UNIVERZITA PARDUBICE

Fakulta chemicko-technologická
Katedra analytické chemie

Licenční studium chemometrie na téma

Statistické zpracování dat

Semestrální práce z 3. soustředění

Předmět: 2.1 Tvorba lineárních regresních modelů při analýze dat

Vedoucí licenčního studia: Prof. RNDr. Milan Meloun, DrSc.

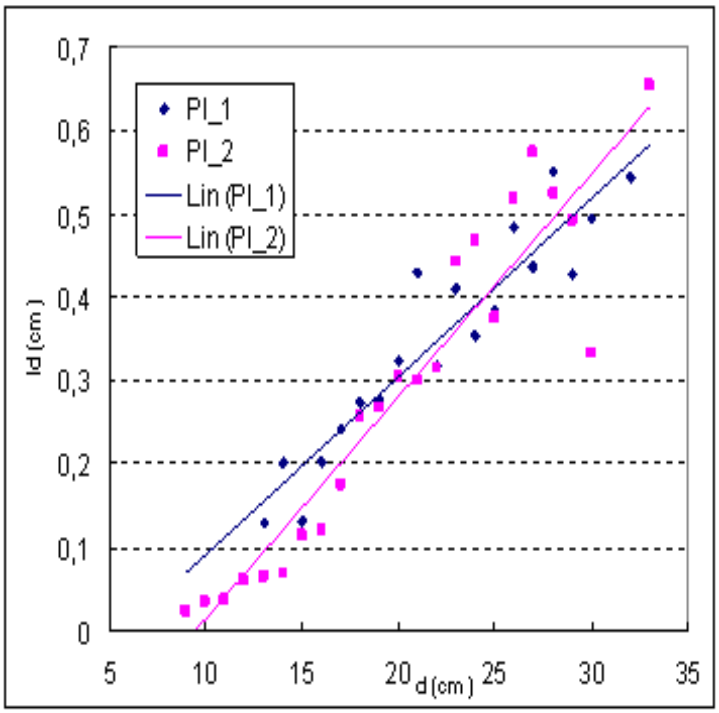
Vypracoval: Ing. Jiří Souček, Ph.D.

Licenční studium Statistické zpracování experimentálních dat.
Předmět: 2.1 Tvorba lineárních regresních modelů při analýze dat
Přednášející: Prof. RNDr. Milan Meloun, DrSc.

Úloha 1. Porovnání dvou regresních přímek u jednoduchého lineárního regresního modelu

Zadání: Porovnejte tloušťkový přírůst dvou sousedních porostů s rozdílným režimem výchovy.

	Plocha_1		Plocha_2	
	id	N	id	N
9			0,02	4
10			0,04	5
11			0,04	9
12			0,06	18
13	0,13	2	0,07	16
14	0,2	8	0,07	23
15	0,13	10	0,11	21
16	0,2	12	0,12	12
17	0,24	18	0,18	10
18	0,28	21	0,26	15
19	0,28	20	0,27	11
20	0,32	19	0,31	11
21	0,43	11	0,3	7
22	0,32	17	0,32	10
23	0,41	9	0,44	7
24	0,35	6	0,47	8
25	0,39	7	0,38	6
26	0,48	4	0,52	2
27	0,44	7	0,57	3
28	0,55	1	0,52	3
29	0,43	1	0,49	2
30	0,5	1	0,33	2
31				
32	0,55	1		
33			0,65	1



Řešení:

1. Návrh modelu

Navržený regresní model přímky je $y = \beta_0 + \beta_1 x$. Bude testována nulová hypotéza H_0: $\beta_0 = 0, \beta_1 = 1$, tj. testování úseku a směrnice přímky.

2. Základní analýza dat

Název sloupce:	pl_1	pl_2
Průměr:	0,3482500875	0,2841210175
Spodní mez:	0,2853417874	0,1990382957
Horní mez:	0,4111583875	0,3692037393
Rozptyl:	0,01703529372	0,03871205638

Směr. odchylka:	0,1305193232	0,1967537964
Šikmost:	-0,1369586018	0,1960590943
Odchylka od 0:	Nevýznamná	Nevýznamná
Špičatost:	1,966948359	1,799552174
Odchylka od 3:	Nevýznamná	Nevýznamná
Polosuma:	0,3397727275	0,338636364
Modus:	0,3637559037	0,3327911208
Medián:	0,353787879	0,301298701
IS spodní:	0,249466871	0,1273505245
IS horní:	0,458108887	0,4752468775
Med. směr. odchylka:	0,0496548687	0,08387599957
Medianový rozptyl:	0,002465605985	0,007035183303

Homogenita i normalita dat přijata

3a. Odhadování parametrů Plocha_1

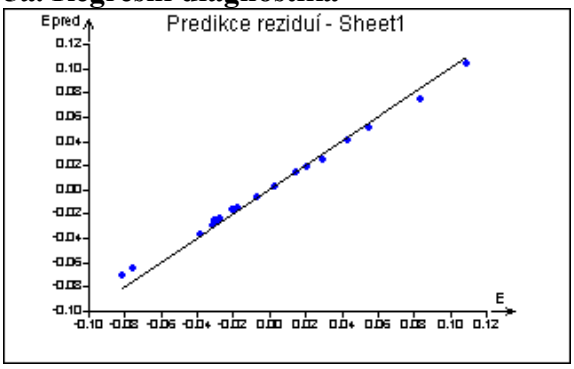
Proměnná	Odhad	Směr.Odch.	Závěr	Pravděpodobnost	Spodní mez	Horní mez
Abs	-0,12472	0,04297	Významný	0,00991	-0,21537	-0,03406
A	0,02145	0,00189	Významný	2,33864E-009	0,01746	0,02543

Metodou nejmenších čtverců byly nalezeny nejlepší odhady úseku β_0 a směrnice β_1. Studentův t-test ukazuje, že absolutní člen (β_0) a směrnice (β_1) jsou statisticky významné.

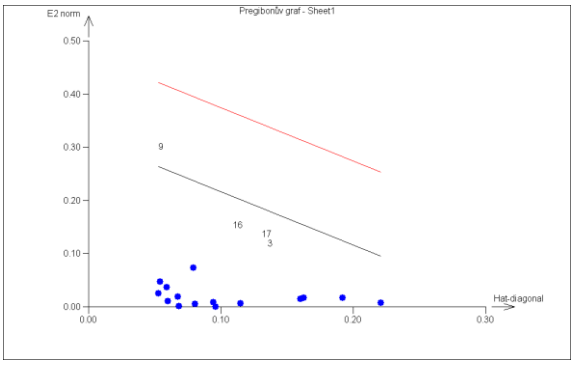
4a. Základní statistické charakteristiky

Vícenásobný korelační koeficient R :	0,9399312648
Koeficient determinace R^2 :	0,8834707826
Predikovaný korelační koeficient Rp :	0,7345757673
Střední kvadratická chyba predikce MEP :	0,002306640436
Akaikeho informační kritérium :	-115,2468206

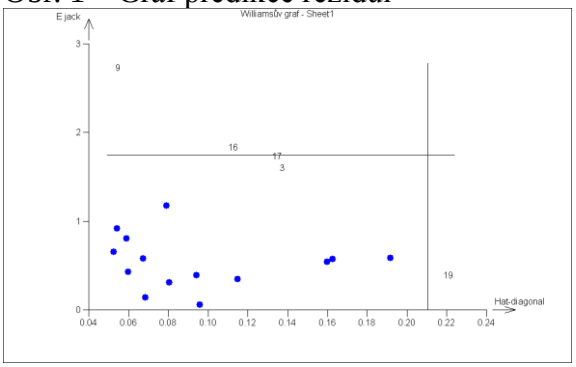
5a. Regresní diagnostika

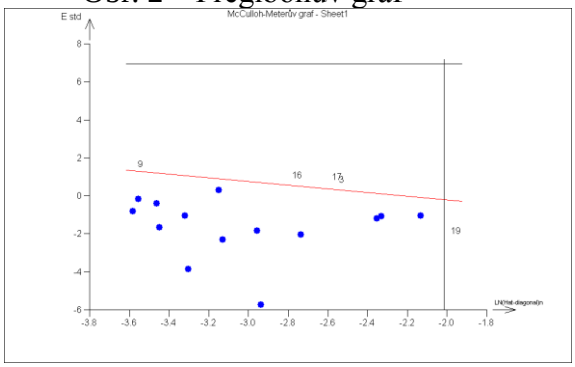


Obr. 1 – Graf predikce reziduí



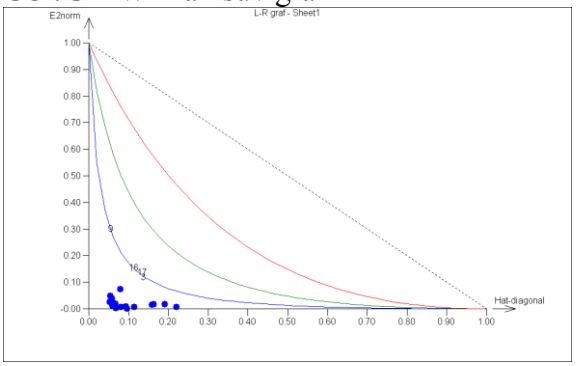
Obr. 2 – Pregibonův graf





Obr. 3 – Williamsův graf

Obr. 4 – McCulloh-Meeterův graf



Obr. 5 – L-R graf

Jednotlivé grafy (obr. 1-5) predikují body 9, 16, 17 a 3 jako odlehlé, bod 19 lze považovat z většiny grafů za extrémní.

Testování regresního tripletu

Fisher-Snedecorův test významnosti modelu

Hodnota kritéria F :	128,8861595
Kvantil F (1-alfa, m-1, n-m) :	4,451321772
Pravděpodobnost :	2,338636625E-009
Závěr :	Model je významný

Cook-Weisbergův test heteroskedasticity

Hodnota kritéria CW :	0,2608736742
Kvantil Chi^2(1-alfa,1) :	3,841458829
Pravděpodobnost :	0,6095205634
Závěr :	Rezidua vykazují homoskedasticitu.

Jarque-Berrův test normality

Hodnota kritéria JB :	1,313590015
Kvantil Chi^2(1-alfa,2) :	5,991464547
Pravděpodobnost :	0,5185104966
Závěr :	Rezidua mají normální rozdělení.

Waldův test autokorelace

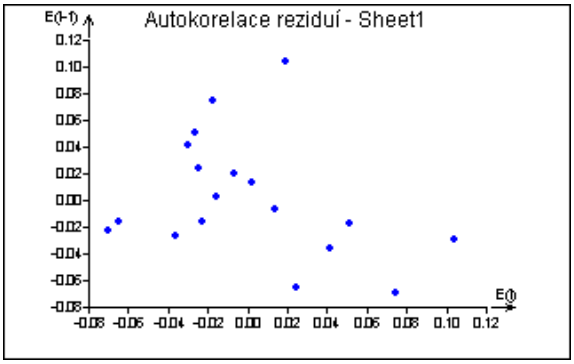
Hodnota kritéria WA :	1,729310882
Kvantil Chi^2(1-alfa,1) :	3,841458829
Pravděpodobnost :	0,1884989283
Závěr :	Autokorelace je nevýznamná.

Durbin-Watsonův test autokorelace

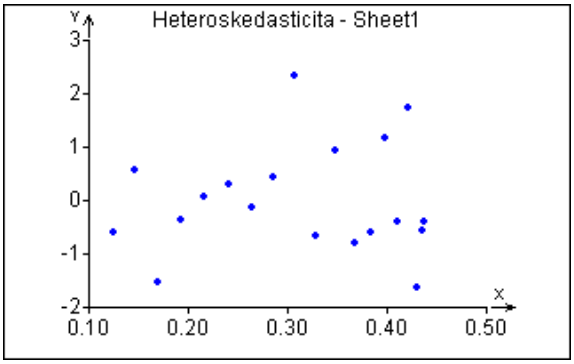
Hodnota kritéria DW :	-1
Kritické hodnoty DW	1,08 1,53
Závěr :	Negativní autokorelace reziduí není prokázána.

Znaménkový test reziduí

Hodnota kritéria Sg :	1,569658616
Kvantil N(1-alfa/2) :	1,959963999
Pravděpodobnost :	0,1164945539
Závěr :	V reziduích není trend.



Obr. 6 – Graf autokorelace reziduí



Obr. 7 – Graf heteroskedasticity

6a. Konstrukce zpřesněného modelu

Hodnoty přírůstů v tloušťkách byly vypočítány z původního souboru dat, např. řádky 9 a 3 navrhované jako odlehlé byly vypočítány jako střední hodnoty z několika stromů. Pouze řádky 16 a 17 je možné vyloučit, tyto jsou tvořeny pouze 1 stromem.

