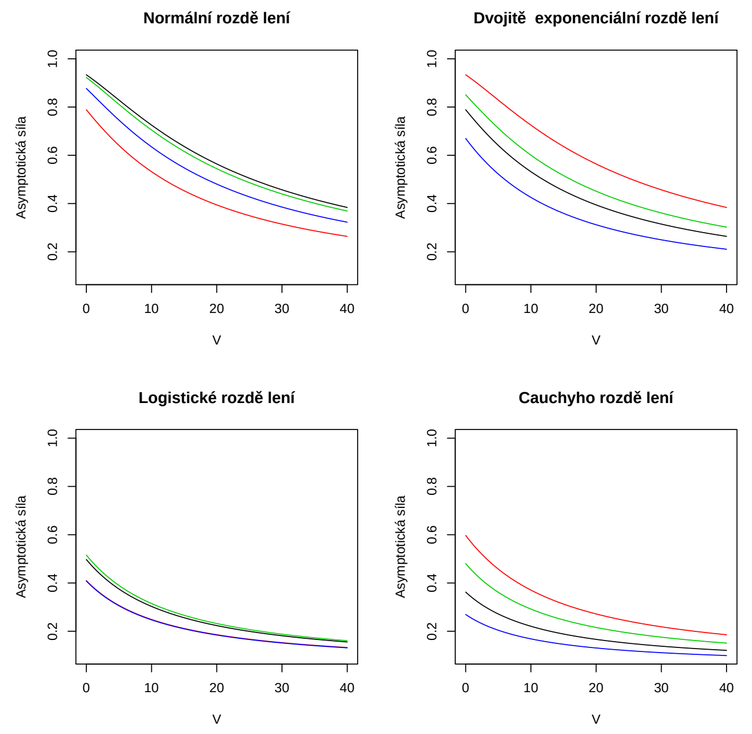
kde $\hat{R}_1, \ldots, \hat{R}_n$ jsou pořadí $Z_1, \ldots, Z_n$. Statistika $\hat{T}_n^2$ má za hypotézy asymptoticky χ^2 rozdělení o p stupních volnosti, zatímco za alternativy asymptoticky necentrální χ^2 rozdělení o p stupních volnosti s parametrem necentrality $\eta^2 = \boldsymbol{\beta}_*'\mathbf{Q}\boldsymbol{\beta}_*\frac{\gamma^2(\varphi,h)}{A^2(\varphi)}$, kde $h(y) = \int f(x)g(y-x)dx$. ARE tohoto testu vzhledem ke stejnému testu bez přítomnosti chyb měření je $\left(\frac{\gamma(\varphi,h)}{\gamma(\varphi,f)}\right)^2$. Případ chyb v regresorech i odezvě je snadnou kombinací předchozích dvou případů.

Simulace

Uvažujme regresní přímku $Y_i = \beta_0 + \beta_1 x_i + e_i$. Nechť x_i pochází z rovnoměrného rozdělení R(-2,10) a volme $n = 100$ a $\beta_0 = 1$ a testujeme hypotézu, že $\beta_1 = 0$ proti obecné alternativě. Budeme porovnávat sílu testu v závislosti na volbě skórové funkce φ (použijeme přibližné skóry) a na rozdělení chyb pozorování. Uvažujme nejprve chybu v regresorech x_i. Asymptotická síla odvozeného testu závisí pouze na asymptotickém rozptylu chyb měření (V), nikoliv na jejich samotném rozdělení. Na obrázku je tato závislost nakreslena pro různé volby chyb modelu e_i a pro různé skórové funkce (β_1 zvoleno pevně 0,1):

$$\varphi(u) = \begin{cases} \Phi^{-1}(u) \\ u - \frac{1}{2} \\ \mathrm{sign}(u - \frac{1}{2}) \\ -1 - \log(1-u), \end{cases}$$

kde $\Phi^{-1}(u)$ je kvantilová funkce standardního normálního rozdělení. Poznamenejme, že případ $V = 0$ odpovídá situaci bez chyb měření.



Zkoumejme nyní vliv chyb měření v odezvě Y_i. V následující tabulce jsou uvedeny odhady síly odvozeného testu na základě 10 000 simulací v závisloti na rozdělení chyb modelu e_i i chyb měření W_i pro různé volby skórové funkce:

$W_i \backslash e_i$	N(0,1)		DExp(0,1)		Logis(0,1)		Cauchy(0,1)	
0	0,91	0,91	0,73	0,80	0,45	0,49	0,33	0,44
	0,77	0,85	0,80	0,60	0,39	0,35	0,54	0,23
N(0,$\frac{1}{3}$)	0,81	0,81	0,63	0,69	0,41	0,44	0,29	0,38
	0,65	0,72	0,60	0,50	0,35	0,32	0,41	0,20
R(-1,1)	0,82	0,82	0,64	0,69	0,41	0,45	0,27	0,36
	0,64	0,72	0,58	0,50	0,35	0,32	0,39	0,20
Exp($\frac{1}{\sqrt{3}}$)	0,83	0,83	0,64	0,70	0,41	0,45	0,28	0,38
	0,68	0,69	0,64	0,49	0,35	0,31	0,43	0,20
Logis(0,$\frac{1}{\pi}$)	0,82	0,82	0,64	0,70	0,42	0,45	0,27	0,36
	0,66	0,72	0,63	0,51	0,36	0,32	0,41	0,19
R(-1,1)	0,82	0,82	0,64	0,69	0,41	0,45	0,27	0,36
	0,64	0,72	0,58	0,50	0,35	0,32	0,39	0,20
R(-2,2)	0,58	0,56	0,45	0,47	033,	0,35	0,20	0,26
	0,38	0,48	0,32	0,35	0,26	0,25	0,23	0,14

Závěr

Přítomnost chyb měření neovlivňuje hladinu uvažovaných pořadových testů, snižuje však jejich výslednou sílu ve srovnání s modelem bez chyb měření. Pro chyby měření s malým rozptylem však toto snížení není nikterak markantní. Sílu testu také nemale ovlivňuje volba skórové funkce φ. Ukazuje se, že použití tzv. Wilcoxonových skórů ($\varphi(u) = u - \frac{1}{2}$) dosahuje i přes svou jednoduchost velmi dobrých výsledků pro širokou třídu uvažovaných chyb modelu i chyb měření, na rozdíl od jiných skórů, které dosahují v modelech bez chyb měření dobrých výsledků, avšak přítomnost chyb měření výslednou sílu znatelně sníží (např. volba $\varphi(u) = \mathrm{sign}(u - \frac{1}{2})$).

Poděkování

Rád bych na tomto místě poděkoval paní profesorce RNDr. Janě Jurečkové, DrSc. za cenné rady a připomínky, které mi pomohly při psaní tohoto posteru a také MFF UK za poskytnutí finančního příspěvku na účast na této konferenci.

Literatura

[1] Jurečková J., Picek J. and Saleh A.K.Md.E. (2009). *Rank tests and regression rank score tests in measurement error models.* Computational Statistics and Data Analysis, doi:10.1016/j.csda.2009.08.020.

[2] Jurečková J., Picek J. (2009): *Rank tests in partially linear and measurement errors models.* Submitted.

[3] Jurečková J., Kalina J., Picek J. and Saleh A.K.Md.E.(2009) *Rank tests of linear hypothesis with measurement errors both in regressors and responses.* KPMS Preprint **66**.