

6.2 Validace nové analytické metody

Vzorová úloha 6.2 Postup validace a regresní diagnostika

Na úloze V6.14 Validace stanovení amonných iontů v pitných vodách provedte ověření časově nenáročné metody stanovení obsahu amonných iontů y soupravou Spektroquant, a to srovnáním se standardní metodou x stanovení amoniaku podle ČSN - ISO 7150-1, která je však náročná na provedení. Pro účely vyhodnocení se předpokládá, že rozptyl obsahu u standardní metody je zanedbatelný. (1) Vyšetřete statistickou významnost úseku b_0 (má být $\beta_0 = 0$). (2) Odstraňte z dat odlehlé hodnoty. (3) K jakým závěrům vede kombinovaný test úseku a směrnice?

Řešení:

1. Návrh modelu: navrhne regresní model (přímky) $y = \beta_0 + \beta_1 x$, u kterého budeme testovat nulovou hypotézu $H_0: \beta_0 = 0, \beta_1 = 1$.

2. Předběžná analýza dat: poloha a proměnlivost proměnných y, x se posuzuje na základě průměru a směrodatné odchylky hodnot každé proměnné. Pearsonův párový korelační koeficient ukazuje vysokou korelaci proměnných y a x .

Proměnná	Průměr	Směrodatná odchylka	Párový korelační koeficient	Spočtená hladina významnosti
y	2.0443E-01	1.9296E-01	1.0000	-----
x	2.1048E-01	2.0848E-01	0.9958	0.000

3. Odhadování parametrů: klasickou metodou nejmenších čtverců (MNČ) byly nalezeny odhady parametrů, úseku β_0 a směrnice β_1 . Studentův t -test ukázal, že úsek (absolutní člen) β_0 je statisticky nevýznamný, zatímco směrnice β_1 je statisticky významná, když $t_{0,95}(20-2) = 2.101$.

Parametr	Odhad	Směrodatná odchylka	$H_0: b_j = 0$ vs. $H_A: b_j \neq 0$ t -kritérium	Spočtená hlad. význam.
b_0	0.010432	0.00566	1.8436	Akceptována 0.081
b_1	0.92170	0.01933	47.681	Zamítnuta 0.000

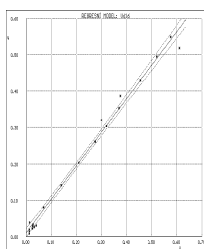
4. Základní statistické charakteristiky: párový korelační koeficient r ukazuje, že navržený lineární regresní model je statisticky významný. Vysoká hodnota koeficientu determinace D (= 99.17 %), představující procento bodů vyhovujících regresnímu modelu, ukazuje, že všechny body výtečně korespondují s modelem přímky. Střední kvadratická chyba predikce MEP a Akaiikovo informační kritérium AIC se užívají k rozlišení mezi několika navrženými modely. Za optimální se považuje model, pro který dosahuje MEP a AIC minimální hodnotu.

Vícenásobný korelační koeficient, r	: 0.99585
Koeficient determinace, D [%]	: 99.171
Predikovaný koeficient determinace, R^2_p	: 0.99414
Střední kvadratická chyba predikce, MEP	: 4.1406E-04
Akaiikovo informační kritérium, AIC	: -166.78

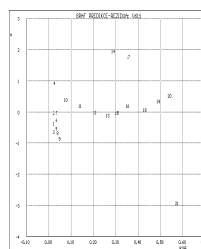
5. Regresní diagnostika: obsahuje pomůcky a postupy pro interaktivní analýzu (a) dat, (b) modelu, (c) metody, což jsou složky tzv. *regresního tripletu*.

Kritika dat: věrohodnost nalezených odhadů parametrů β_0, β_1 lze posoudit na základě grafu regresního modelu (obr. 6.2-1a).

