# TESTY NORMALITY ZA PRÍTOMNOSTI RUŠIVEJ REGRESIE

**RADKA SABOLOVÁ**

sabolova@karlin.mff.cuni.cz

Katedra pravděpodobnosti a matematické statistiky, MFF UK, Praha

**Uvažujeme lineárny model $Y_i = \theta + \mathbf{x}_i'\beta + \sigma e_i$, $i = 1, \dots, n$, kde $\mathbf{x}_i$ sú pevné regresory, $\theta$, $\beta$ a $\sigma > 0$ neznáme parametre a $e_i$ sú nezávislé rovnako rozdelené so spojitou distribučnou funkciou $F$. Chceme testovať hypotézu, že chyby $e_1, \dots, e_n$ pochádzajú z normovaného normálneho rozdelenia. Popíšeme modifikáciu známeho Shapiro-Wilkovho testu ([2]), ktorá porovnáva maximálne vierohodný odhad parametra $\sigma^2$ a jeho najlepší lineárny nevychýlený odhad (BLUE) – tieto dva odhady sú asymptoticky ekvivalentné, ak chyby $e_1, \dots, e_n$ sú normálne rozdelené. V príspevku sú ďalej prezentované výsledky simulácií, v ktorých sa skúmala sila tohoto testu.**

## MODEL

Nech nezávislé pozorovania $Y_1, \dots, Y_n$ spĺňajú model

$$Y_i = \theta + \mathbf{x}_i'\beta + \sigma e_i, \qquad i = 1, \dots, n, \tag{1}$$

kde $\mathbf{x}_i' \in \mathbb{R}^p$, $i = 1, \dots, n$ sú pevné regresory, $\theta \in \mathbb{R}$, $\beta \in \mathbb{R}^p$ a $\sigma > 0$ sú neznáme parametre. Chyby $e_i$ sú nezávislé rovnako rozdelené z rozdelenia s distribučnou funkciou $F$, ktoré je centrované a má jednotkový rozptyl. Chceme testovať hypotézu

$$H_0 : F \equiv \Phi \text{ proti } H_1 : F \equiv F_1 \neq \Phi.$$

Najrpv pripomenieme „klasický“ Shapiro-Wilkov test, teda test pre prípad, keď $\beta = 0$, neskôr uvedieme modifikáciu tohto testu pre lineárnu regresiu.

## SHAPIRO-WILKOV TEST

Pre hodnotu $\beta = 0$ bol v [3] odvodený test založený na podiele dvoch odhadov parametra $\sigma$ - a to maximálne vierohodného odhadu pre $Y_1, \dots, Y_n$ normálne rozdelené

$$\hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y}_n)^2$$

a BLUE odhadu

$$L_n = \sum_{i=1}^{n} a_{ni} Y_{n:i},$$

kde $(Y_{n:1}, \dots, Y_{n:n})'$ je príslušná poriadková štatistika a

$$\mathbf{a}' = (\mathbf{M}_n'\mathbf{V}_n^{-1}\mathbf{M}_n)^{-1}(\mathbf{M}_n'\mathbf{V}_n^{-1}),$$

pričom $\mathbf{M}_n$ označuje vektor stredných hodnôt poriadkovej štatistiky a $\mathbf{V}_n$ príslušnú variančnú maticu. Odhad $L_n$ bol však ešte mierne upravený na $L_{n0} = \sum_{i=1}^{n} a_{ni,0} Y_{n:i}$, kde

$$\mathbf{a}_{n0}' = \frac{\mathbf{M}_n'\mathbf{V}_n^{-1}}{(\mathbf{M}_n'\mathbf{V}_n^{-1}\mathbf{V}_n^{-1}\mathbf{M}_n)^{1/2}}. \tag{2}$$

Testovú štatistiku zapíšeme v tvare

$$\boxed{W_n = n\left\{1 - \frac{L_{n0}^2}{\sum_{i=1}^{n}(Y_i - \overline{Y}_n)^2}\right\}.}$$

## MODIFIKÁCIA TESTU

Predpokladáme, že matica $\mathbf{X}_n$ rozmerov $n \times p$ spĺňa

$$\mathbf{X}_n'\mathbf{1}_n = \mathbf{0},$$

jej hodnosť je $p < n - 1$ a $\max_{1 \le i \le n} h_{n,ii} = O(n^{-1})$, kde

$$\mathbf{H}_n = \mathbf{X}_n(\mathbf{X}_n'\mathbf{X}_n)^{-1}\mathbf{X}_n' = [h_{n,ij}]_{i,j=1}^{n}.$$