Proměnná	Odhad	Směr.odch.	Závěr	Pravděp.	Spodní mez	Horní mez
Abs	-0,1268	0,0406	Významný	0,0069	0,2132	-0,0403
D	0,0215	0,0019	Významný	6,3616E-009	0,0176	0,0255

Statistické charakteristiky (v závorce jsou původní hodnoty):

Vícenásobný korelační koeficient R :	0,9490544605	(0,9399312648)
Koeficient determinace R^2 :	0,9007043691	(0,8834707826)
Predikovaný korelační koeficient Rp :	0,771846858	(0,7345757673)
Střední kvadratická chyba predikce MEP :	0,001822138	(0,002306640436)
Akaikeho informační kritérium :	-106,6557374	(-115,2468206)

Vyloučením řádků 16 a 17 došlo k dílčímu zpřesnění modelu, snížily se hodnoty kritérií MEP a AIC a současně zvýšily parametry r a R^2. Předpoklady metody nejmenších čtverců jsou splněny.

Testování regresního tripletu

Fisher-Snedecorův test významnosti modelu

Hodnota kritéria F :	136,0640484
Kvantil F (1-alfa, m-1, n-m) :	4,543077165
Pravděpodobnost :	6,361639049E-009
Závěr :	Model je významný

Cook-Weisbergův test heteroskedasticity

Hodnota kritéria CW :	0,03235399191
Kvantil Chi^2(1-alfa,1) :	3,841458829
Pravděpodobnost :	0,8572529362
Závěr :	Rezidua vykazují homoskedasticitu.

Jarque-Berrův test normality

Hodnota kritéria JB :	3,299986634
Kvantil Chi^2(1-alfa,2) :	5,991464547
Pravděpodobnost :	0,1920511921

Závěr :	Rezidua mají normální rozdělení.

Waldův test autokorelace

Hodnota kritéria WA :	0,781821757
Kvantil Chi^2(1-alfa,1) :	3,841458829
Pravděpodobnost :	0,3765845602
Závěr :	Autokorelace je nevýznamná.

Durbin-Watsonův test autokorelace

Hodnota kritéria DW :	-1
Kritické hodnoty DW	1,02 1,54
Závěr :	Rezidua nejsou autokorelována.

Znaménkový test reziduí

Hodnota kritéria Sg :	1,173556889
Kvantil N(1-alfa/2) :	1,959963999
Pravděpodobnost :	0,2405725575
Závěr :	V reziduích není trend.

7a. Zhodnocení kvality modelu

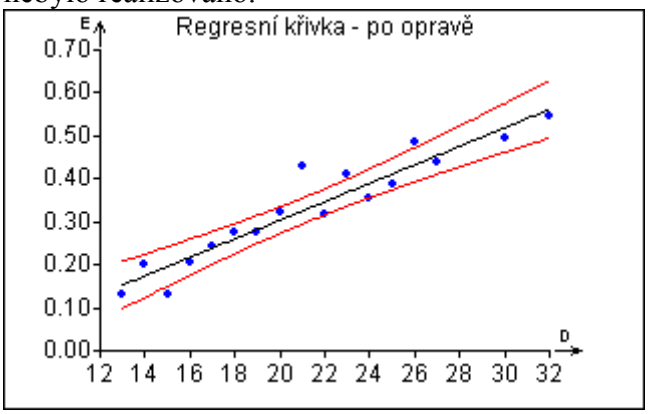
Nalezený regresní model má tvar

$$y = -0{,}1268\ (0{,}0406) + 0{,}0215\ (0.0019)x$$

Intervalový odhad parametrů úseku (β_0) a směrnice (β_1) je:

Proměnná	Spodní mez	Horní mez
β_0	-0,2131844689	-0,04032579674
β_1	0,01759691184	0,02546560857

Graf regresní křivky po opravě nadále naznačuje odlehlé body (tloušťky 15, 21, 23 a 26 cm), tyto hodnoty však byly vypočteny jako střední průměr z více stromů. Jejich další vyloučení nebylo realizováno.



Obr. 7 Regresní křivka zpřesněného modelu (po vyloučení 2 odlehlých bodů)

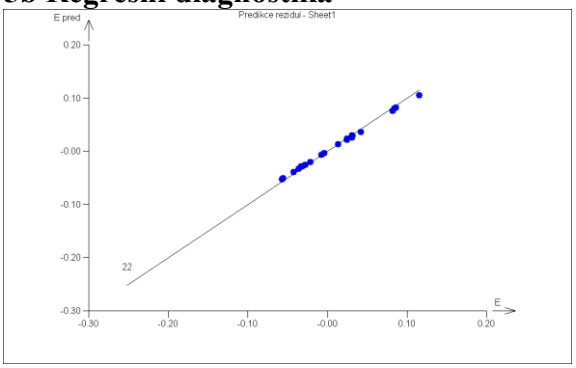
3b – Odhad parametrů – plocha 2

Proměnná	Odhad	Směr.Odch.	Závěr	Pravděpodobnost	Spodní mez	Horní mez
Abs	-0,2533	0,0433	Významný	8,3489E-006	-0,3434	-0,1633
E	0,02676	0,0020	Významný	1,4295E-011	0,0225	0,0310

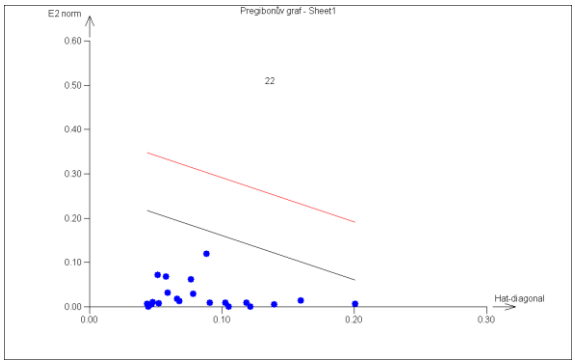
4b Základní statistické výpočty

Vícenásobný korelační koeficient R :	0,9438804212
Koeficient determinace R^2 :	0,8909102495
Predikovaný korelační koeficient Rp :	0,7441654341
Střední kvdratická chyba predikce MEP :	0,005085913808
Akaikeho informační kritérium :	-122,7677267

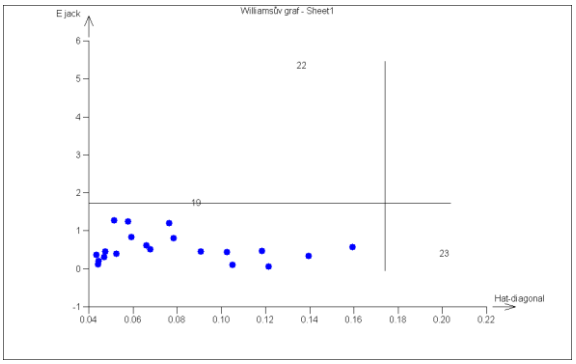
5b Regresní diagnostika



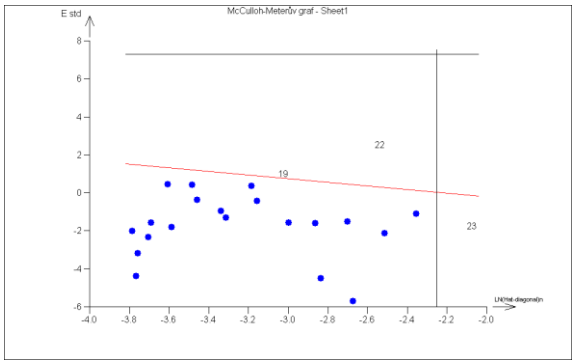
Obr. 8 – Graf predikce reziduí



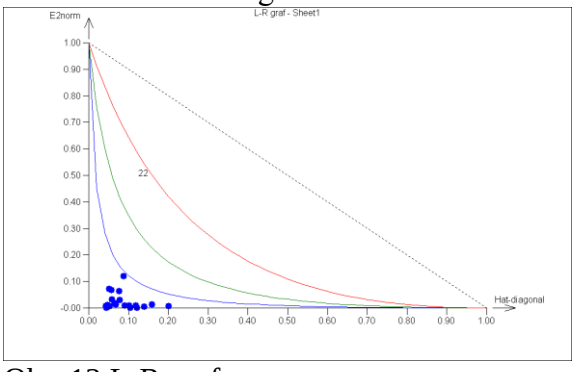
Obr. 9 – Pregibonův graf



Obr. 10 Wiliamsův graf



Obr. 11 McCuloh-Meeterův graf



Obr. 12 L-R graf

Jednotlivé grafy vlivných bodů (obr. 8-12) predikují bod 22 jako výrazně odlehlý, bod 19 je indikován grafy na hranici odlehlosti. Bod 23 je naznačován jako bod extrémní. Hodnota bodu 22 vznikla jako střední hodnota ze 2 stromů, při dalším výpočtu bude bod vyloučen.

Testování regresního tripletu

Fisher-Snedecorův test významnosti modelu

Hodnota kritéria F :	171,5020444
Kvantil F (1-alfa, m-1, n-m) :	4,324793743
Pravděpodobnost :	1,429525141E-011
Závěr :	Model je významný

Cook-Weisbergův test heteroskedasticity

Hodnota kritéria CW :	8,444493729
Kvantil Chi^2(1-alfa,1) :	3,841458829
Pravděpodobnost :	0,003661503544
Závěr :	Rezidua vykazují heteroskedasticitu!

Jarque-Berrův test normality

Hodnota kritéria JB :	19,93414218
Kvantil Chi^2(1-alfa,2) :	5,991464547
Pravděpodobnost :	4,691978616E-005
Závěr :	Rezidua nemají normální rozdělení!

Waldův test autokorelace

Hodnota kritéria WA :	0,7157524089
Kvantil Chi^2(1-alfa,1) :	3,841458829
Pravděpodobnost :	0,3975407404
Závěr :	Autokorelace je nevýznamná.

Durbin-Watsonův test autokorelace

Hodnota kritéria DW :	-1
Kritické hodnoty DW	1,17 1,54
Závěr :	Rezidua nejsou autokorelována.

Znaménkový test reziduí

Hodnota kritéria Sg :	1,273771206
Kvantil N(1-alfa/2) :	1,959963999
Pravděpodobnost :	0,2027445127
Závěr :	V reziduích není trend.

6b Konstrukce zpřesněného modelu

Proměnná	Odhad	Směr.Odch.	Závěr	Pravděpodobnost	Spodní mez	Horní mez
Abs	-0,2897	0,0292	Významný	3,6787E-009	-0,3507	-0,2287
G	0,0291	0,0014	Významný	5,9952E-015	0,0262	0,0321

Metodou nejmenších čtverců byly nalezeny nejlepší odhady úseku β_0 a směrnice β_1. Studentův t-test ukazuje, že absolutní člen (β_0) a směrnice (β_1) jsou statisticky významné.

Statistické charakteristiky (v závorce jsou původní hodnoty):

Vícenásobný korelační koeficient R :	0,9772997133	(0,9438804212)
Koeficient determinace R^2 :	0,9551147296	(0,8909102495)
Predikovaný korelační koeficient Rp :	0,894120789	(0,7441654341)
Střední kvdratická chyba predikce MEP :	0,002100841994	(0,005085913808)
Akaikeho informační kritérium :	-135,8770202	(-122,7677267)

Zpřesnění modelu bylo úspěšné – snížily se rozhodující kritéria *MEP* a *AIC* a zvýšily se parametry *r*, R^2, předpoklady metody nejmenších čtverců jsou splněny.

Testování regresního tripletu

Fisher-Snedecorův test významnosti modelu

Hodnota kritéria F :	425,5804722
Kvantil F (1-alfa, m-1, n-m) :	4,351243503
Pravděpodobnost :	5,971535855E-015
Závěr :	Model je významný

Cook-Weisbergův test heteroskedasticity

Hodnota kritéria CW :	0,9035106234
Kvantil Chi^2(1-alfa,1) :	3,841458829
Pravděpodobnost :	0,3418421249
Závěr :	Rezidua vykazují homoskedasticitu.

Jarque-Berrův test normality

Hodnota kritéria JB :	1,23155097
Kvantil Chi^2(1-alfa,2) :	5,991464547
Pravděpodobnost :	0,540221799
Závěr :	Rezidua mají normální rozdělení.

