SEMESTRÁLNÍ PRÁCE

Tvorba lineárních regresních modelů při analýze dat

Ing. Pavel Bouchalík

1. ZADÁNÍ

Úloha 1. Porovnání dvou regresních přímek u jednoduchého lineárního regresního modelu (včetně testování úseku a směrnice, s vyšetřením vlivných bodů a jejich event. odstraněním, posouzením míry spolehlivosti navrženého modelu). Test shodnosti dvou (nebo i více) přímek, test jejich paralelity a společného úseku.

Úloha 2. Určení stupně polynomu metodou MNČ a RH křivkové závislosti (porovnání obou metod vede k odstranění multikolinearity, testování statistické významnosti nalezených parametrů, vyšetření regresního tripletu metodou regresní diagnostiky, zdůvodnění a výklad všech užitých diagnostik a statistik).

Úloha 3. Validizace nové analytické metody (vyšetřením regresního tripletu testujte a diskutujte statistickou významnost jednotlivých parametrů v modelu stejně jako i jejich fyzikální smysl, zdůvodnění a výklad všech užitých diagnostik a statistik).

Úloha 4. Vícerozměrný lineární regresní model o alespoň 4 proměnných (vyšetřením regresního tripletu nalezněte nejlepší model, využijte regresní diagnostiku a pomocí parciálních regresních a reziduálních grafů diskutujte významnost jednotlivých parametrů v modelu stejně jako i jejich fyzikální smysl).

2. ZPRACOVANÉ ÚLOHY

2.1 Porovnání dvou regresních přímek

2.1.1 Zadání příkladu

Obsah tuhé fáze v suspenzi povrchově upraveného materiálu TiO_2 se určuje stanovením měrné hmotnosti suspenze. Byla připravena kalibrační řada jednotlivých vzorků s obsahem sušiny od 25 do 26,4% obsahu sušiny. V jednotlivých vzorcích byla stanovena hustota pyknometricky a na laboratorním hustoměru. Určete zda mezi sušinou a měrnou hmotností existuje stejná závislost pro oba způsoby stanovení hustoty.

data:

sušina [%]	5	10	15	20	25	30	35	40
ρ_T1 [kg/m^3]	1037,5	1080,1	1126,3	1191,7	1251,7	1292,2	1358,9	1432,9
ρ_T2 [kg/m^3]	1037,1	1079,5	1125,6	1183,8	1253,6	1290,9	1357,3	1430,9

2.1.2 Model sušina vs. ρ_T

2.1.2.1 Návrh modelu

Nejprve provedeme určení modelu sušina vs. ρ_T a provedeme kritiku dat a ověření předpokladů metody nejmenších čtverců.

tabulka 1: Parametry modelu sušina vs. ρ_T1

Parametr	Odhad	Směrodatná odchylka	t-kritérium	Hypotéza H_0	Hladina významnosti
β_0	968,58	8,4090	115,18	Zamítnuta	0,000
β	11,237	0,33305	33,740	Zamítnuta	0,000

tabulka 2: Statistické charakteristiky regrese

Vícenásobný korelační koeficient, R	0,99738
Koeficient determinace, R^2	0,99476
Predikovaný korelační koeficient, Rp^2	0,99400
Střední kvadratická chyba predikce, MEP	199,44
Akaikeho informační kritérium, AIC	39,759

Jak hodnota úseku tak i hodnota směrnice vyhází statisticky významně odlišné od nuly (hodnota kritického kvantitu Studentova rozdělení je větší než hodnota testovacího kritéria-primární hypotéza o $\beta i=0$ je

zamítnuta). Hodnota R je 0,997, což znamená, že cca 99,7% bodů leží na přímce určené metodou nejmenších čtverců.

2.1.2.2 Kritika dat

V hodnotách Jacknife reziduí není indikována přítomnost odlehlých bodů. Z hodnot Cookovi a Atkinsonovi vzdálenosti je patrná přítomnost podezřelých bodů 1, 3, 6 a 8.

tabulky 3,4: Analýzy reziduí

Bod	Standardizované reziduum	Jacknife reziduum	Predikované reziduum	Diagonální Prvky
i	eS[i]	eJ[i]	eP[i]	H[i,i]
1	1,5448	1,8172	21,829	0,41667
2	-0,092556	-0,084552	-1,1721	0,27381
3	-1,1078	-1,1339	-13,191	0,17857
4	-0,16105	-0,14734	-1,8644	0,13095
5	0,2182	0,19998	2,5260	0,13095
6	-1,3791	-1,5234	-16,422	0,17857
7	-0,32336	-0,29779	-4,095	0,27381
8	1,8006	2,4296	25,443	0,41667

Bod	Zobecněné diag. prvky	Cookova vzdálenost	Atkinsonova vzdálenost	Vliv na Predikci
i	Hm[i,i]	D[i]	A[i]	DF[i]
1	0,64869	085233*	2,6601*	1,5358*
2	0,27485	0,001615	0,089926	-0,051919
3	0,34659	0,1334*	0,9157	-0,52868
4	0,13471	0,0019542	0,099063	-0,057194
5	0,13785	0,0035871	0,13446	0,077629
6	0,43896	0,20674*	1,2302	-0,71027
7	0,28646	0,019712	0,31672	-0,18286
8	0,73189	1,1580*	3,5492*	2,0491*

Z grafu predikovaných reziduí je patrná přítomnost odlehlých bodů č. 1 a 8. V L-R grafu jsou body 1 a 8 vyhodnoceny jako body na rozhraní mezi odlehlými body a extrémními body. Ve Williamsově grafu je bod 8 vyhodnocen jako odlehlý bod. U McCulloh-Meeterova grafu jsou body 1 a 8 vyhodnoceny jako podezřelé na odlehlé body. Z provedených grafických diagnostik lze usoudit, že ve výběru jsou přítomny dva odlehlé body.

2.1.2.3 Ověření předpokladů metody nejmenších čtverců

Použití metody nejmenších čtverců je podmíněno splněním určitých kritérií, které prověřují níže uvedené statistické testy (výjimkou je multikolinearita).

Fischer-Snedocorův test významnosti regrese, F 1138,4
Tabulkový kvantit, F(1-α, m-1, n-m) 5,9874
Závěr: Navržený model je přijat jako významný.
Spočtená hladina významnosti 0

Scottovo kriterium multikolinearity, M -4,2943E-15
Závěr: Navržený model je korektní.

Cook-Weisbergův test heteroskedasticity, Sf 11,889
Tabulkový kvantit, Chi^2(1-α, 1) 3,8415
Závěr: Rezidua vykazují heteroskedasticitu
Spočtená hladina významnosti 0,001

Jarque-Berraův test normality reziduí, L(e) 0,41491
Tabulkový kvantit, Chi^2(1-α, 2) 5,9915
Závěr: Normalita je přijata.

Spočtená hladina významnosti	0,658
Waldův test autokorelace, Wa	0,015805
Tabulkový kvantit, Chi^2(1-α, 1)	3,8415
Závěr: Rezidua nejsou autokorelována	
Spočtená hladina významnosti	0,9
Znaménkový test, Dt	-0,20597
Tabulkový kvantil, N(1-α/2)	1,6449
Závěr: Rezidua nevykazují trend.	
Spočtená hladina významnosti	0,418

Rezidua vykazují heteroskedasticitu, což je porušení jednoho z předpokladů použití metody nejmenších čtverců. Heteroskedasticitu se pokusíme odstranit vypuštěním odlehlých bodů a to konkrétně bodu č. 1 a 8.

