TESTOVÁNÍ RŮZNÝCH TYPŮ REZIDUÍ V REGRESNÍ DIAGNOSTICE

JAVŮREK M., TAUFER I.
Univerzita Pardubice, Fakulta elektrotechniky a informatiky, Katedra řízení procesů, Milan.Javurek@upce.cz

1 Úvod

Regrese je jednou z nejběžnějších a nejoblíbenějších metod aproximace experimentálních závislostí. Principem je optimalizace (tj. minimalizace) účelové funkce, nejčastěji v podobě metody nejmenších čtverců, která nám vyjadřuje těsnost proložení regresní a experimentální závislosti. Prokládaná funkce může být známa v analytickém tvaru, kdy parametry mají přímo fyzikální význam, nebo používáme různé typy matematických závislostí. Základní dělení regresních metod je podle parametrů prokládaných závislostí, tj. známe regresi lineární a nelineární. Zatímco regrese lineární se vyhodnocuje podle jednoznačných vzorců, tzv. normálních rovnic, průběh a výsledky regrese nelineární jsou ovlivněny řadou faktorů, jako jsou např. počáteční odhad parametrů, použitá optimalizační metoda, podmíněnost jednotlivých parametrů atd. Z tohoto důvodu nemusí být nalezené parametry správné ani přesné, zejména tehdy, pokud neznáme ani jejich přibližné hodnoty. Nelineární regrese nám nabízí poměrně málo nástrojů pro verifikaci nalezených parametrů. Pokud provedeme výpočtů více s různými modely, používají se pro jejich porovnání Akaikeho informační kritérium (AIC), střední kvadratická chyba predikce (MEP), hodnota účelové funkce (RSC) atd. [MELOUN, 2004]. Pokud ovšem máme výpočet jenom jeden, jedním z mála dostupných nástrojů je analýza souboru reziduí. Je zcela paradoxní, že většina i komerčních programů, např. STATISTICA, tuto analýzu nezahrnuje, pouze v algoritmickém režimu poskytne hodnoty parametrů s jejich směrodatnými odchylkami. Další verifikace se však neprovádí a nemůžeme zhodnotit kvalitu proložení regresní a experimentální závislosti.

Pokud jsou splněny podmínky aplikace regresní metody (data nevykazují heteroskedasticitu, supernormalitu, multikolinearitu, autokorelaci, odlehlé body, model je významný), měl by soubor reziduí vykazovat normální rozdělení, což lze prokázat na základě spočtených hodnot centrálních momentů, Pearsonovým, znaménkovým a dalšími testy. Rezidua však lze definovat různými způsoby a jejich vyjadřovací schopnost je různá.

2 Definice různých typů reziduí

2.1 Klasická rezidua $\hat{e}_i$ [MELOUN, 2004]

Tato rezidua jsou definována jako rozdíl vypočtených a experimentálních hodnot. Jsou korelovaná, nemají konstantní rozptyl a nemusí správně indikovat vychýlené body.

2.2 Normovaná rezidua $\hat{e}_{Ni}$ [MELOUN, 2004]

V tomto případě normování spočívá ve vydělení hodnoty klasického rezidua hodnotou směrodatné odchylky celého souboru. Soubor reziduí má mít normální rozdělení s nulovou střední hodnotou a jednotkovou směrodatnou odchylkou. Hodnoty větší než trojnásobek směrodatné odchylky jsou brány jako odlehlé.

Z matematické analýzy však vyplývá, že rozptyl $D(\hat{e}_{Ni}) = (1-H_{ii})$ není konstantní ani jednotkový, takže doporučované vylučování hodnot přesahujících interval trojnásobku směrodatné odchylky nemusí být správné.

2.3 Standardizovaná rezidua $\hat{e}_{Si}$ [MELOUN, 2004]

Také mají mít normální rozdělení s konstantním rozptylem, jsou definována:

$$\hat{e}_{Si} = \frac{\hat{e}_i}{\hat{\sigma}\sqrt{1-H_{ii}}} \qquad (1)$$

kde $\hat{\sigma}$ je směrodatná odchylka, H_{ii} jsou diagonální prvky projekční matice $H = X (X^T X)^{-1} X^T$.

Jejich vlastnosti jsou téměř shodné s klasickými.

