Univerzita Pardubice

Fakulta chemicko – technologická

Katedra analytické chemie

Licenční studium chemometrie

Statistické zpracování dat

Tvorba lineárních regresních modelů při analýze dat

Zdravotní ústav se sídlem v Ostravě
Odbor hygienických laboratoří Karviná

V Karviné dne 10.9.2005 **Ing. Miluše Galuszková**

Předmět:

2.1 Tvorba lineárních regresních modelů při analýze dat

Přednášející: ***Prof.RNDr. Milan Meloun, DrSc.***

Obsah

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.

Zadání

Z důvodu restrukturalizačních změn je nutné sloučit dvě laboratoře. V rámci mezilaboratorních porovnávacích zkoušek analýz půd je nutné rozhodnout zda obě laboratoře poskytují shodné výsledky nebo je nutné přezkoumat metody.
x je vztažná hodnota prvků v mg/kg uvedná organizátorem mezilaboratorní porovnávací zkoušky
y_1, y_2 jsou výsledky z laboratoře L1 a L2

Data:

	1	2	3	4	5	6	7	8	9	10	11	12
x	6,36	0,325	0,220	51,00	38,30	16,50	41,58	14,30	34,15	9,70	10,1	8,16
y_1	6,40	0,320	0,210	54,40	40,5	18,30	41,20	14,60	37,00	8,70	10,00	7,50
y_2	6,40	0,320	0,180	51,00	40,34	17,41	41,50	13,60	34,20	9,78	10,24	10,15
	13	14	15	16	17	18	19	20	21	22	23	24
x	6,20	0,300	0,210	26,37	21,90	11,00	39,89	9,065	29,04	3,60	0,750	2,70
y_1	6,40	0,300	0,200	26,10	21,80	11,60	37,50	9,00	27,00	3,60	0,720	2,70
y_2	5,90	0,290	0,160	26,37	22,00	12,00	38,00	9,00	28,00	3,20	0,750	2,90

Program: **ADSTAT**
Modul: **Lineární regrese**
Řešení:
Za předpokladu, že obě laboratoře poskytují shodné výsledky, budou obě regresní přímky statisticky nevýznamně odlišné. K testování shody přímek použijeme testační kritérium F_{CH}.

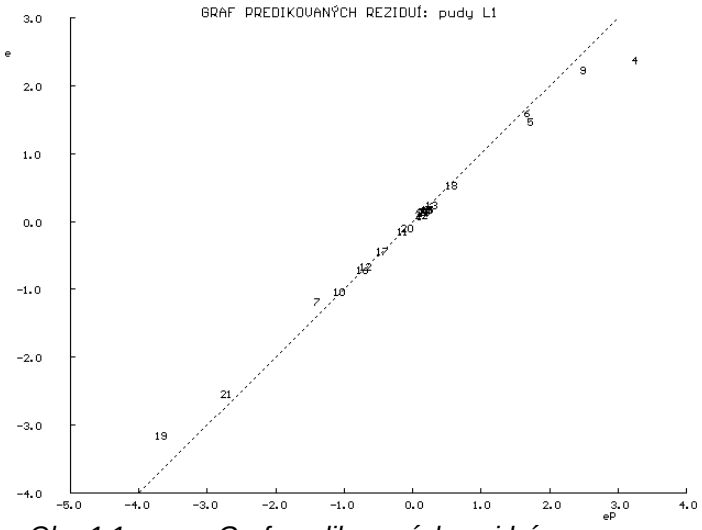
1. Laboratoř L1

Vyšetření vlivných bodů pomocí diagnostických grafů

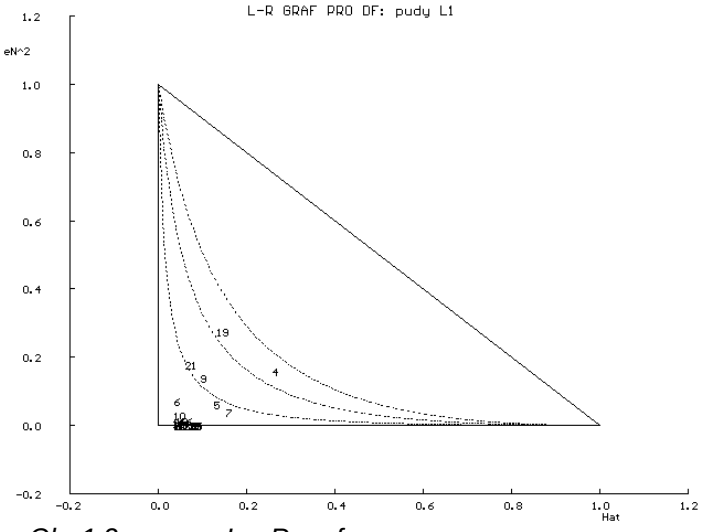
Program: ADSTAT
Modul: LINEÁRNÍ REGRESE
Regresní diagnostika
Název: půdy L1

ZVOLENÁ STRATEGIE REGRESNÍ ANALÝZY:
Omezení, P : 1.0000E-34
Transformace : Ne
Váhy : Ne
Absolutní člen zahrnut: Ano

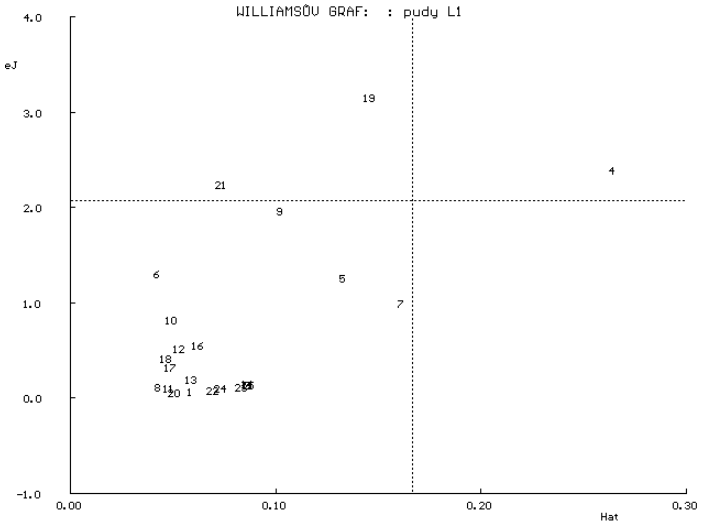
PODMÍNKY A KVANTILY PRO STATISTICKÉ TESTY:
Hladina významnosti, alfa : 0.050
Počet bodů, n : 24
Počet parametrů, m : 2
Kvantil Studentova rozdělení t(1-alpha/2,n-m) : 2.074
Kvantil rozd. Chí-kvadrát Chi-square(1-alpha,m) : 5.991



Obr..1.1 Graf predikovaných reziduí



Obr.1.2. L – R graf



Obr.1.3. Williamsův graf

Závěr vyšetření vlivných bodů pomocí diagnostických grafů:
V datech se vyskytují odlehlé body č.4,21,19 Body odstraníme ze souboru dat.

ZVOLENÁ STRATEGIE REGRESNÍ ANALÝZY :

Omezení, P	: 1.0000E-34
Transformace	: Ne
Váhy	: Ne
Absolutní člen zahrnut	: **Ano**

PODMÍNKY A KVANTILY PRO STATISTICKÉ TESTY:

Hladina významnosti, alfa	: 0.050
Počet bodů, n	: **21**
Počet parametrů, m	: 2
Kvantil Studentova rozdělení t (1-alpha/2,n-m)	: 2.093

navržený model: $y = \beta_0 + \beta_1 x$

ODHADY PARAMETRŮ A TESTY VÝZNAMNOSTI:

			Test H0: B[j] = 0 vs. HA: B[j] <> 0		
Parametr	**Odhad**	**Směrodatná odchylka**	**t-kritérium**	**Hypotéza H0**	**Hladina významnosti**
B[0]	-1.7720E-01	2.5439E-01	-6.9654E-01	Akceptována	0.495
B[1]	1.0347E+00	1.4372E-02	7.1993E+01	Zamítnuta	0.000

Porovnáním t-kritéria s kritickou hodnotou t = 2.093 (Studentův t-test) jsme zjistili, že absolutní člen (úsek) je statisticky nevýznamný a směrnice je statisticky významná.