Waldův test autokorelace

Hodnota kritéria WA :	0,8044115833
Kvantil Chi^2(1-alfa,1) :	3,841458829
Pravděpodobnost :	0,3697776431
Závěr :	Autokorelace je nevýznamná

Durbin-Watsonův test autokorelace

Hodnota kritéria DW :	-1
Kritické hodnoty DW	1,15 1,54
Závěr :	Pozitivní autokorelace reziduí není prokázána.

Znaménkový test reziduí

Hodnota kritéria Sg :	1,529260071
Kvantil N(1-alfa/2) :	1,959963999
Pravděpodobnost :	0,1261999841
Závěr :	V reziduích není trend.

7b. Zhodnocení kvality modelu

Nalezený regresní model má tvar

$$y = -0{,}290\ (0{,}029) + 0{,}029\ (0.0014)x$$

Intervalový odhad parametrů úseku (β_0) a směrnice (β_1) je:

Proměnná	Spodní mez	Horní mez
β_0	-0,3506865721	-0,2287317812
β_1	0,02616871934	0,03205612837

8. Porovnání regresních přímek

Plocha 1: y = -0,1268 (0,0406) + 0,0215 (0.0019)x
Reziduální součet čtverců: 0,02532539372
Reziduální směr. odchylka: 0,04108965297
β_0 (-0,2131844689 -0,04032579674)
β_1 (0,01759691184 0,02546560857)

Plocha 2: y = -0,290 (0,029) + 0,029 (0,0014)x
Reziduální součet čtverců: 0,03812046819
Reziduální směr. odchylka: 0,04365802801
β_0 (-0,3506865721 -0,2287317812)
β_1 (0,02616871934 0,03205612837)

Plocha 1+2 y = 0,238 (0,026) + 0,027 (0,001)x
Reziduální součet čtverců: 0,08252821028
Reziduální směr. odchylka: 0,287277
β_0 (-0,2905919203 -0,1847404387)
β_1 (0,02410040821 0,02907698512

POROVNÁNÍ DVOU LINEÁRNÍCH ZÁVISLOSTÍ

1. *Test shody rozptylů*
   Testační statistika F = max (σ_1^2, σ_2^2) / min (σ_1^2, σ_2^2) =
   = 0,03812046819/ 0,02532539372= 1,505227

Tabulková hodnota $F_{0.95}$ (22, 17) = 2,20
Testační statistika je menší než tabulková hodnota F-rozdělení, tj. **rozptyly jsou shodné**.

2. *Chowův test*

Byla testována hypotéza H_0: $\beta_1 = \beta_2$ oproti H_A: $\beta_1 \neq \beta_2$
Testační kritérium testu:

F_c = [(RSC – RSC1 – RSC2).(n – 2m)] / [(RSC1 + RSC2).(m)]

Kde RSC……..reziduální součet čtverců pro složený model
RSCi…….reziduální součet čtverců pro i-tý model
n = n1 + n2
m = 2 (pro lineární závislost)

V prvním kroku byly zjištěny parametry složeného modelu

Do vzorce pro Chowův test bylo tedy dosazeno:

RSC1 = 0,0253
RSC2 = 0,0381
RSC = 0,0825
n = 17 + 22 = 39
m = 2

F_c = 5,272

Tabulková hodnota $F_{0.95}$ (2, 35) = 19,465

Hodnota testačního kritéria F_c je menší než tabulkový kvantil F-rozdělení, tj. **lineární závislosti jsou shodné**.

3. Závěr

Závěr: H_0 je přijata, regresní přímky mají statisticky shodné směrnice.

Licenční studium Statistické zpracování experimentálních dat.
Předmět: 2.1 Tvorba lineárních regresních modelů při analýze dat
Přednášející: Prof. RNDr. Milan Meloun, DrSc.
Úloha 2. Určení stupně polynomu

Zadání: Určete stupeň polynomu pro proložení křivky metodou nejmenších čtverců pro vztah mezi objemovým přírůstem (m^3/ha/rok) a věkem smrkového porostu 1 výškové bonity (34 m).

Vstupní data:

	2	2	3	3	4	4	5	5	6	6	7	7	8	8	9	9	10	10	11	11	12	12	13
Věk	0	5	0	5	0	5	0	5	0	5	0	5	0	5	0	5	0	5	0	5	0	5	0
Obj.	1	2	2	2	2	2	2	1	1	1	1	1	1	1	1	1							
přírůst	9	3	3	3	2	1	0	9	8	6	5	5	3	2	1	1	10	10	8	8	8	7	7

Data: prirust.vts

Program: QCExpert, Lineární regrese

Řešení:

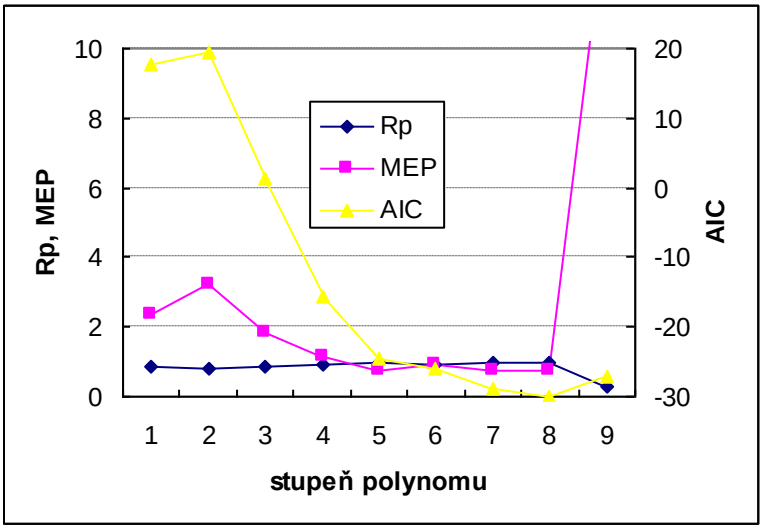
Návrh modelu

Určení stupně polygonu

Pro nalezení stupně polygonu se používají hodnoty predikovaného koeficientu determinace Rp, hodnota Střední kvadratické chyby predikce MEP a Akaikeho informační kriterium (AIK). Hodnoty MEP a AIK při dosažení optimálního stupně polygonu nabývají minimálních hodnot, hodnota Rp hodnot maximálních.

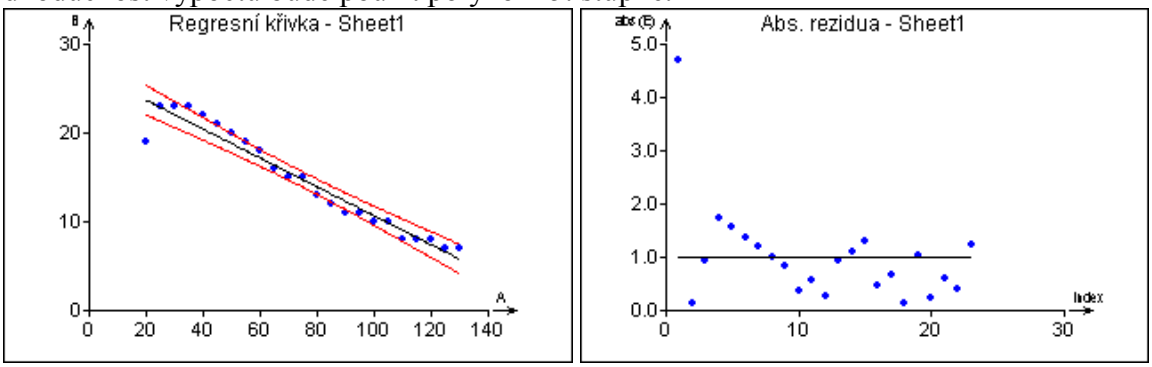
Tab. 1 Hodnoty statistických charakteristik regrese pro různé stupně polynomů

	Stupeň polynomu								
	1	2	3	4	5	6	7	8	9
Predikovaný korel. koeficient Rp :	0,852	0,803	0,88 3	0,929	0,953	0,943	0,952	0,951	0,291
Střední kvadr. chyba predikce MEP :	2,385	3,210	1,86 4	1,122	0,739	0,900	0,761	0,775	14,279
Akaikeho inform. kritérium (AIC):	17,58 0	19,52 7	1,42 0	- 15,561	- 24,607	- 26,082	- 28,977	- 29,969	- 27,166

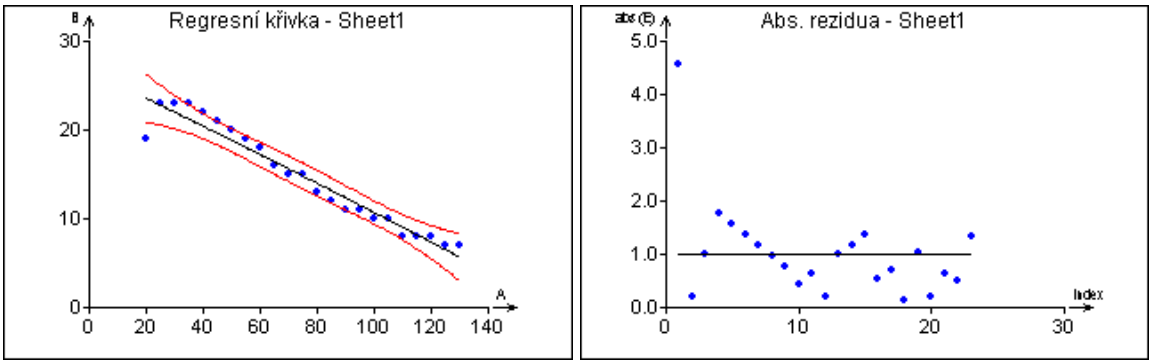


Obr. 1 Průběh statistických charakteristik regrese pro různé stupně polynomů

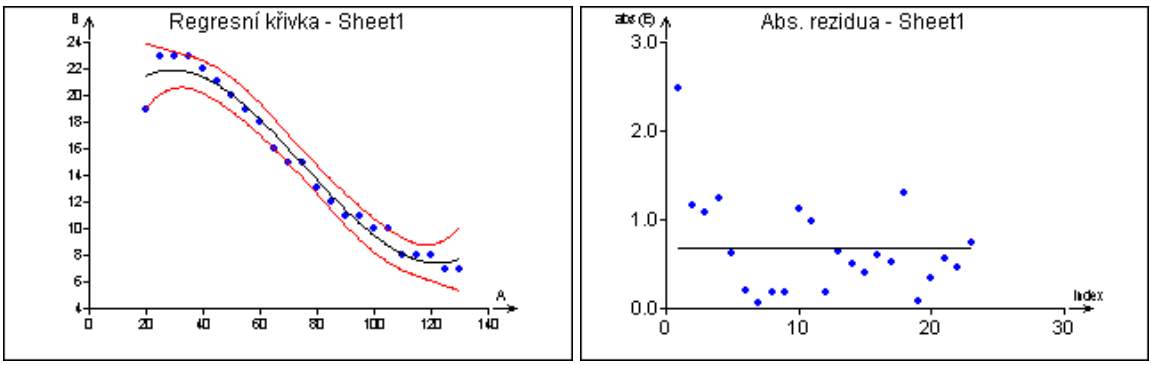
Z tabulky i obrázku je patrné, že charakteristiky Rp a MEP dosahují svých extrémů pro stupeň polynomu m = 5, charakteristika AIK má své minimum až u stupně polynomu 8. Pro jednoduchost výpočtu bude použit polynom 5. stupně.



