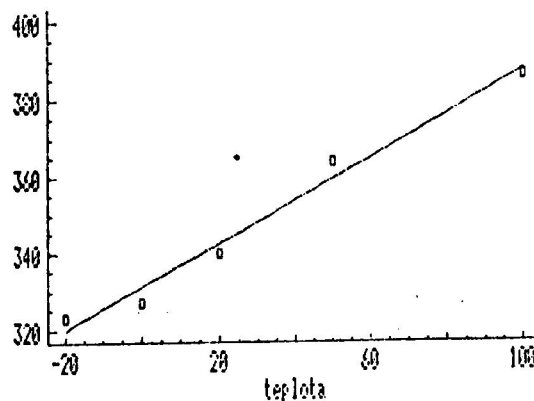


TAB. 12.2. Měření rychlosti zvuku v závislosti na teplotě

teplota [°C]	-20	0	20	50	100
– rychlost [m/s]	323	327	340	364	386



OBR. 12.2. Měření rychlosti zvuku v závislosti na teplotě

Příklad 12.2. Byla provedena měření rychlosti zvuku ve vzduchu v závislosti na teplotě. Výsledky jsou v tab. 12.2 a na obr. 12.2.

Data ukazují, že závislost rychlosti zvuku na teplotě lze v uvažovaném rozsahu pokládat prakticky za lineární. Výpočtem dostaneme

$$\bar{x} = 30, \quad \bar{Y} = 348, \quad b_1 = 0,561, \quad b_0 = 331,159, \quad s^2 = 18,955, \quad T_1 = 12,096.$$

Jelikož $|T_1| \geq t_3(0,05) = 3,182$, zamítneme hypotézu, že $\beta_1 = 0$. Tím je prokázáno, že rychlost zvuku závisí na teplotě.

Lineární charakter závislosti však opravdu platí jen ve sledovaném intervalu teplot. Ve fyzikálních tabulkách zjistíme, že třeba při teplotě 500 °C je rychlost zvuku 557 m/s, tímco náš model by dal hodnotu 612 m/s. $\diamond$

data file: csu2002.sta [14 cases with 3 variables]

0 CASE NAME	1 VEK	2 PLAT	3 DETI
Praha	41.500	19879.00	8.400
Stredoces	39.500	15534.00	9.300
Jihocesky	39.000	14029.00	9.200
Plzensky	39.700	14840.00	8.900
Karlovarsk	38.300	13535.00	9.700
Ustecky	38.300	14269.00	10.000
Liberecky	38.600	14081.00	9.700
Kralovehra	39.600	13753.00	9.000
Pardubicky	39.000	13539.00	9.200
Vysocina	38.700	13441.00	9.200
Jihomoravs	39.500	14059.00	9.000
Olomoucky	39.000	13373.00	8.900
Zlinsky	39.000	13843.00	8.800
Moravskosl	38.500	14923.00	9.100

Klasičtím způsobem najítom širší, která by vyplechala závislost platů na věku, tj: $Y_i = f(x_i) + e_i$, kde Y_i je průměrný plat i -tému kraji, x_i je průměrný věk i -tému kraji.

Pozor! Měda' se používá metoda lineární regrese, protože hodnoty x_i nejsou navzájem různé a matice X tedy nemá plnou hodnotnost.

Program STATISTICA používá přibližně klasické metody, které se hodí i pro nelineární regresi.

Klasičtím: 1) položit přímku $Y_i = b_0 + b_1 x_i$

2) polynom 2. stupně $Y_i = b_0 + b_1 x_i + b_2 x_i^2$

3) polynom 3. stupně $Y_i = b_0 + b_1 x_i + b_2 x_i^2 + b_3 x_i^3$

podle hodnot rezidualů fce $S_e = \sum_i (Y_i - \hat{Y}_i)^2$ a koeficientů

determinace R je nejhorší přímka a nejlepší polynom 2. stupně, přičemž polynom 3. stupně je jen nepatrně horší.

Proložení přímky :

STAT. NONLIN. ESTIMAT.	Model: plat=b0+b1*vek (csu2002.sta) Dep. var: PLAT Loss: (OBS-PRED)**2 Final loss: 11674538.325 R=.82298 Variance explained: 67.730%	
N=14	B0	B1
Estimate	-51078.8	1674.940

STAT. NONLIN. ESTIMAT.	Observed Va	Predicted V	Residual Va
Case	PLAT Observed.	PLAT Predictd	PLAT Residual
Praha	19879.00	18431.14	1447.86
Stredoces	15534.00	15081.26	452.74
Jihocesky	14029.00	14243.79	-214.80
Plzensky	14840.00	15416.25	-576.25
Karlovarsk	13535.00	13071.34	463.66
Ustecky	14269.00	13071.34	1197.66
Liberecky	14081.00	13573.82	507.18
Kralovehra	13753.00	15248.76	-1495.76
Pardubicky	13539.00	14243.79	-704.80
Vysocina	13441.00	13741.31	-300.31
Jihomoravs	14059.00	15081.26	-1022.26
Olomoucky	13373.00	14243.79	-870.80
Zlinsky	13843.00	14243.79	-400.80
Moravskosl	14923.00	13406.33	1516.67