ZVOLENÁ STRATEGIE REGRESNÍ ANALÝZY -**zpřesněný model**:

Omezení, P	: 1.0000E-34
Transformace	: Ne
Váhy	: Ne
Absolutní člen zahrnut	: **NE**

PODMÍNKY A KVANTILY PRO STATISTICKÉ TESTY:

Hladina významnosti, alfa	: 0.050
Počet bodů, n	: **21**
Počet parametrů, m	: 2
Kvantil Studentova rozdělení t (1-alpha/2,n-m)	: **2.086**

ODHADY PARAMETRŮ A TESTY VÝZNAMNOSTI:

			Test H0: B[j] = 0 vs. HA: B[j] <> 0		
Parametr	**Odhad**	**Směrodatná odchylka**	**t-kritérium**	**Hypotéza H0**	**Hladina významnosti**
B[0]	2.2337E-08	---	---	---	---
B[1]	1.0207E+00	1.0071E-02	1.020E+02	Zamítnuta	0.000

Navržený zpřesněný model: **$y_1 = 1{,}0207\ (\pm 0{,}01007)x$**

ANALÝZA KLASICKÝCH REZIDUÍ:

Reziduální součet čtverců, RSC	**1.3346E+01**
Průměr absolutních hodnot reziduí, Me	5.3847E-01
Průměr relativních reziduí, Mer	4.8003E+00
Odhad reziduálního rozptylu, s^2(e)	6.5894E-01
Odhad směrodatné odchylky reziduí, s(e)	8.1175E-01
Odhad šikmosti reziduí, g1(e)	6.3837E-01
Odhad špičatosti reziduí, g2(e)	3.6105E+00

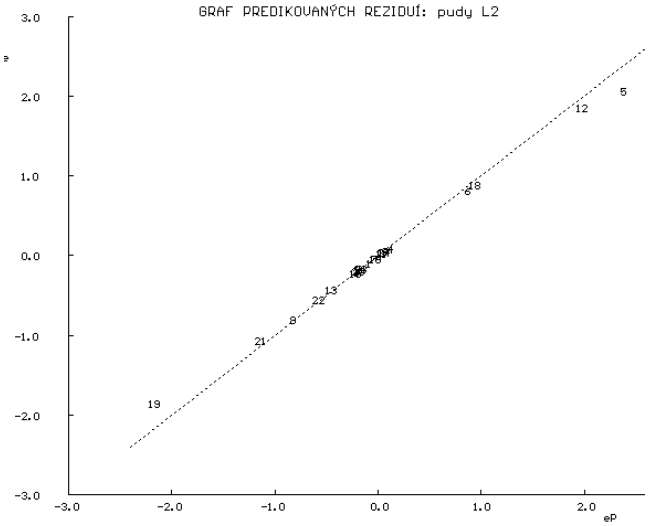
Laboratoř L2

Vyšetření vlivných bodů pomocí diagnostických grafů

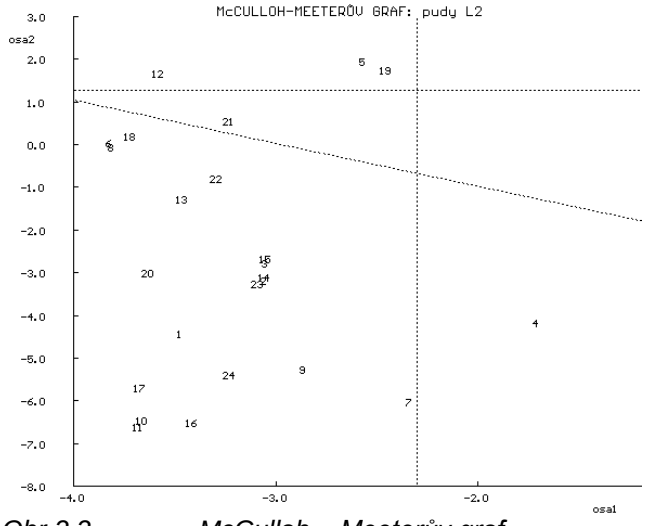
Program: ADSTAT
Modul: LINEÁRNÍ REGRESE
Regresní diagnostika
Název: půdy L2

ZVOLENÁ STRATEGIE REGRESNÍ ANALÝZY:
Omezení, P : 1.0000E-34
Transformace : Ne
Váhy : Ne
Absolutní člen zahrnut: Ano

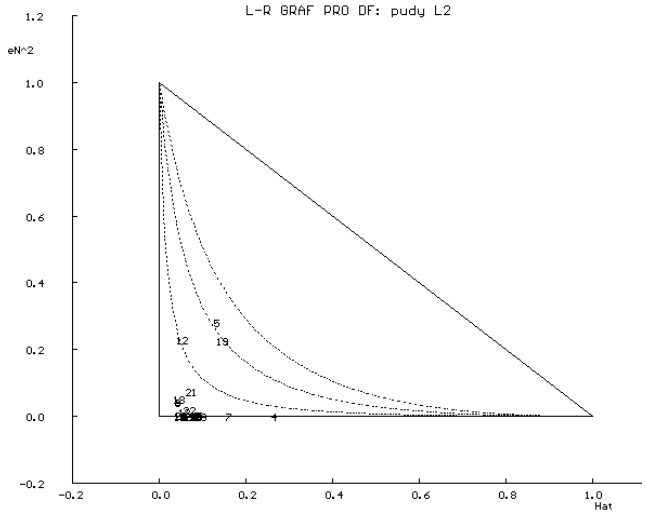
PODMÍNKY A KVANTILY PRO STATISTICKÉ TESTY:
Hladina významnosti, alfa : 0.050
Počet bodů, n : 24
Počet parametrů, m : 2
Kvantil Studentova rozdělení t(1-alpha/2,n-m) : 2.074
Kvantil rozd. Chí-kvadrát Chi-square(1-alpha,m) : 5.991



Obr.2.1. Graf predikovaných reziduí



Obr.2.2. McCulloh – Meeterův graf



Obr.2.3. L – R graf

Závěr vyšetření vlivných bodů pomocí diagnostických grafů:
V datech se vyskytují odlehlé body č.5,12,19 které odstraníme ze souboru dat.

ZVOLENÁ STRATEGIE REGRESNÍ ANALÝZY:

Omezení, P : 1.0000E-34
Transformace : Ne
Váhy : Ne
Absolutní člen zahrnut : Ano
PODMÍNKY A KVANTILY PRO STATISTICKÉ TESTY:
Hladina významnosti, alfa : 0.050
Počet bodů, n **: 21**
Počet parametrů, m : 2
Kvantil Studentova rozdělení t (1-alpha/2,n-m) : 2.093

navržený model: $y = \beta_0 + \beta_1 x$
ODHADY PARAMETRŮ A TESTY VÝZNAMNOSTI:

			Test H0: B[j] = 0 vs. HA: B[j] <> 0		
Parametr	**Odhad**	**Směrodatná odchylka**	**t-kritérium**	**Hypotéza H0**	**Hladina významnosti**
B[0]	3.1360E-02	1.3492E-01	2.3244E-01	Akceptována	0.819
B[1]	9.9719E-01	6.7099E-03	1.4862E+02	Zamítnuta	0.000

Porovnáním t-kritéria s kritickou hodnotou t = 2.093 (Studentův t-test) jsme zjistili, že absolutní člen (úsek) je statisticky nevýznamný a směrnice je statisticky významná.

ZVOLENÁ STRATEGIE REGRESNÍ ANALÝZY -**zpřesněný model**:
Omezení, P : 1.0000E-34
Transformace : Ne
Váhy : Ne
Absolutní člen zahrnut : **NE**
PODMÍNKY A KVANTILY PRO STATISTICKÉ TESTY:
Hladina významnosti, alfa : 0.050
Počet bodů, n : **21**
Počet parametrů, m : 2
Kvantil Studentova rozdělení t (1-alpha/2,n-m) : **2.086**

ODHADY PARAMETRŮ A TESTY VÝZNAMNOSTI:

			Test H0: B[j] = 0 vs. HA: B[j] <> 0		
Parametr	**Odhad**	**Směrodatná odchylka**	**t-kritérium**	**Hypotéza H0**	**Hladina významnosti**
B[0]	2.2337E-08	---	---	---	---
B[1]	9.9829E-01	4.6803E-03	2.1330E+02	Zamítnuta	0.000

Navržený zpřesněný model: **y_2 = 0.99829 (±0,0046803)x**

ANALÝZA KLASICKÝCH REZIDUÍ:

Reziduální součet čtverců, RSC	**3.7197E+00**
Průměr absolutních hodnot reziduí, Me	2.5509E-01
Průměr relativních reziduí, Mer	5.2236E+00
Odhad reziduálního rozptylu, s^2(e)	1.8572E-01
Odhad směrodatné odchylky reziduí, s(e)	4.3095E-01
Odhad šikmosti reziduí, g1(e)	2.29507E-01
Odhad špičatosti reziduí, g2(e)	4.6087+00

3. Laboratoř L1 + L2

Program: ADSTAT
Modul: LINEÁRNÍ REGRESE
Regresní diagnostika
Název: půdy L1+L2

ZVOLENÁ STRATEGIE REGRESNÍ ANALÝZY:
Omezení, P : 1.0000E-34
Transformace : Ne
Váhy : Ne
Absolutní člen zahrnut: Ano

PODMÍNKY A KVANTILY PRO STATISTICKÉ TESTY:
Hladina významnosti, alfa : 0.050
Počet bodů, n : 42
Počet parametrů, m : 2
Kvantil Studentova rozdělení t(1-alpha/2,n-m) : 2.021
Kvantil rozd. Chí-kvadrát Chi-square(1-alpha,m) : 5.991

navržený model: $y = \beta_0 + \beta_1 x$
ODHADY PARAMETRŮ A TESTY VÝZNAMNOSTI:

			Test H0: B[j] = 0 vs. HA: B[j] <> 0		
Parametr	**Odhad**	**Směrodatná odchylka**	**t-kritérium**	**Hypotéza H0**	**Hladina významnosti**
B[0]	-4.6498E-02	1.5366E-01	-3.0259E-01	Akceptována	0,764
B[1]	1.0128E+00	8.1123E-03	1.2485E+02	Zamítnuta	0.000

Porovnáním t-kritéria s kritickou hodnotou t = 2.021 (Studentův t-test) jsme zjistili, že absolutní člen (úsek) je statisticky nevýznamný a směrnice je statisticky významná.