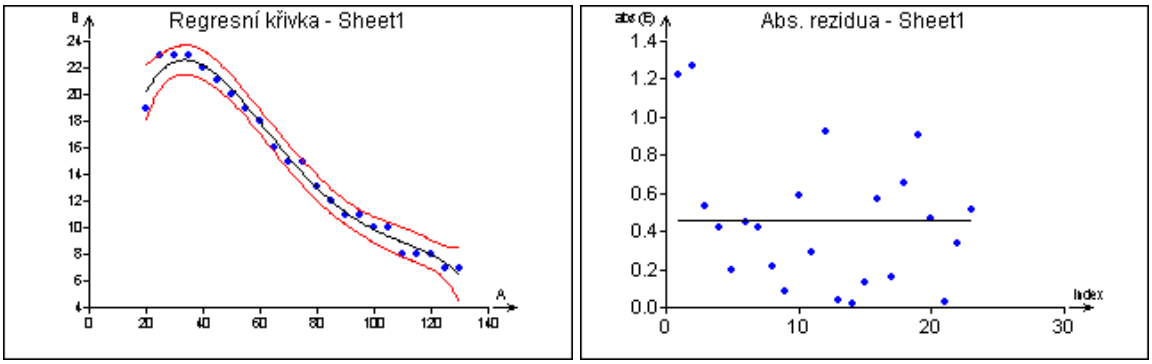
Obr. 2 Těsnost proložení a predikce reziduí pro polynom 1. stupně



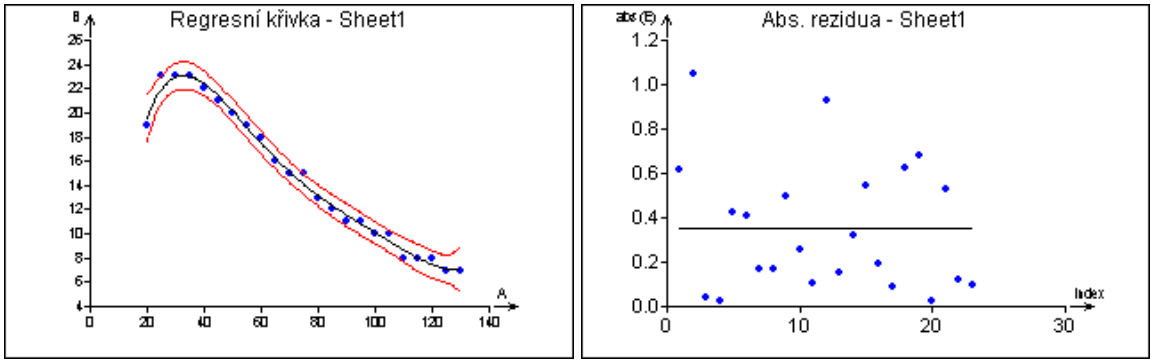
Obr. 3 Těsnost proložení a predikce reziduí pro polynom 2. stupně



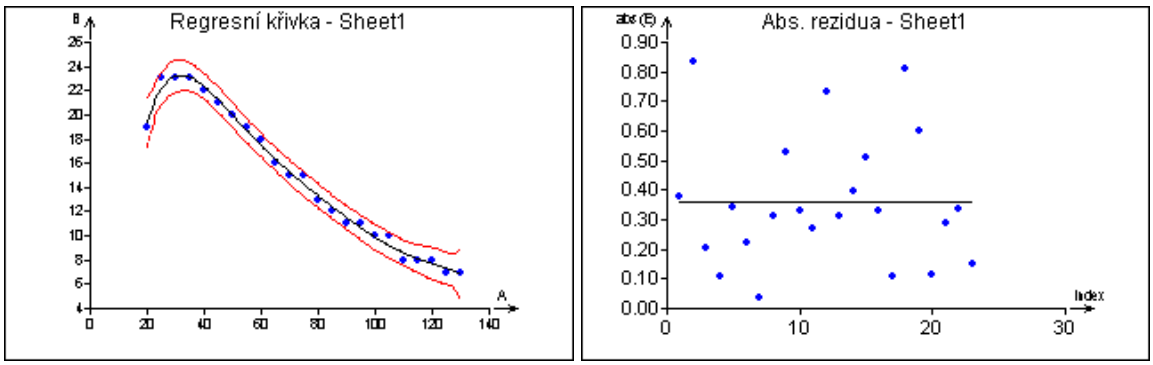
Obr. 4 Těsnost proložení a predikce reziduí pro polynom 3. stupně



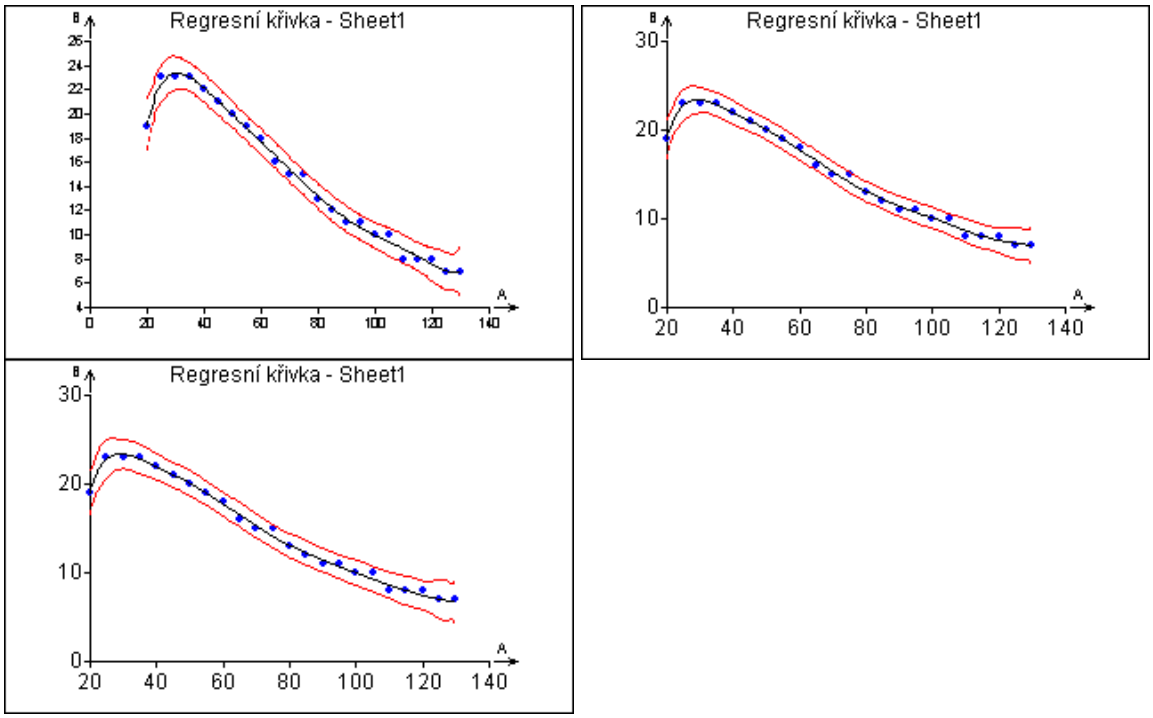
Obr. 5 Těsnost proložení a predikce reziduí pro polynom 4. stupně



Obr. 6 Těsnost proložení a predikce reziduí pro polynom 5. stupně



Obr. 7 Těsnost proložení a predikce reziduí pro polynom 6. stupně



Obr. 8 Těsnost proložení pro polynom 7., 8., a 9. stupně

Předběžná analýza dat

Základní analýza

Proměnná	Průměr	Směr.Odch.	Kor.vs.Y	Významnost
A	75	33,912	-0,970	1,932E-014
A^2	6725	5184,713	-0,954	1,722E-012
A^3	669375	678530,19	-0,910	1,711E-009
A^4	70938125	86020748,97	-0,860	1,474E-007
A^5	7828359375	1,08E+010	-0,811	2,626E-006

Párové korelace(Xi, Xj)

A - A^2	**0,981**	2,220E-016
A - A^3	**0,942**	1,973E-011
A - A^4	**0,899**	5,682E-009
A - A^5	**0,857**	1,727E-007
A^2 - A^3	**0,989**	0
A^2 - A^4	**0,964**	1,339E-013
A^2 - A^5	**0,936**	5,784E-011
A^3 - A^4	**0,993**	0
A^3 - A^5	**0,977**	1,332E-015
A^4 - A^5	**0,995**	0

Indikace multikolinearity

Proměnná	Vlas. čísla kor. m.	Podmíněnost kappa	VI faktor	Vícenás. kor.
Abs	1,942E-07	1,000E+00	**1,000E+00**	0,000E+00
A	1,000E+00	**5,150E+06**	**2,395E+04**	1,000E+00
A^2	3,930E-03	**2,024E+04**	**6,108E+05**	1,000E+00
A^3	1,813E-01	**9,340E+05**	**2,402E+06**	1,000E+00
A^4	4,407E-05	2,270E+02	**1,921E+06**	1,000E+00
A^5	4,815E+00	**2,480E+07**	**2,155E+05**	1,000E+00

Hodnoty Vl. faktoru výrazně převyšují hodnotu 10, to potvrzuje silnou multikolinearitu dat.

Odhady parametrů

Proměnná	Odhad	Směr.Odch.	Závěr	Pravděpodobnost	Spodní mez	Horní mez
Abs	-1,272E+01	5,578E+00	Významný	3,581E-02	-2,449E+01	-9,471E-01
A	2,947E+00	5,106E-01	Významný	2,261E-05	1,869E+00	4,024E+00
A^2	-8,585E-02	1,687E-02	Významný	9,080E-05	-1,214E-01	-5,027E-02
A^3	1,094E-03	2,556E-04	Významný	5,067E-04	5,546E-04	1,633E-03
A^4	-6,623E-06	1,803E-06	Významný	1,883E-03	-1,043E-05	-2,819E-06
A^5	1,552E-08	4,794E-09	Významný	4,841E-03	5,406E-09	2,564E-08

Klasickou metodou nejmenších čtverců byly nalezeny nejlepší odhady jednotlivých proměnných. Studentův t-test prokázal statistickou významnost všech proměnných.

Základní statistické charakteristiky

Vícenásobný korelační koeficient R :	0,996708206
Koeficient determinace R^2 :	0,9934272479
Predikovaný korelační koeficient Rp :	0,9528505999

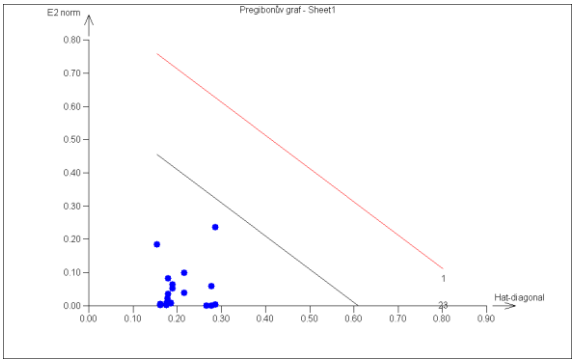
Střední kvdratická chyba predikce MEP : 0,7390530167
Akaikeho informační kritérium : -24,60745508
Regresní diagnostika

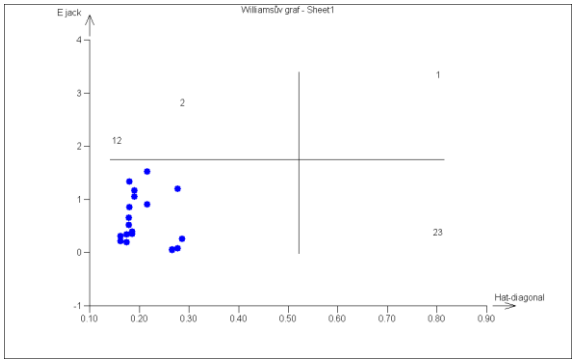
Analýza klasických reziduí

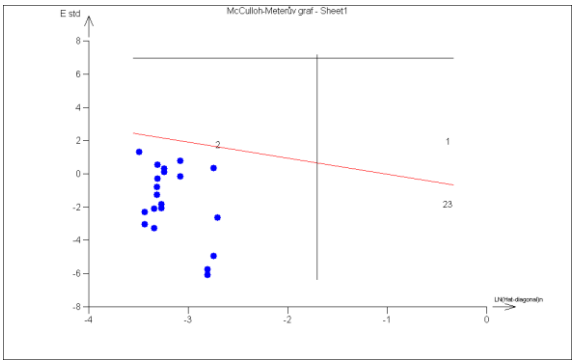
Index	Y naměřené	Y vypočítané	Směr. odch. Y	Reziduum	Reziduum [%Y]
1	19	19,617	0,470	-0,617	-3,248
2	23	21,948	0,281	1,052	4,572
3	23	22,962	0,277	0,038	0,164
4	23	23,021	0,271	-0,021	-0,093
5	22	22,424	0,244	-0,424	-1,928
6	21	21,410	0,223	-0,410	-1,953
7	20	20,166	0,220	-0,166	-0,830
8	19	18,832	0,227	0,168	0,886
9	18	17,506	0,229	0,494	2,744
10	16	16,252	0,222	-0,252	-1,578
11	15	15,105	0,211	-0,105	-0,698
12	15	14,073	0,206	0,927	6,178
13	13	13,150	0,211	-0,150	-1,156
14	12	12,316	0,222	-0,316	-2,635
15	11	11,545	0,229	-0,545	-4,955
16	11	10,811	0,227	0,189	1,722
17	10	10,092	0,220	-0,092	-0,919
18	10	9,379	0,223	0,621	6,206
19	8	8,680	0,244	-0,680	-8,506
20	8	8,025	0,271	-0,025	-0,318
21	8	7,473	0,277	0,527	6,583
22	7	7,118	0,281	-0,118	-1,682
23	7	7,092	0,470	-0,092	-1,320

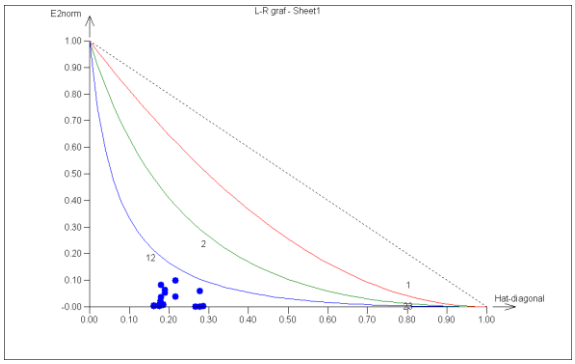
Reziduální součet čtverců : 4,682657236
Průměr absolutních reziduí : 0,3491465806
Reziduální směr. odchylka : 0,5248337124
Reziduální rozptyl : 0,2754504256
Šikmost reziduí : 0,5445418783
Špičatost reziduí : 2,907428813

Odhady šikmosti a špičatosti reziduí se blíží hodnotám pro normální rozdělení.