Polynom 2. stupně:

STAT. NONLIN. ESTIMAT.	Observed Va	Predicted V	Residual Va
Case	PLAT Observd.	PLAT Predictd	PLAT Residual
Praha	19879.00	19910.91	-31.914
Stredoces	15534.00	14306.13	1227.867
Jihocesky	14029.00	13922.43	106.569
Plzensky	14840.00	14573.57	266.427
Karlovarsk	13535.00	14069.00	-534.005
Ustecky	14269.00	14069.00	199.995
Liberecky	14081.00	13908.51	172.492
Kralovehra	13753.00	14431.71	-678.713
Pardubicky	13539.00	13922.43	-383.431
Vysocina	13441.00	13887.57	-446.569
Jihomoravs	14059.00	14306.13	-247.133
Olomoucky	13373.00	13922.43	-549.431
Zlinsky	13843.00	13922.43	-79.431
Moravskosl	14923.00	13945.73	977.273

STAT. NONLIN. ESTIMAT.	Model: plat=b0+b1*vek+b2*vek**2 (csu2002.sta) Dep. var: PLAT Loss: (OBS-PRED)**2 Final loss: 4077344.7850 R=.94196 Variance explained: 88.730%		
N=14	B0	B1	B2
Estimate	1237953.	-63131.2	813.9950

Polynom 3. stupně :

STAT. NONLIN. ESTIMAT.	Observed V	Predicted V	Residual V
Case	PLAT Observed.	PLAT Predictd	PLAT Residual
Praha	19879.00	19912.57	-33.575
Stredoces	15534.00	14302.24	1231.760
Jihocesky	14029.00	13932.40	96.597
Plzensky	14840.00	14560.89	279.107
Karlovarsk	13535.00	14055.83	-520.831
Ustecky	14269.00	14055.83	213.169
Liberecky	14081.00	13913.46	167.545
Kralovehra	13753.00	14423.54	-670.540
Pardubicky	13539.00	13932.40	-393.403
Vysocina	13441.00	13895.59	-454.587
Jihomoravs	14059.00	14302.24	-243.240
Olomoucky	13373.00	13932.40	-559.403
Zlinsky	13843.00	13932.40	-89.403
Moravskosl	14923.00	13946.20	976.803

STAT. NONLIN. ESTIMAT.	Model: plat=b0+b1*vek+b2*vek**2+b3*vek**3 (csu2002.sta) Dep. var: PLAT Loss: (OBS-PRED)**2 Final loss: 4095641.7442 R=.94170 Variance explained: 88.679%			
N=14	B0	B1	B2	B3
Estimate	1954.956	30142.35	-1531.14	19.64448

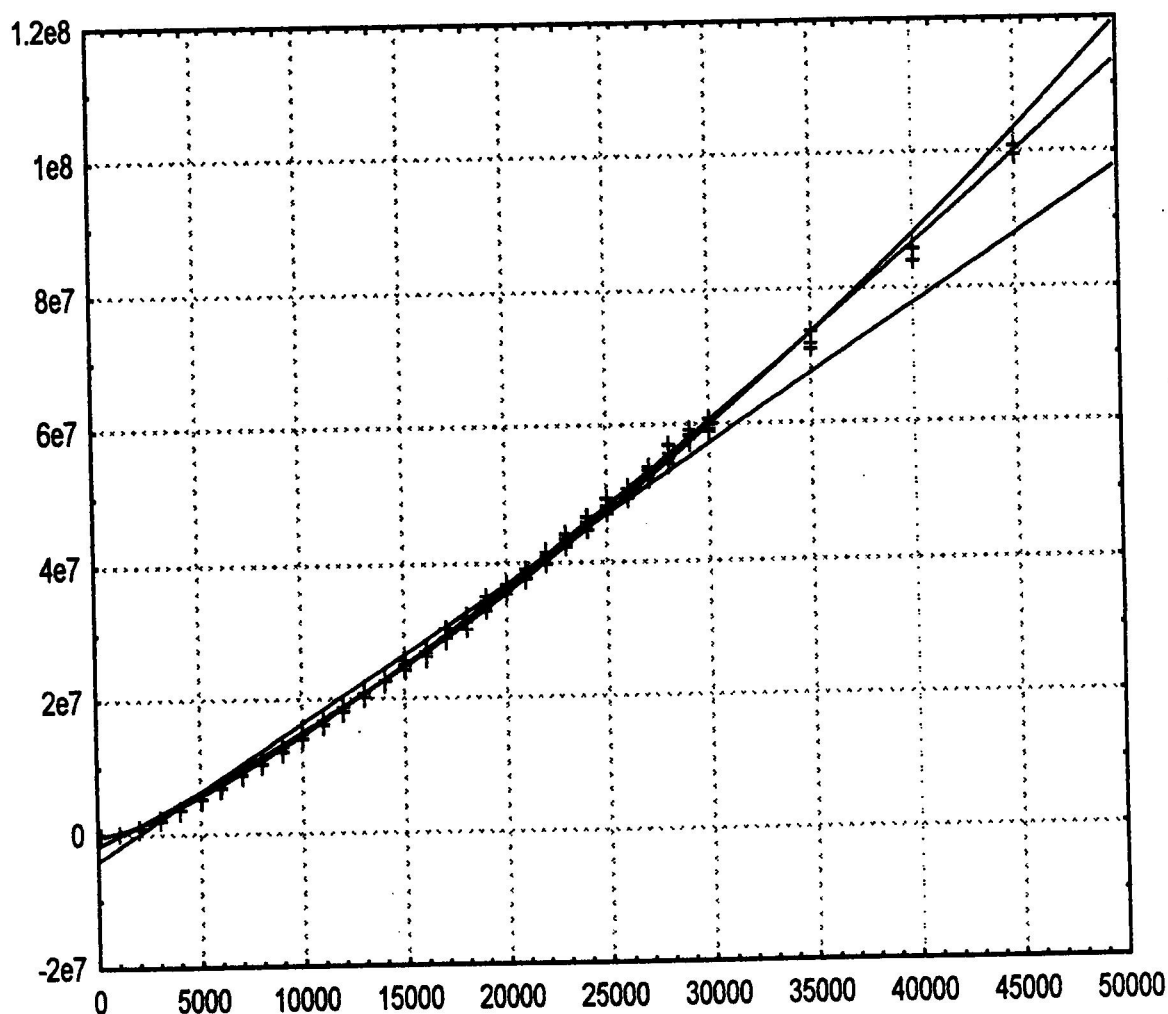
Příklad (M. Černý, diplomová práce):