2.1.2.4 Konstrukce zpřesněného modelu

Po vypuštění odlehlých bodů č. 1 a 8 došlo k odstranění problému s heteroskedasticitou a zlepšení hodnoty Akaikeho kriteria a MEP (střední kvadratická chyba predikce).

tabulka 5: Parametry zpřesněného modelu

Parametr	Odhad	Směrodatná odchylka	t-kritérium	Hypotéza H_0	Hladina významnosti
β_0	965,88	7,8573	122,93	Zamítnuta	0,000
β	11,153	0,32649	34,159	Zamítnuta	0,000

tabulka 6: Statistické charakteristiky regrese

Vícenásobný korelační koeficient, R	0,99829
Koeficient determinace, R^2	0,99658
Predikovaný korelační koeficient, Rp^2	0,99651
Střední kvadratická chyba predikce, MEP	63,411
Akaikeho informační kritérium, AIC	24,621

Fischer-Snedocorův test významnosti regrese, F	1166,8
Tabulkový kvantit, F(1-α, m-1, n-m)	7,7086
Závěr: Navržený model je přijat jako významný.	
Spočtená hladina významnosti	0
Scottovo kriterium multikolinearity, M	-3,1178E-15
Závěr: Navržený model je korektní.	
Cook-Weisbergův test heteroskedasticity, Sf	2,5095
Tabulkový kvantit, Chi^2(1-α, 1)	3,8415
Závěr: Rezidua vykazují heteroskedasticitu	
Spočtená hladina významnosti	0,113
Jarque-Berraův test normality reziduí, L(e)	0,68691
Tabulkový kvantit, Chi^2(1-α, 2)	5,9915
Závěr: Normalita je přijata.	
Spočtená hladina významnosti	0,709
Waldův test autokorelace, Wa	2,5588
Tabulkový kvantit, Chi^2(1-α, 1)	3,8415
Závěr: Rezidua nejsou autokorelována	
Spočtená hladina významnosti	0,110
Znaménkový test, Dt	0,88388
Tabulkový kvantil, N(1-α/2)	1,6449
Závěr: Rezidua nevykazují trend.	

Spočtená hladina významnosti 0,188

Výsledný tvar zpřesněného modelu je **y=965,88(7,8573)+11,153(0,32649).x**. V závorkách jsou uvedeny směrodatné odchylky.

2.1.3 Model sušina vs. ρT2

2.1.3.1 Návrh modelu

tabulka 7: Parametry modelu sušina vs. $\rho_T 2$

Parametr	Odhad	Směrodatná odchylka	t-kritérium	Hypotéza H_0	Hladina významnosti
β_0	967,45	8,8785	108,97	Zamítnuta	0,000
β	11,217	0,35164	31,9	Zamítnuta	0,000

tabulka 8: Statistické charakteristiky regrese

Vícenásobný korelační koeficient, R	0,99706
Koeficient determinace, R^2	0,99414
Predikovaný korelační koeficient, Rp^2	0,99348
Střední kvadratická chyba predikce, MEP	215,98
Akaikeho informační kritérium, AIC	40,629

Jak úsek tak i směrnice vychází statisticky významně odlišné od nuly. Vysoká hodnota R svědčí o tom, že přímka určená metodou nejmenších čtverců dobře vystihuje závislost.

2.1.3.2 Kritika dat

V hodnotách Jacknife reziduí není opět indikována přítomnost odlehlých bodů. Z hodnot Cookovi a Atkinsonovi vzdálenosti je patrná přítomnost podezřelých bodů č. 1a 8.

tabulky 9, 10: Analýzy reziduí

Bod	Standardizované reziduum	Jacknife reziduum	Predikované reziduum	Diagonální Prvky
i	eS[i]	eJ[i]	eP[i]	H[i,i]
1	1,5589	1,8450	23,257	0,41667
2	-0,012382	-0,011303	-0,16556	0,27381
3	-0,9787	-0,97461	-12,304	0,17857
4	-0,75257	-0,72192	-9,1986	0,13095
5	0,5384	0,50381	6,5808	0,13095
6	-1,2654	-1,3491	-15,909	0,17857
7	-0,28370	-0,26074	-3,7934	0,27381
8	1,6958	2,1454	25,300	0,41667

Bod	Zobecněné diag. prvky	Cookova vzdálenost	Atkinsonova vzdálenost	Vliv na Predikci
i	Hm[i,i]	D[i]	A[i]	DF[i]
1	0,65294	0,86793*	2,7007*	1,5593*
2	0,27383	0,000028903	0,012022	-0,0069406
3	0,30971	0,10412	0,78707	-0,45441
4	0,21299	0,042672	0,48538	-0,2824
5	0,17294	0,021840	0,33874	0,19557
6	0,39779	0,17405*	1,0895	-0,62902
7	0,28355	0,015174	0,27731	-0,16010
8	0,69627	1,0271*	3,1405*	1,8132*

Z grafů předikovaných reziduí je patrná přítomnost dvou odlehlých bodů č. 1 a 8. Z Williamsova grafu je patrná přítomnost jednoho odlehlého bodu č. 8. V McCulloh-Meeterově grafu jsou vyhodnoceny jako odlehlé

body č. 1 a 8. V L-R grafu jsou vyhodnoceny jako body na rozhraní mezi odlehlými a extrémními body č. 1a 8. V následující fázi je nutné vypustit body č. 1a 8.

2.1.3.3 Ověření předpokladů metody nejmenších čtverců

Fischer-Snedocorův test významnosti regrese, F	1017,6
Tabulkový kvantit, F(1-α, m-1, n-m)	5,9874
Závěr: Navržený model je přijat jako významný.	
Spočtená hladina významnosti	0
Scottovo kriterium multikolinearity, M	-2,3461E-15
Závěr: Navržený model je korektní.	
Cook-Weisbergův test heteroskedasticity, Sf	11,918
Tabulkový kvantit, Chi^2(1-α, 1)	3,8415
Závěr: Rezidua vykazují heteroskedasticitu	
Spočtená hladina významnosti	0,001
Jarque-Berraův test normality reziduí, L(e)	0,69588
Tabulkový kvantit, Chi^2(1-α, 2)	5,9915
Závěr: Normalita je přijata.	
Spočtená hladina významnosti	0,706
Waldův test autokorelace, Wa	0,051146
Tabulkový kvantit, Chi^2(1-α, 1)	3,8415
Závěr: Rezidua nejsou autokorelována	
Spočtená hladina významnosti	0,821
Znaménkový test, Dt	-0,20597
Tabulkový kvantil, N(1-α/2)	1,6449
Závěr: Rezidua nevykazují trend.	
Spočtená hladina významnosti	0,418

Z výsledku Cook-Weisbergova testu je patrné, že není splněna podmínka homoskedasticity reziduí. Pokusíme se model zpřesnit vypuštěním odlehlých dodů č. 1 a 8.

