

Úloha M608 **Závislost obsahu lipoproteinů v krevním séru na třech faktorech**
(*Lineární regresní modely*)

Zadání : Při kvantitativní analýze lidského krevního séra ovlivňují hodnotu obsahu vysokohusotního lipoproteinu (y) tři proměnné, a to obsah celkového cholesterolu (x_1), obsah celkového triglyceridu (x_2) a konečně tzv. pre-beta komponenty (x_3), které jsou buď přítomné ($x_3 = 1$) nebo nepřítomné ($x_3 = 0$), str. 141 v cit⁶⁵. (1) Navrhněte regresní model a rozhodněte, zda (2) x_1 , x_2 , x_3 *samostatně* ovlivňují v predikci model jako celek, (3) x_1 , x_2 , x_3 společně ovlivňují v predikci model jako celek, (4) testujte i členy x_1x_2 a x_2x_3 . Testy proveďte na hladině významnosti $\alpha = 0,05$. (5) Jsou v datech vlivné body? Je třeba odstranit vybočující hodnoty?

Data :

Obsah cholesterolu x_1	Obsah triglyceridu x_2	Přítomnost pre-beta komponenty x_3	Obsah lipoproteinů y	Obsah cholesterolu x_1	Obsah triglyceridu x_2	Přítomnost pre-beta komponenty x_3	Obsah lipoproteinů y
287	111	0	47	192	115	1	57
236	135	0	38	349	408	1	42
255	98	0	47	263	103	1	54
135	63	0	39	223	102	1	60
121	46	0	44	316	274	0	33
171	103	0	64	288	130	0	55
260	227	0	58	256	149	0	36
237	157	0	49	318	180	0	36
261	266	0	55	270	134	0	42
397	167	0	52	262	154	0	41
295	164	0	49	264	86	0	42
261	119	1	47	325	148	0	39
258	145	1	40	388	191	0	27
280	247	1	42	260	123	0	31
339	168	1	63	284	135	0	39
161	68	1	40	326	236	1	56
324	92	1	59	248	92	1	40
171	56	1	56	285	153	1	58
265	240	1	76	361	126	1	43
280	306	1	67	248	226	1	40
248	93	1	57	280	176	1	46

Řešení :

- Ø Program ADSTAT : Lineární regrese – Lineární regrese MNČ
– Regresní diagnostika

Při řešení úlohy bylo využito lineárního regresního modelu ve tvaru :

$$E(y/x) = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \beta_3 \cdot x_3$$

V S T U P

1. ZVOLENÁ STRATEGIE REGRESNÍ ANALÝZY :

Omezení, P	:	1.0000 · 10 ⁻³⁴
Transformace	:	Ne
Váhy	:	Ne
Absolutní člen zahrnut	:	Ano

2. PODMÍNKY A KVANTILY PRO STATISTICKÉ TESTY :

Hladina významnosti, α	:	0.050
Počet bodů, n	:	42
Počet parametrů, m	:	4
Kvantil Studentova rozdělení t (1 - α / 2, n - m)	:	2.024
Kvantil rozdělení χ^2 (1 - α , m)	:	9.488

V Ý S T U P

1. PŘEDBĚŽNÁ STATISTICKÁ ANALÝZA:

Proměnná	Průměr	Směrodatná odchylka	Párový korelační koeficient
y	47.7620	10.6080	1.0000
x ₁	267.810	60.4030	-0.1001
x ₂	155.050	73.7500	0.0680
x ₃	0.47619	0.50549	0.3992

Párové korelační koeficienty mezi dvojicemi vysvětlujících proměnných

x ₁ versus x ₂	:	0.51276
x ₁ versus x ₃	:	4.6407 · 10 ⁻⁰³
x ₂ versus x ₃	:	0.11125

2. INDIKACE MULTIKOLINEARITY:

Č _j	Vlastní čísla korelační matice I _j	Čísla podmíněnosti K _j	Variance inflation factor VIF _j
1	0.47626	3.2034	1.3618
2	0.99808	1.5286	1.3789
3	1.52570	1.0000	1.0164

Maximální číslo podmíněnosti K : 3.20340

Protože $K > 1000$ indikuje silnou multikolaritu a $VIF_j > 10$ indikuje silnou multikolaritu nebyla v našem případě multikolarita indikována ani jedním z kritérií, neboť jednotlivá $K_j < 1000$ a také jednotlivá $VIF_j < 10$.