Tento příklad nespočívá s DM, je to experimentální analýza složitosti algoritmu na výpočet tranzitivního uzávěru grafu. Používám ho jako ukázkou analýzy reziduí.

Následující obrázek ukazuje srovnání tří regresních modelů

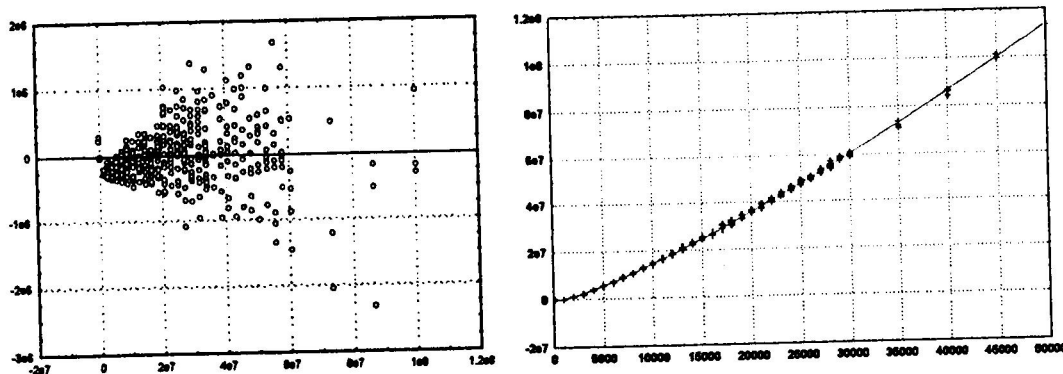
$$\begin{aligned}y &= b_1 + b_2x + b_3x^2 \quad (\text{regresní parabola}), \\y &= b_1 + b_2x + b_3x \ln x \quad (\text{model "x ln x"}), \\y &= b_1 + b_2x \quad (\text{regresní přímka})\end{aligned}$$

prokládající redundanci pro $p=0.01$. Z obrázku nelze první dva modely příliš rozlišit (kvadratický roste v obrázku nepatrně rychleji, viz graf 9.24). Tabulka 9.3. ukazuje, že v obou případech je koeficient determinace tak blízký 100%, že oba modely se jeví jako vhodné pro naše data. Lineární model se jeví jako nevhodný (nejspodnější přímka v grafu 9.24); ovšem to, že by redundance mohla růst pomaleji než kvadraticky, vyloučit nemůžeme.

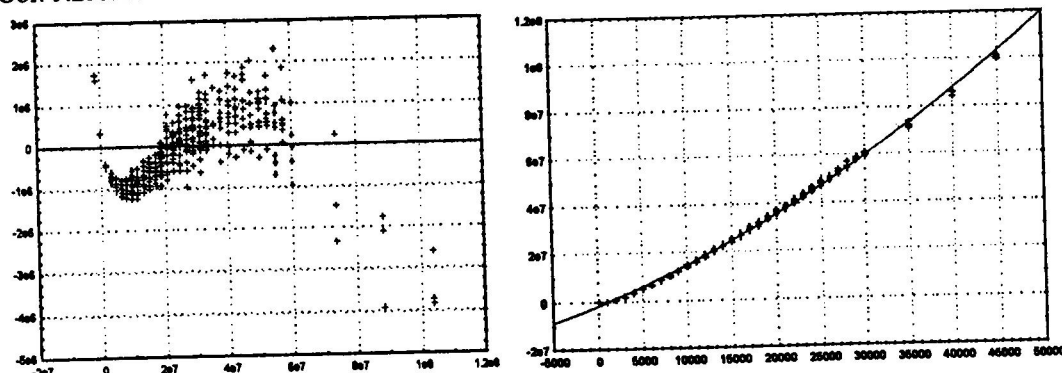


Obr. 9.24. Model kvadratický (horní, modrá křivka), $x \ln x$ (střední, zelená křivka) a lineární (spodní červená přímka) pro redundanci Simonova algoritmu $p = 0.01$