ZVOLENÁ STRATEGIE REGRESNÍ ANALÝZY -**zpřesněný model**:
Omezení, P : 1.0000E-34
Transformace : Ne
Váhy : Ne
Absolutní člen zahrnut : **NE**

PODMÍNKY A KVANTILY PRO STATISTICKÉ TESTY:

Hladina významnosti, alfa : 0.050
Počet bodů, n : **42**
Počet parametrů, m : 2
Kvantil Studentova rozdělení t (1-alpha/2,n-m) : **2.020**

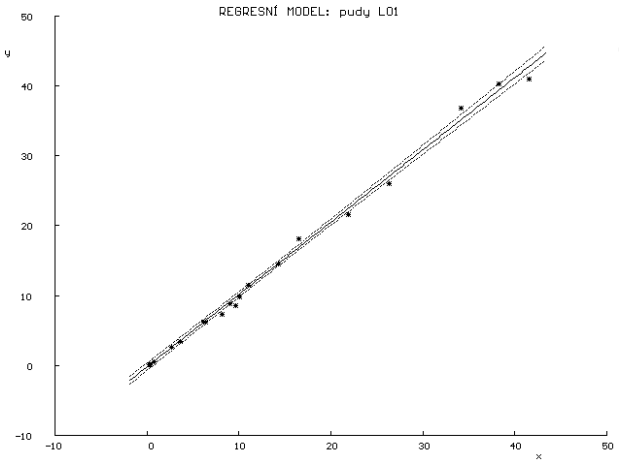
ODHADY PARAMETRŮ A TESTY VÝZNAMNOSTI:

			Test H0: B[j] = 0 vs. HA: B[j] <> 0		
Parametr	**Odhad**	**Směrodatná odchylka**	**t-kritérium**	**Hypotéza H0**	**Hladina významnosti**
B[0]	2.2337E-08	---	---	---	---
B[1]	1.0111E+00	5.7262E-03	1.7262E+02	Zamítnuta	0.000

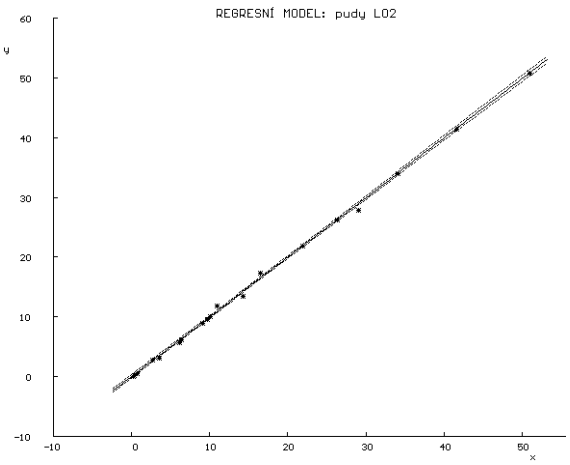
Navržený zpřesněný model: **y = 1.0111 (±0,005726)x**

ANALÝZA KLASICKÝCH REZIDUÍ:

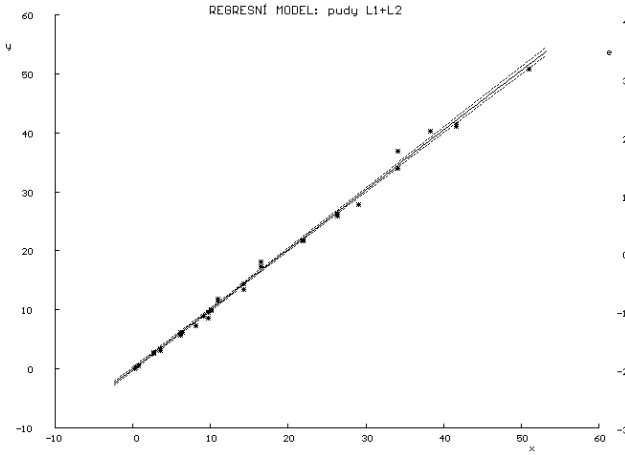
Reziduální součet čtverců, RSC	**2.0260E+01**
Průměr absolutních hodnot reziduí, Me	4.2443E-01
Průměr relativních reziduí, Mer	4.8725E+00
Odhad reziduálního rozptylu, s^2(e)	4.9357E-01
Odhad směrodatné odchylky reziduí, s(e)	7.0254E-01
Odhad šikmosti reziduí, g1(e)	1.5096E-01
Odhad špičatosti reziduí, g2(e)	6.5224E+0



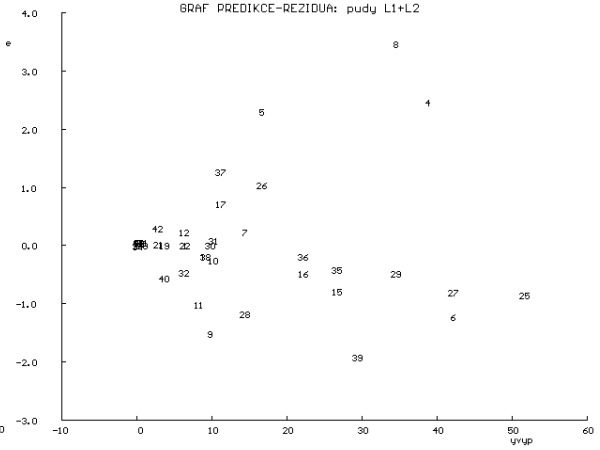
Obr.3.1 Regresní model pro L1



Obr.3.2 Regresní model pro L21



Obr.3.3 Regresní model pro L1+L2



Obr.3.2 Graf predikce – rezidua pro L1+L2

Řešení

K testování shody přímek použijeme testační kritérium F_{CH}.

ANALÝZA KLASICKÝCH REZIDUÍ:

n = 42 , m= 2	RSC
L1	13,346
L2	3,7197
L1+L2	20,260

$$F_{CH} = \frac{(RSC_{1,2} - RSC_1 - RSC_2).(n-2m)}{(RSC_1 + RSC_2).(m)}$$

$$F_{CH} = \frac{(20{,}260 - 13{,}346 - 3{,}7197).(42-4)}{(13{,}346 + 3{,}7197).(2)}$$

F_{CH} =4.092
$F_{(1-\alpha)(m,\ n-2m)}$ = 3,2448

$F_{exp.} > F_{tab.}$ nulová hypotéza H0: $\beta_1 = \beta_2$, proti HA: $\beta_1 <> \beta_2$ je zamítnuta.

Závěr:

Na základě regresní analýzy lze konstatovat, že analýza půd v laboratoři L1a L2 je odlišná, proto bude nutné přezkoumat postupy analýz.

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).

Zadání:

V obytném prostoru ve výšce 50 cm byla od 13 hod do 8 hod proměřována teplota v ^{o}C(x) a relativní vlhkost v % (y).

Nalezněte optimální polynomický model.

Data:

	1	2	3	4	5	6	7	8	9	10	11	12
x	20,9	20,5	20,4	21,6	22,9	22,9	22,0	20,8	19,6	18,8	18,2	17,7
y	36,0	36,6	36,9	34,7	32,6	31,9	33,0	34,7	36,1	38,0	39,1	40,1
	13	14	15	16	17	18	19	20	21	22	23	24
x	17,4	17,1	16,9	16,6	16,4	16,3	16,1	16,0	15,9	15,7	15,7	15,6
y	40,5	41,1	41,7	42,3	42,7	42,5	42,9	43,1	43,4	43,6	43,4	43,4
	25	26	27	28	29	30	31					
x	15,5	15,4	16,0	19,3	20.7	21,9	22,0					
y	43,4	43,3	42,7	38,0	35,1	33,4	33,2					

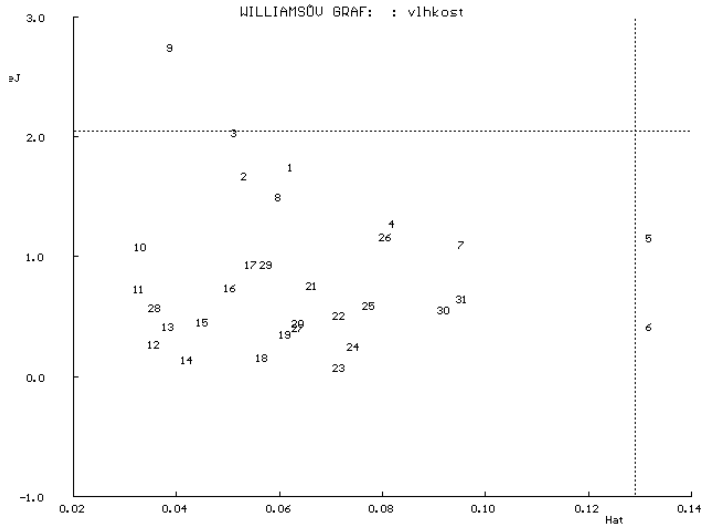
Program: **ADSTAT**
Modul: **Lineární regrese**
Řešení: Vyšetření vlivných bodů pomocí diagnostických grafů
Určení stupně polynomu a nalezení nejlepších odhadů
Určení odhadů parametrů metodou racionálních hodností