3. ODHADY PARAMETRŮ A TESTY VÝZNAMNOSTI:

Parametr	Odhad	Směrodatná odchylka	Test $H_0: \beta_j = 0$ vs. $H_A: \beta_j \neq 0$		
			t – kritérium	hypotéza H_0 je	Hladina význ.
β_0	48.86500	7.249900	6.74010	Zamítnuta	0.000
β_1	-0.027312	0.030148	-0.90594	Akceptována	0.371
β_2	0.015039	0.024846	0.60529	Akceptována	0.549
β_3	8.148300	3.112200	2.61810	Zamítnuta	0.013

Vliv vysvětlujících proměnných x_1 a x_2 je zamítnut, význam má pouze x_3 .

4. STATISTICKÉ CHARAKTERISTIKY REGRESE:

Vícenásobný korelační koeficient, R	:	0.42152
Koeficient determinace, R^2	:	0.17768
Predikovaný korelační koeficient, R_p^2	:	0.00000
Střední kvadratická chyba predikce, MEP	:	115.02000
Akaikeho informační kritérium, AIC	:	197.15000

5. ANALÝZA KLASICKÝCH REZIDUÍ:

Reziduální součet čtverců, RSC	:	3 793.90000
Průměr absolutních hodnot reziduí, Me	:	8.20180
Průměr relativních reziduí, Mer	:	17.88800
Odhad reziduálního rozptylu, $s^2(e)$:	99.83900
Odhad směrodatné odchylky reziduí, $s(e)$:	9.99190
Odhad šikmosti reziduí, $g_1(e)$:	0.29118
Odhad špičatosti reziduí, $g_2(e)$:	2.20520

6. TESTOVÁNÍ REGRESNÍHO TRIPLETU (DATA + MODEL + METODA) :

Fisher-Snedocorův test významnosti regrese, F	:	2.736900
Tabulkový kvantil, F (1 - α , m - 1, n - m)	:	2.851700
Závěr : Navržený model není přijat jako významný.		
Spočtená hladina významnosti	:	0.057000
Scottovo kritérium multikolinearity, M	:	0.010393
Závěr : Navržený model je korektní.		
Cook-Weisbergův test heteroskedasticity, Sf	:	111.150000
Tabulkový kvantil, χ^2 (1 - α , 1)	:	3.841500
Závěr : Rezidua vykazují heteroskedasticitu.		
Spočtená hladina významnosti	:	0.000000
Jarque-Berraův test normality reziduí, L (e)	:	1.699000
Tabulkový kvantil, χ^2 (1 - α , 2)	:	5.991500
Závěr : Normalita je přijata.		
Spočtená hladina významnosti	:	0.428000
Waldův test autokorelace, Wa	:	0.938630
Tabulkový kvantil, χ^2 (1 - α , 1)	:	3.841500
Závěr : Rezidua nejsou autokorelována.		
Spočtená hladina významnosti	:	0.333000
Znaménkový test, Dt	:	-0.331230
Tabulkový kvantil, N (1 - α / 2)	:	1.644900
Závěr : Rezidua nevykazují trend.		
Spočtená hladina významnosti	:	0.370000