Ve prospěch modelu s členem $x \ln x$ ovšem hovoří grafy residuí. Na následujícím obrázku ukazuje první dvojice grafů residua a model $x \ln x$, zatímco druhý graf zobrazuje residua při proložení parabolou; je vidět, že parabola malé hodnoty podhodnocuje a velké hodnoty nadhodnocuje, zatímco residua modelu $x \ln x$ mají sice nekonzistentní rozptyl (což lze očekávat), ale jejich střední hodnota je blízko nuly a nejsou korelovaná, což potvrzuje i Durbin-Watson¹² (pro model $x \ln x$ je 1.79, zatímco pro parabolu 0.31) a pořadová korelace¹³ (pro model $x \ln x$ je 0.11, pro parabolu 0.84).



Obr. 9.25a. Residua a model $x \ln x$



Obr.9.25b. Residua a model $y = b_1 + b_2x + b_3x^2$

V případě, kdybychom chtěli srovnat ještě model

$$y = b_1 + b_2x + b_3x \ln x + b_4x \ln \ln x,$$

bude již matice Z (viz kap. 8.2) natolik špatně podmíněná, že výsledky mohou být zkreslené numerickými chybami; i přesto uveďme, že vychází p-value t-testu koeficientu b_4 v řádu 0.2. Nemůžeme dost dobře z našich naměřených dat mezi modelem " $x \ln x$ " a " $x \ln \ln x$ " rozlišit.

Následující tabulka ukazuje celkový přehled regresních modelů pro redundanci Simonova algoritmu. Sloupec "p-val b_i " obsahuje p-value t-testů na regresní parametry b_1 , b_2 a b_3 , je-li

¹² viz kap. 8.4.

¹³ pořadová korelace je neparametrický test na korelaci v residuích, viz [4]. Hodnoty blízke 1 ukazují na silnou korelaci, hodnoty blízke nule na nekorelovanost

b_3 přítomen, "D. W." je hodnota Durbin-Watsonovy statistiky, "corr" je pořadová korelace v residuích, " R^2 " je koeficient determinace.

		b_1	b_2	b_3	p-val b_i	D.W.	corr	R^2
$p=0.01$	parabola	-1771191	1491.83	0.018814	0, 0, 0	0.31	0.84	99.87%
	$x \ln x$	-222457.59	-3182.46	504.7382	0, 0, 0	1.76	0.11	99.97%
	přímka	-4068100	2045.54		0, 0	0.07	0.98	99.28%
$p=0.05$	parabola	-174 449.403	279.506	0.0019443	0, 0, 0	1.22	0.38	99.91%
	$x \ln x$	-1 359.271	-230.829	54.87015	0, 0, 0	1.85	0.072	99.94%
	přímka	-442 822.305	993.9976		0, 0	0.29	0.86	99.62%
$p=0.1$	parabola	-66 038.826	116.630	0.0007104	0, 0, 0	1.61	0.179	99.88%
	$x \ln x$	1857.658	-75.44398	20.58461965	0, 0, 0	1.96	-0.03	99.9%
	přímka	-175432.489	139.958 661		0, 0	0.5	0.78	99.65%
$p=0.15$	parabola	-25505.203	63.076	0.0004740462	0, 0, 0	1.75	0.1	99.86%
	$x \ln x$	-2166.58	-26.8073	9.9796962	0.49, 0, 0	2.12	0.07	99.88%
	přímka	-64851	75.3871433		0, 0	0.83	0.57	99.71%
$p=0.25$	parabola	-13510.585	29.11927793	0.000139797	0, 0, 0	1.83	0.07	99.79%
	$x \ln x$	-1478.339	-6.1136943	3.97747859	0.42, 0.001, 0	2.06	-0.03	99.81%
	přímka	-30777.9381	33.2341103		0, 0	1.16	0.41	99.66%
$p=0.45$	parabola	-3595.23	8.646089	0.0000439721	0, 0, 0	1.66	0.15	99.55%
	$x \ln x$	-425.72	-1.398681	1.09221955	0.59, 0.1, 0	1.76	0.1	99.57%
	přímka	-8215.02	9.8526532		0, 0	1.76	0.1	99.45%
$p=0.65$	parabola	-1098.98	2.8078	0.0000149727	0, 0, 0	2.05	0.02	99.40%
	$x \ln x$	-78.83	-0.5270	0.36362218175	0.79, 0.11, 0	2.12	0.06	99.42%
	přímka	-2672.04	3.218710		0, 0	1.74	0.12	99.29%
$p=0.99$	parabola	-2.909	0.012202	0.00000018515	0.4, 0, 0	2.09	-0.04	95.34%
	$x \ln x$	4.888	-0.002109	0.003716638	0.3, 0, 0	2.09	-0.04	95.33%
	přímka	-20.37	0.01704		0, 0	1.93	0.03	94.91%