Vyšetření vlivných bodů pomocí diagnostických grafů

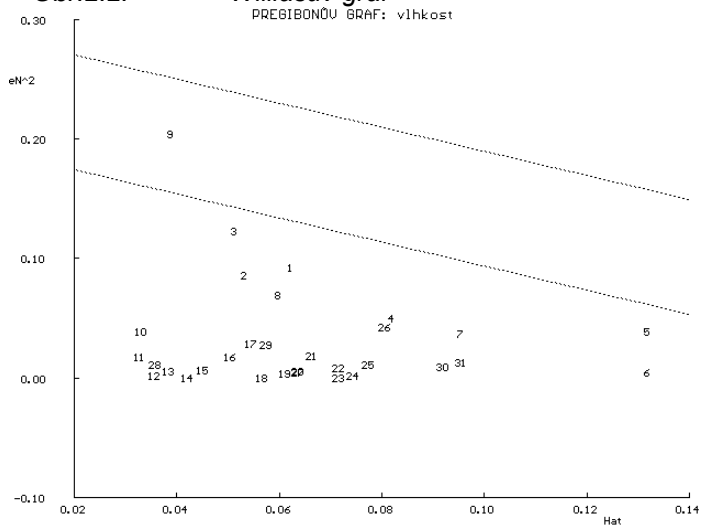
Program: ADSTAT
Modul: LINEÁRNÍ REGRESE
Regresní diagnostika
Název: Vlhkost

ZVOLENÁ STRATEGIE REGRESNÍ ANALÝZY:
Omezení, P : 1.0000E-34
Transformace : Ne
Váhy : Ne
Absolutní člen zahrnut: Ano

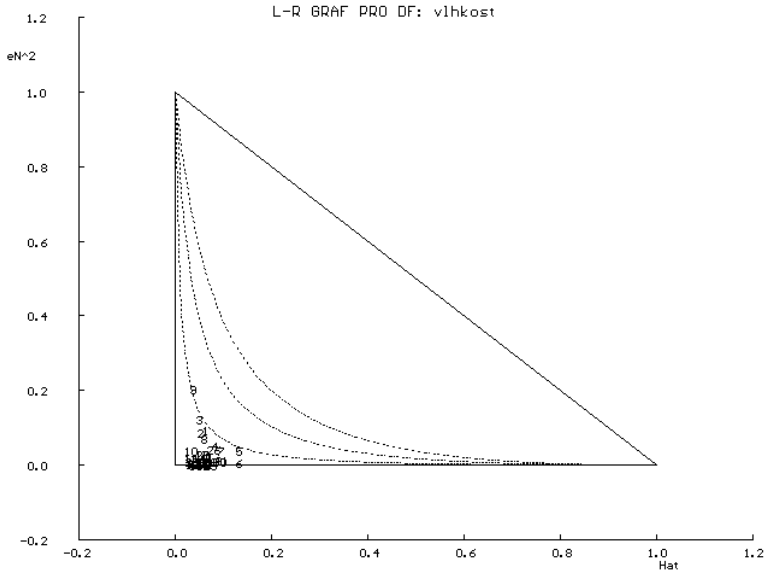
PODMÍNKY A KVANTILY PRO STATISTICKÉ TESTY:
Hladina významnosti, alfa : 0.050
Počet bodů, n : 31
Počet parametrů, m : 2
Kvantil Studentova rozdělení t(1-alpha/2,n-m) : 2.045
Kvantil rozd. Chí-kvadrát Chi-square(1-alpha,m) : 5.991



Obr.1.1. Williasův graf



Obr.1.2. Pregibonův graf



Obr.1.3. L – R graf

Závěr vyšetření vlivných bodů pomocí diagnostických grafů:
V datech se vyskytuje odlehlý body č.9 který odstraníme ze souboru dat.

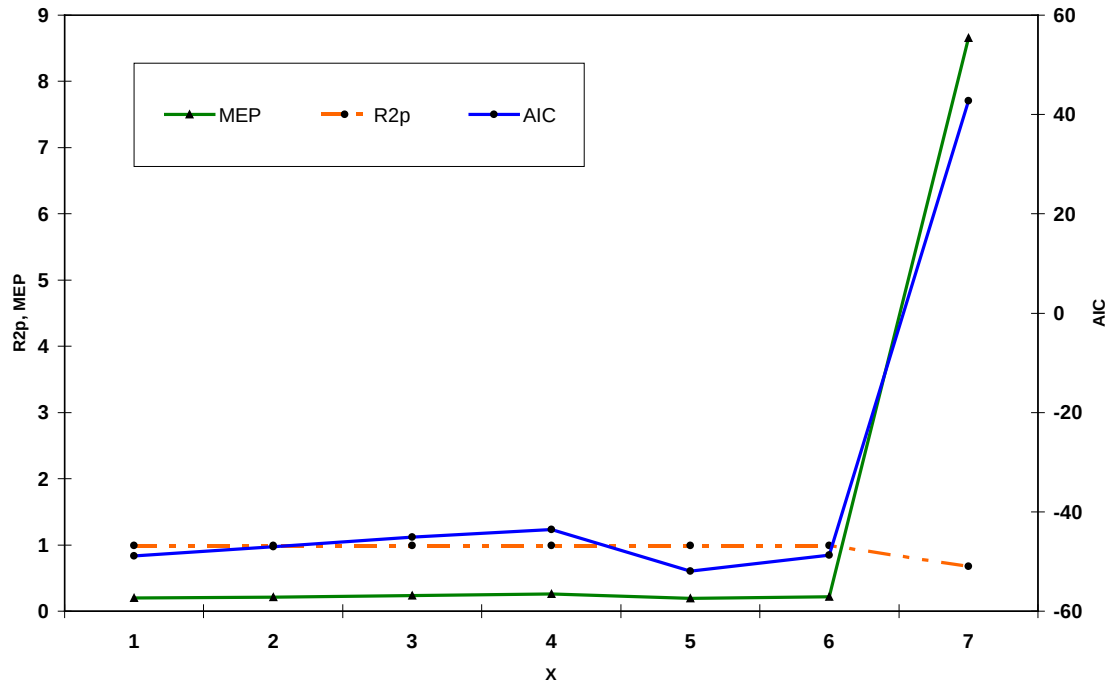
Určení stupně polynomu a nalezení nejlepších odhadů

ZVOLENÁ STRATEGIE REGRESNÍ ANALÝZY:
Omezení, P : 1.0000E-34
Transformace : polynom
stupeň polynomu: :1.....7
Váhy : Ne
Absolutní člen zahrnut : Ano

PODMÍNKY A KVANTILY PRO STATISTICKÉ TESTY:
Hladina významnosti, alfa : 0.050
Počet bodů, n : 30
Počet parametrů, m : 2
Kvantil Studentova rozdělení t(1-alpha/2,n-m) : 2.048
Kvantil rozd. Chí-kvadrát Chi-square(1-alpha,m) : 5.991
Jméno výstupního souboru :vlhkost1.txt

K určení stupně polynomu porovnáváme v tabulce uvedené statistické charakteristiky pro stupně polynomu m=1-7.

stupeň polynomu (m)	MEP	$\hat{R}_p^2$	AIC
1	0,197	0,994	-48,91
2	0,214	0,993	-47,06
3	0,237	0,993	-45,06
4	0,259	0,992	-43,59
5	0,193	0,994	-51,94
6	0,219	0,993	-48,75
7	8,655	0,678	+42,79



Obr. 2.1 Graf závislosti MEP, AIC a $\hat{R}_p^2$ na stupni polynomu m

Závěr hledání stupně polynomu:
Úloha má dvě řešení. Na obr. 2.1 jsou znázorněna minima MEP a AIC pro stupeň polynomu 5 i těsně pro stupeň polynomu 1 a maximum $\hat{R}_p^2$ je nejvyšší pro stupeň polynomu 5 i 1.
Na obr. 3.1. a 3.2. jsou znázorněny regresní modely pro optimální stupně polynomu m =1 a m=5.