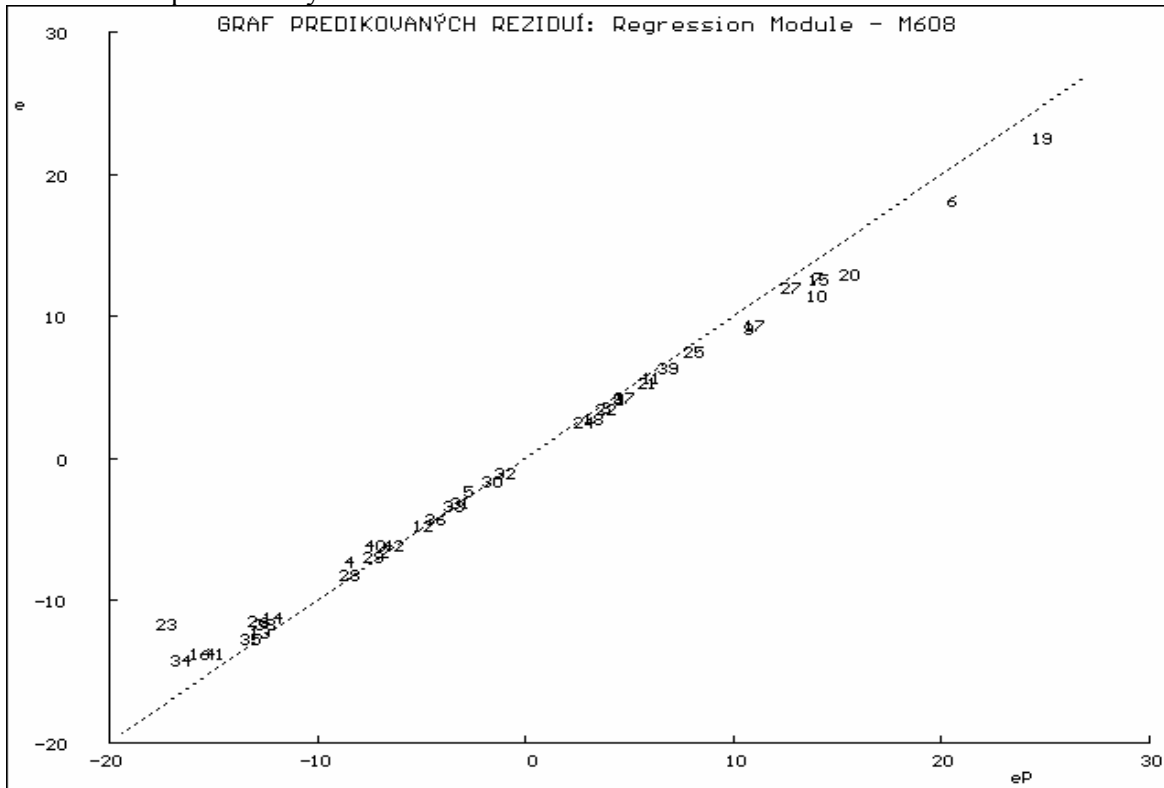
7. INDIKACE VLIVNÝCH BODŮ :

Podezřelé body stanovené podle výpočtů	:	č. 6, 19, 20, 23, 34
--	---	----------------------