Tab. 9.3. Srovnání regresních modelů pro redundanci Simonova algoritmu

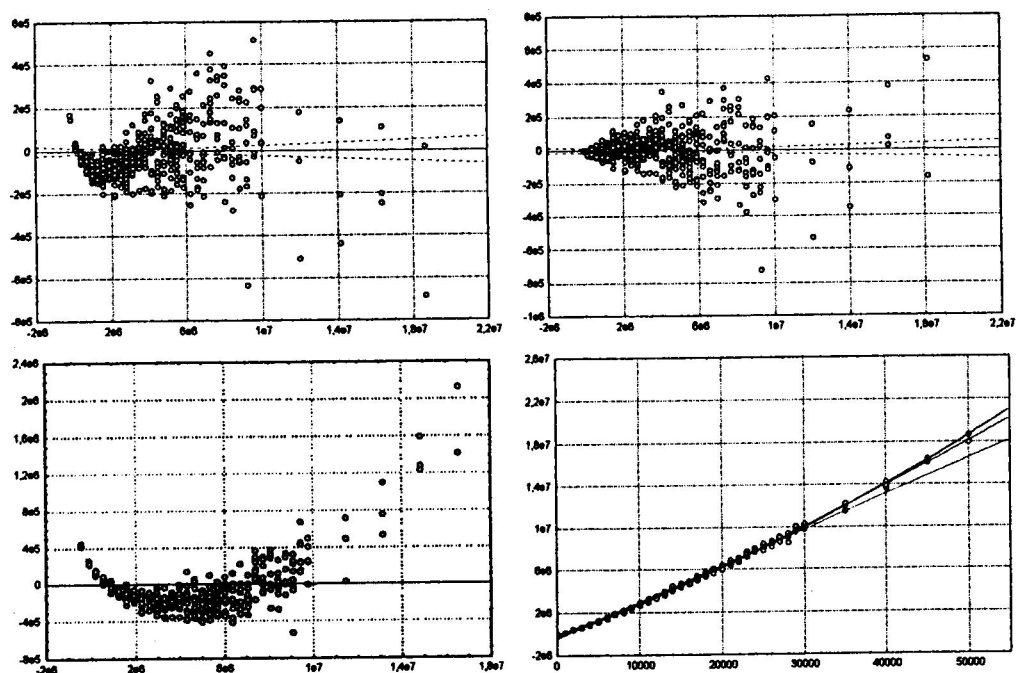
Další obrázky ilustrují údaje v tabulce. Pro každé p (s výjimkou 0.01, neboť tento případ byl diskutován v předchozím textu) uvádíme graf residuí pro všechny tři modely (parabola, $x \ln x$ a přímka) spolu obrázkem srovnání těchto modelů. Následuje vždy po řadě parabola, vedle ní model $x \ln x$, pod nimi přímka a srovnání všech tří modelů. V těch případech, kdy se jeví model jako smysluplný, uvádíme v obrázcích residuí i 95%-ní intervaly spolehlivosti pro střední hodnotu. Ve všech obrázcích srovnání modelů platí, že modrá křivka, jež ve v pravé části grafu nejvýše, je parabola, zelená střední křivka je $x \ln x$ a nejnižší položena je červená přímka.

pro D.W. ≈ 2 je autokorelační koeficient nejnižší

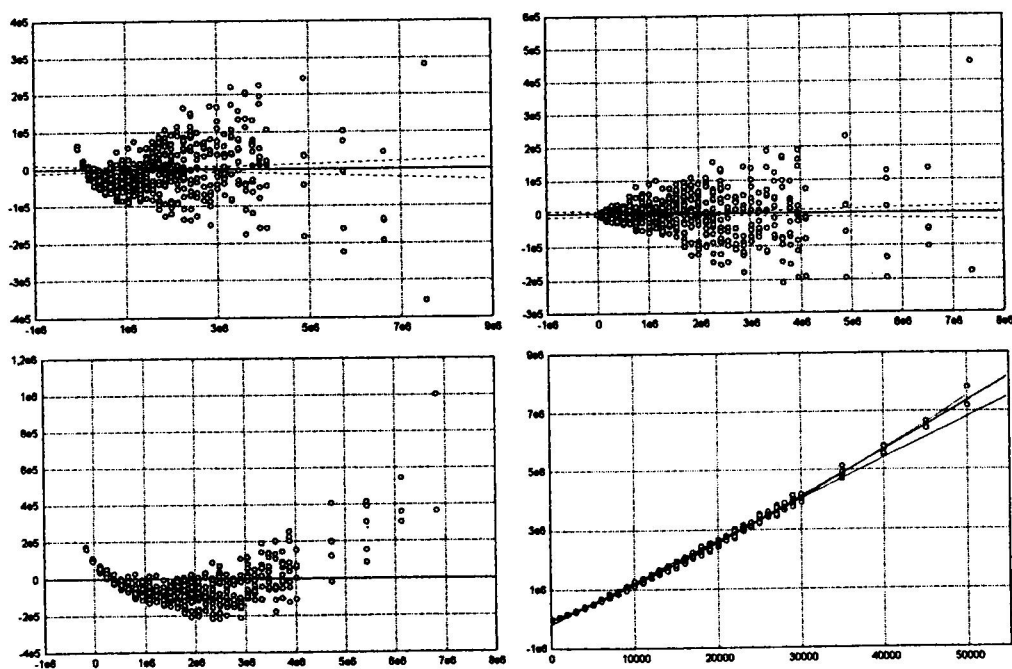
pro D.W. ≈ 0 nebo ≈ 4 se hypotéza o nezávislosti residuí namíká