Ďalej označme

$$\bar{\mathbf{e}}_n = \frac{1}{n}\mathbf{1}'_n\mathbf{e}_n,$$

$$\mathbf{H}_{\mathbf{n_0}} = \frac{1}{n}\mathbf{1}'_n\mathbf{1}_n.$$

Maximálne vierohodné odhady parametrov $\theta$, $\beta$, $\sigma$ za platnosti hypotézy sú

$$\hat{\theta}_n = \overline{Y}_n = \theta + \bar{\mathbf{e}}_n,$$

$$\hat{\beta}_n = (\mathbf{X}_n'\mathbf{X}_n)^{-1}\mathbf{X}_n'\mathbf{Y}_n,$$

$$\hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{\theta}_n - \mathbf{x}_i'\hat{\beta}_n)^2 = \frac{1}{n}\sigma^2\mathbf{e}_n'[\mathbf{1}_n - \mathbf{H}_{\mathbf{n_0}} - \mathbf{H}_n]\mathbf{e}_n.$$

Označme reziduá

$$\tilde{\mathbf{r}}_n = \mathbf{Y}_n - \hat{\theta}_n\mathbf{1}_n - \mathbf{X}_n\hat{\beta}_n = \sigma[\mathbf{1}_n - \mathbf{H}_{\mathbf{n_0}} - \mathbf{H}_n]$$

Testová štatistika bude založená na štandardizovaných reziduách

$$\mathbf{r}_n = \mathbf{D}_n^{-1/2}\tilde{\mathbf{r}}_n,$$

kde

$$\mathbf{D}_n = \operatorname{diag}(1 - \frac{1}{n} - h_{n,11}, \dots, 1 - \frac{1}{n} - h_{n,nn}),$$

a má tvar

$$\boxed{\widehat{W}_n = n\left\{1 - \frac{(\sum_{i=1}^{n} a_{n,i}^0 r_{n:i})^2}{\sum_{i=1}^{n} r_{n,i}^2}\right\},}$$

kde $a_{n,i}^0$ je ako v (2) a $r_{n:i}$ je opäť príslušná poriadková štatistika.
Dá sa dokázať (viď [2]), že za platnosti hypotézy platí

$$\widehat{W}_n - W_n = o_P\left(\frac{1}{n}\right) \qquad \text{pre } n \to \infty.$$

## SIMULÁCIE

Uvažujeme model (1) pre $\beta \neq \mathbf{0}$ – teda budeme skúmať silu modifikovaného testu. V simuláciách použijeme Mayerovu maticu, a to v počte replikácií 1, 2, 5, 10 a 25, čiže rozsah výberu bude 27, 54, 135, 270 a 675.
Koeficienty $a_{n,i}^0$ sú tabelované len pre $n \le 50$ (viď napr. [3]). Pre $n > 50$ použijeme nasledujúcu aproximáciu (viď [1])

$$\hat{a}_i^* = 2M_i, \qquad i = 2, 3, \dots, n-1$$

$$\hat{a}_1^2 = \hat{a}_n^2 = \frac{\Gamma(\frac{1}{2}(n+1))}{\sqrt{2}\Gamma(\frac{1}{2}n + 1)}.$$

Koeficienty $\hat{a}_2^*, \dots, \hat{a}_{n-1}^*$ je ešte potrebné upraviť

$$\hat{a}_i = \hat{a}_i^* \sqrt{\frac{1 - 2\hat{a}_1^2}{\sum_{i=2}^{n-1} \hat{a}_i^{*2}}}.$$

Kritické hodnoty sme odhadli na základe $10^6$ výberov daného rozsahu, kde chyby pochádzali z normálneho rozdelenia. Pri testovaní hypotézy sme opäť zvolili $10^6$ opakovaní, pričom chyby sme generovali nielen z normálneho rozdelenia (N(0,1), N(0,5)), ale aj Laplaceovho rozdelenia (Lap(0,1), Lap(0,5)), logistického rozdelenia (Log(0,1), Log(0,5)) a t-rozdelenia ($t_1$, $t_5$, $t_{10}$).
V nasledujúcich tabuľkách sú uvedené percentá prípadov, v ktorých bola zamietnutá nulová hypotéza, na hladinách $\alpha = 0{,}01$, $\alpha = 0{,}05$ a $\alpha = 0{,}1$.

| $n = 1 \times 27$ | $\alpha = 0{,}01$ | $\alpha = 0{,}05$ | $\alpha = 0{,}1$ |
|---|---|---|---|
| N(0,1) | 1,00 | 4,96 | 9,96 |
| N(0,5) | 1,02 | 5,00 | 10,00 |
| $t_1$ | 83,78 | 89,58 | 91,99 |
| $t_5$ | 10,11 | 19,01 | 25,86 |
| $t_{10}$ | 3,52 | 9,59 | 15,45 |
| Log(0,1) | 4,32 | 11,25 | 17,54 |
| Log(0,5) | 4,40 | 11,28 | 17,56 |
| Lap(0,1) | 13,22 | 25,29 | 33,58 |
| Lap(0,5) | 13,27 | 25,30 | 33,59 |