8. GRAFICKÁ INDIKACE :

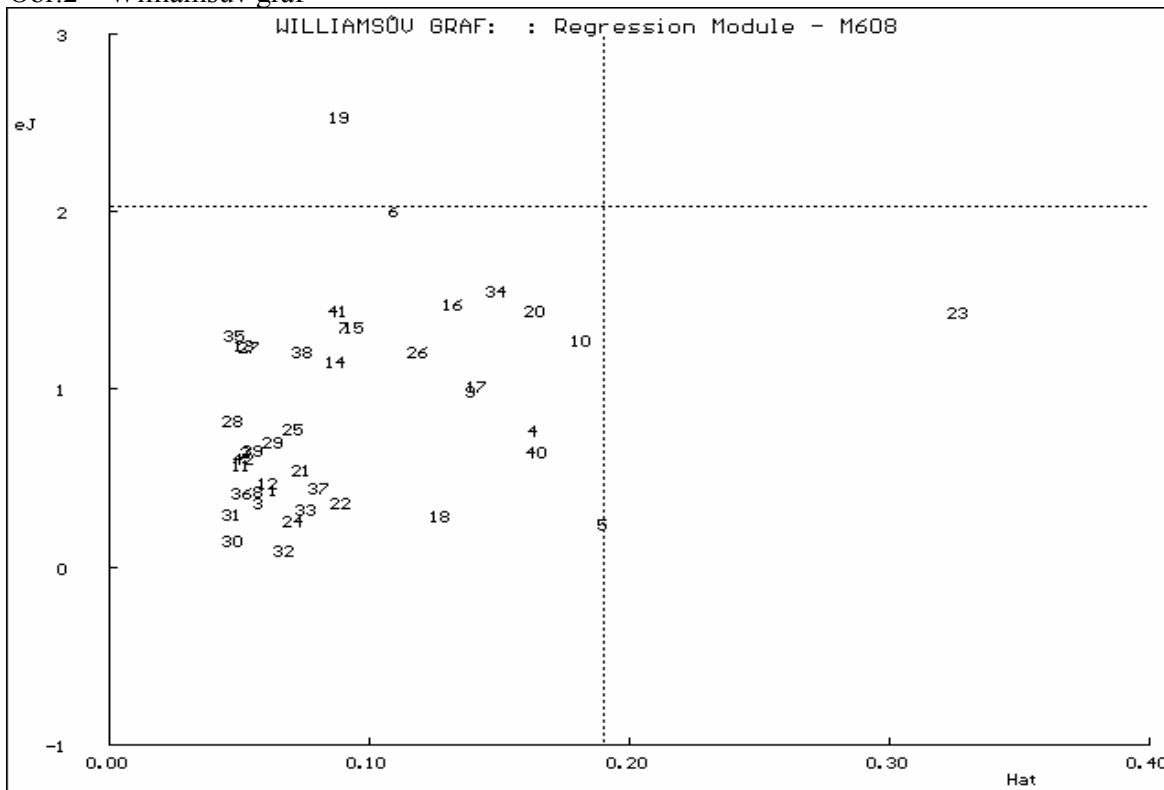
Graf predikovaných reziduí	:	O – body 6, 19, 20,34; E – 23
Williamsův graf	:	O – 19, E – 23
Pregibonův graf	:	Silně vlivný 23
McCulloh–Meeterův graf	:	O – bod 19; E – 23; O, E – 6
L–R graf	:	Silně vlivné body č. 6, 19, 23

Obr.1 – Graf predikovaných reziduí



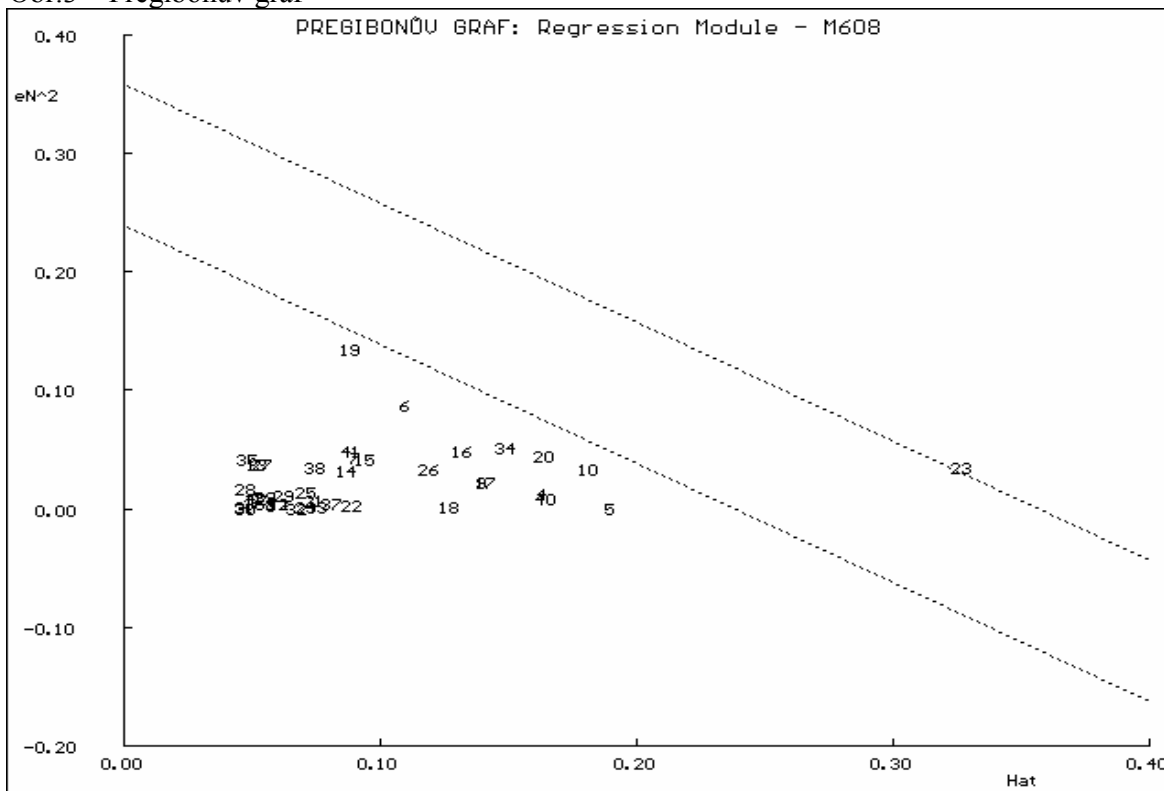
Graf ukazuje možnost extrému (E) u bodu č. 23, vybočující body (O) č. 6, 19, 20, 34.