2.4 Jackknife rezidua $\hat{e}_{Ji}$ [MELOUN, 2004]

Pokud v (1) použijeme místo celkové směrodatné odchylky její odhad získaný při vynechání i-tého bodu:

$$\hat{e}_{Ji} = \sqrt{\frac{n-m-1}{n-m-\hat{e}_{Si}}} \qquad (2)$$

kde n je počet měření, m je počet určovaných parametrů.

Tato rezidua mají za předpokladu normality chyb Studentovo rozdělení s n-m-1 stupni volnosti. Tato rezidua se používají k identifikaci odlehlých bodů.

2.5 Predikovaná rezidua $\hat{e}_{Pi}$ [MELOUN, 2004]

Jsou definována:

$$\hat{e}_{Pi} = y_i - x_i \, \mathbf{b}_{(i)} = \frac{\hat{e}_i}{1-H_{ii}} \qquad (3)$$

kde x je nezávisle proměnná(é) a y je závisle proměnná veličina, $\mathbf{b}_{(i)}$ jsou odhady parametrů získané metodou nejmenších čtverců ze všech bodů kromě i-tého

3 Další diagnostické nástroje

3.1 Cookova vzdálenost Di [MELOUN, 2004]

Je to vlastně eukleidovská vzdálenost mezi vektorem predikce závisle proměnné získaném MNČ a tímtéž vektorem při vynechání i-tého bodu. Cookova vzdálenost vyjadřuje vliv i-tého bodu pouze na odhady parametrů. Je definována:

$$D_i = \frac{\hat{e}_{Si}}{m}\frac{H_{ii}}{1-H_{ii}} \qquad (4)$$

3.2 Atkinsonova vzdálenost [MELOUN, 2004]

Používá se ke zvýraznění citlivosti regrese na extrémní body. Je definována:

$$A_i = |\hat{e}_{Ji}|\sqrt{\frac{n-m}{m}\frac{H_{ii}}{1-H_{ii}}} \qquad (5)$$

3.3 Věrohodnostní vzdálenosti [MELOUN, 2004]

Tato veličina je rozdíl logaritmů věrohodnostní funkce při použití všech bodů a při vynechání i-tého bodu. Pokud je hodnota větší než kvantil $X^2_{1-\alpha}(m+1)$ rozdělení, je daný bod považován za vlivný.

3.4 Souhrnné charakteristiky vlastností celého souboru reziduí

Pro zjištění platnosti základních předpokladů aplikace MNČ z vlastností celého souboru reziduí (použita vždy klasická rezidua) byly použity také tyto charakteristiky:

– Cookův – Weisbergův test heteroskedasticity,
– Jarque – Berrův test normality,
– Waldův test autokorelace,
– Znaménkový test.

Popis těchto testů je poněkud složitý a zdlouhavý, proto viz [MELOUN, 2004].

Pokračování na další straně

Tab. 1 – Charakteristiky souboru reziduí výchozích dat

Původní data					
Typ reziduí	**Klasická**	**Normovaná**	**Standardiz.**	**Jackknife**	**Predikovaná**
Účelová fce	0,0030	1,7373	29,9600	32,2588	0,0035
Aritm. průměr	0,0000	-0,0430	0,3816	-0,0058	0,0000
Směrodatná odchylka	0,0101	0,2368	0,9236	1,0369	0,0108
Koef. šikmosti	-0,2084	-3,0775	-1,2577	-0,2908	-0,1922
Koef. špičatosti	0,0298	11,6552	2,3077	0,3215	-0,0634
R-faktor	0,0006	0,0136	0,0563	0,0584	0,0006
AIC	-227,0000	–	–	–	–
MEP	0,0001	–	–	–	–
Heterosked.	ano	–	–	–	–
Normalita	ano	–	–	–	–
Autokorelace	ne	–	–	–	–
Znam. test	negat.	–	–	–	–