2.1.3.4 Konstrukce zpřesněného modelu

Po vypuštění bodů č. 1 a 8 došlo k odstranění problémů s heteroskedasticitou a zároveň ke zlepšení hodnot statistických ukazatelů regrese.

tabulka 11: Parametry modelu sušina vs. $\rho_T 2$

Parametr	Odhad	Směrodatná odchylka	t-kritérium	Hypotéza H_0	Hladina významnosti
β_0	963,80	8,9572	107,60	Zamítnuta	0,000
β	11,170	0,3722	30,010	Zamítnuta	0,000

tabulka 12: Statistické charakteristiky regrese

Vícenásobný korelační koeficient, R	0,99779
Koeficient determinace, R^2	0,99558
Predikovaný korelační koeficient, Rp^2	0,99563
Střední kvadratická chyba predikce, MEP	79,661
Akaikeho informační kritérium, AIC	26,194

Fischer-Snedocorův test významnosti regrese, F	900,62
Tabulkový kvantit, F(1-α, m-1, n-m)	7,7086
Závěr: Navržený model je přijat jako významný.	
Spočtená hladina významnosti	0

Scottovo kriterium multikolinearity, M	-5,4279E-15
Závěr: Navržený model je korektní.	
Cook-Weisbergův test heteroskedasticity, Sf	3,1420
Tabulkový kvantit, Chi^2(1-α, 1)	3,8415
Závěr: Rezidua vykazují homoskedasticitu	
Spočtená hladina významnosti	0,076
Jarque-Berraův test normality reziduí, L(e)	0,44997
Tabulkový kvantit, Chi^2(1-α, 2)	5,9915
Závěr: Normalita je přijata.	
Spočtená hladina významnosti	0,799
Waldův test autokorelace, Wa	3,5786
Tabulkový kvantit, Chi^2(1-α, 1)	3,8415
Závěr: Rezidua nejsou autokorelována	
Spočtená hladina významnosti	0,059
Znaménkový test, Dt	0,45644
Tabulkový kvantil, N(1-α/2)	1,6449
Závěr: Rezidua nevykazují trend.	
Spočtená hladina významnosti	0,324

Konečný tvar zpřesněného modelu je tedy: y=963,80(8,9572)+11,170(0,3722).x. V závorkách jsou uvedeny hodnoty směrodatné odchylky příslušného parametru.

2.1.4 Porovnání dvou lineárních závislostí

Nyní provedeme vlastní porovnání zkoumaných přímek. V první fázi provedeme test shody rozptylů a to pomocí klasického F-testu. Testovací kritérium má následující tvar: $F = \frac{\sigma^2_{12}}{\sigma^2_{22}}$, kde σ^2_{12}, resp. σ^2_{22} jsou hodnoty reziduálního rozptylu obou výběrů po vypuštění odlehlých bodů pro první testovaný výběr a druhý testovaný výběr. Hodnotu testovací statistiky porovnáváme s hodnotou F rozdělení F(0,05;6;6).

F test = 0,9959 F (0,05;6;6) = 4,84

Podmínka shodnosti rozptylů je tedy splněna. Nyní můžeme provést **Chowův test shodnosti přímek.** V tomto případě porovnáváme hodnotu testovacího kritéria F_{Ch} s hodnotou kvantitu F rozdělení F (0,05;2;8). V případě, že je hodnota testovacího kritéria F_{Ch} menší než hodnota kritického kvantitu F rozdělení, hypotézu o shodnosti přímek na hladině významnosti 0,05 přijímáme.

$$F_{Ch} = \frac{(RSC - RSC_1 - RSC_2)(n - 2m)}{(RSC_1 + RSC_2)m} = 0{,}08144,\ F(0{,}05;2;8) = 357{,}63$$

RSC_1 - je reziduální součet čtverců odchylek prvního rozdělení
RSC_2 - je reziduální součet čtverců odchylek druhého rozdělení
n - je celkový počet pro obě přímky
m - je počet parametrů regresního modelu, v tomto případě m = 2

Jelikož je hodnota testovacího kritéria menší než hodnota kritického kvantitu F rozdělení, je možné přijat nulovou hypotézu a lze tedy říci, že obě lineární závislosti jsou shodné.

2.2 Validace nové analytické metody

2.2.1 Zadání

Bylo provedeno stanovení hustoty (suspenze TiO_2 [kg/m^3]) v suspenzi pigmentového oxidu titaničitého dvěma různými způsoby. Prověřte pomocí regrese, zda oba způsoby stanovení poskytují shodné výsledky.

data:

ρ_T1 [kg/m^3]	1037,5	1079,5	1126,3	1191,7	1253,6	1292,2	1358,9	1430,9
ρ_T2 [kg/m^3]	1037,1	1080,1	1125,6	1183,8	1251,7	1290,9	1357,3	1432,9

2.2.2 Návrh modelu

Pomocí metody nejmenších čtverců určíme parametry modelu $y = \beta.x + \beta_0$, kde y2 je hustota ρ_T2 a x je hustota ρ_T1. V ideálním případě bychom měli získat přímku s nulovým úsekem a jednotkovou směrnicí.

V tabulce č.1 je proveden odhad parametrů modelu a je testována hypotéza H_0 o statistické nevýznamnosti parametrů regrese ($\beta = 0$). Hodnota kritického kvantitu Studentova rozdělení t (0,05;6) je 2,447 a vhledem k tomu, že hodnota testovacího kritéria je v případě úseku menší než je hodnota kritického kvantitu můžeme primární hypotézu akceptovat (úsek se statisticky významně neodlišuje od nuly).

tabulka 13: Parametry modelu

Parametr	Odhad	Směrodatná odchylka	t-kritérium	Hypotéza H_0	Hladina významnosti
β_0	-4,4787	10,549	-0,4245	Akceptována	0,686
β	1,0025	0,0085895	116,71	Zamítnuta	0,000

tabulka 14: Statistické charakteristiky regrese

Vícenásobný korelační koeficient, R	0,99978
Koeficient determinace, R^2	0,99956
Predikovaný korelační koeficient, Rp^2	0,99962
Střední kvadratická chyba predikce, MEP	12,645
Akaikeho informační kritérium, AIC	19,951

Hodnoty uvedené v tab. č.2 (vysoká hodnota střední kvadratické chyby predikce a Akaikeho kritéria) svědčí o dobré shodě obou metod stanovení hustoty. Hodnota R^2 je 0,99956, což znamená, že cca 99,96% bodů leží na přímce určené metodou nejmenších čtverců.

2.2.3 Kritika dat

V další fázi je posouzeno, zda data neobsahují odlehlé body, které by mohli vést ke zkreslení při návrhu regresního vztahu.

tabulka 15: Reziduální odchylky modelu

Bod	Standardizované reziduum	Jacknife reziduum	Predikované reziduum	Diagonální Prvky
i	eS[i]	eJ[i]	eP[i]	H[i,i]
1	0,5937	0,55863	2,3587	0,37960
2	0,88572	0,86723	3,2587	0,27655
3	0,3342	0,30796	1,1642	0,19303
4	-2,2034	-4,6039*	-7,399	0,13161
5	-0,19951	-0,18274	-0,67046	0,13285
6	-0,02745	-0,02506	-0,093884	0,16285
7	-0,20477	-0,18703	-0,74656	0,2676
8	1,2441	1,3184	5,2781	0,4559

V tabulce č.13 je jedna hodnota označena hvězdičkou, což znamená, že ve výběru je jeden podezřelý bod, ale v grafech odlehlých bodů nebyl potvrzen jako odlehlý bod.