(a) *Analýza klasických reziduí* není příliš spolehlivá a nemusí indikovat silně odlehlé hodnoty. Grafická analýza \hat{e} vs. \hat{y}_p (obr. 6.2-1b) je schopna indikovat podezřelé body, trend a heteroskedasticitu. Míry polohy a rozptýlení klasických reziduí by měly dosahovat hodnot, blízkých experimentálnímu šumu. *Odhad směrodatné odchylky $s(e)$* se totiž blíží svou velikostí experimentální chybě, kterou je zatížena závisle proměnná. Odhad šikmosti a špičatosti nedokazují Gaussovo normální rozdělení reziduí, normalitu.



Obr. 6.2-1a Graf regresního modelu, ADSTAT.



Obr. 6.2-1b Analýza klasických reziduí, ADSTAT.

Bod	Měřená hodnota	Predikovaná hodnota	Směrodatná odchylka	Klasické reziduum	Relativní reziduum
i	$y_{exp, i}$	$y_{vyp, i}$	$s(y_{vyp, i})$	e_i	$e_{r, i}$
1	1.5000E-02	2.1493E-02	5.4943E-03	-6.4929E-03	-4.3286E+01
2	2.1000E-02	2.1493E-02	5.4943E-03	-4.9287E-04	-2.3470E+00
3	1.0000E-02	2.1493E-02	5.4943E-03	-1.1493E-02	-1.1493E+02
4	4.0000E-02	2.3336E-02	5.4674E-03	1.6664E-02	4.1659E+01
5	2.4000E-02	3.3475E-02	5.3219E-03	-9.4750E-03	-3.9479E+01
6	3.0000E-02	3.4397E-02	5.3089E-03	-4.3967E-03	-1.4656E+01
7	3.5000E-02	3.5318E-02	5.2959E-03	-3.1838E-04	-9.0966E-01
8	2.8000E-02	3.9927E-02	5.2317E-03	-1.1927E-02	-4.2596E+01
9	3.2000E-02	4.7300E-02	5.1310E-03	-1.5300E-02	-4.7814E+01
10	8.2000E-02	7.4952E-02	4.7793E-03	7.0485E-03	8.5957E+00
11	1.4300E-01	1.3947E-01	4.1622E-03	3.5294E-03	2.4681E+00
12	2.0400E-01	2.0399E-01	3.9329E-03	1.0337E-05	5.0673E-03
13	2.6200E-01	2.6390E-01	4.1260E-03	-1.9002E-03	-7.2528E-01
14	3.2200E-01	2.8694E-01	4.2968E-03	3.5057E-02	1.0887E+01
15	3.0500E-01	3.0538E-01	4.4666E-03	-3.7676E-04	-1.2353E-01
16	3.5500E-01	3.5146E-01	4.9977E-03	3.5382E-03	9.9666E-01
17	3.8800E-01	3.5607E-01	5.0579E-03	3.1930E-02	8.2293E+00
18	4.3100E-01	4.2981E-01	6.1490E-03	1.1936E-03	2.7693E-01
19	4.9600E-01	4.8972E-01	7.1601E-03	6.2830E-03	1.2667E+00
20	5.5000E-01	5.4041E-01	8.0697E-03	9.5895E-03	1.7435E+00
21	5.2000E-01	5.7267E-01	8.6668E-03	-5.2670E-02	-1.0129E+01
Reziduální součet čtverců, RSC			: 6.1717E-03		
Průměr absolutních hodnot reziduí, M_e			: 1.0937E-02		
Průměr relativních reziduí, M_{erel}			: 1.8720E+01		
Odhad reziduálního rozptylu, $s^2(e)$: 3.2482E-04		
Odhad směrodatné odchylky reziduí, $s(e)$: 1.8023E-02		
Odhad šikmosti reziduí, $g_1(e)$: -6.8328E-01		
Odhad špičatosti reziduí, $g_3(e)$: 5.7550E+00		