Obr.2 – Williamsův graf



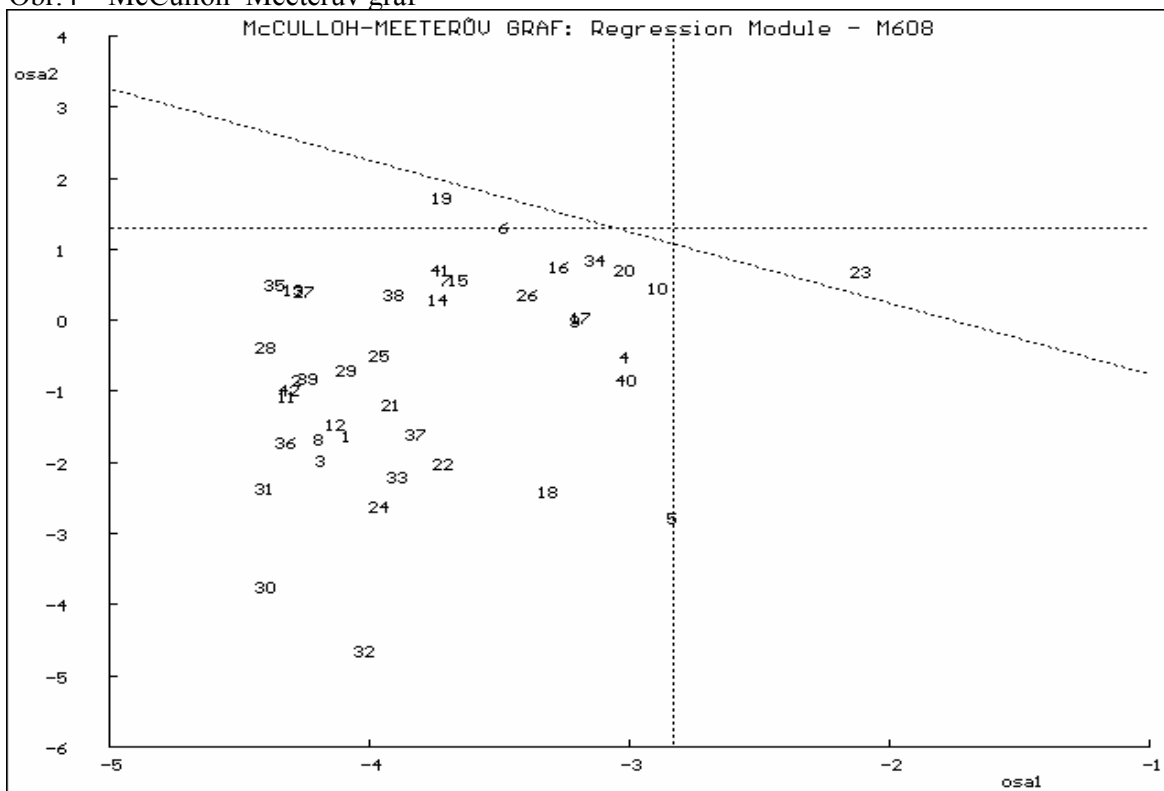
Graf indikuje extrémní hodnotu bod č. 23 a podezřelý č.5, vybočující bod č. 19 a podezřelý č 6.