2.2.4 Ověření předpokladů metody nejmenších čtverců

Použití metody nejmenších čtverců je podmíněno splněním určitých kritérií. Mezi ně patří normalita reziduí, homoskedasticita a v reziduích nesmí být trend. Důležitým hodnotícím prvkem je i Scottovo kriterium multikolinearity. To nepatří mezi základní předpoklady použití metody nejmenších čtverců, ale vysoké hodnoty multikolinearity indikují přeurčenost modelu a způsobují nárůst hodnot směrodatných odchylek parametrů modelu.

Fischer-Snedocorův test významnosti regrese, F	13622
Tabulkový kvantit, F(1-α, m-1, n-m)	5,9874
Závěr: Navržený model je přijat jako významný.	
Spočtená hladina významnosti	0
Scottovo kriterium multikolinearity, M	3,2782E-14
Závěr: Navržený model je korektní.	
Cook-Weisbergův test heteroskedasticity, Sf	0,0349617
Tabulkový kvantit, Chi^2(1-α, 1)	3,8415
Závěr: Rezidua vykazují homoskedasticitu	
Spočtená hladina významnosti	0,85167
Jarque-Berraův test normality reziduí, L(e)	3,34398
Tabulkový kvantit, Chi^2(1-α, 2)	5,9915
Závěr: Normalita je přijata.	
Spočtená hladina významnosti	0,18787
Waldův test autokorelace, Wa	0,006182
Tabulkový kvantit, Chi^2(1-α, 1)	3,8415
Závěr: Rezidua nejsou autokorelována	
Spočtená hladina významnosti	0,85167

Z výše uvedených výsledků testů je patrné, že jsou splněny všechny předpoklady použití metody nejmenších čtverců.

2.2.5 Závěr provedené validace

Navržený model je korektní a rovněž je korektní použití metody nejmenších čtverců. V případě úseku je 95% interval spolehlivosti úseku (-30,29; 21,33) a interval spolehlivosti směrnice je (0,98; 1,02). Interval spolehlivosti úseku obsahuje nulu a tedy úsek je statisticky nevýznamný a interval spolehlivosti směrnice obsahuje jedničku a tedy se statisticky významně neodlišuje od jedničky. Obě použité metody vedou ke stejným výsledkům.

2.3 Vícerozměrný lineární regresní model

2.3.1 Zadání

Při rozkladu ilmenitu kyselinou sírovou a následném rozpouštění rozkladné hmoty se provádí redukce přítomného železa. Železo je možno odstranit z roztoku ve formě zelené skalice. V průběhu roku došlo ke kolísání spotřeby železných odstřižků vztažených na 1tunu ilmenitu. Ověřte zda není skrytá závislost mezi obsahem TiO_2 a železem dvojmocným a trojmocným přítomném v ilmenitu. Proveďte regresní diagnostiku a diskutujte význam a fyzikální smysl jednotlivých parametrů.

data:

Měsíc	odstřižky/ilm. [kg/t]	TiO_2 [%]	Fe^{2+} [%]	Fe^{3+} [%]
leden	78,051	56,11	15,94	14,92
únor	77,976	54,83	14,52	16,53
březen	79,383	55,45	14,95	15,64
duben	77,476	54,67	14,06	15,66
květen	80,618	54,99	14,29	15,04
červen	73,420	54,90	12,62	16,90
červenec	77,438	55,14	13,30	16,42
srpen	80,789	55,62	14,27	15,64
září	80,064	55,78	14,31	15,58
říjen	80,592	55,82	14,35	15,50
listopad	87,221	55,96	14,45	15,43
prosinec	79,810	55,85	14,87	15,19

1) Navržený model

$y=c+b_1*x_1+b_2*x_2+b_3*x_3$

y měrná spotřeba Fe odstřižků na redukci Fe^{3+} vztažená na tunu ilmenitu
x_1 obsah TiO_2 v ilmenitu vyjádřený v hmotnostních procentech
x_2 obsah Fe^{2+} v ilmenitu vyjádřený v hmotnostních procentech
x_3 obsah Fe^{3+} v ilmenitu vyjádřený v hmotnostních procentech
c konstanta

2.3.2 Předběžná analýza dat

tabulka 16: Párové korelační koeficienty

Proměnná	Průměr	Směrodatná odchylka	Párový korelační koeficient	Spočítaná hladina významnosti
y	15,704	0,6096	1	-
x_1	79,451	3,2309	-0,5649	0,056
x_2	55,427	0,49789	-0,6117	0,035
x_3	14,330	0,82134	-0,7625	0,004

tabulka 17: Párové korelační koeficienty mezi dvojicemi nezávisle proměnných

Parametr	x_1	x_2	x_3
x_1	1		
x_2	0,56976	1	
x_3	0,38848	0,60053	1

Z hodnot párových korelačních koeficientů je patrná vysoká korelace mezi y a x_2, x_3. Významná korelace je mezi nezávisle proměnnými x_1 vs. x_2 a x_2 vs. x_3. Z hodnot korelačních koeficientů lze usoudit, že půjde o významný, ale přeurčený model s rizikem multikolinearity.

2.3.3 Odhady parametrů a testy významnosti

tabulka 18: Odhady parametrů modelu a test významnosti parametrů
H_0: $b_i=0$ vs. H_a: $b_i<>0$

Parametr	Odhad	Směrodatná odchylka	t-kriterium	Hypotéza H_0 je	Hladina významnosti
konstanta	32,595	16,617	1,9616	Akceptována	0,085
B[1]	-0,052708	0,04668	-1,129	Akceptována	0,292
B[2]	-0,11449	0,34910	-0,32795	Akceptována	0,751
B[3]	-0,44367	0,18874	-2,3507	Zamítnuta	0,047

Testujeme hypotézu H_0. b_i je rovno nule. Vzhledem k tomu, že tabulkový kvantit t (0,05, 4) je 2,306 a hodnoty testovacího kritéria pro úsek, b_1 a b_2 jsou menší, je primární hypotéza pro tyto parametry akceptována (příslušné směrnice a úsek se statisticky významně neodlišují od nuly). Směrnice b_3 je významná.

2.3.4 Statistické charakteristiky regrese

Vícenásobný korelační koeficient, R	0,81905
Koeficient determinace, R^2	0,67084
Predikovaný korelační koeficient, Rp^2	0,42221
Střední kvadratická chyba predikce, MEP	0,27989
Akaikeho informační kritérium, AIC	-18,259

Koeficient determinace má hodnotu 0,67084 cca 67,08% hodnot vyhovuje zvolenému regresnímu modelu. Lze tedy očekávat, že zvolený model je statisticky významný.

2.3.5 Charakteristiky reziduí

Reziduální součet čtverců, RSC	1,3454
Průměr absolutních hodnot reziduí, Me	0,23632
Průměr relativních reziduí, Mer	1,5050
Odhad reziduálního rozptylu, s^2 (e)	0,16817
Odhad směrodatné odchylky reziduí, s (e)	0,41008
Odhad šikmosti reziduí, g1 (e)	0,26396
Odhad špičatosti reziduí, g2 (e)	3,7603

Z hodnot šikmosti a špičatosti lze usuzovat, že bude splněn jeden z předpokladů použití metody nejmenších čtverců – normalita.