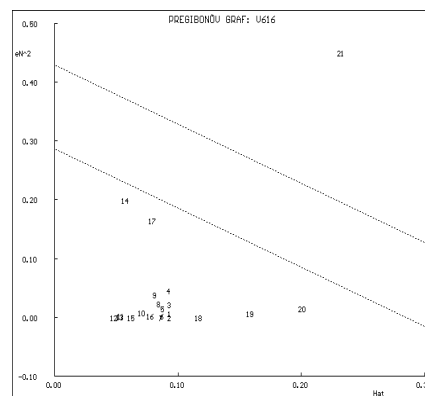
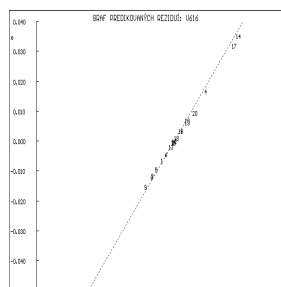
(b) **Analýza ostatních reziduí:** Jackknife rezidua indikují odlehlé body, z diagonálních prvků H_{ii} projekční matice H a diagonálních prvků H_{mii} zobecněné projekční matice H_m pouze extrémy. Ostatní druhy reziduí a kritéria v tabulce indikují obecně vlivné body (značeno hvězdičkou u hodnoty). Jackknife rezidua $e_{j,i}$ ukazují, že body č. 14 a 21 jsou odlehlé, stejně tak i Cookova vzdálenost D_i ; Atkinsonova vzdálenost A_i na č. 14, 17, 21; kritérium DF_i na č. 21, věrohodnostní vzdálenosti $LD(b)_i$, $LD(s^2)_i$ na č. 21 a $LD(b, s^2)_i$. Diagonální prvky H_{ii} projekční matice H ukazují na extrémy č. 20, 21, a diagonální prvky zobecněné H_{mii} projekční matice H_m pak na extrémy č. 21.

INDIKACE VLIVNÝCH BODU: (* indikuje odlehlý nebo vlivný bod)			
Bod Standardizované reziduum	Jackknife reziduum	Predikované reziduum	Diagonální prvky