Obr.3 – Pregibonův graf



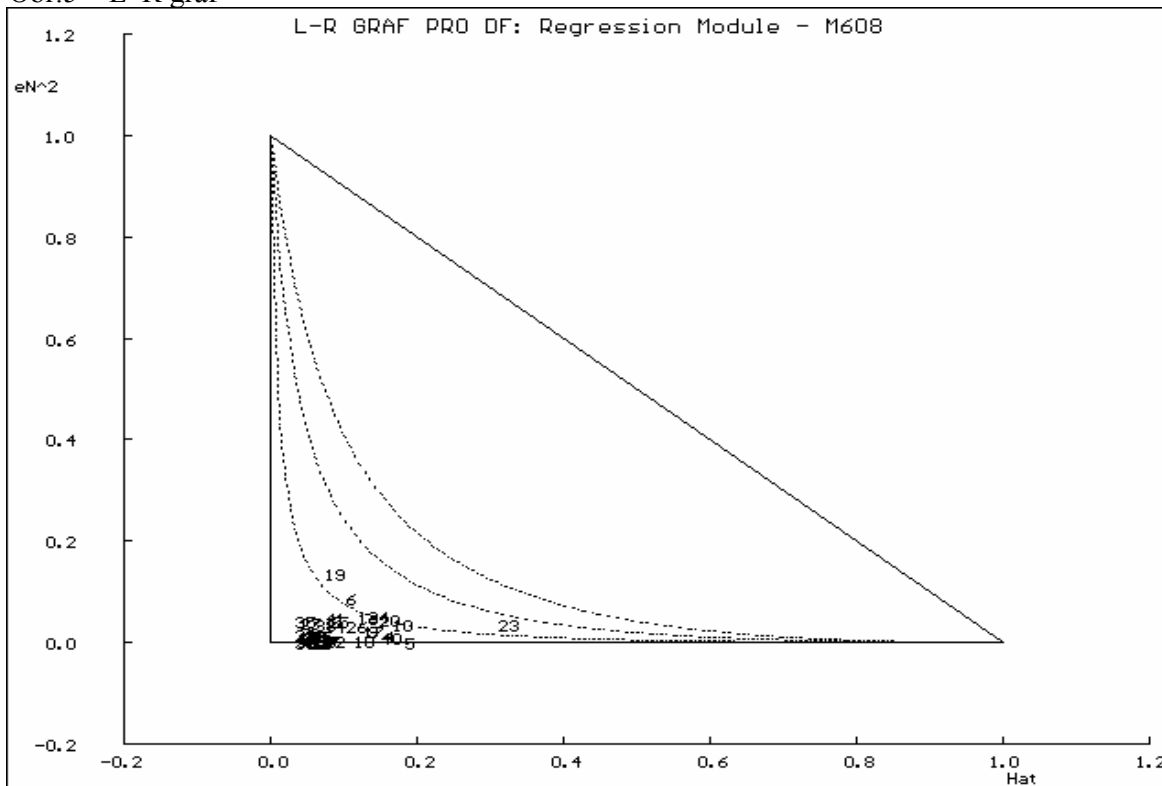
Graf indikuje silně vlivný bod č. 23

Obr.4 – McCulloh–Meeterův graf



Graf indikuje extrém v bodě č. 23, vybočující bod č. 19. Obojí vlivný (E, O) může být bod č. 6

Obr.5 – L-R graf



Jako silně vlivné body se jeví bod č.6, 19 a 23.

Závěr : Testováním modelu závislosti obsahu vysokohustotního lipoproteinu (y) na obsahu celkového cholesterolu (x_1), triglyceridu (x_2) a pre-beta komponenty (x_3) byl zjištěn lineární regresní model ve tvaru :

$$y = 48.87 (\pm 7.25) - 0.03 (\pm 0.03) \cdot x_1 + 0.02 (\pm 0.02) \cdot x_2 + 8.15 (\pm 3.11) \cdot x_3$$

V datech byl indikován extrémní bod č.23 a odlehlý bod č.19. Po jejich vyloučení však nedošlo k výrazné změně ve výsledcích. V modelu nebyla prokázána multikolinearita.

V modelu je významně odlišný od nuly pouze absolutní člen a koeficient x_3 . Jako statisticky významný se jeví parametr x_3 , což je obsah pre-beta komponenty.