Tab. 2 – Různé varianty výpočtu pro klasická rezidua

Klasická rezidua						
Násobek charakt. původ. souboru	**1. bod +1 s**	**1. bod + 2s**	**1. bod + 3s**	**1. bod +1s, 2. bod -1s**	**1. bod +2s, 2. bod -2s**	**1. bod +3s, 2. bod -3s**
Účelová fce	0,9596	3,1891	11,1415	1,0754	3,6223	12,1170
Aritm. průměr	-0,8662	2,4045	5,1910	0,0000	2,8989	0,6816
Směrodat. odchylka	0,9796	1,7858	3,3903	1,0370	1,9032	3,4809
Koef. šikmosti	1,2973	-12,1966	-19,1828	2,0439	-9,8283	-17,4348
Koef. špičatosti	4,7594	350,2452	655,3301	7,9841	317,6662	606,9347
R-faktor	0,9796	1,7858	3,3378	1,0370	1,9032	3,4809
Další kriteria						
AIC	-273,2000	-273,2000	-199,7000	-269,8000	-233,4000	-197,1000
MEP	0,0001	0,0004	0,0014	0,0001	0,0005	0,0016
Heterosked.	ano	ano	ano	ano	ano	ano
Normalita	ano	ano	ne	ano	ano	ne
Autokorelace	ne	ne	ne	ne	ne	ne
Znam. test	negat.	negat.	negat.	negat.	negat.	negat.

Zmíněné diagnostické nástroje (kromě 3.4) jsou zpravidla používány pouze pro odhalení významných (odlehlých, vychýlených) bodů, což z hlediska celkového pohledu na kvalitu proložení regresního modelu nemá tak velký význam. Pokud model není vhodný, ani hledání těchto bodů nemá smysl. Pro posouzení těsnosti proložení se ve většině případů používají klasická rezidua, ale jejich informační efektivita je velice malá, při výpočtu statistických momentů nepostihneme nejen odlehlé body, ale ani trendy v proložení. Zde se především uplatní testy uvedené v 3.4, dále ještě testování vhodnosti celého modelu F-testem či Studentův test významnosti jednotlivých parametrů. Dva poslední testy zde nebyly uvažovány, model byl znám. Velmi vypovídající je také Hamiltonův R-faktor:

Tab. 3 – Různé varianty výpočtu pro normovaná rezidua

Normovaná rezidua						
Násobek charakt. původ. souboru	**1. bod +1 s**	**1. bod + 2s**	**1. bod + 3s**	**1. bod +1s, 2. bod -1s**	**1. bod +2s, 2. bod -2s**	**1. bod +3s, 2. bod -3s**
Účelová fce	0,4125	29,8402	112,5996	0,8572	33,3538	121,3035
Aritm. průměr	0,4781	-3,6503	-7,2397	0,6204	-3,3425	-6,7443
Směrodat. odchylka	0,6469	5,5122	10,8850	0,9342	5,8382	11,1266
Koef. šikmosti	0,7344	-1,6724	-1,6877	1,2930	-1,5192	-1,6224
Koef. špičatosti	0,8122	2,3804	2,4561	1,6985	2,1565	2,3071
R-faktor	0,6422	5,4626	10,6110	0,9259	5,7752	11,0136
Další kriteria						
AIC	-273,2000	-273,2000	-199,7000	-269,8000	-233,4000	-197,1000
MEP	0,0001	0,0004	0,0014	0,0001	0,0005	0,0016
Heterosked.	ano	ano	ano	ano	ano	ano
Normalita	ano	ano	ne	ano	ano	ne
Autokorelace	ne	ne	ne	ne	ne	ne
Znam. test	negat.	negat.	negat.	negat.	negat.	negat.

Tab. 4 – Různé varianty výpočtu pro standardizovaná rezidua

Standardizovaná rezidua						
Násobek charakt. původ. souboru	**1. bod +1 s**	**1. bod + 2s**	**1. bod + 3s**	**1. bod +1s, 2. bod -1s**	**1. bod +2s, 2. bod -2s**	**1. bod +3s, 2. bod -3s**
Účelová fce	17,2142	1,0595	1,0595	1,0034	1,0487	1,0606
Aritm. průměr	-0,0003	0,0145	0,0165	-0,0024	0,0112	0,0139
Směrodat. odchylka	1,0810	1,1068	1,1312	1,0838	1,1080	1,1143
Koef. šikmosti	0,2081	-2,0917	-3,2168	0,3467	-1,6840	-2,9227
Koef. špičatosti	0,0453	4,7456	8,6059	0,1070	4,2868	7,9575
R-faktor	0,9991	1,0229	1,0293	1,0017	1,0241	1,0298
Další kriteria						
AIC	-273,2000	-273,2000	-199,7000	-269,8000	-233,4000	-197,1000
MEP	0,0001	0,0004	0,0014	0,0001	0,0005	0,0016
Heterosked.	ano	ano	ano	ano	ano	ano
Normalita	ano	ano	ne	ano	ano	ne
Autokorelace	ne	ne	ne	ne	ne	ne
Znam. test	negat.	negat.	negat.	negat.	negat.	negat.

$$R = \sqrt{\frac{RSC}{\sum y_i^2}} \qquad (6)$$

Pokud hodnota R-faktoru není větší než nejistota měření, lze proložení považovat za dobré. Použití klasických reziduí by však mohlo být nahrazeno jiným typem, který je citlivější a lépe postihne odchylky v proložení regresní závislosti. V dalším se tedy pokusíme vybrat vhodnější z nabízených typů reziduí.