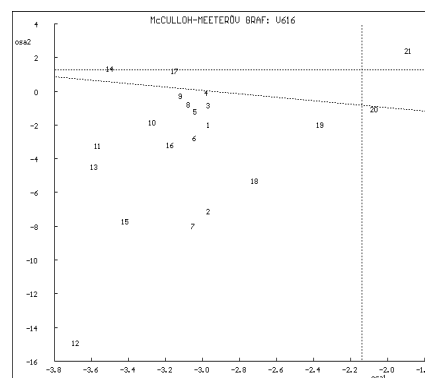
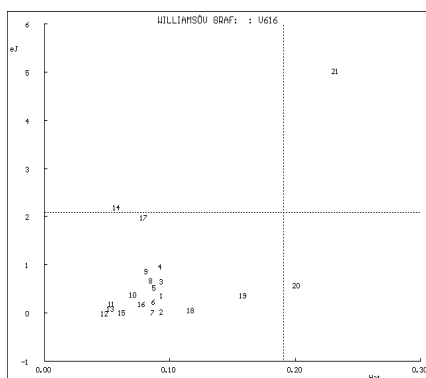
<i>i</i>	e_{Si}	e_{Ji}	e_{Pi}	H_{ii}
1	-3.7826E-01	-3.6957E-01	-7.1581E-03	9.2936E-02
2	-2.8713E-02	-2.7948E-02	-5.4336E-04	9.2936E-02
3	-6.6955E-01	-6.5952E-01	-1.2670E-02	9.2936E-02
4	9.7031E-01	9.6874E-01	1.8353E-02	9.2027E-02
5	-5.5026E-01	-5.3990E-01	-1.0380E-02	8.7194E-02
6	-2.5528E-01	-2.4889E-01	-4.8144E-03	8.6768E-02
7	-1.8481E-02	-1.7989E-02	-3.4847E-04	8.6345E-02
8	-6.9154E-01	-6.8173E-01	-1.3024E-02	8.4263E-02
9	-8.8560E-01	-8.8034E-01	-1.6650E-02	8.1052E-02
10	4.0561E-01	3.9651E-01	7.5816E-03	7.0320E-02
11	2.0127E-01	1.9611E-01	3.7282E-03	5.3333E-02
12	5.8773E-04	5.7205E-04	1.0854E-05	4.7619E-02
13	-1.0831E-01	-1.0546E-01	-2.0053E-03	5.2408E-02
14	2.0029E+00	2.1949E+00*	3.7170E-02	5.6839E-02
15	-2.1578E-02	-2.1003E-02	-4.0142E-04	6.1418E-02
16	2.0433E-01	1.9910E-01	3.8329E-03	7.6894E-02
17	1.8458E+00	1.9831E+00	3.4659E-02	7.8758E-02
18	7.0453E-02	6.8582E-02	1.3508E-03	1.1640E-01
19	3.7988E-01	3.7116E-01	7.4605E-03	1.5783E-01
20	5.9505E-01	5.8466E-01	1.1994E-02	2.0048E-01*
21	-3.3331E+00	-5.0342E+00*	-6.8513E-02	2.3124E-01*
Bod	Zobecněné diag.	Cookova	Atkinsonova	Vliv na
	prvky	vzdálenost	vzdálenost	predikci
<i>i</i>	H_{mii}	D_i	A_i	DF_i
1	9.9767E-02	7.3300E-03	3.6461E-01	-1.1830E-01
2	9.2975E-02	4.2236E-05	2.7573E-02	-8.9460E-03
3	1.1434E-01	2.2966E-02	6.5067E-01	-2.1111E-01
4	1.3702E-01	4.7713E-02	9.5058E-01	3.0841E-01
5	1.0174E-01	1.4461E-02	5.1431E-01	-1.6687E-01
6	8.9900E-02	3.0958E-03	2.3647E-01	-7.6720E-02
7	8.6361E-02	1.6139E-05	1.7045E-02	-5.5300E-03
8	1.0731E-01	2.2002E-02	6.3739E-01	-2.0680E-01
9	1.1898E-01	3.4587E-02	8.0583E-01	-2.6145E-01
10	7.8370E-02	6.2219E-03	3.3611E-01	1.0905E-01
11	5.5351E-02	1.1411E-03	1.4347E-01	4.6548E-02
12	4.7619E-02	8.6356E-09	3.9426E-04	1.2791E-04
13	5.2994E-02	3.2441E-04	7.6440E-02	-2.4800E-02
14	2.5598E-01	1.2088E-01*	1.6608E+00	5.3883E-01
15	6.1441E-02	1.5234E-05	1.6559E-02	-5.3726E-03
16	7.8922E-02	1.7389E-03	1.7711E-01	5.7462E-02
17	2.4395E-01	1.4563E-01*	1.7872E+00	5.7985E-01
18	1.1663E-01	3.2694E-04	7.6724E-02	2.4892E-02
19	1.6423E-01	1.3522E-02	4.9524E-01	1.6068E-01
20	2.1538E-01	4.4394E-02	9.0237E-01	2.9277E-01
21	6.8074E-01*	1.6709E+00*	8.5100E+00*	-2.7610E+00*
Bod	Věrohodnostní vzdálenosti			
<i>i</i>	$LD(b)_i$	$LD(s^2)_i$	$LD(b,s^2)_i$	
1	1.6197E-02	1.7607E-02	3.3156E-02	
2	9.3364E-05	2.4550E-02	2.4639E-02	
3	5.0706E-02	6.4629E-03	5.5981E-02	
4	1.0521E-01	4.3416E-05	1.0573E-01	
5	3.1943E-02	1.1124E-02	4.2062E-02	
6	6.8422E-03	2.1276E-02	2.7816E-02	
7	3.5677E-05	2.4575E-02	2.4609E-02	
8	4.8581E-02	5.6554E-03	5.3172E-02	
9	7.6317E-02	4.6137E-04	7.6411E-02	
10	1.3749E-02	1.6655E-02	2.9869E-02	

11	2.5223E-03	2.2505E-02	2.4912E-02
12	1.9089E-08	2.4593E-02	2.4593E-02
13	7.1711E-04	2.3980E-02	2.4663E-02
14	2.6552E-01	3.9713E-01	7.1973E-01
15	3.3675E-05	2.4569E-02	2.4601E-02
16	3.8434E-03	2.2442E-02	2.6111E-02
17	3.1948E-01	2.4454E-01	6.1812E-01
18	7.2270E-04	2.4333E-02	2.5021E-02
19	2.9870E-02	1.7552E-02	4.6238E-02
20	9.7905E-02	9.3426E-03	1.0458E-01
21	3.4023E+00	9.7291E+00*	1.8199E+01*