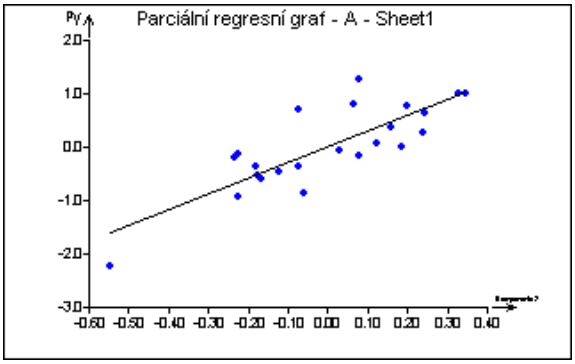


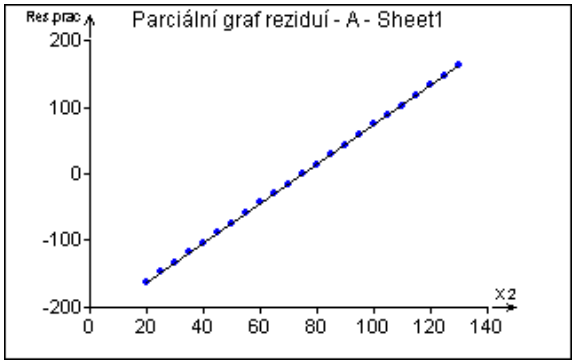
Jednotlivé grafy indikují body 1 a 23 jako extrémní, body 2 a 12 jako odlehlé. Extremita krajového bodu 1 (věk 20 let) je dána průběhem křivky. Objemový přírůst porostu po dosažení odpovídajících dimenzí středních stromů pro výpočet objemu se prudce zvyšuje. Zvyšování objemového přírůstu ovlivňuje postupné zapojování stromů do výpočtu i mimořádně vysoké hodnoty přírůstu v tomto věku. Přírůst však poměrně rychle dosahuje kulminace a pozvolna klesá. Extremitu hodnoty u bodu 23 (věk 130 let) lze vysvětlit středním věkem porostu. Navazující body 2 a 12 jsou považovány za odlehlé, protože jsou výrazně ovlivňovány extrémními hodnotami krajových bodů.

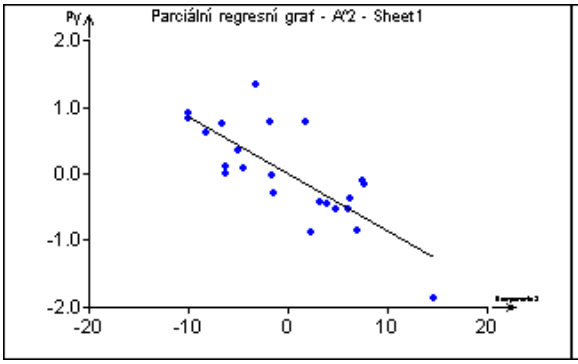
Model:

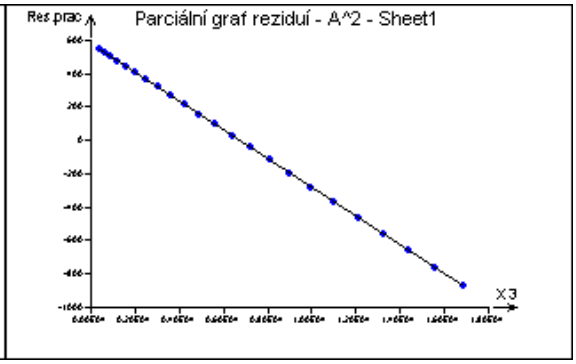
Parciální regresní i parciální reziduální grafy ukazují lineární závislost všech nezávisle proměnných. Navržený model má tvar:

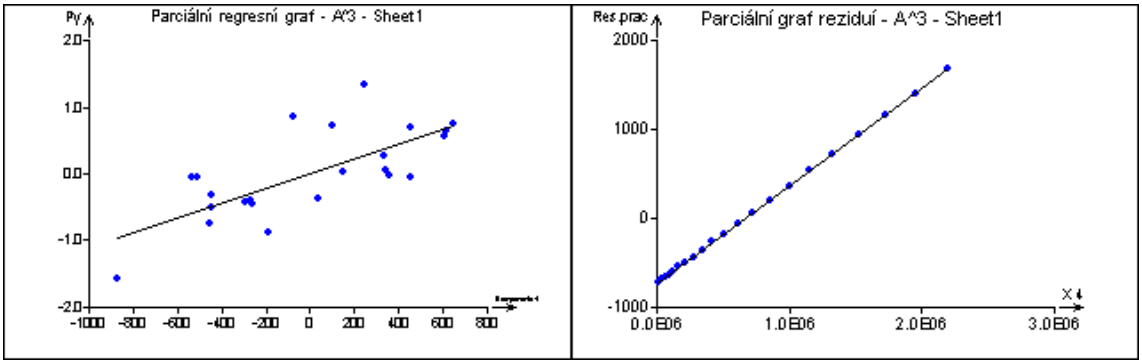
$$\beta 0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 x^4 + \beta_5 x^5$$

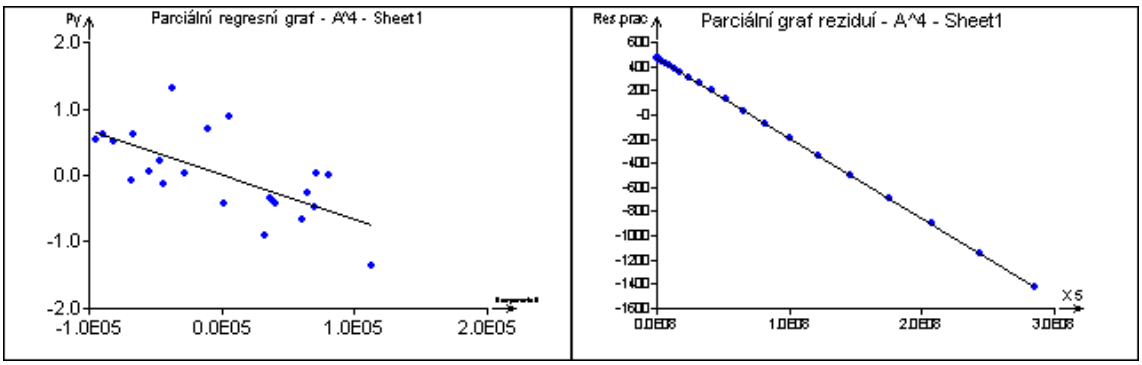


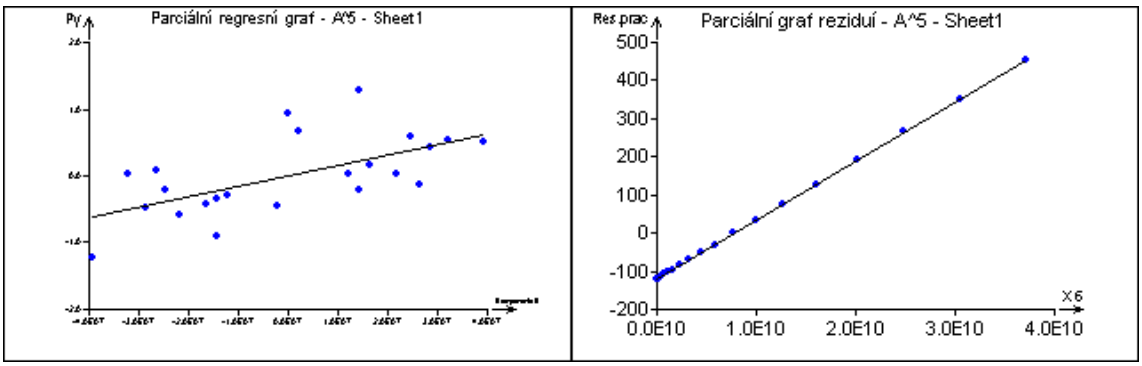












Testování regresního tripletu
Metoda:
Fisher-Snedecorův test významnosti modelu

Hodnota kritéria F :	513,8871169
Kvantil F (1-alfa, m-1, n-m) :	2,809996175
Pravděpodobnost :	6,405059932E-018
Závěr :	Model je významný

Scottovo kritérium multikolinearity

Hodnota kritéria SC :	0,9331050932
Závěr :	Model je nekorektní!

Cook-Weisbergův test heteroskedasticity

Hodnota kritéria CW :	0,1763607256
Kvantil Chi^2(1-alfa,1) :	3,841458829
Pravděpodobnost :	0,6745196118
Závěr :	Rezidua vykazují homoskedasticitu.

Jarque-Berrův test normality

Hodnota kritéria JB :	2,095622899
Kvantil Chi^2(1-alfa,2) :	5,991464547
Pravděpodobnost :	0,3507044443
Závěr :	Rezidua mají normální rozdělení.

Waldův test autokorelace

Hodnota kritéria WA :	0,7340241203
Kvantil Chi^2(1-alfa,1) :	3,841458829
Pravděpodobnost :	0,3915819162
Závěr :	Autokorelace je nevýznamná

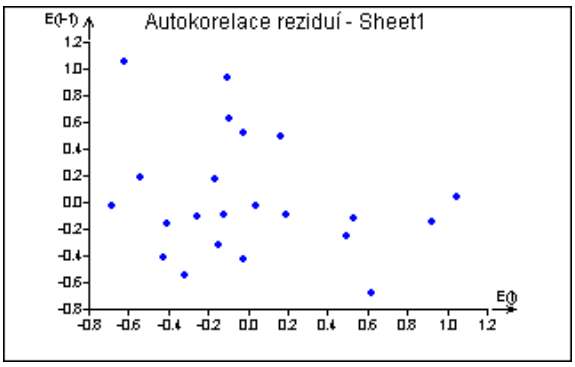
Durbin-Watsonův test autokorelace

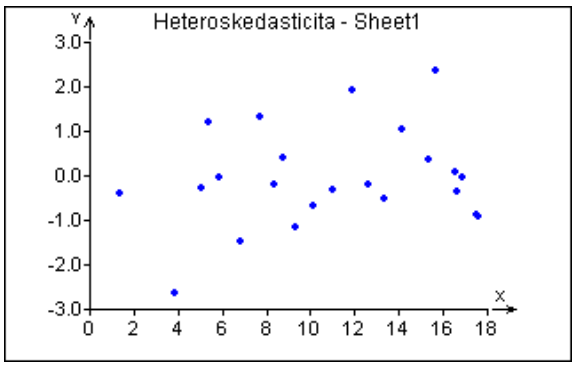
Hodnota kritéria DW :	-1
Kritické hodnoty DW	0,9 1,92
Závěr :	Negativní autokorelace reziduí není prokázána.

Znaménkový test reziduí

Hodnota kritéria Sg :	0,9762689171
Kvantil N(1-alfa/2) :	1,959963999
Pravděpodobnost :	0,3289312189
Závěr :	V reziduích není trend.

Všechny testy s výjimkou Scottova kritéria multikolinearity potvrdily správnost odhadů lineárních regresních parametrů. Graf autokorelace reziduí i graf heteroskedasticity vykazují náhodný mrak bodů, body netvoří výrazný klín.