Stupeň polynomu 1:

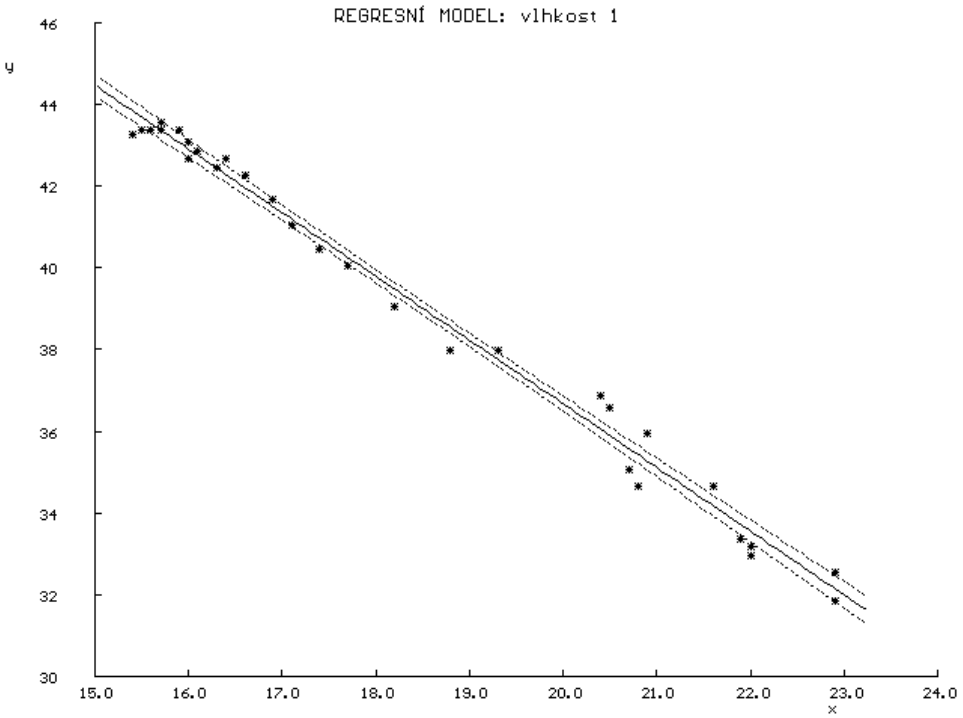
Odhady parametrů a testy významnosti

stupeň polynomu 1			Test H0: B[j] = 0 vs. HA: B[j] <> 0		
Parametr	**Odhad**	**Směrodatná odchylka**	**t-kritérium**	**Hypotéza H0**	**Hladina významnosti**
B[0]	6.7850E+01	5.7049E-01	1.1893E+02	Zamítnuta	0,000
B[1]	-1.5586E+00	3.0645E-02	-5.0859E+01	Zamítnuta	0,000

Navržený model: **y = 67,85 (±0,5705) - 1,5586 (±0,03065)x**

Statistické charakteristiky regrese

Vícenásobný korelační koeficient, R	9.9463E-01
Koeficient determinace, R^2	9.8929E-01
Predikovaný korelační koeficient, Rp^2	9.9382E-01
Střední kvadratická chyba predikce, MEP	1.9717E-01
Akaikeho informační kritérium, AIC	-4.8912E+01



Obr. 3.1. Regresní model – stupeň polynomu 1

Stupeň polynomu 5:

Odhady parametrů a testy významnosti

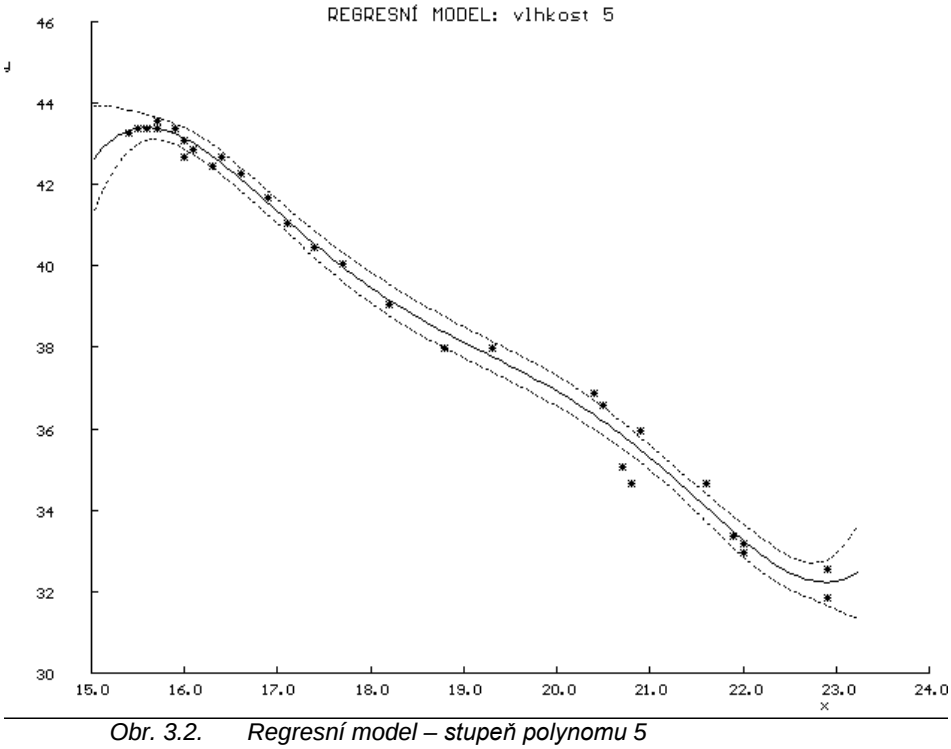
stupeň polynomu 5			Test H0: B[j] = 0 vs. HA: B[j] <> 0		
Parametr	**Odhad**	**Směrodatná odchylka**	**t-kritérium**	**Hypotéza H0**	**Hladina významnosti**
B[0]	-1.6220E+04	5.0757E+03	-3.1956E+00	Zamítnuta	0,004
B[1]	4.3163E+03	1.3500E+03	3.1973E+00	Zamítnuta	0,004
B[2]	-4.5563E+02	1.4298E+02	-3.1867E+00	Zamítnuta	0,004
B[3]	2.3922E+01	7.5381E+00	3.1735E+00	Zamítnuta	0,004
B[4]	-6.2495E-01	1.9783E-01	-3.1591E+00	Zamítnuta	0,004
B[5]	6.4993E-03	2.0675E-03	3.1435E+00	Zamítnuta	0,004

Navržený model:

$$y=-16220(\pm5076)+4316(\pm1350)x_1 -445,6(\pm143,0)x_2 +23,92(\pm7,538)x_3 -0,625(\pm0,1980)x_4+0,006499(\pm0,02068)x_5$$

Statistické charakteristiky regrese

Vícenásobný korelační koeficient, R	9.9629E-01
Koeficient determinace, R^2	9.9258E-01
Predikovaný korelační koeficient, Rp^2	9.9397E-01
Střední kvadratická chyba predikce, MEP	1.9259E-01
Akaikeho informační kritérium, AIC	-5.1939E+01



Obr. 3.2. Regresní model – stupeň polynomu 5

Určení odhadů parametrů metodou racionálních hodností

Odhady parametrů j a testy významnosti jsou uvedeny v tabulkách a jsou statisticky významné. Z tohoto důvodu nemusíme hledat jinou hodnotu omezení na vlastní čísla P a použijeme odhady metodou nejmenších čtverců pro $P = 10^{-34}$.

Závěr

Při zvolení stupně polynomu m=1 získáme jednoduchý model. Zadání úlohy je získat maximální těsnost proložení, proto řešením je stupeň polynomu m = 5.

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).

Zadání

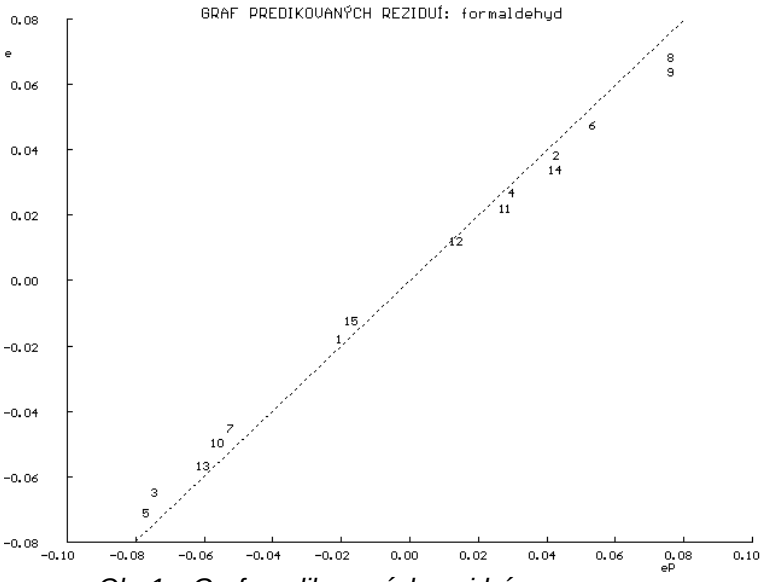
Ve zdravotnických pomůckách se po sterilizaci formaldehydem provádí stanovení reziduálního obsahu formaldehydu. Vzorky se změří, vypočte se smáčecí plocha a připraví se 24 hodinový výluh do fyziologického roztoku. Uvolněný formaldehyd reaguje v kyselém prostředí s kyselinou chromotropovou (**data x**), intenzita zbarvení se měří fotometricky. Výsledky formaldehydu se uvádí v µg na cm^2. Metoda je náročná na dodržování reakčních podmínek i na zručnost laborantky provádějící analýzu. Laboratoř uvažuje začít používat pro stanovení formaldehydu metodu s pararosanilinem **(data y)**. Obě metody využívají stejné zařízení spektrofotometr HITACHI.
Může laboratoř provádět stanovení formaldehydu ve fyziologickém roztoku méně náročnou metodou s pararosanilinem?