(0 znamena pozitivní autokorelaci, 4 negativní)

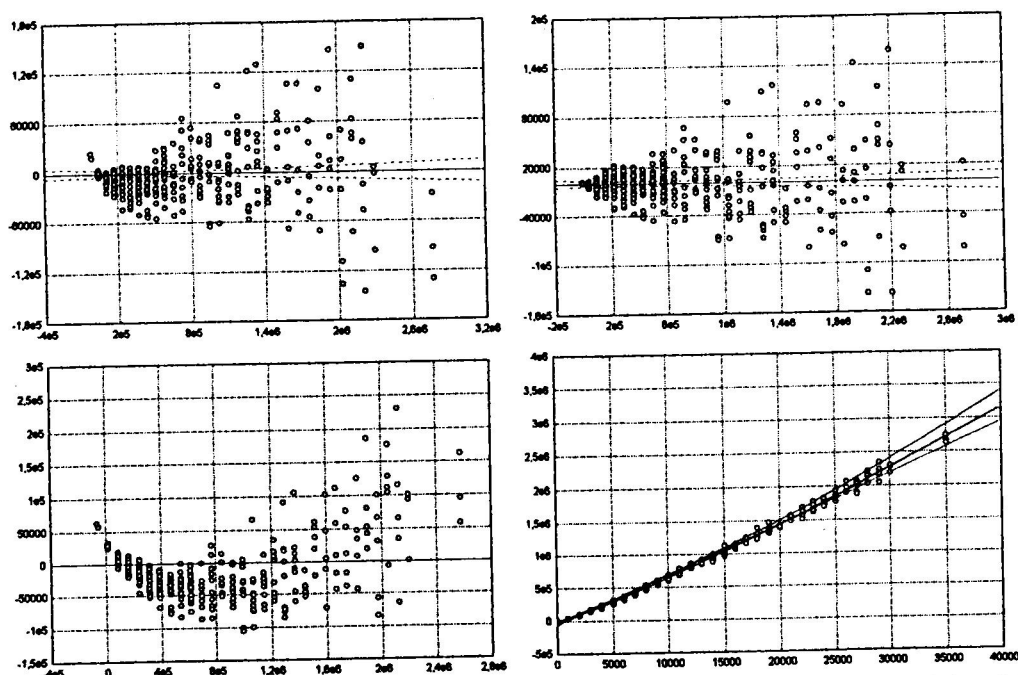
existuje pro ka tabulované kritické hodnoty



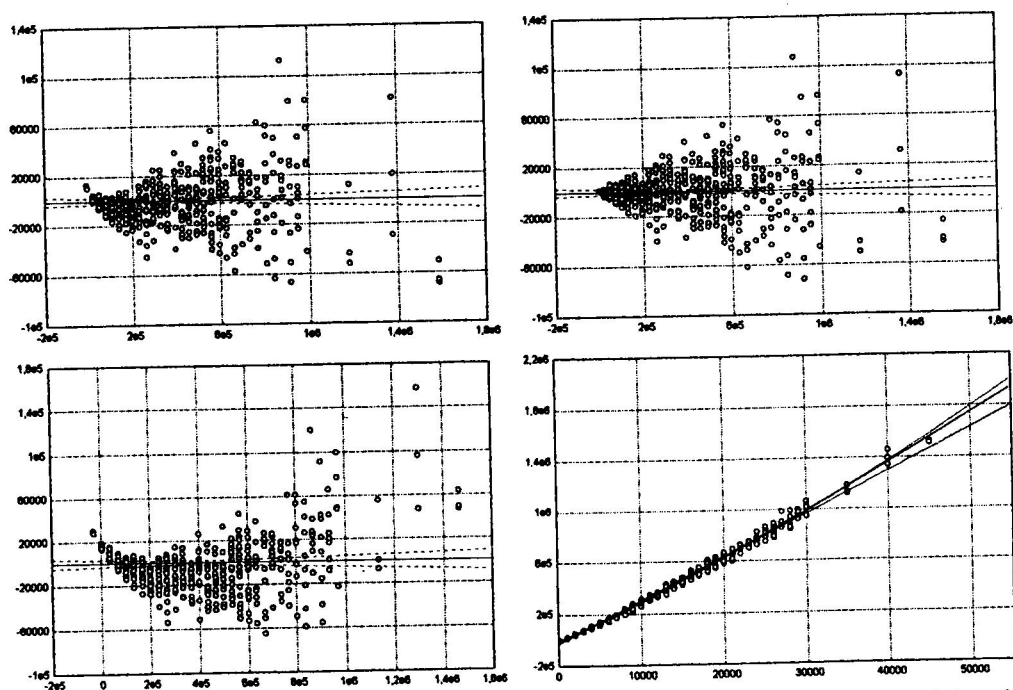
Obr. 9.26. Simonův algo, $p = 0.05$, residua pro parabolu (a), $x \ln x$ (b), přímku (c), srovnání modelů (d)