Zhodnocení kvality modelu:

Nalezený model polynomické závislosti objemového přírůstu na věku pro smrkový porost výškové bonity 34 m má tvar:

$$\underline{\mathbf{Y= -12,\ 716\ (5,578) + 2,947\ (0,511)X - 0,086\ (0,017)X^2 + 0,001\ (2,566E\text{-}04)X^3 - 6,623\ (1,803E\text{-}06)X^4 + 1,552E\text{-}08\ (4,794E\text{-}09)X^5}}$$

Licenční studium Statistické zpracování experimentálních dat.
Předmět: 2.1 Tvorba lineárních regresních modelů při analýze dat
Přednášející: Prof. RNDr. Milan Meloun, DrSc.
Úloha 3. Validace nové analytické metody

Zadání:
Otestujte nový přístroj pro měření výšek stromů se standardním výškoměrem.
Data: validace.vts

1. Návrh modelu
Navržený regresní model přímky je $y = \beta_0 + \beta_1 x$. Bude testována nulová hypotéza H_0: $\beta_0 = 0, \beta_1 = 1$, tj. testování úseku a směrnice přímky.

2. Základní analýza dat

Název sloupce :	vyskomerA	vyskomerB
Průměr :	22,088	21,99
Spodní mez :	19,92587939	19,83315675
Horní mez :	24,25012061	24,14684325
Rozptyl :	57,87903673	57,59683673
Směr. odchylka :	7,607827333	7,589257983
Šikmost	0,4010482992	0,4139081882
Odchylka od 0 :	Nevýznamná	Nevýznamná
Špičatost :	1,787430299	1,809879168
Odchylka od 3 :	Nevýznamná	Nevýznamná
Polosuma	24,9	25,25
Modus :	13,21227451	13,2545098
Medián :	19,05	19
IS spodní :	14,53862579	14,64242263
IS horní :	23,56137421	23,35757737
Medianová směr. odchylka :	2,244939194	2,168407176
Medianový rozptyl :	5,039751986	4,701989682
Normalita :	Přijata	Přijata
Homogenita :	Přijata	Přijata
Počet vybočujících bodů :	0	0

Data jsou nezávislá, normalita i homogenita dat byla přijata

Charakteristika proměnných

Proměnná	Průměr;	Směr.Odch.	Kor.vs.Y	Významnost
VyskomerB	21,99	7,589	1,000	
vyskomerA	22,088	7,608	0,996	0

Odhad parametrů

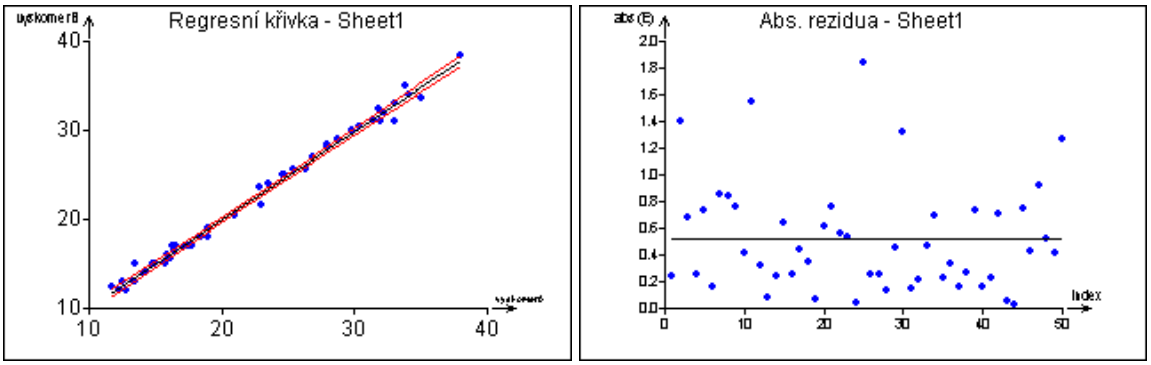
Proměnná	Odhad	Směr.Odch.	Závěr	Pravděpodobnost		Spodní mez	Horní mez
Abs	0,040	0,293	Nevýznamný	0,893	-	0,549	0,628
vyskomerA	0,994	0,013	Významný	0		0,969	1,019

Statistické charakteristiky regrese

Vícenásobný korelační koeficient R :	0,9961979923
Koeficient determinace R^2 :	0,9924104398

Predikovaný korelační koeficient Rp : 0,9833167095
Střední kvadratická chyba predikce MEP : 0,4728236906
Akaikeho informační kritérium : -38,38583464

Regresní diagnostika
Kritika dat



Obr. 1 Graf regresního modelu a analýza klasických reziduí

Graf regresního modelu ukazuje poměrně těsné proložení i několik odlehlých hodnot, většina hodnot v grafu reziduí tvoří shluk bodů, některé hodnoty budou odlehlé. Odhady šikmosti i špičatosti leží blízko hodnot odpovídající normalitě. Experimentální chyba měření původního výškoměru je 0,5 m, reziduální směrodatná odchylka je sice vyšší, ale blíží se experimentální chybě.

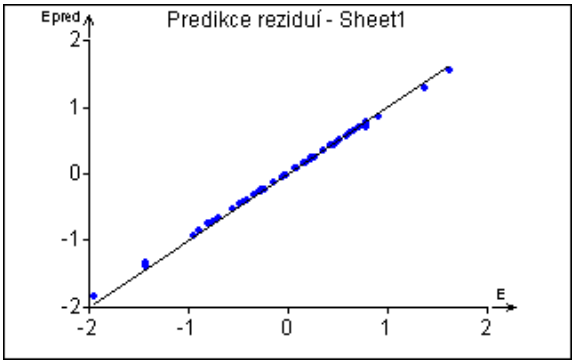
Reziduální součet čtverců : 21,41959833
Průměr absolutních reziduí : 0,5147225586
Reziduální směr. odchylka : 0,6680131973
Reziduální rozptyl : 0,4462416318
Šikmost reziduí : 0,1549549585
Špičatost reziduí : 3,412474538

Indikace vlivných dat

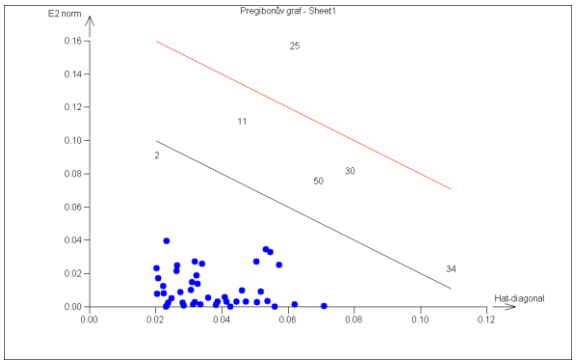
Index	Standard	Jackknife	Predik.	Diag(Hii)	Diag(H*ii)	Cookova vzdál.
11	2,367	2,492	1,619	0,046	**0,157**	0,057
25	-2,835	-3,074	-1,955	0,062	**0,219**	-0,094
30	-2,061	-2,136	-1,435	0,079	**0,160**	-0,088
34	1,106	1,109	0,783	**0,109**	**0,132**	0,068
50	1,973	2,037	1,366	0,069	**0,145**	0,073

Z celkových 50 měření bylo 5 měření indikováno jako odlehlé, tyto body jsou indikovány prvky projekční matice rozšířené o závisle proměnnou, 1 měření (index 34) je indikováno projekční maticí.

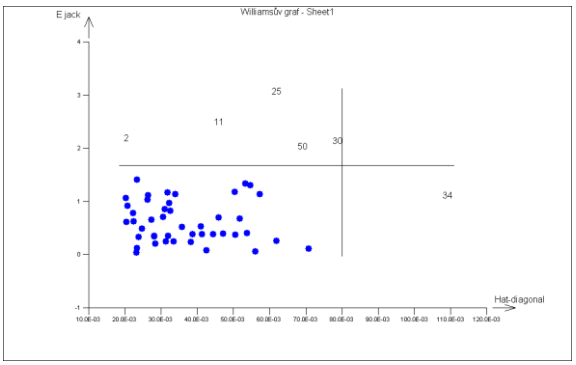
Jednotlivé grafy vlivných bodů indikují odlehlost bodů 25, 11, 50 a 30, bod 34 grafy diagnostikují jako odlehlý. Oproti předchozímu zjišťování vlivných bodů je nově indikován bod 2, ten je však indikován pouze Williamsovým grafem.



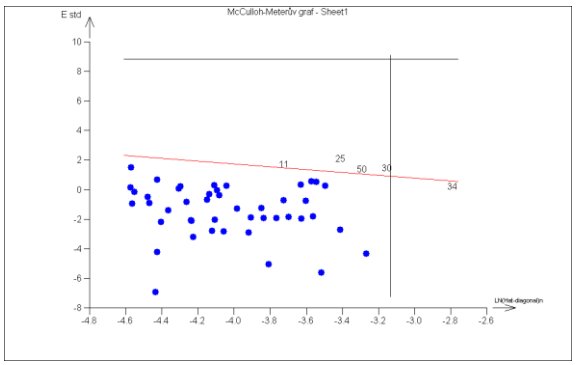
Obr. 2 Graf předikovaných reziduí



Obr. 3 Pregibonův graf



Obr. 4. Williamsův graf



Obr. 5 McCulloh-Meet graf

Testování regresního tripletu

Fisher-Snedecorův test významnosti modelu

Hodnota kritéria F :	6276,477142
Kvantil F (1-alfa, m-1, n-m) :	4,042652129
Pravděpodobnost :	1,53406667E-052
Závěr :	Model je významný

Cook-Weisbergův test heteroskedasticity

Hodnota kritéria CW :	2,810124205
Kvantil Chi^2(1-alfa,1) :	3,841458829
Pravděpodobnost :	0,09367112312
Závěr :	Rezidua vykazují homoskedasticitu.

Jarque-Berrův test normality

Hodnota kritéria JB :	1,645739746
Kvantil Chi^2(1-alfa,2) :	5,991464547
Pravděpodobnost :	0,4391694836
Závěr :	Rezidua mají normální rozdělení.

Waldův test autokorelace

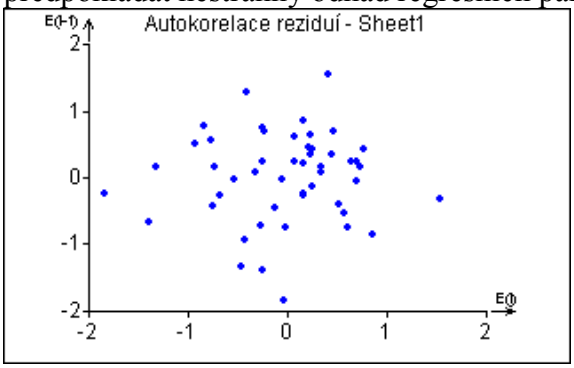
Hodnota kritéria WA :	0,1174696707
Kvantil Chi^2(1-alfa,1) :	3,841458829
Pravděpodobnost :	0,7317952058
Závěr :	Autokorelace je nevýznamná.

Durbin-Watsonův test autokorelace

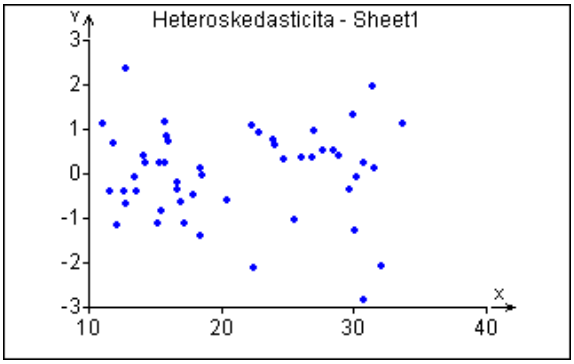
Hodnota kritéria DW :	-1
Kritické hodnoty DW	1,46 1,63
Závěr :	Rezidua nejsou autokorelována.

Znaménkový test reziduí	
Hodnota kritéria Sg :	1,536067766
Kvantil N(1-alfa/2) :	1,959963999
Pravděpodobnost :	0,1245217609
Závěr :	V reziduích není trend.

Jednotlivé testy potvrdily splnění základních předpokladů metody nejmenších čtverců, lze předpokládat nestranný odhad regresních parametrů.



Obr. 6 Graf autokorelace



Obr. 7 Graf heteroskedasticity

6. Konstrukce zpřesněného modelu

Z dat byly odstraněny silně odlehlé body č. 11, 25, 30 a 50 a byly určeny nové odhady parametrů zpřesněného modelu.