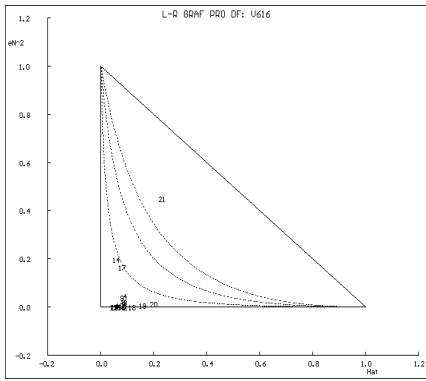
(c) **Grafy vlivných bodů** (obr. 6.2-2) jsou schopny indikovat přítomnost odlehlých hodnot a extrémů. *Graf predikovaných reziduí* ukazuje na odlehlé body č. 21, 14, 17. *Pregibonův graf* ukazuje na silně vlivný bod č. 21. *Williamsův graf* indikuje č. 14 a 21 jako odlehlé body a jako extrémý č. 20, 21. *McCullohův-Meeterův graf* dokazuje odlehlé body č. 14, 17, 21 a extrémý č. 20, 21. Konečně *L-R graf* dokazuje odlehlé body č. 14, 17, 21 a současně extrém č. 20. Lze uzavřít, že body č. 14, 21 jsou většinou diagnostik indikovány jako odlehlé.



Obr. 6.2-2 Grafy vlivných bodů, vlevo, graf predikovaných reziduí, a vpravo, Pregibonův graf, *ADSTAT*.

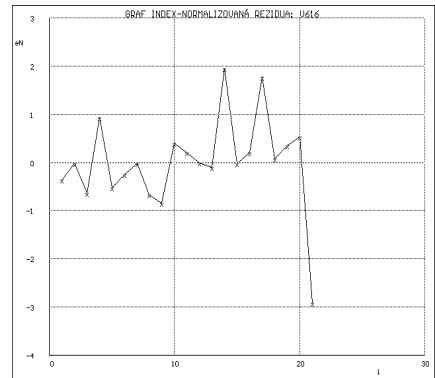
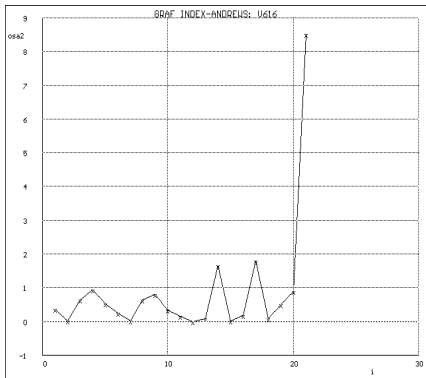


Obr. 6.2-2 Grafy vlivných bodů, vlevo, Williamsův graf, a vpravo, McCullohův-Meeterův graf, *ADSTAT*.

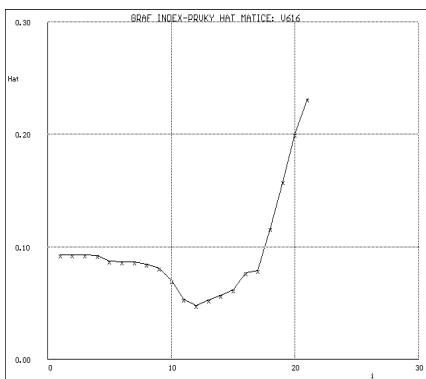


Obr. 6.2-2 Grafy vlivných bodů, L-R graf, *ADSTAT*.

(d) **Indexové grafy** (obr. 6.2-3) upozorňují pouze na podezřelé body. *Andrewsův indexový graf* a *graf normovaných reziduí* ukazují na podezřelé body č. 14, 17 a 21. *Indexový graf prvků H projekční matice* pak na podezřelé extrémní č. 21.