4 Testování

Testování jednotlivých typů reziduí bylo prováděno na simulovaném příkladu lineární závislosti. Pro zadané hodnoty nezávisle proměnné a zadané parametry přímky byly vypočteny hodnoty závisle proměnné. Tyto byly zatíženy chybami s normálním rozdělením. Získaná data byla vyhodnocena lineární regresí a spočteny

Tab. 5 – Různé varianty výpočtu pro Jackknife rezidua

Jackknife rezidua						
Násobek charakt. původ. souboru	**1. bod +1 s**	**1. bod + 2s**	**1. bod + 3s**	**1. bod +1s, 2. bod -1s**	**1. bod +2s, 2. bod -2s**	**1. bod +3s, 2. bod -3s**
Účelová fce	1,0026	2,3131	9,0801	1,0122	2,0193	6,2412
Aritm. průměr	1,0198	-21,0248	-70,0348	1,7396	-16,2757	-52,5175
Směro-dat. odchylka	1,0013	1,5164	3,0392	1,0060	1,4182	2,4812
Koef. šikmosti	1,2606	-14,4374	-18,0003	1,8302	-12,6178	-17,8036
Koef. špičatosti	1,4778	64,8356	89,5911	1,6265	56,6179	86,6017
R-faktor	1,0013	1,5209	3,0133	1,0061	1,4210	2,4982
Další kriteria						
AIC	-273,2000	-273,2000	-199,7000	-269,8000	-233,4000	-197,1000
MEP	0,0001	0,0004	0,0014	0,0001	0,0005	0,0016
Heteros-ked.	ano	ano	ano	ano	ano	ano
Normalita	ano	ano	ne	ano	ano	ne
Autokore-lace	ne	ne	ne	ne	ne	ne
Znam. test	negat.	negat.	negat.	negat.	negat.	negat.

Tab. 6 – Různé varianty výpočtu pro predikovaná rezidua

Predikovaná rezidua						
Násobek charakt. původ. souboru	**1. bod +1 s**	**1. bod + 2s**	**1. bod + 3s**	**1. bod +1s, 2. bod -1s**	**1. bod +2s, 2. bod -2s**	**1. bod +3s, 2. bod -3s**
Účelová fce	0,9561	3,4940	12,5057	1,0829	3,9852	13,6282
Aritm. průměr	0,0956	-8,0159	-17,0309	0,8011	-6,6097	-14,9265
Směro-dat. odchylka	0,9778	1,8691	3,5919	1,0406	1,9962	3,6915
Koef. šikmosti	1,3179	-14,1426	-21,2923	2,3294	-11,3779	-19,3415
Koef. špičatosti	-1,0926	-181,0766	-318,7822	-4,1581	-162,9381	-294,3271
R-faktor	0,9778	1,8692	3,5363	1,0406	1,9963	3,6916
Další kriteria						
AIC	-273,2000	-273,2000	-199,7000	-269,8000	-233,4000	-197,1000
MEP	0,0001	0,0004	0,0014	0,0001	0,0005	0,0016
Heteros-ked.	ano	ano	ano	ano	ano	ano
Normalita	ano	ano	ne	ano	ano	ne
Autokore-lace	ne	ne	ne	ne	ne	ne
Znam. test	negat.	negat.	negat.	negat.	negat.	negat.

centrální momenty souboru reziduí. V dalším byla hodnota závisle proměnné prvního bodu zvýšena postupně o jedno-, dvoj- a trojnásobek směrodatné odchylky z původního souboru a opět provedeno vyhodnocení jednotlivých souborů dat lineární regresí a spočteny centrální momenty souborů reziduí. Tyto hodnoty byly vztahovány k hodnotám centrálních momentů původního souboru – tj. do jaké míry se změna jednoho bodu projeví změnou charakteristik souboru reziduí. V dalším byly přidány ke změnám prvního bodu také změny druhého bodu v opačném směru než u prvního bodu.