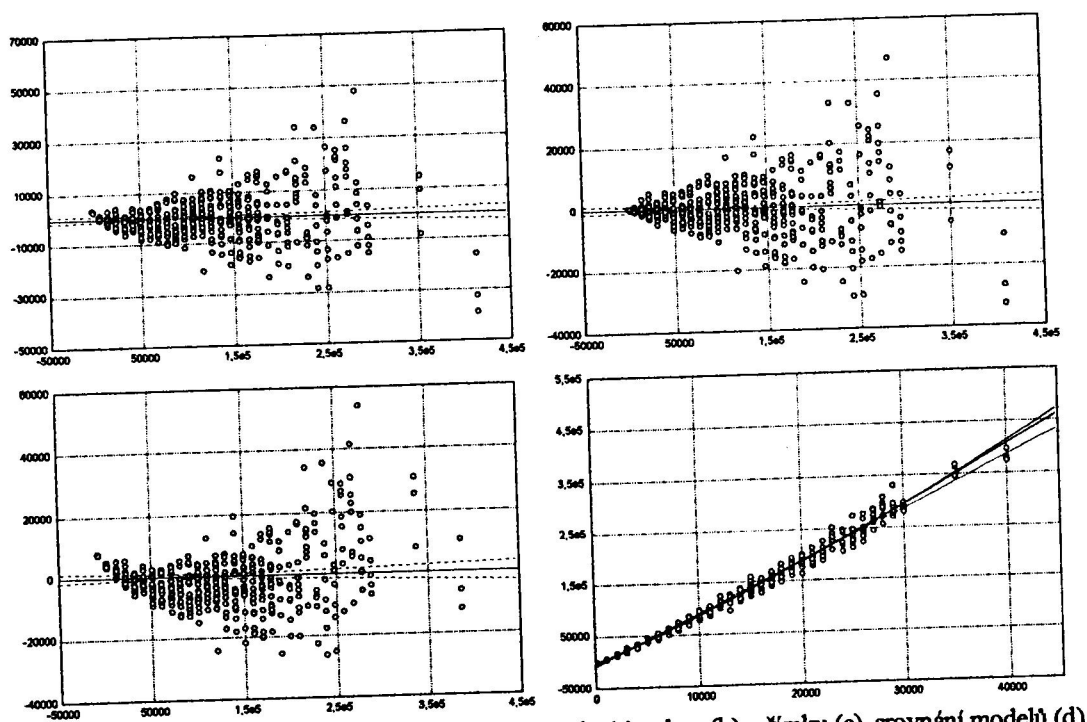
Obr. 9.27. Simonův algo, $p = 0.1$, residua pro parabolu (a), $x \ln x$ (b), přímku (c), srovnání modelů (d)



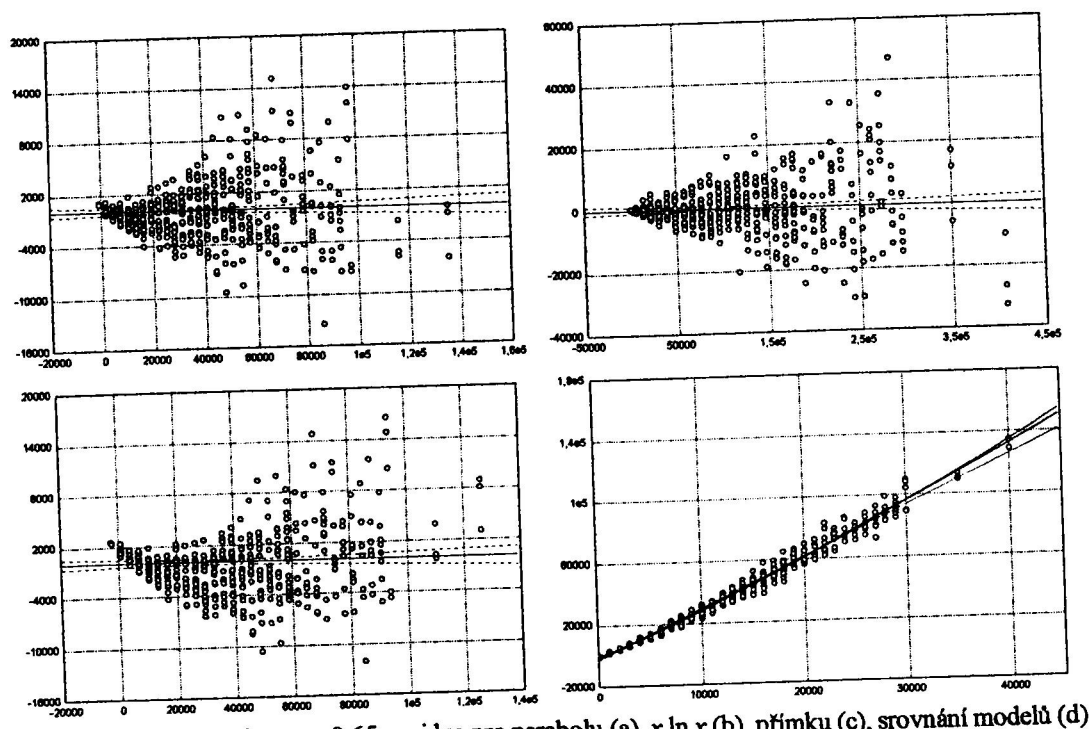
Obr. 9.28. Simonův algo, $p = 0.15$, residua pro parabolu (a), $x \ln x$ (b), přímku (c), srovnání modelů (d)