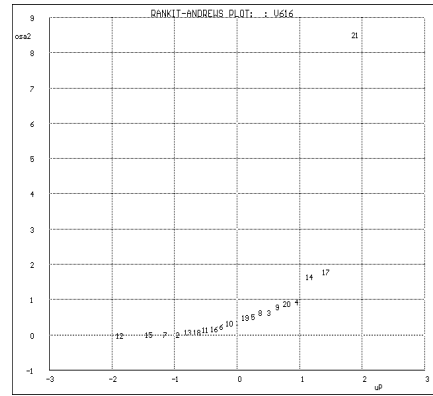
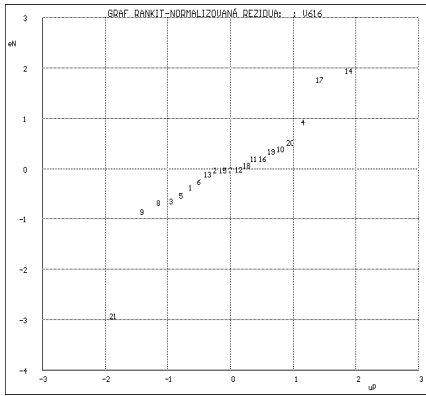


Obr. 6.2-3 Indexové grafy, vlevo: Andrewsův graf, a vpravo: graf normovaných reziduí, *ADSTAT*.



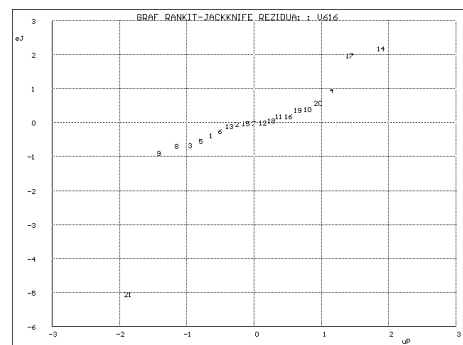
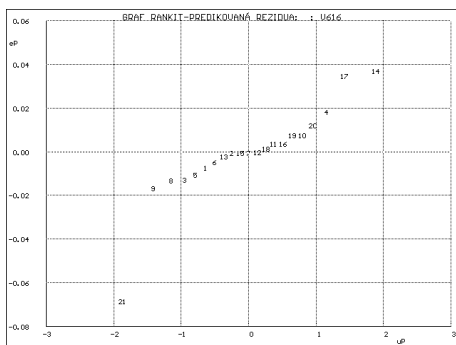
Obr. 6.2-3 Graf prvků H-projekční matice, *ADSTAT*.

(e) **Rankitové grafy** (obr. 6.2-4) ukazují vedle normality rozdělení dotyčných reziduí i na vlivné (zde odlehlé) body.



Obr. 6.2-4 Rankitové grafy, vlevo, graf normovaných reziduí, a vpravo, Andrewsův graf, *ADSTAT*.

Graf normovaných reziduí ukazuje na č. 21 a na č. 17 a 14 jako na odlehlé body. *Andrewsův graf* představuje č. 21 jako odlehlý bod. *Graf předikovaných reziduí* a *graf Jackknife reziduí* č. 21, 14, 17 jako odlehlé body.



Obr. 6.2-4 Rankitové grafy, vlevo, graf předikovaných reziduí, a vpravo, graf Jackknife reziduí, *ADSTAT*.

Model: *Parciální regresní grafy* a *parciální reziduální grafy* jsou určeny pro vícerozměrné lineární regresní modely a nemají proto smysl u jednorozměrného regresního modelu. Vhodnost modelu se posuzuje přímo v grafu obsahujícím data a průběh modelové funkce. Je patrné, že v tomto případě je přímka akceptovatelná a data nevykazují nelineární průběh.

Metoda: do této části patří vyšetření splnění základních předpokladů metody nejmenších čtverců (MNC), za kterých by měla metoda vést k nejlepším lineárním nestranným odhadům regresních parametrů:

Fisherův-Snedecorův test významnosti regrese potvrdil, že navržený model je přijat jako významný.

Scottovo kritérium multikolinearity nemá smysl u jednorozměrného regresního modelu.