Kromě popsaného testování byly ještě prováděny výpočty se soubory, u kterých byly již popsané změny realizovány uprostřed závislosti – to se však při vyhodnocení regresního procesu téměř neprojevilo a nebylo to dále testováno. To souvisí s různou podmíněností parametrů v celém průběhu regresní závislosti – viz např. [MELOUN, 1984].

5 Závěr

Z provedeného testování jednoznačně vyplývá, že nejvhodnější pro posuzování těsnosti proložení jsou normovaná rezidua (tab. 3). Na statistických charakteristikách souborů normovaných reziduí se nejvíce projevují prováděné změny v analyzovaných datech. Ovšem tyto změny se týkají kritérií, jako jsou hodnota účelové funkce, první a druhý centrální moment a R-faktor. Parametry charakterizující tvar pravděpodobnostního rozdělení se však zásadně nemění, což opět mluví ve prospěch tohoto typu reziduí.

Podobně lze hodnotit predikovaná rezidua (tab. 6), ale zde je variabilita menší než u normovaných reziduí.

Za predikovaná rezidua lze zařadit Jackknife rezidua (tab. 5), zde se variabilita projevovala ještě o něco méně. Tento typ je však velmi užitečný při tipování významných bodů.

U klasických reziduí (tab. 2) se první skupina kritérií mění málo – je tedy problematické vyhodnotit změny v proložení – a druhá skupina (3. a 4. centrální moment) se mění velice výrazně, z čehož lze usuzovat naprostá nevhodnost tohoto typu pro posouzení kvality proložení – tj. snižuje se normalita souboru reziduí.

Variabilita se téměř neprojevila u standardizovaných reziduí (tab. 4), jejich použití tudíž postrádá jakýkoli smysl.

Hodnoty kritérií pro porovnání kvality proložení mezi jednotlivými soubory (tj. AIC a MEP) jasně kopírují zhoršující se podmínky výpočtu. To je však také jediné, co nám tyto charakteristiky říkají. Zajímavé však je, že hodnota AIC i MEP je pro první dvě varianty výpočtu lepší než u základního souboru – to je dáno tím, že první bod v základním souboru má nižší experimentální hodnotu závisle proměnné než predikovanou, tudíž při změnách se regrese zpočátku zlepšuje.

Zbylé charakteristiky (test heteroskedasticity, normality rozdělení, autokorelace a znaménkový) mají vypovídací schopnost poměrně malou a jejich užití může být pouze orientační. Úspěšně mohou být nahrazeny mapou rozložení reziduí okolo nulové hodnoty, kde trendy, jako je heteroskedasticita, normalita, autokorelace či střídání znamének, mohou být posouzeny objektivněji pouhým pohledem.

Literatura

[1] Meloun, M., Javůrek, M., Miltiparametric Curve Fitting VIII. The Reliability of Dissociation Constants Estimated by Analysis of Abstorbance-pH Curves. *Talanta* 32(10), (1985), pp. 973–986, ISSN 0039-9140

[2] Meloun, M., Militký, J. Statistická analýza experimentálních dat. Praha: *Academia* 2004, ISBN 80-200-1254-0, 953 s.

Problematika je řešena v rámci výzkumného záměru MŠM 0021627505 „Řízení, optimalizace a diagnostika složitých systémů“, programu vědeckovýzkumné spolupráce ČR a SR KONTAKT MŠMT č. MEB 0810003 „Identifikace a řízení složitých nelineárních soustav s využitím metod umělé inteligence“ a projektu Univerzity Pardubice SGFEI06/2011 „Artificial Intelligence Control Toolbox pro MATLAB“.

Abstract

TESTS OF VARIOUS TYPES OF RESIDUALS IN REGRESSION DIAGNOSTICS

Summary: *Approximation of experimental data by means of an analytical or general mathematical dependence is performed most frequently by the regression method using the least squares approach. The quality of curve fitting is evaluated on the basis of analysis of resulting set of residuals which, however, can be defined in various ways. This paper deals with suitability tests of the individual types from the standpoint of curve fitting quality of the regression dependence.*

Key words: *analysis of residuals, various types of residuals, regression diagnostics*