Odhad zpřesněného modelu

Proměnná	Odhad	Směr.Odch.	Závěr	Pravděpod.	Spodní mez	Horní mez
Abs	-0,282	0,243	Nevýznamný	0,252	-0,772	0,208
vyskomerA	1,009	0,011	Významný	0	0,987	1,031

Statistické charakteristiky regrese (v závorkách původní hodnoty)

Vícenásobný korelační koeficient R :	0,9975210287	(0,996197)
Koeficient determinace R^2 :	0,9950482028	(0,992410)
Predikovaný korelační koeficient Rp :	0,9892619014	(0,983317)
Střední kvadratická chyba predikce MEP :	0,2779723486	(0,472827)
Akaikeho informační kritérium :	-58,73614776	(-38,38584)

Testování regresního tripletu

Fisher-Snedecorův test významnosti modelu

Hodnota kritéria F :	8841,6627
Kvantil F (1-alfa, m-1, n-m) :	4,06170646
Pravděpodobnost :	2,309724015E-052
Závěr :	Model je významný

Cook-Weisbergův test heteroskedasticity

Hodnota kritéria CW :	0,0002572251139
Kvantil Chi^2(1-alfa,1) :	3,841458829
Pravděpodobnost :	0,9872038852
Závěr :	Rezidua vykazují homoskedasticitu.

Jarque-Berrův test normality

Hodnota kritéria JB :	2,154567559
Kvantil Chi^2(1-alfa,2) :	5,991464547
Pravděpodobnost :	0,3405191958
Závěr :	Rezidua mají normální rozdělení.

Waldův test autokorelace

Hodnota kritéria WA :	0,0005161840938
Kvantil Chi^2(1-alfa,1) :	3,841458829
Pravděpodobnost :	0,9818738734
Závěr :	Autokorelace je nevýznamná

Durbin-Watsonův test autokorelace

Hodnota kritéria DW :	-1
Kritické hodnoty DW	1,43 1,62
Závěr :	Rezidua nejsou autokorelována.

Znaménkový test reziduí

Hodnota kritéria Sg :	0,09800224174
Kvantil N(1-alfa/2) :	1,959963999
Pravděpodobnost :	0,9219305133
Závěr :	V reziduích není trend.

Opravený model má tvar

$$\mathbf{Y = -0,282\ (0,243) + 1,009\ (0,011)X}$$

Intervalový odhad parametrů úseku (abs, β_0) a směrnice (β_1):

Proměnná	Spodní mez	Horní mez
β_0	-0,772	0,208
β_1	0,987	1.031

Interval spolehlivosti úseku regresní přímky obsahuje nulu, lze tedy tento úsek považovat za nulový. Interval spolehlivosti směrnice obsahuje jedničku, směrnici lze tedy považovat za jednotkovou.

Závěr:

Úsek regresní přímky lze považovat za nulový, směrnice není významně odlišná od jedničky. Výsledky měření výšek novým typem výškoměru se statisticky významně neliší od standardního postupu měření výšek.

Licenční studium Statistické zpracování experimentálních dat.
Předmět: 2.1 Tvorba lineárních regresních modelů při analýze dat
Přednášející: Prof. RNDr. Milan Meloun, DrSc.
Úloha 4. Vícerozměrný lineární regresní model

Zadání: Zjistěte vztah mezi středními měsíčními teplotami a nadmořskou výškou jednotlivých klimatických stanic na území Německa.

Řešení:
Návrh modelu
Navržený regresní model přímky je
$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8 + \beta_9 x_9 + \beta_{10} x_{10} + \beta_{11} x_{11} + \beta_{12} x_{12}$

y = nadmořská výška
x1= střední teplota v lednu
x2 = střední teplota v únoru
x3 = střední teplota v březnu
x4 = střední teplota v dubnu
x5 = střední teplota v květnu
x6 = střední teplota v červnu
x7 = střední teplota v červenci
x8 = střední teplota v srpnu
x9 = střední teplota v září
x10 = střední teplota v říjnu
x11 = střední teplota v listopadu
x12 = střední teplota v prosinci

Základní analýza dat

Proměnná	Průměr	Směr.Odch.	Kor.vs.Y	Významnost
leden	-0,676	1,157	-0,715	0,000
unor	0,267	1,025	-0,640	0,000
brezen	3,416	1,075	-0,536	0,000
duben	7,222	1,166	-0,466	0,001
kveten	11,953	1,179	-0,546	0,000
cerven	15,171	1,136	-0,592	0,000
cervenec	16,837	1,139	-0,439	0,002
srpen	16,357	1,094	-0,594	0,000
zari	13,198	0,933	-0,591	0,000
rijen	8,831	0,876	-0,768	0,000
listopad	3,802	1,062	-0,865	0,000
prosinec	0,584	1,180	-0,824	0,000

Pearsonovy párové korelační koeficienty závislostí ukazují korelace jednotlivých proměnných. Velmi silný lineární vztah existuje mezi sousedními měsíci, v letních měsících je silný lineární vztah mezi vyšším počtem měsíců.

Párové korelační koeficienty

	leden	unor	brezen	duben	kveten	cerven	cervenec	srpen	zari	rijen	listopad
unor	0,927										
brezen	0,735	0,916									
duben	0,500	0,744	0,938								
kveten	0,467	0,687	0,886	0,974							
cerven	0,452	0,654	0,846	0,940	0,977						
cervenec	0,390	0,633	0,840	0,936	0,941	0,968					
srpen	0,503	0,695	0,852	0,918	0,936	0,969	0,969				
zari	0,624	0,797	0,900	0,902	0,894	0,917	0,932	0,971			
rijen	0,877	0,895	0,820	0,693	0,694	0,713	0,676	0,792	0,870		
listopad	0,929	0,879	0,732	0,563	0,585	0,604	0,521	0,660	0,742	0,953	
prosinec	0,976	0,893	0,707	0,499	0,502	0,507	0,420	0,557	0,654	0,908	0,974

Studentův t-test statistické významnosti jednotlivých parametrů ukázal nevýznamnost absolutního členu a zimních měsíců, statisticky významné jsou střední teploty v měsících dubnu, květnu, červnu, červenci a listopadu.

Proměnná	Odhad	Směr.Odch.	Závěr	Pravděpodobnost	Spodní mez	Horní mez
Abs	650,24	383,05	Nevýznamný	0,10	-126,62	1427,10
leden	81,82	74,63	Nevýznamný	0,28	-69,55	233,19
unor	-45,77	83,28	Nevýznamný	0,59	-214,67	123,13
brezen	168,27	83,96	Nevýznamný	0,05	-2,00	338,54
duben	-207,49	93,03	**Významný**	0,03	-396,16	-18,82
kveten	177,89	76,53	**Významný**	0,03	22,69	333,10
cerven	-288,18	93,61	**Významný**	0,00	-478,04	-98,33
cervenec	225,84	79,72	**Významný**	0,01	64,17	387,51
srpen	-138,21	88,12	Nevýznamný	0,13	-316,93	40,51
zari	103,02	83,16	Nevýznamný	0,22	-65,63	271,67
rijen	111,35	72,78	Nevýznamný	0,13	-36,25	258,95
listopad	-241,32	74,30	**Významný**	0,00	-392,01	-90,63
prosinec	-95,09	85,80	Nevýznamný	0,28	-269,11	78,92

Základní statistické charakteristiky

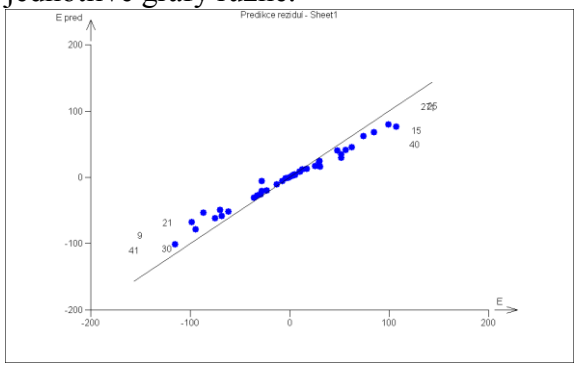
Vícenásobný korelační koeficient R :	0,9705995364
Koeficient determinace R^2 :	0,94206346
Predikovaný korelační koeficient Rp :	0,7873147911
Střední kvadratická chyba predikce MEP :	5847,131734
Akaikeho informační kritérium :	418,4112511

Indikace vlivných bodů

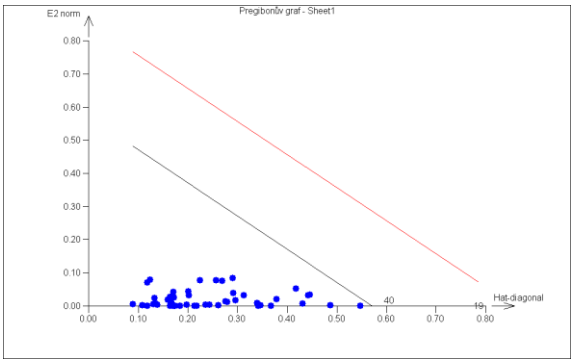
Index	Standardní	Jackknife	Predikované	Diag(Hii)	Diag(H*ii)	Cookova vzdál.	Atkinsonova vzdál.	Andrews-Pregibon st.	Vliv na Y^	Vliv na parametry LD(b)	Vliv na rozptyl LD(s)	Celkový vliv LD(b,s)
1	1,921	1,99	143,7	0,269	0,344	0,054	**2,020**	0,656	**1,214**	1,818	0,193	2,215
9	-1,797	-1,86	-150,6	0,418	0,470	-0,099	**2,618**	0,530	-1,573	3,056	0,135	3,529
15	1,485	1,51	127,6	0,446	0,480	0,092	**2,256**	0,520	**1,356**	2,359	0,045	2,566
16	-0,040	-0,04	-3,8	**0,547**	0,547	-0,004	0,073	0,453	-0,044	0,003	0,010	0,013
19	-0,204	-0,20	-28,2	**0,786**	**0,786**	-0,058	0,642	**0,214**	-0,386	0,208	0,009	0,213
25	1,938	2,02	143,8	0,257	0,335	0,052	**1,977**	0,665	**1,188**	1,740	0,202	2,139
27	1,886	1,96	137,0	0,225	0,301	0,042	**1,754**	0,699	**1,054**	1,382	0,175	1,699
40	1,235	1,25	125,8	**0,606**	**0,622**	0,146	**2,567**	**0,378**	**1,542**	3,090	0,013	3,276
41	-2,056	-2,16	-156,2	0,291	0,374	-0,065	**2,299**	0,626	-1,381	2,303	0,276	2,893

Diagnostika regresního tripletu

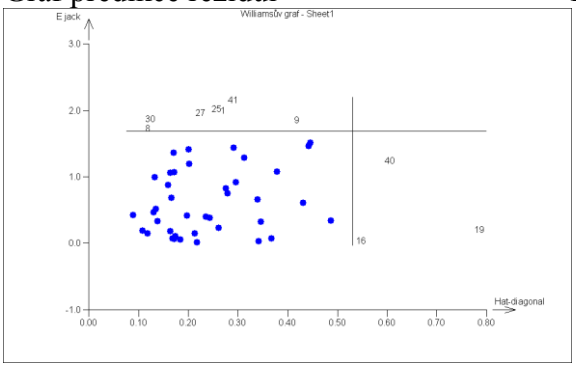
Grafy vlivných bodů ukazují lokality č. 16, 19 a 40 jako extrémní, odlehlé lokality ukazují jednotlivé grafy různě.