2.3.6 Testování regresního tripletu

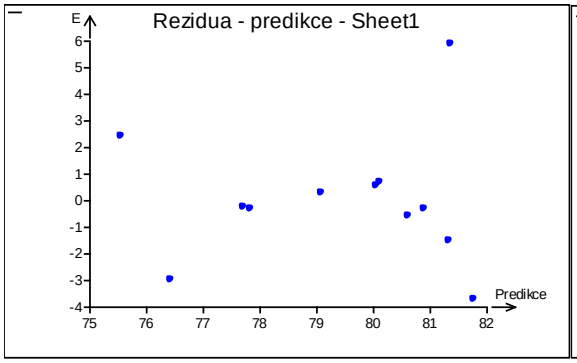
Použití metody nejmenších čtverců je podmíněno splněním určitých podmínek. Jejich prověření je provedeno v následujícím bloku testů a doloženo pomocí níže uvedených grafů.

Fischer-Snedocorův test významnosti regrese, F	5,4349
Tabulkový kvantit, F(1-α, m-1, n-m)	4,0662
Závěr: Navržený model je přijat jako významný.	
Spočtená hladina významnosti	0,025
Scottovo kriterium multikolinearity, M	0,40480
Závěr: Navržený model není korektní.	
Cook-Weisbergův test heteroskedasticity, Sf	0,26105
Tabulkový kvantit, Chi^2(1-α, 1)	3,8415
Závěr: Rezidua vykazují homoskedasticitu	
Spočtená hladina významnosti	0,609
Jarque-Berraův test normality reziduí, L(e)	0,4284
Tabulkový kvantit, Chi^2(1-α, 2)	5,9915
Závěr: Normalita je přijata.	
Spočtená hladina významnosti	0,807
Waldův test autokorelace, Wa	0,22171
Tabulkový kvantit, Chi^2(1-α, 1)	3,8415
Závěr: Rezidua nejsou autokorelována	
Spočtená hladina významnosti	0,638

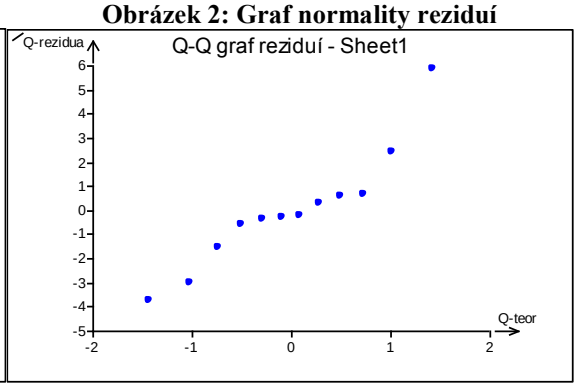
Znaménkový test, Dt:	0,30277
Tabulkový kvantit, N(1-α/2)	1,6449
Závěr: Rezidua nevykazují trend	
Spočtená hladina významnosti	0,381

Z výše uvedených testů je patrné, že jsou splněny všechny podmínky použití metody nejmenších čtverců. Zvolený model je statisticky významný, ale z hlediskamultikolinarity není příliš korektní. To je důsledek přeurčenosti modelu. Tento problém odstraníme vypuštěním nevýznamných parametrů.
Na Obr. 1 rezidua náhodně oscilují kolem nulové hodnoty. Obdobně tomu je i u grafu homoskedasticity Obr. 3 a autokorelace reziduí Obr. 4. Z Q-Q grafu je dobře patrná normalita reziduí (body vykazují hladkou lineární závislost, Obr. 2).

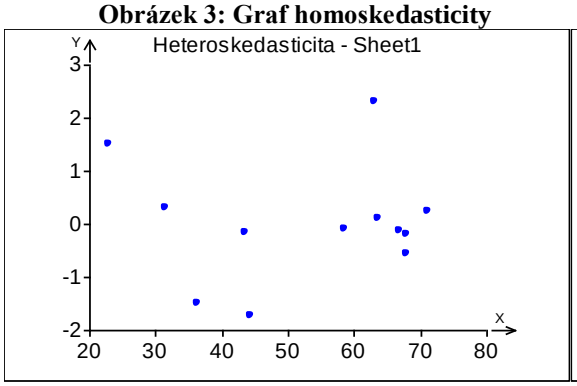
Obrázek 1: Graf reziduí



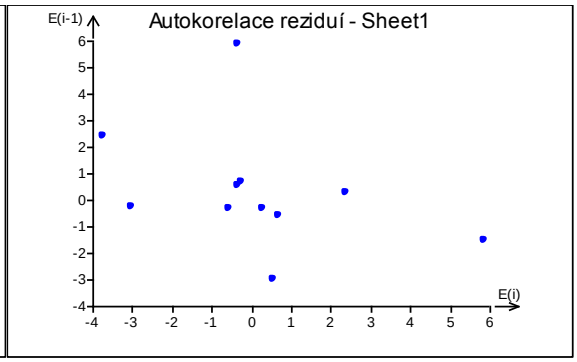
Obrázek 2: Graf normality reziduí



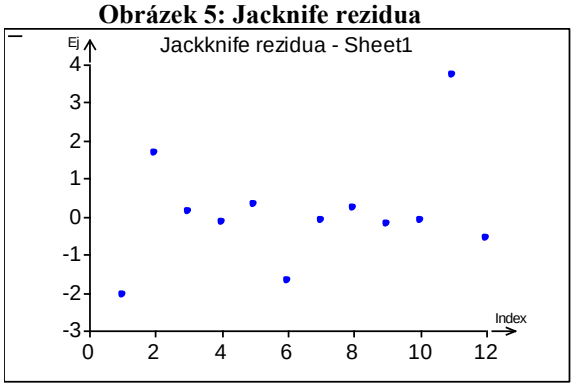
Obrázek 3: Graf homoskedasticity



Obrázek 4: Autokorelace reziduí



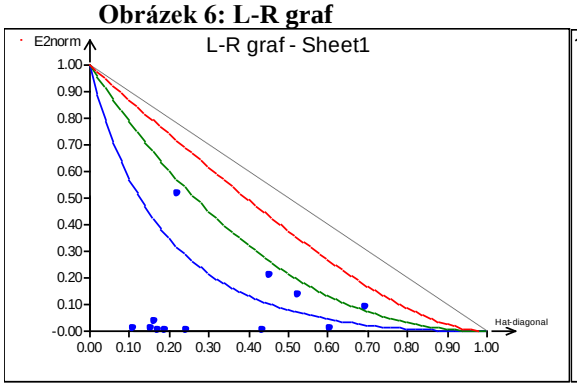
Obrázek 5: Jacknife rezidua



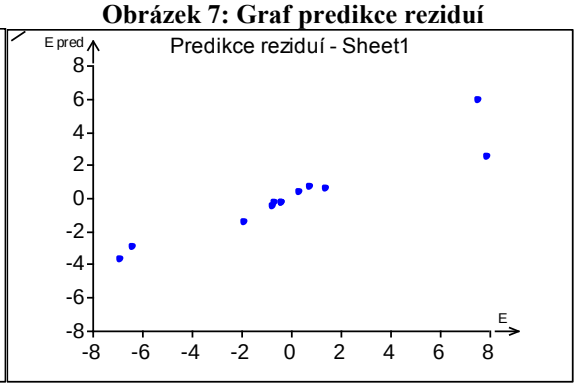
2.3.7 Identifikace odlehlých bodů

Z grafu Jacknife reziduí Obr. 5 je patrná přítomnost odlehlých bodů. Přítomnost podezřelých bodů naznačuje rovněž graf na Obr. 7, 8. Nicméně ostatní grafy Obr. 6, 9 přítomnost odlehlých bodů neindikují. V tabulce č. 19 (Jacknife) jsou indikovány dva odlehlé body, ale z grafů byly vyhodnoceny jako extrémní body. Ve výběru jsou podezřelé body 2, 5 a 11 viz. Cookova a Atkinsonova vzdálenost a parametr DF, ale k jejich vyloučení není dostatek argumentů.