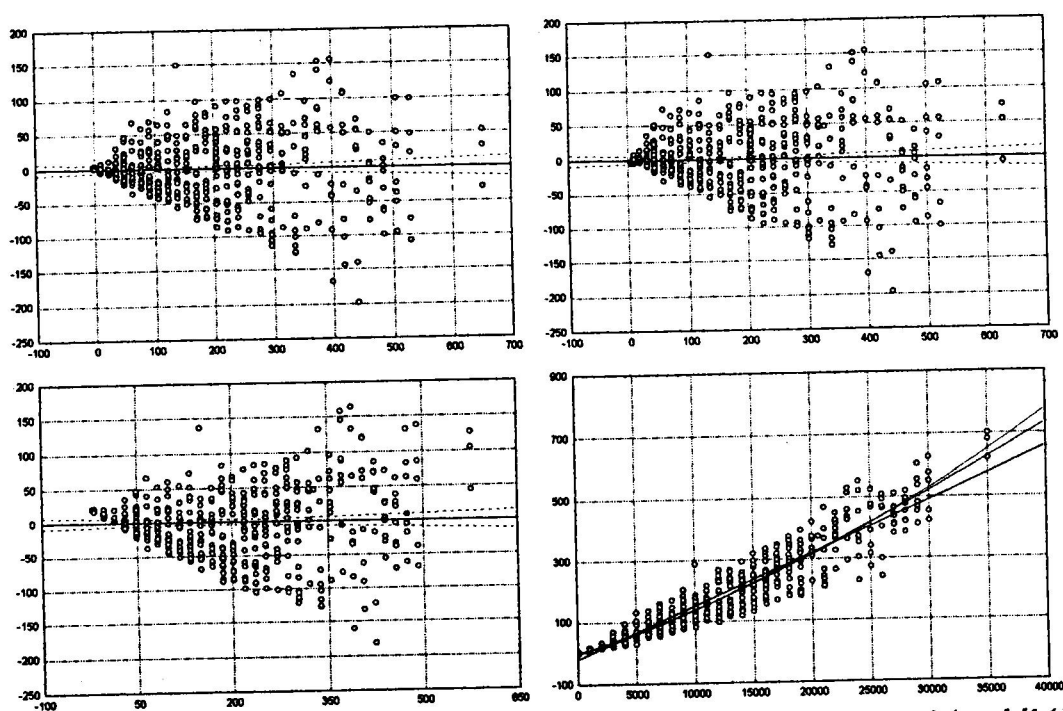
Obr. 9.29. Simonův algo, $p = 0.25$, residua pro parabolu (a), $x \ln x$ (b), přímku (c), srovnání modelů (d)



Obr. 9.30. Simonův algo, $p = 0.45$, residua pro parabolu (a), $x \ln x$ (b), přímku (c), srovnání modelů (d)



Obr. 9.31. Simonův algo, $p = 0.65$, residua pro parabolu (a), $x \ln x$ (b), přímku (c), srovnání modelů (d)



Obr. 9.32. Simonův algo, $p = 0.99$, residua pro parabolu (a), $x \ln x$ (b), přímku (c), srovnání modelů (d)

Z předchozí tabulky a obrázků se ukazuje, že ve všech případech se model $x \ln x$ jeví jako vhodný, pro menší hodnoty p je dokonce výrazně vhodnější než trend kvadratický; lineární trend se ve všech případech jeví jako nejméně vhodný. Usuzujeme z toho, že – alespoň podle provedených měření – redundance není lineární, avšak zdá se být významně subkvadratická (o tom viz i kap. 9.10).

Ve svém článku [35] Simon uvádí výsledek citovaný v kapitole 5, že pro model $G_{n,p}$ má jeho algoritmus složitost v průměrném případě $O(n^2)$, je-li $p \geq (\ln n)/n$ a $O(n^2 \ln \ln n)$ v ostatních případech. Simonův algoritmus totiž běží v čase $O(e^* + ke^R)$; je zřejmé, že člen e^* zlepšit nelze, neboť vypočtený transitivní uzávěr musí být alespoň zapsán na výstup, avšak z našich simulací vychází, že člen ke^R , který Simon odhaduje jako n^2 , resp. $n^2 \ln \ln n$, by bylo možné zlepšit dokonce až na $n \ln n$; naše simulace tedy napovídají, že pro pevné p pracuje Simonův algoritmus nad grafem $G_{n,p}$ v průměrném čase $O(e^* + n \ln n) = O(e^*)$.