Data:

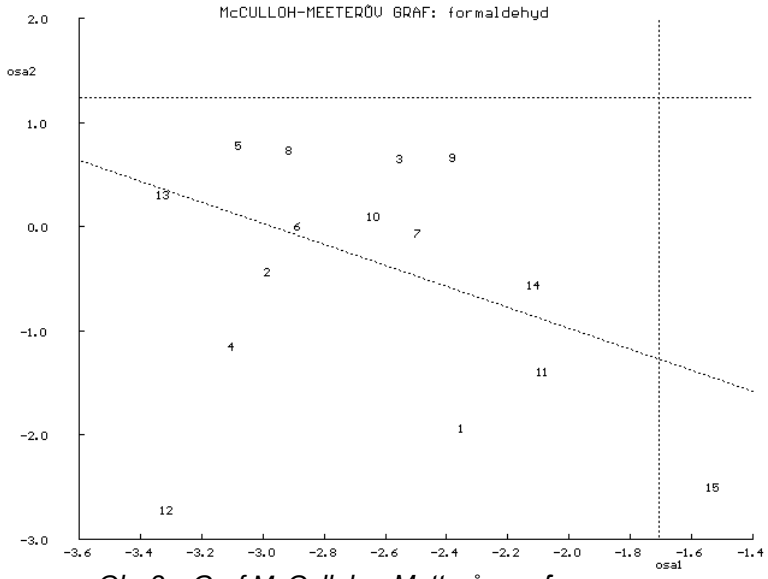
formaldehyd ve sterilních materiálech µg/cm^2								
x	0,58	0,22	0,14	0,44	0,24	0,20	0,13	0,48
y	0,55	0,26	0,08	0,46	0,17	0,25	0,09	0,54
x	0,11	0,53	0,06	0,32	0,35	0,62	0,72	
y	0,18	0,47	0,09	0,33	0,29	0,64	0,69	

Program: **ADSTAT**
Modul: **Lineární regrese**
Řešení: **1.Regresní diagnostika**
Vyšetření vlivných bodů pomocí diagnostických grafů
2.Metoda nejmenších čtverců
Vyčíslení intervalu spolehlivosti úseku a intervalu spolehlivosti směrnice

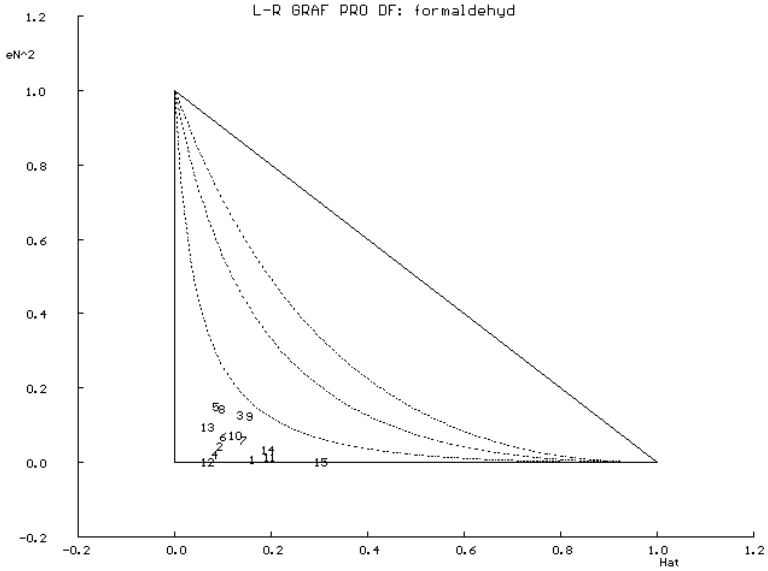
1.Vyšetření vlivných bodů pomocí diagnostických grafů



Obr.1 Graf predikovaných reziduí



Obr.2 Graf McCulloh – Metterův graf



Obr.3 L – R graf

Závěr vyšetření vlivných bodů pomocí diagnostických grafů:
V datech se nevyskytují odlehlé body. Podezřelé body a extrém (bod 15) neodstraníme také z důvodu, že jsou zdrojem potřebné informace. Všechna data použijeme pro další statistické šetření.

2.Vyčíslení intervalu spolehlivosti úseku a intervalu spolehlivosti směrnice
Metoda nejmenších čtverců

ZVOLENÁ STRATEGIE REGRESNÍ ANALÝZY:

Omezení, P	: 1.0000E-34
Transformace	: Ne
Váhy	: Ne
Absolutní člen zahrnut	: Ano

PODMÍNKY A KVANTILY PRO STATISTICKÉ TESTY:

Hladina významnosti, alfa	: 0.050
Počet bodů, n	: 15
Počet parametrů, m	: 2
Kvantil Studentova rozdělení t (1-alpha/2,n-m)	: 2.160

ODHADY PARAMETRŮ A TESTY VÝZNAMNOSTI:

			Kvantil Studentova rozdělení $t_{1-\alpha/2(n-m)} = 2.160$		
			Test H0: B[j] = 0 vs. HA: B[j] <> 0		
Parametr	**Odhad**	**Směrodatná odchylka**	**t-kritérium**	**Hypotéza H0**	**Hladina významnosti**
B[0]	9.9949E-03	2.5485E-02	3.9219E-01	Akceptována	0.701
B[1]	9.6110E-01	6.4153E-02	1.4982E+01	Zamítnuta	

Navržený model: y = 0,00999 (±0,02549) + 0,09611 (±0,06415)x

Testování úseku

Pokud interval spolehlivosti úseku obsahuje nulu, lze úsek považovat za nulový.

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

$$0{,}0099949 - 2{,}160 \cdot 0{,}02548 \leq \beta_0 \leq 0{,}0099949 + 2{,}160 \cdot 0{,}02548$$

$$\mathbf{-0{,}0451 \leq \beta_0 \leq 0{,}06504}$$

Závěr testování úseku:
Interval spolehlivosti úseku obsahuje 0, hypotéza H0: B[0] =0, úsek považujeme za nulový, nová metoda není zatížená systematickou chybou.

Testování směrnice

Pokud interval spolehlivosti směrnice obsahuje jedničku, lze směrnici považovat za jednotkovou.

$$b_1 - t_{1-\alpha/2(n-m)} \cdot \sqrt{D(b_1)} \leq \beta_1 \leq b_1 + t_{1-\alpha/2(n-m)} \cdot \sqrt{D(b_1)}$$

$$0{,}96119 - 2{,}160 \cdot 0{,}064153 \leq \beta_1 \leq 0{,}96119 + 2{,}160 \cdot 0{,}064153$$

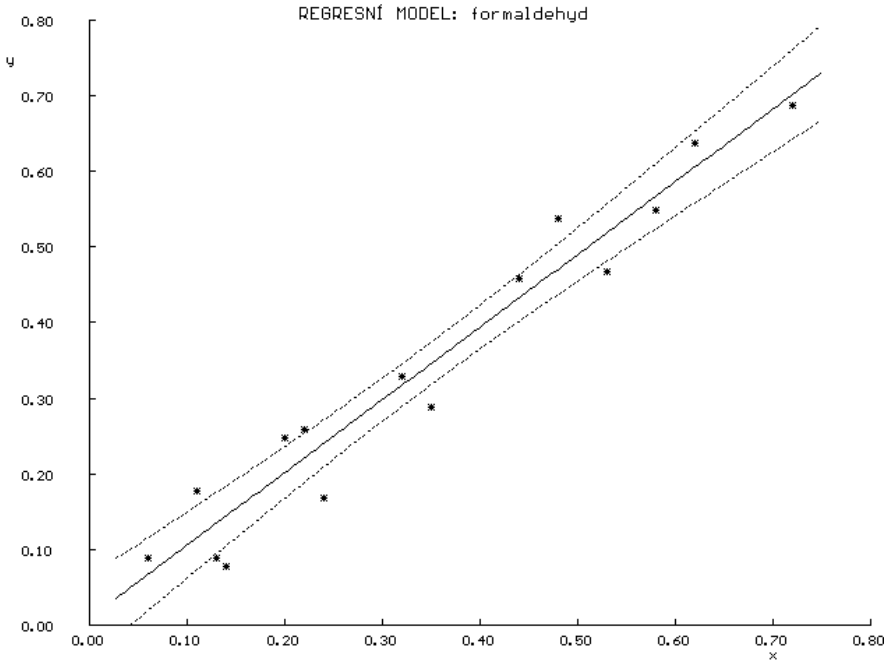
$$\mathbf{0{,}82262 \leq \beta_1 \leq 1{,}09976}$$

Závěr testování směrnice:
Interval spolehlivosti směrnice obsahuje 1, směrnici považujeme za jednotkovou.