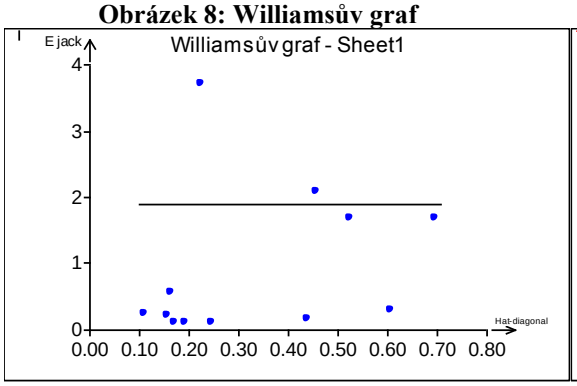
Obrázek 6: L-R graf



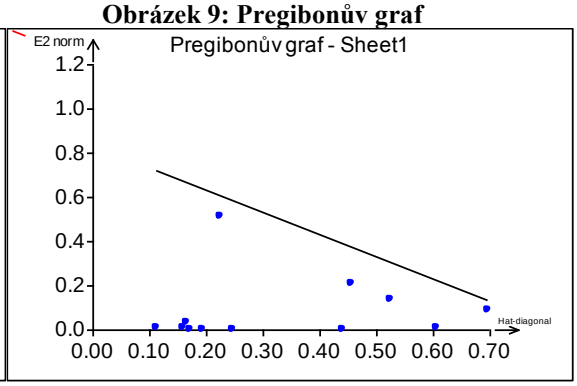
Obrázek 7: Graf predikce reziduí



Obrázek 8: Williamsův graf



Obrázek 9: Pregibonův graf



tabulka 19: Identifikace vlivných bodů (*indikuje odlehlý nebo vlivný bod)

Bod	Standardizované reziduum	Jacknife reziduum	Predikované reziduum	Diagonální Prvky
i	eS[i]	eJ[i]	eP[i]	H[i,i]
1	-0,27585	-0,25927	-0,195968	0,66581
2	2,3037	3,7142*	1,1680	0,34584
3	0,59696	0,57127	0,26835	0,16780
4	-1,0710	-1,0824	-0,54388	0,34788
5	-1,9241	-2,4555*	-0,92866	0,27812
6	0,24829	0,23315	0,17576	0,66441
7	0,33503	0,31561	0,15738	0,23787
8	0,004888	0,0045731	0,0021331	0,11664
9	-0,080169	-0,075022	-0,035828	0,15798
10	-0,2411	-0,22635	-0,10849	0,16943
11	1,070	1,0812	0,77105	0,67617*
12	-0,55523	-0,52968	-0,25024	0,17206

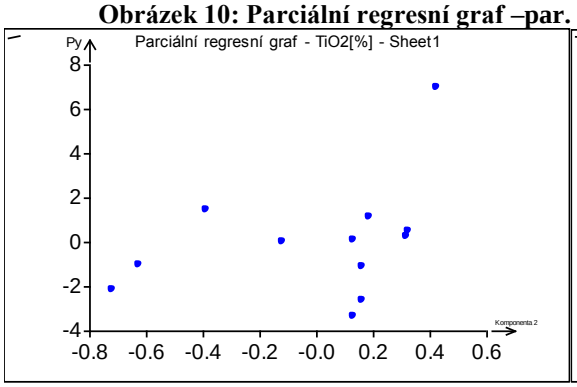
tabulka 20: Identifikace odlehlých bodů

Bod	Zobecněné diag. prvky	Cookova vzdálenost	Atkinsonova vzdálenost	Vliv na Predikci
i	Hm[i,i]	D[i]	A[i]	DF[i]
1	0,66899	0,0379	0,51784	-0,36595
2	0,7798	0,70142*	3,8192*	2,7006*
3	0,2048	0,017963	0,36277	0,25652
4	0,44138	0,15298	1,1181	-0,79060
5	0,61217	0,35657*	2,1554*	-1,5241*
6	0,66699	0,030512	0,46394	0,32806
7	0,24857	0,0087583	0,24936	0,17632
8	0,11664	0,0000007889	0,00235	0,0016617
9	0,15866	0,00030146	0,045956	-0,032496
10	0,17547	0,0029645	0,14458	-0,10223
11	0,72251	0,59761*	2,2095*	1,5623*
12	0,20396	0,016016	0,3414	-0,24146

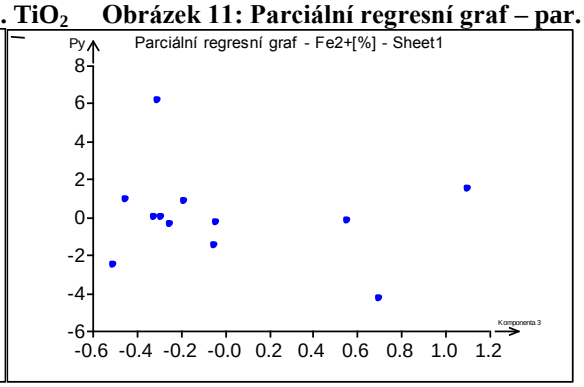
2.3.8 Upravený model

Z předchozího textu je patrné, že je nutné vypustit parametry 1, 2 a úsek, které se jeví jako nevýznamné viz. Obr. 10 a 11. Hodnoty parametrů modelu jsou po úpravě následující viz. tab. 21.

Obrázek 10: Parciální regresní graf –par. TiO_2



Obrázek 11: Parciální regresní graf – par. Fe^{2+}



tabulka 21: Odhady parametrů upraveného modelu a testy významnosti

Parametr	Odhad	Směrodatná odchylka	t-kriterium	Hypotéza H_0 je	Hladina významnosti
B[1]	0,1972	0,0040064	49,221	Zamítnuta	0,000

2.3.9 Regresní parametry upraveného regresního modelu jsou následující:

tabulka 22: Parametry původního a upraveného regresního modelu

Model	Původní	Upravený
Vícenásobný korelační koeficient, R	0,81905	0,000
Koeficient determinace, R^2	0,67084	0,000
Predikovaný korelační koeficient, Rp^2	0,42221	0,000
Střední kvadratická chyba predikce, MEP	0,27989	1,3232
Akaikeho informační kritérium, AIC	-18,259	3,3053

Z výsledků níže uvedených testů je patrné, že jsou splněny podmínky použití metody nejmenších čtverců. Vypuštěním přebytečných parametrů byl odstraněn problém s multikolinearitou (došlo k významnému snížení Scottova kriteria multikolinearity).