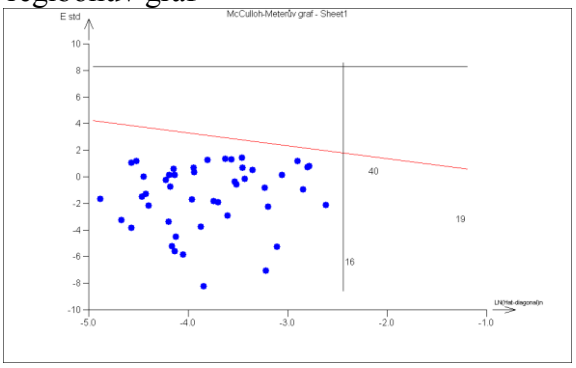
Graf predikce reziduí



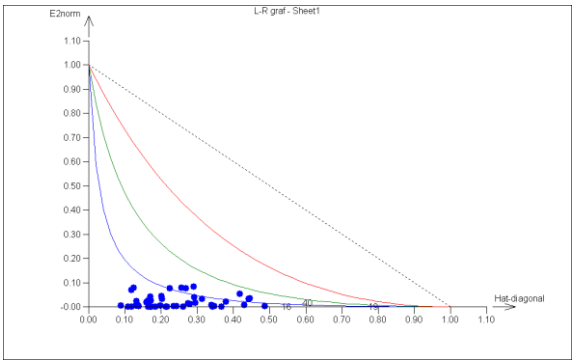
Pregibonův graf



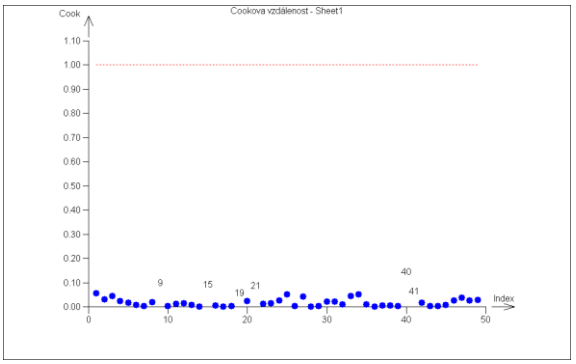
Williamsův graf



McCullh-Meterův graf

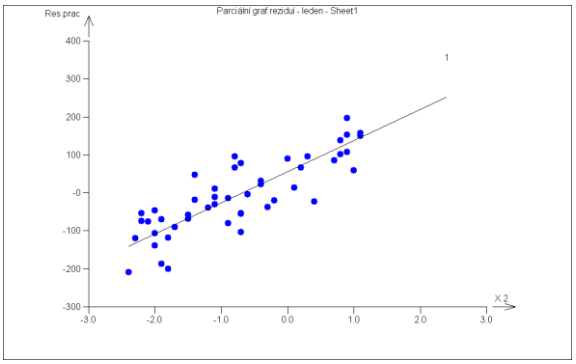


LR graf

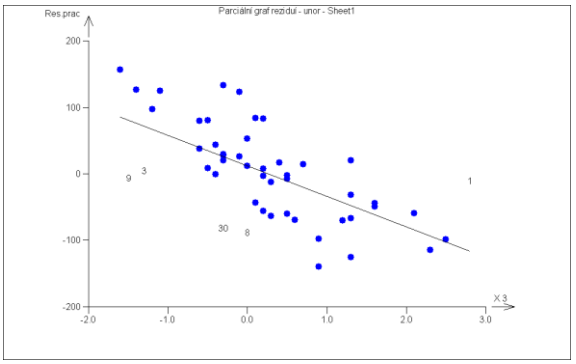


Graf Cookovy vzdálenosti

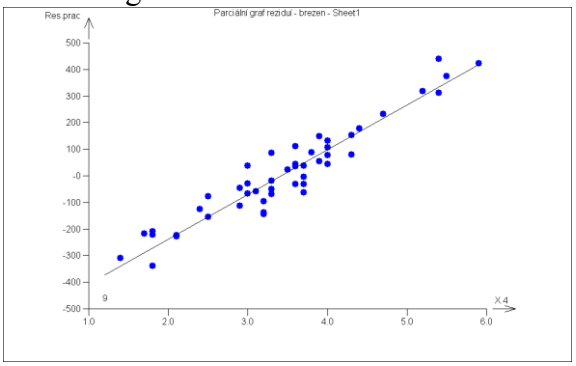
Parciální reziduální grafy naznačují linearitu závislosti jednotlivých nezávisle proměnných. V letních měsících rezidua vykazují výrazně nižší odchylky od linearity



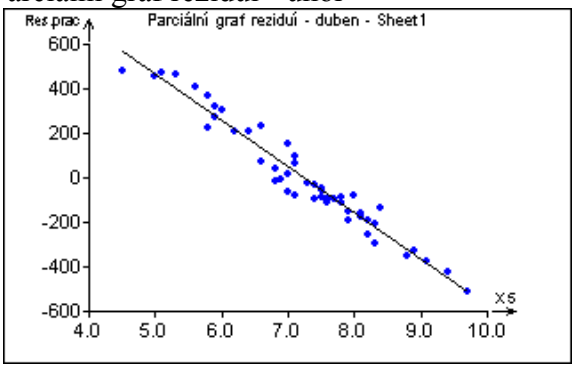
Parciální graf reziduí – leden



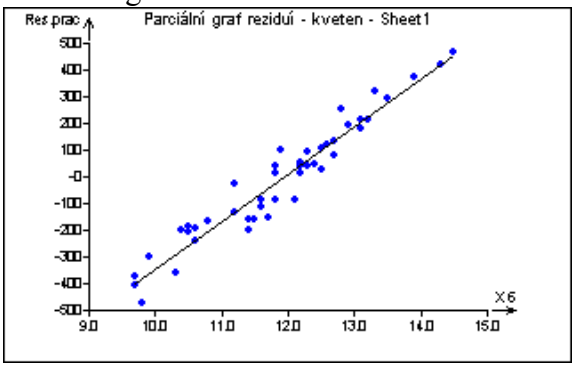
Parciální graf reziduí - únor



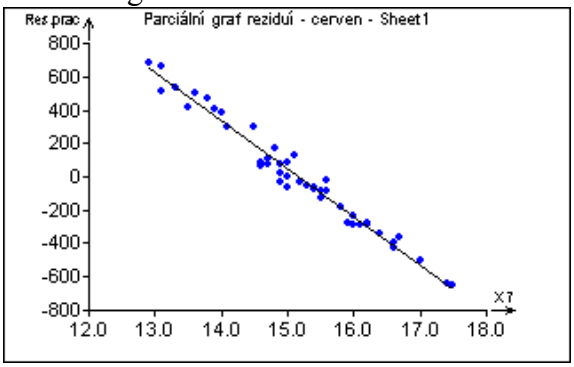
Parciální graf reziduí – březen



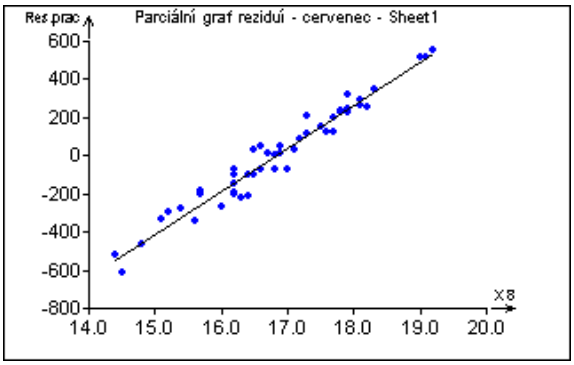
Parciální graf reziduí - duben



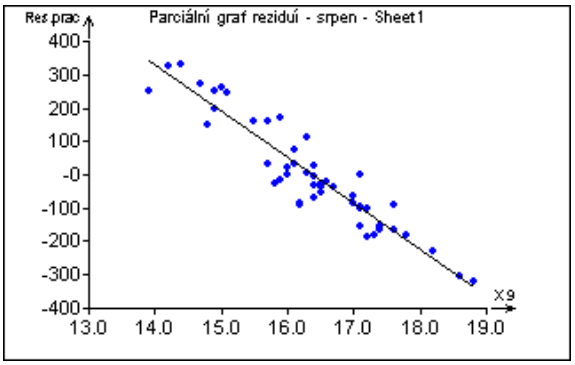
Parciální graf reziduí – květen



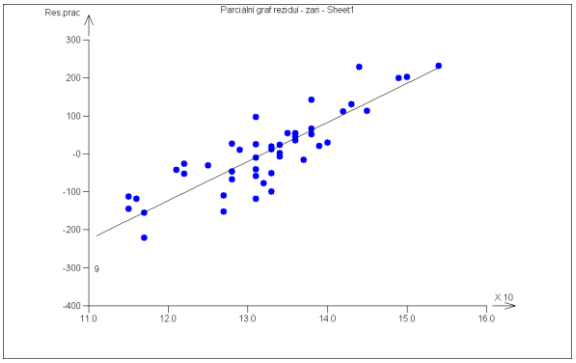
Parciální graf reziduí - červen



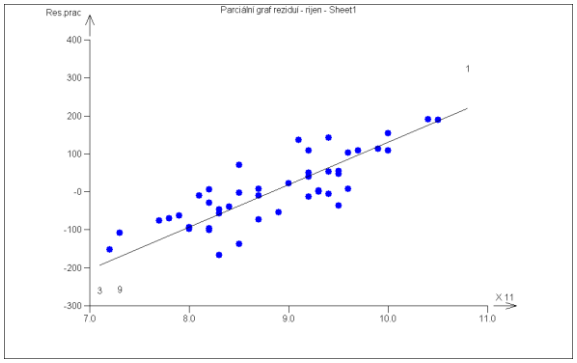
Parciální graf reziduí – červenec



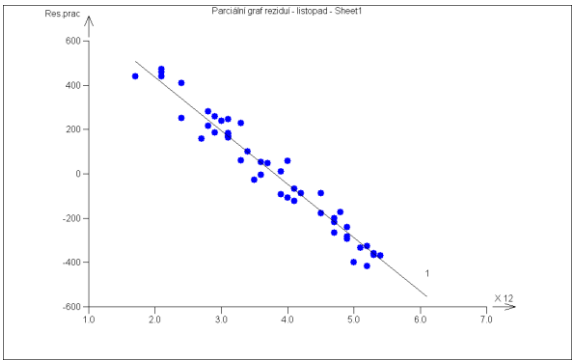
Parciální graf reziduí - srpen



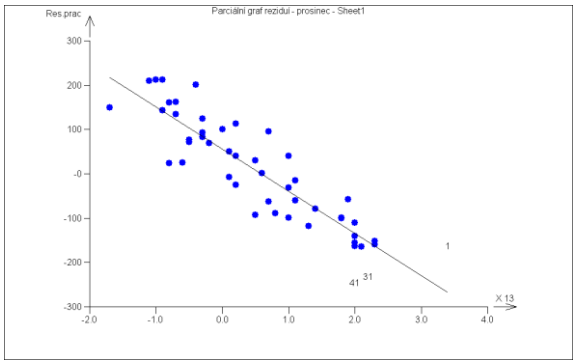
Parciální graf reziduí – září



Parciální graf reziduí - říjen



Parciální graf reziduí – listopad



Parciální graf reziduí - prosinec

Testování regresního tripletu

Fisher-Snedecorův test významnosti modelu

Hodnota kritéria F :	48,78079324
Kvantil F (1-alfa, m-1, n-m) :	2,032703133
Pravděpodobnost :	1,37270487E-018
Závěr :	Model je významný

Scottovo kritérium multikolinearity

Hodnota kritéria SC :	0,8419946525
Závěr :	Model je nekorektní!

Cook-Weisbergův test heteroskedasticity

Hodnota kritéria CW :	0,7481924519
Kvantil Chi^2(1-alfa,1) :	3,841458829
Pravděpodobnost :	0,3870491153
Závěr :	Rezidua vykazují homoskedasticitu.

Jarque-Berrův test normality

Hodnota kritéria JB :	0,5119648005
Kvantil Chi^2(1-alfa,2) :	5,991464547
Pravděpodobnost :	0,7741555936
Závěr :	Rezidua mají normální rozdělení.

Waldův test autokorelace

Hodnota kritéria WA :	0,4450197415

Kvantil Chi^2(1-alfa,1) : 3,841458829
Pravděpodobnost : 0,5047095376
Závěr : Autokorelace je nevýznamná

Durbin-Watsonův test autokorelace
Hodnota kritéria DW : -1
Kritické hodnoty DW 1,29 1,78
Závěr : Pozitivní autokorelace reziduí není prokázána.

Znaménkový test reziduí
Hodnota kritéria Sg : 0,5092331307
Kvantil N(1-alfa/2) : 1,959963999
Pravděpodobnost : 0,610588823
Závěr : V reziduích není trend.

Odhady parametrů

Proměnná	Odhad	Směr.Odch.	Závěr
Abs	650,2401064	383,0469253	Nevýznamný
Leden	81,8186548	74,63494727	Nevýznamný
Unor	-45,76842511	83,28106289	Nevýznamný
Brezen	168,2679373	83,95570728	Nevýznamný
Duben	-207,4915147	93,028789	Významný
Kveten	177,8932087	76,52654929	Významný
Cerven	-288,1844088	93,61320217	Významný
Červenec	225,8436176	79,71543008	Významný
Srpen	-138,2081028	88,12341246	Nevýznamný
Zari	103,0224456	83,15583435	Nevýznamný
Rijen	111,3510102	72,77679511	Nevýznamný
Listopad	-241,3218694	74,30116382	Významný
Prosinec	-95,09417464	85,80244437	Nevýznamný

Nalezený model má tvar (v závorce odhad směrodatné odchylky parametru):

$$\mathbf{y = - 207{,}492\ (93{,}029)x_4 + 177{,}893\ (76{,}527)x_5 - 288{,}184\ (93{,}613)x_6 + 225{,}844\ (79{,}715)x_7 - 241{,}322\ (74{,}301)x_{11}}$$