Cookův-Weisbergův test heteroskedasticity dokazuje, že rezidua vykazují heteroskedasticitu (nekonstantnost rozptylu).

Jarqueův-Berraův test normality reziduí ukazuje, že klasická rezidua nevykazují Gaussovo rozdělení.

Waldův test autokorelace ukazuje, že klasická rezidua nejsou autokorelována. To by bylo totiž vážným upozorněním ke zhodnocení provedeného experimentu, že došlo k narušení podmínek. Mnohdy však může způsobit heteroskedasticitu i jeden odlehlý bod.

Znaménkový test prokazuje, že znaménko klasických reziduí se dostatečně střídá, a proto rezidua nevykazují žádný trend.

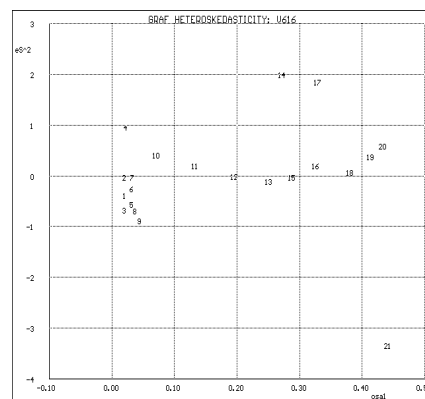
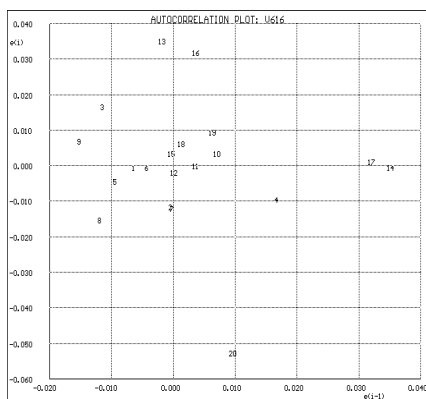
TESTOVÁNÍ REGRESNÍHO TRIPLETU (DATA + MODEL + METODA):

Fisherův-Snedecorův test významnosti regrese, F_{exp}	: 2273.5
Tabulkový kvantil, $F_{1-\alpha}(m-1, n-m)$: 4.3807
Závěr: Navržený model je přijat jako významný.	
Spočtená hladina významnosti	: 0.000
Scottovo kritérium multikolinearity, M	: 3.0004E-15
Závěr: Navržený model je korektní.	

Cookův-Weisbergův test heteroskedasticity, S_f	: 902.50
Tabulkový kvantil, $\chi^2_{1-\alpha}(1)$: 3.8415
Závěr: Rezidua vykazují heteroskedasticitu.	
Spočtená hladina významnosti	: 0.000
Jarqueův-Berraův test normality reziduí, $L(e)$: 8.2754
Tabulkový kvantil, $\chi^2_{1-\alpha}(2)$: 5.9915
Závěr: Normalita není přijata.	
Spočtená hladina významnosti	: 0.016
Waldův test autokorelace, W_a	: 0.5898
Tabulkový kvantil, $\chi^2_{1-\alpha}(2)$: 3.8415
Závěr: Rezidua nejsou autokorelována.	
Spočtená hladina významnosti	: 0.000
Znamékový test, D_t	: -0.8870
Tabulkový kvantil, $N_{1-\alpha/2}$: 1.6449
Závěr: Rezidua nevykazují trend.	
Spočtená hladina významnosti	: 0.188

Graf autokorelace (obr. 6.2-5) vykazuje náhodný mrak bodů reziduí.

Graf heteroskedasticity (obr. 6.2-5) vykazuje trend, klín, což odpovídá heteroskedasticitě, nekonstantnosti rozptylu.



Obr. 6.2-5 Vlevo, graf autokorelace, a vpravo, graf heteroskedasticity, *ADSTAT*.

6. Konstrukce zpřesněného modelu: (a) Po odstranění bodů č. 14, 17, 21 byly nalezeny nové odhady parametrů zpřesněného modelu.