STATISTICKÉ CHARAKTERISTIKY REGRESE:

Vícenásobný korelační koeficient, R	9.7224E-01
Koeficient determinace, R^2	9.4525E-01
Predikovaný korelační koeficient, Rp^2	9.6419E-01
Střední kvadratická chyba predikce, MEP	2.7764E-03
Akaikeho informační kritérium, AIC	-8.8057E+01

Koeficient determinace D = R^2 = 94,5 % , to je 94,5 % bodů, které vyhovují regresnímu modelu. MEP a AIC se využívají pro rozhodnutí mezi několika modely. Optimální model nabývá minimálních hodnot MEP a AIC a maximálních R^2p.



Obr.4 Regresní model

Závěr validace:

Interval spolehlivosti úseku obsahuje 0, metoda není zatížená systematickou chybou.

Interval spolehlivosti směrnice obsahuje 1, metoda nenadhodnocuje ani nepodhodnocuje.

Nová metoda stanovení formaldehydu ve sterilních materiálech s činidlem pararosanilinem je úspěšně zvalidována.

Úloha 4. Vícerozměrný lineární regresní model o alespoň 4 proměnných

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

Data:

hod	Poř.č.	SO_2 (µg/m^3)	CO (µg/m^3)	NO_2 (µg/m^3)	Benzen (µg/m^3)	Prach PM10 (µg/m^3)
11:00	1	17	314	12	2,20	73
12:00	2	18	290	11	2,95	55
13:00	3	18	313	10	1,83	46
14:00	4	18	344	10	1,96	55
15:00	5	18	358	11	2,03	48
16:00	6	18	341	8	1,87	47
17:00	7	17	301	8	60,5	54
18:00	8	16	379	9	0,05	69
19:00	9	14	327	9	0,05	72
20:00	10	13	305	9	2,07	65
21:00	11	14	536	21	1,05	90
22:00	12	14	534	14	2,21	106
23:00	13	13	542	13	2,96	87
00:00	14	13	448	11	2,71	65
1:00	15	13	354	10	3,28	55
2:00	16	13	395	10	2,38	53
3:00	17	13	351	8	2,35	53
4:00	18	13	311	6	3,20	43
5:00	19	13	346	8	2,57	51
6:00	20	13	390	10	2,08	58
7:00	21	13	516	10	1,64	42
8:00	22	16	590	21	1,60	38
9:00	23	24	454	22	5,17	46
10:00	24	31	496	25	5,40	37

hod	Poř.č.	Směr větru (stupeň)	Rychlost větru (m/s)	Teplota (^{0}C)	Relativní vlhkost (%)	**O_3 (µg/m^3)**
11:00	1	26	1,8	9,2	70	**38**
12:00	2	25	1,9	10,9	62	**40**
13:00	3	22	2,1	13,2	56	**44**
14:00	4	18	2,0	14,0	54	**44**
15:00	5	2	1,6	14,6	49	**46**
16:00	6	20	1,7	14,7	48	**40**
17:00	7	27	1,4	13,8	54	**41**
18:00	8	31	0,6	11,9	56	**28**
19:00	9	60	0,4	8,1	57	**17**
20:00	10	60	0,1	5,5	61	**15**
21:00	11	65	0,3	4,0	54	**9**
22:00	12	52	0,1	2,8	59	**7**
23:00	13	45	0,1	1,7	60	**5**
00:00	14	55	0,1	0,8	57	**8**
1:00	15	48	0,3	0,1	59	**8**
2:00	16	48	0,1	-0,7	62	**4**
3:00	17	49	0,1	-1,0	60	**5**
4:00	18	51	0,8	-1,3	60	**7**
5:00	19	218	1,2	-1,6	58	**5**
6:00	20	220	1,8	-1,9	61	**4**
7:00	21	218	0,5	-1,1	60	**6**
8:00	22	215	0,3	2,5	63	**10**
9:00	23	30	0,4	6,9	61	**20**
10:00	24	31	0,7	10,0	59	**29**

Zadání

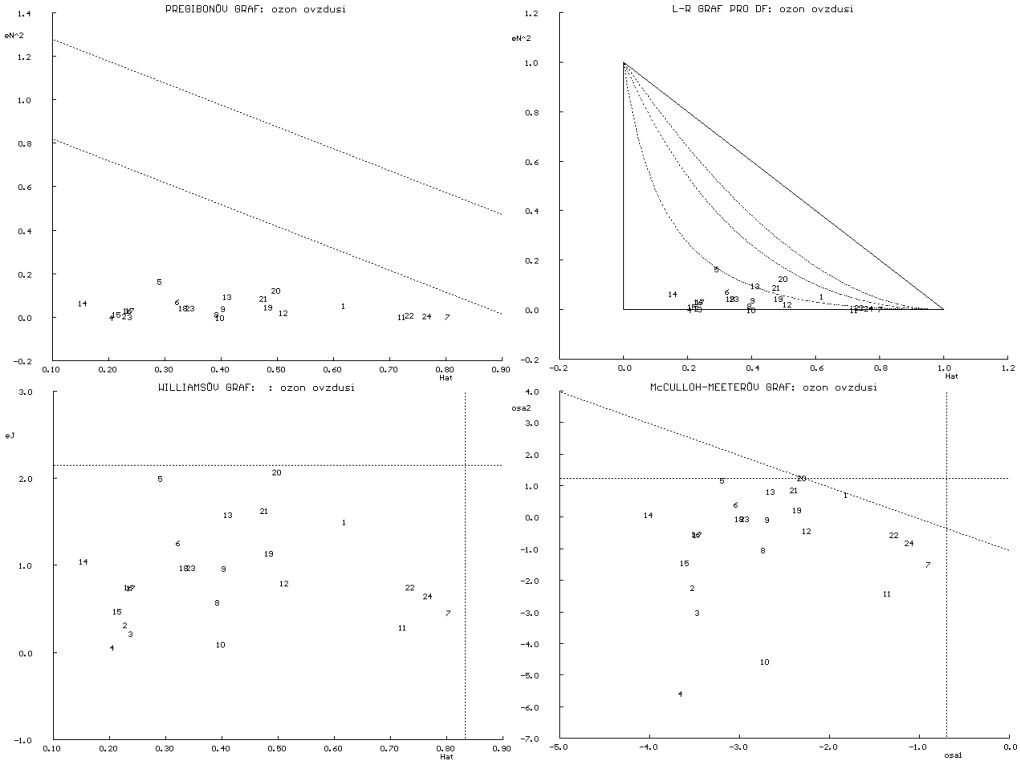
Automatickými analyzátory byly 24 hodin proměřovány na jednom stanovišti škodliviny v ovzduší. Zároveň byly proměřeny i další faktory uvedené v tabulce. Navrhněte regresní model pro závislost koncentrace ozonu na jednotlivých parametrech uvedených v tabulce. Jsou v datech vlivné body? Který z parametrů je významný?

Program: **ADSTAT**
Modul: **Lineární regrese**
Řešení: 1.1.Vyšetření vlivných bodů pomocí diagnostických grafů
1.2.Návrh modelu a odhad parametrů
2.Konstrukce zpřesněného modelu

1.1. Vyšetření vlivných bodů pomocí diagnostických grafů

Program: ADSTAT
Modul: LINEÁRNÍ REGRESE Regresní diagnostika
Název: OZON

ZVOLENÁ STRATEGIE REGRESNÍ ANALÝZY:
Omezení, P : 1.0000E-34
Transformace : Ne
Váhy : Ne
Absolutní člen zahrnut: Ano
PODMÍNKY A KVANTILY PRO STATISTICKÉ TESTY:
Hladina významnosti, alfa : 0.050
Počet bodů, n : 24
Počet parametrů, m : 10
Kvantil Studentova rozdělení t(1-alpha/2,n-m) : 2.145
Kvantil rozd. Chí-kvadrát Chi-square(1-alpha,m) : 18.307



Závěr vyšetření vlivných bodů pomocí diagnostických grafů: V datech se nevyskytují odlehlé body vhodné k odstranění.