| $n = 2 \times 27$ | $\alpha = 0{,}01$ | $\alpha = 0{,}05$ | $\alpha = 0{,}1$ |
|---|---|---|---|
| N(0,1) | 0,99 | 4,98 | 9,97 |
| N(0,5) | 1,02 | 5,01 | 9,99 |
| $t_1$ | 98,70 | 99,35 | 99,57 |
| $t_5$ | 17,97 | 27,91 | 34,72 |
| $t_{10}$ | 5,01 | 11,37 | 17,12 |
| Log(0,1) | 6,42 | 13,94 | 20,19 |
| Log(0,5) | 6,40 | 13,91 | 20,17 |
| Lap(0,1) | 24,93 | 39,76 | 48,59 |
| Lap(0,5) | 24,98 | 39,74 | 48,64 |

| $n = 5 \times 27$ | $\alpha = 0{,}01$ | $\alpha = 0{,}05$ | $\alpha = 0{,}1$ |
|---|---|---|---|
| N(0,1) | 0,99 | 5,00 | 9,98 |
| N(0,5) | 0,99 | 5,03 | 10,02 |
| $t_1$ | 100,00 | 100,00 | 100,00 |
| $t_5$ | 27,90 | 37,90 | 44,17 |
| $t_{10}$ | 4,73 | 9,76 | 14,32 |
| Log(0,1) | 6,63 | 13,21 | 18,62 |
| Log(0,5) | 6,68 | 13,25 | 18,64 |
| Lap(0,1) | 46,90 | 62,65 | 70,56 |
| Lap(0,5) | 46,84 | 62,69 | 70,60 |

| $n = 10 \times 27$ | $\alpha = 0{,}01$ | $\alpha = 0{,}05$ | $\alpha = 0{,}1$ |
|---|---|---|---|
| N(0,1) | 1,01 | 4,97 | 9,94 |
| N(0,5) | 1,00 | 4,99 | 10,01 |
| $t_1$ | 100,00 | 100,00 | 100,00 |
| $t_5$ | 36,86 | 48,97 | 53,06 |
| $t_{10}$ | 3,37 | 6,92 | 10,32 |
| Log(0,1) | 5,62 | 11,31 | 16,17 |
| Log(0,5) | 5,67 | 11,41 | 16,25 |
| Lap(0,1) | 71,53 | 83,14 | 88,00 |
| Lap(0,5) | 71,54 | 83,17 | 87,98 |

| $n = 25 \times 27$ | $\alpha = 0{,}01$ | $\alpha = 0{,}05$ | $\alpha = 0{,}1$ |
|---|---|---|---|
| N(0,1) | 1,00 | 4,98 | 9,91 |
| N(0,5) | 0,98 | 5,01 | 9,99 |
| $t_1$ | 100,00 | 100,00 | 100,00 |
| $t_5$ | 60,96 | 70,38 | 75,08 |
| $t_{10}$ | 2,12 | 4,76 | 7,23 |
| Log(0,1) | 6,56 | 13,16 | 18,42 |
| Log(0,5) | 6,54 | 13,18 | 18,40 |
| Lap(0,1) | 97,87 | 99,13 | 99,47 |
| Lap(0,5) | 97,85 | 99,12 | 99,48 |

## ZÁVER

Z numerických simulácií je zrejmé, že tento modifikovaný test dokáže v modeli lineárnej regresie dobre rozlíšiť medzi normálnym a iným rozdelením chýb, hoci pravdepodobnosti zamietnutia boli nižšie v prípade logistického rozdelenia ako pri použití zvyšných uvažovaných rozdelení. Z uvedených rozdelení bola hypotéza najčastejšie zamietaná v prípade výberu z $t_1$, teda Cauchyho rozdelenia. S rastúcim počtom stupňov voľnosti Studentovho $t$-rozdelenia podľa očakávania klesalo percento prípadov zamietnutia hypotézy. S rastúcim rozsahom výberu sa menila aj sila testu - a to najvýraznejšie pre Laplaceovo rozdelenie, kde pre najväčší uvažovaný rozsah výberu bola rovná takmer 1.

## POĎAKOVANIE

Na tomto mieste by som sa rada poďakovala prof. RNDr. Jane Jurečkovej, DrSc. za cenné rady a nápady, ktoré mi pomohli pri písaní tohto posteru a MFF UK za príspevok umožňujúci účasť na tejto konferencii.

### Literatúra

**[1]** Jurečková J. a Picek J. (2007). *Shapiro-Wilk-type Test of Normality under Nuisance Regression and Scale.* Computational Statistics & Data Analysis **51**, 5184 – 5191.

**[2]** Sen P. K., Jurečková J. a Picek J. (2003). *Goodness-of-Fit Test of Shapiro-Wilk Type with Nuisance Regression and Scale.* Austrian Journal of Statistics **1 & 2**, 163 – 177.

**[3]** Shapiro S. S. a Wilk M. B. (1965). *An Analysis of Variance for Normality (Complete Samples).* Biometrika **52**, 591 – 611.