Fischer-Snedocorův test významnosti regrese, F 0,0024227
Tabulkový kvantit, F(1-α, m-1, n-m) 4,84443
Závěr: Navržený model je přijat jako významný.
Spočtená hladina významnosti 0,000

Scottovo kriterium multikolinearity, M	0,00000
Závěr: Navržený model je korektní.	
Cook-Weisbergův test heteroskedasticity, Sf	60,5000
Tabulkový kvantit, Chi^2(1-α, 1)	3,8415
Závěr: Rezidua vykazují heteroskedasticitu	
Spočtená hladina významnosti	0,0000
Jarque-Berraův test normality reziduí, L(e)	0,87393
Tabulkový kvantit, Chi^2(1-α, 2)	5,9915
Závěr: Normalita je přijata.	
Spočtená hladina významnosti	0,646
Waldův test autokorelace, Wa	0,13147
Tabulkový kvantit, Chi^2(1-α, 1)	3,8415
Závěr: Rezidua nejsou autokorelována	
Spočtená hladina významnosti	0,717
Znaménkový test, Dt:	0,11498
Tabulkový kvantit, N(1-α/2)	1,6449
Závěr: Rezidua nevykazují trend	
Spočtená hladina významnosti	0,454

Výsledný model má tedy tvar: $y = 0{,}1972(0{,}0040064)x_1$. Kde x_1 je obsah Fe^{3+}. Z toho plyne logická závislost, že čím je v ilmenitu vyšší obsah Fe^{3+} tím je vyšší poměr spotřeby železných odstřižků vztaženo na 1[t] ilmenitu.

2.4 Určení stupně polynomu

2.4.1 Zadání

V laboratorním měřítku byla provedena zkouška rychlosti dehydratace sádrovce $CaSO_4.2H_2O$ při teplotě 105°C jako závislost úbytku hmotnosti na čase. Proložte tuto závislost vhodným typem polynomu.

data:

čas [min]	0	2	4	6	10	11	14	24	28	31	48
Δm [%]	0	0,75	2,31	4,8	9,45	10,31	13,1	15,68	15,8	15,8	15,98

2.4.2 Návrh modelu s absolutním členem

Při určení stupně polynomu budeme postupovat tak, že zkusmo vygenerujeme pro stupeň polynomu 2, 3, 4, 5 a 6 statistické charakteristiky regrese (AIC, Rp^2, MEP a R^2). Pro optimální stupeň polynomu vychází hodnota parametru AIC (Akaikeho informační kriterium) a MEP (střední kvadratická chyba predikce nejmenší). Pomocí těchto ukazatelů provedeme návrh modelu.

tabulka 23: Určení stupně polynomu

Stupeň polynomu	2	3	4	5	6	7
Vícenásobný korelační koeficient, R	0,98627	0,99131	0,99805	0,99992	0,99992	0,99993
Koeficient determinace, R^2	0,973	0,98269	0,9961	0,99985	0,99985	0,99986
Predikovaný korelační koeficient, Rp	0,7699	0,000	0,000	0,98611	0,98127	0,000
Střední kvadratická chyba predikce, MEP	15,073	257,74	1334,8	1,0194	1,3712	448,11
Akaikeho informační kritérium, AIC	6,0766	3,5296	-9,363	-39,772	-37,798	-36,745

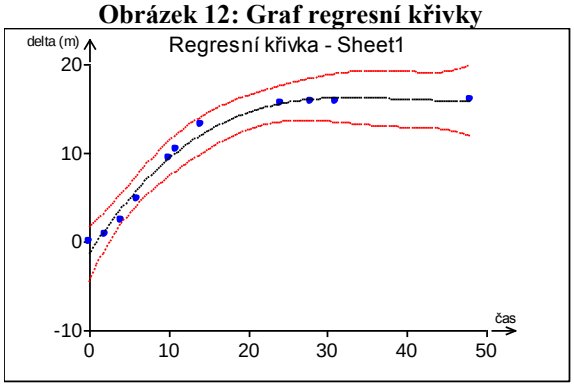
Z tab. 23 a Obr 12 je patrné, že nejvýhodnější bude proložit uvedenou závislost polynomem 3. stupně.

tabulka 24: Odhady parametrů s absolutním členem

Proměnná	Odhad	Směr. odchylka	Závěr	Pravděpodobnost	Spodní mez	Horní mez
Abs	-1,4466	0,733242	Nevýznamný	0,089111	-3,18045	0,287237
čas	1,408948	0,166127	Významný	6,27E-05	1,016119	1,801777
čas^2	-0,03673	0,008691	Významný	0,003907	-0,05728	-0,01618
čas^3	0,00031	0,000119	Významný	0,034951	2,91E-05	0,000592

Součástí tabulky 24 je testování významnosti parametrů modelu. Absolutní člen vychází jako nevýznamný (hodnota pravděpodobnosti je větší než 0,05 a 95% interval spolehlivosti zahrnuje nulu).

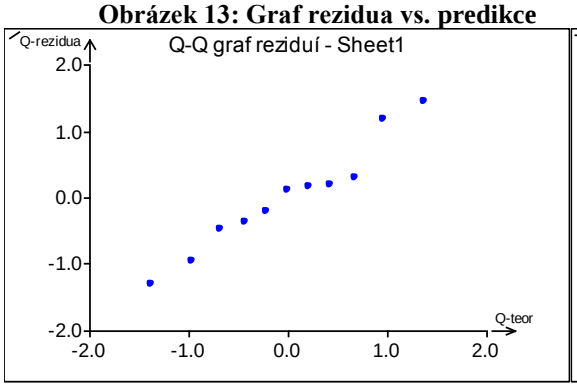
Obrázek 12: Graf regresní křivky



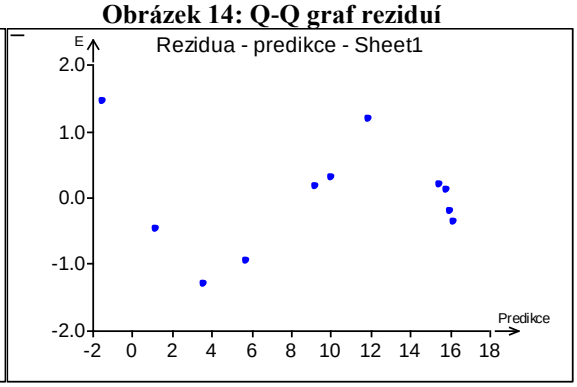
2.4.3 Kritika dat

Z obrázku 14 je patrné, že bude splněna podmínka normality reziduí. Body na obrázku 14 (graf rezidua vs. Predikce) lze proložit křivku z čehož vyplývá, že dojde patrně k problému s multikolinearitou.

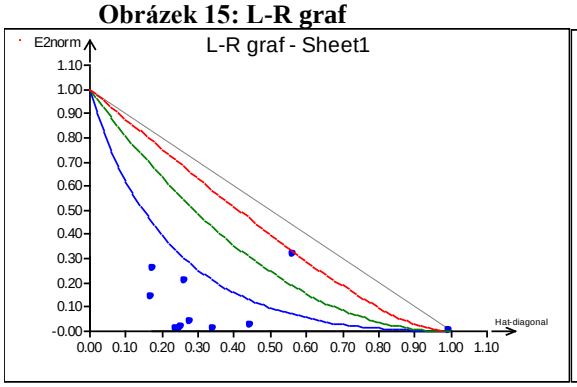
Obrázek 13: Graf rezidua vs. predikce



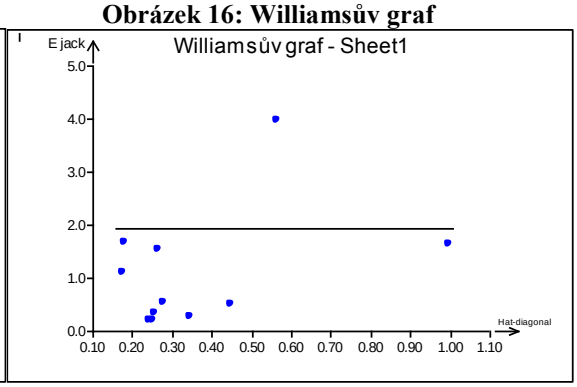
Obrázek 14: Q-Q graf reziduí



Obrázek 15: L-R graf



Obrázek 16: Williamsův graf



Z obrázku 16 je patrná přítomnost odlehlého bodu s číslem 11. Tento bod leží v L-R grafu na obrázku 15 těsně na hranici, za kterou jsou body považovány za odlehlé.