Parametr	Odhad	Směrodatná odchylna	$H_0: b_j = 0$ vs. $H_A: b_j \neq 0$ t -kritérium	Spočtená hlad. význam.
b_0	0.00639	0.002412	2.6490	Zamítnuta
b_1	0.94034	0.009356	100.51	Zamítnuta

Zpřesněný model (v závorce je uveden odhad směrodatné odchylny parametru)

$$y = 0.00639 (0.00241) + 0.9403 (0.0094) x$$

je doložen statistickými charakteristikami: střední kvadratická chyba predikce *MEP* a Akaiikovo informační kritérium *AIC* dosáhly nižších hodnot, čímž dokazují kvalitnější model než předešlý.

Vícenásobný korelační koeficient, r	: 0.99921
Koeficient determinace, 100 % D	: 99.842
Predikovaný koeficient determinace, R^2_p	: 0.9990
Střední kvadratická chyba predikce, <i>MEP</i>	: 6.1534E-05
Akaiikovo informační kritérium, <i>AIC</i>	: -174.03

Rezidua nyní vykazují normální rozdělení a nevykazují trend, stále však vykazují heteroskedasticitu, a proto lze doporučit užití metody vážených nejmenších čtverců.

(b) Užitím statistické váhy ($w_i = 1/y_i^2$) kompenzujeme heteroskedasticitu v datech. Obdržíme nové správnější odhady parametrů.

Parametr	Odhad	Směrodatná odchylna	$H_0: b_j = 0$ vs. $H_A: b_j \neq 0$ t -kritérium	Spočtená hlad. význam.
b_0	0.002129	0.001974	1.0782	Akceptována

b_1	0.94617	0.073078	12.947	Zamítnuta	0.000
-------	---------	----------	--------	-----------	-------

Opravený model má tvar, (v závorce je vždy uveden odhad směrodatné odchylky parametru):

$$y = 0.00213 (0.00197) + 0.9462 (0.0731) x.$$

Jelikož došlo ke snížení rozhodujících kritérií, *střední kvadratické chyby predikce MEP* a *Akaikova informačního kritéria AIC*, lze považovat tyto odhady za lepší než předešlé. Pearsonův korelační koeficient r , a tím pádem i koeficient determinace D vychází nepatrně horší než u předešlého odhadu bez statistické váhy.

Vícenásobný korelační koeficient, r	: 0.95544
Koeficient determinace, 100 % D	: 91.287
Predikovaný koeficient determinace, R^2_p	: 0.93371
Střední kvadratická chyba predikce, MEP	: 5.1179E-05
Akaikovo informační kritérium, AIC	: -181.63

7. Zhodnocení kvality modelu: nalezený model má tvar (v závorce je vždy uveden odhad směrodatné odchylky parametru)

$$y = 0.00213 (0.00197) + 0.9462 (0.0731) x$$

a intervalový odhad parametrů úseku β_0 a směrnice β_1 bude

$$b_0 - t_{1-\alpha/2}(18) \sqrt{D(b_0)} \leq \beta_0 \leq b_0 + t_{1-\alpha/2}(18) \sqrt{D(b_0)}$$

a po dosazení

$$0.00213 - 2.12 \times 0.00197 \leq \beta_0 \leq 0.00213 + 2.12 \times 0.00197$$

vyjde

$$-0.00205 \leq \beta_0 \leq 0.00630.$$

Tento interval spolehlivosti úseku regresní přímky zahrnuje nulu, takže lze úsek β_0 považovat za nulový.

Analogicky dosazením do intervalu spolehlivosti směrnice obdržíme nerovnost

$$0.9462 - 2.12 \times 0.0731 \leq \beta_1 \leq 0.9462 + 2.12 \times 0.0731$$

a po vyčíslení

$$0.7912 \leq \beta_1 \leq 1.1012.$$

Jelikož tento interval obsahuje jedničku, lze považovat směrnici β_1 za jednotkovou.

Lze uzavřít, že úsek regresní přímky lze považovat za nulový $\beta_2 = 0$ a směrnice β_1 není významně odlišná od jedničky. Výsledky nové metody se proto statisticky významně neliší od metody standardní.