2.4.4 Testování regresního modelu s absolutním členem

Použití metody nejmenších čtverců je podmíněno splnění určitých kritérií. Jejich výpočet a ověření je uveden v následujícím textu.

Statistické charakteristiky regrese:

Vícenásobný korelační koeficient R:	0,99131
Koeficient determinace R^2:	0,98269
Predikovaný korelační koeficient Rp:	0,000
Střední kvadratická chyba predikce MEP:	257,74
Akaikeho informační kritérium:	3,5296

Testování regresního tripletu

Fisher-Snedecorův test významnosti modelu

Hodnota kritéria F :	142,6608
Kvantil F (1-alfa, m-1, n-m) :	4,346831
Pravděpodobnost :	1,2231E-006

Závěr : Model je významný

Scottovo kritérium multikolinearity

Hodnota kritéria SC :	0,61969

Závěr : Model vykazuje multikolinearitu!

Cook-Weisbergův test heteroskedasticity

Hodnota kritéria CW :	1,4859
Kvantil Chi^2(1-alfa,1) :	3,8414
Pravděpodobnost :	0,2228

Závěr : Rezidua vykazují homoskedasticitu.

Jarque-Berrův test normality

Hodnota kritéria JB :	0,2385
Kvantil Chi^2(1-alfa,2) :	5,99146
Pravděpodobnost :	0,88758

Závěr : Rezidua mají normální rozdělení.

Waldův test autokorelace

Hodnota kritéria WA :	0,33098
Kvantil Chi^2(1-alfa,1) :	3,84146
Pravděpodobnost :	0,22284

Závěr : Autokorelace je nevýznamná

Durbin-Watsonův test autokorelace

Hodnota kritéria DW :	1,19608	
Kritické hodnoty DW	0	2

Závěr : Pozitivní autokorelace reziduí není prokázána.

Znaménkový test reziduí

Hodnota kritéria Sg :	0,61237
Kvantil N(1-alfa/2) :	1,95996
Pravděpodobnost :	0,54029

Závěr : V reziduích není trend.

Z výše uvedených statistických charakteristik modelu je patrné, že prakticky všechny důležité parametry jsou v případě tohoto modelu splněny. Výjimkou je pouze vysoká hodnota Scottova kritéria. Hodnota tohoto ukazatele, ale nedosahuje kritické hodnoty, která by vyžadovala zamítnutí modelu.

2.4.5 Návrh modelu bez absolutního členu

Absolutní člen vyšel v tabulce 24 jako nevýznamný. Model s absolutním členem vykazuje rovněž vysokou hodnotu Scottova kritéria, což může být způsobeno přeurčeností modelu. Z těchto důvodů provedeme výpočet parametrů regrese pro model bez absolutního členu a následně provedeme testování regresního tripletu.

tabulka 25: Odhad parametrů modelu bez absolutního členu

Proměnná	Odhad	Směr. odchylka	Závěr	Pravděpodobnost	Spodní mez	Horní mez
čas	1,143277	0,113519	Významný	8,05E-06	0,881501	1,405053
čas^2	-0,024998	0,007387	Významný	0,009616	-0,04202	-0,00795
čas^3	0,000167	0,00011	Nevýznamný	0,166754	-8,64E-05	0,000421

Z tabulky 25 je patrné, že po odstranění úseku se jeví parametr čas^3 jako nevýznamný. Jeho vypuštění, ale vede k výraznému zhoršení modelu – kvality proložení křivky a ukazatelů MEP a R^2 a Rp viz. tabulka 26.

tabulka 26: Statistické parametry regrese bez absolutního členu

Stupeň polynomu	3
Vícenásobný korelační koeficient, R	0,9874
Koeficient determinace, R^2	0,974959
Predikovaný korelační koeficient, Rp	1,134662
Střední kvadratická chyba predikce, MEP	77,6809
Akaikeho informační kritérium, AIC	5,341414

2.4.6 Testování modelu bez absolutního členu

Testování regresního tripletu

Fisher-Snedecorův test významnosti modelu
Hodnota kritéria F : 155,739
Kvantil F (1-alfa, m-1, n-m) : 4,45897
Pravděpodobnost : 3,932E-007
Závěr : Model je významný

Scottovo kritérium multikolinearity
Hodnota kritéria SC : 0,91544
Závěr : Model je nekorektní!

Cook-Weisbergův test heteroskedasticity
Hodnota kritéria CW : 0,72770
Kvantil Chi^2(1-alfa,1) : 3,8414
Pravděpodobnost : 0,39363
Závěr : Rezidua vykazují homoskedasticitu.

Jarque-Berrův test normality
Hodnota kritéria JB : 1,19381
Kvantil Chi^2(1-alfa,2) : 5,991464
Pravděpodobnost : 0,5505132
Závěr : Rezidua mají normální rozdělení.

Waldův test autokorelace
Hodnota kritéria WA : 3,02031
Kvantil Chi^2(1-alfa,1) : 3,841459
Pravděpodobnost : 0,393630
Závěr : Autokorelace je nevýznamná

Durbin-Watsonův test autokorelace
Hodnota kritéria DW : 0,837657
Kritické hodnoty DW 0 2
Závěr : Pozitivní autokorelace reziduí není prokázána.

Znaménkový test reziduí

Hodnota kritéria Sg :	0,612372
Kvantil N(1-alfa/2) :	1,959964
Pravděpodobnost :	0,5402914

Závěr : V reziduích není trend.

Použitím polynomu třetího stupně bez absolutního členu vedlo k výraznému zhoršení Scottova kritéria a model je z tohoto hlediska nekorektní.

2.4.7 Závěr

Volbou polynomu třetího stupně s absolutním členem bylo dosaženo nejlepších hodnot statistických hodnot ukazatelů regrese. Modle je zatížen multikolinearitou, ale hodnota Scottova kritéria není vyšší než 0,8. Z tohoto důvodu nebyla použita metoda Racionálních hodností. Ve výběru je indikována přítomnost podezřelého bodu č. 11, ale k jeho vypuštění není dostatek argumentů. Absolutní člen vyhází jako statisticky nevýznamný. Jeho vypuštění vedlo k vysoké hodnotě multikolinearity a proto byl zachován. Tvar zvoleného modelu je následující (v závorkách jsou uvedeny směrodatné odchylky, **t** je čas a **y** je pokles hmotnosti v hmotnostních procentech):

$$y = -1{,}45(0{,}733)+1{,}4089(0{,}166).t-0{,}0367(0{,}0367).t^2+0{,}0003(0{,}00031).t^3$$