1.2. NÁVRH REGRESNÍHO MODELU:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8 + \beta_9 x_9$$

INDIKACE MULTIKOLINEARITY:

Č[j]	Vlastní čísla korel. matice l[j]	Čísla podmíněnosti K[j]	Variance inflation factor VIF[j]	Vícenás.korel koef pro X[j]
1	6.8460E-02	4.1903E+01	8.0254E+00	0.9356
2	1.1097E-01	2.5850E+01	5.1947E+00	0.8986
3	1.1797E-01	2.4318E+01	7.0346E+00	0.9262
4	4.2334E-01	6.7762E+00	1.8539E+00	0.6787
5	7.3418E-01	3.9073E+00	2.3099E+00	0.7530
6	9.9033E-01	2.8967E+00	3.4179E+00	0.8411
7	1.4299E+00	2.0063E+00	2.9280E+00	0.8115
8	2.2562E+00	2.0063E+00	2.9280E+00	0.8115
9	2.8687E+00	1.0000E+00	1.7276E+00	0.6490
Maximální číslo podmíněnosti K 4.1903E+01 (K[j], K > 1000 indikuje silnou multikolinearitu) (VIF[j] > 10 indikuje silnou multikolinearitu)				

Maximální číslo podmíněnosti K = 41.903 je menší než 1000 a VIF[j] jsou menší než 10, proto v datech není indikována silná multikolinearita

ODHADY PARAMETRŮ A TESTY VÝZNAMNOSTI:

Kvantil Studentova rozdělení **t(1-alpha/2,n-m)= 2.145**			Test H0: B[j] = 0 vs. HA: B[j] <> 0		
Parametr	**Odhad**	**Směrodatná odchylka**	**t-kritérium**	**Hypotéza H0**	**Hladina významnosti**
B[0]	-9.0571E+00	9.8549E+00	-9.1905E-01	Akceptována	0.374
B[1]	-4.6100E-01	3.1628E-01	-1.4576E+00	Akceptována	0.167
B[2]	1.4681E-02	1.1681E-02	1.2568E+00	Akceptována	0.229
B[3]	-3.4763E-02	2.4777E-01	-1.4031E-01	Akceptována	0.890
B[4]	5.1627E-01	4.4923E-01	1.1492E+00	Akceptována	0.270
B[5]	-1.3842E-01	4.1903E-02	-3.3033E+00	Zamítnuta	0.005
B[6]	-3.5677E-02	1.2477E-02	-2.8593E+00	Zamítnuta	0.013
B[7]	6.7476E+00	1.1025E+00	6.1205E+00	Zamítnuta	0.000
B[8]	2.1305E+00	1.8905E-01	1.1270E+01	Zamítnuta	0.000
B[9]	3.8329E-01	1.3564E-01	2.8258E+00	Zamítnuta	0.013

STATISTICKÉ CHARAKTERISTIKY REGRESE:

Vícenásobný korelační koeficient, R	9.9377E-01
Koeficient determinace, R^2	9.8758E-01
Predikovaný korelační koeficient, Rp^2	9.7966E-01
Střední kvadratická chyba predikce, MEP	9.7358E+00
Akaikeho informační kritérium, AIC	4.6387E+01

TESTOVÁNÍ REGRESNÍHO TRIPLETU (DATA + MODEL + METODA):

Fisher-Snedocor–v test významnosti regrese,F	: 1.2369E+02
Tabulkový kvantil, F(1-alpha,m-1,n-m)	: 2.6458E+00
Závěr	: Navržený model je přijat jako významný.
Spočtená hladina významnosti	: 0.000
Scottovo kriterium multikolinearity, M	: 6.9981E-01
Závěr	: Navržený model není korektní.
Cook-Weisberg–v test heteroskedasticity, Sf	: 9.5927E+00
Tabulkový kvantil, Chi^2(1-alpha,1)	: 3.8415E+00
Závěr:	: Rezidua vykazují heteroskedasticitu.
Spočtená hladina významnost	: 0.002
Jarque-Berraův test normality reziduí, L(e)	: 9.5675E-01
Tabulkový kvantil, Chi^2(1-alpha,2)	: 5.9915E+00
Závěr	: Normalita je přijata.
Spočtená hladina významnosti	: 0.620
Waldův test autokorelace, Wa	: 7.9684E+00
Tabulkový kvantil, Chi^2(1-alpha,1)	: 3.8415E+00
Závěr	: Rezidua jsou autokorelována.
Spočtená hladina významnosti	: 0.005
Znaménkový test, Dt	: 6.2614E-01
Tabulkový kvantil, N(1-alpha/2)	: 1.6449E+00
Závěr	: Rezidua nevykazují trend.
Spočtená hladina významnosti	: 0.266

Závěr:
Na základě Studentova t -testu (porovnání t – kritéria a kritickou hodnotou t(1-alpha/2,n-m)= 2.145) jsme zjistili, že absolutní člen a směrnice $\beta_1, \beta_2, \beta_3, \beta_4$ jsou statisticky nevýznamné a proto budou odstraněny z modelu.

2. Konstrukce zpřesněného modelu

NÁVRH ZPŘESNĚNÉHO REGRESNÍHO MODELU
$\mathbf{y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5}$

Program:	ADSTAT	
Modul:	LINEÁRNÍ REGRESE	Regresní diagnostika
Název:	OZON1	

ZVOLENÁ STRATEGIE REGRESNÍ ANALÝZY:

Omezení, P	: 1.0000E-34
Transformace	: Ne
Váhy	: Ne
Absolutní člen zahrnut	: **NE**

PODMÍNKY A KVANTILY PRO STATISTICKÉ TESTY:

Hladina významnosti, alfa	: 0.050
Počet bodů, n	: 24
Počet parametrů, m	: 6
Kvantil Studentova rozdělení t(1-alpha/2,n-m)	: 2.093
Kvantil rozd. Chí-kvadrát Chi-square(1-alpha,m)	: 11.070

INDIKACE MULTIKOLINEARITY:

Č[j]	Vlastní čísla korel. matice l[j]	Čísla podmíněnosti K[j]	Variance inflation factor VIF[j]	Vícenás.korel koef pro X[j]
1	2.7538E-02	1.3289E+02	1.1062E+00	0.3099
2	9.6464E-02	3.7935E+01	1.9153E+00	0.6913
3	3.1963E-01	1.1449E+01	1.9332E+00	0.6948
4	8.9701E-01	4.0795E+00	2.6718E+00	0.7910
5	3.6594E+00	1.0000E+00	1.0000E+00	0.0000
Maximální číslo podmíněnosti K		1.3289E+02		
(K[j], K > 1000 indikuje silnou multikolinearitu)				
(VIF[j] > 10 indikuje silnou multikolinearitu)				

ODHADY PARAMETRŮ A TESTY VÝZNAMNOSTI:

Kvantil Studentova rozdělení **t(1-alpha/2,n-m)= 2.093**			Test H0: B[j] = 0 vs. HA: B[j] <> 0		
Parametr	**Odhad**	**Směrodatná odchylka**	**t-kritérium**	**Hypotéza H0**	**Hladina významnosti**
B[0]	2.2337E-08	-----	-----	-----	-----
B[1]	-9.9328E-02	2.8971E-02	-3.4285E+00	Zamítnuta	0.167
B[2]	-3.0278E-02	9.3315E-03	-3.2447E+00	Zamítnuta	0.229
B[3]	6.4155E+00	8.9495E-01	7.1686E+00	Zamítnuta	0.890
B[4]	1.8733E+00	1.2787E-01	1.4649E+01	Zamítnuta	0.270
B[5]	2.0153E-01	3.7917E-02	5.3150E+00	Zamítnuta	0.005

$$y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5$$

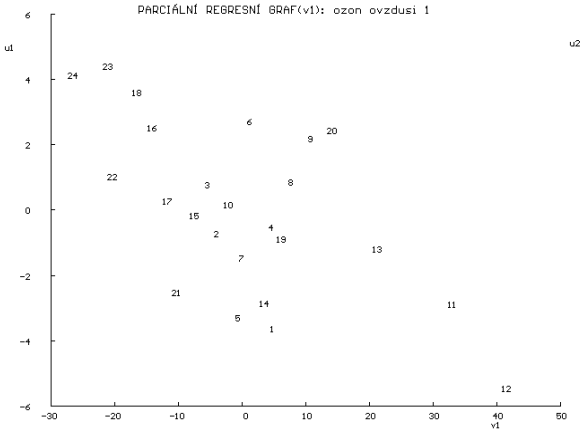
Vyčíslený regresní model má tvar:

$$y = -0{,}09933(\pm 0{,}02897)x_1 - 0{,}03028(\pm 0{,}00933)x_2 + 6{,}416(\pm 0{,}895)x_3 + 1{,}873(\pm 0{,}1279)x_4 + 0{,}2015(\pm 0{,}03792)x_5$$

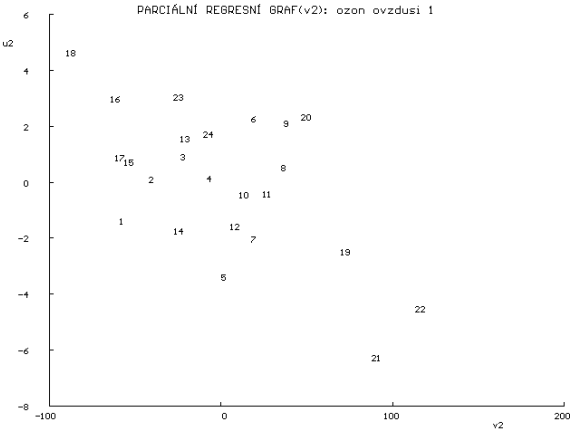
STATISTICKÉ CHARAKTERISTIKY REGRESE:

Vícenásobný korelační koeficient, R	9.9155E-01
Koeficient determinace, R^2	9.8318E-01
Predikovaný korelační koeficient, Rp^2	9.8467E-01
Střední kvadratická chyba predikce, MEP	7.3540E+00
Akaikeho informační kritérium, AIC	4.3669E+01

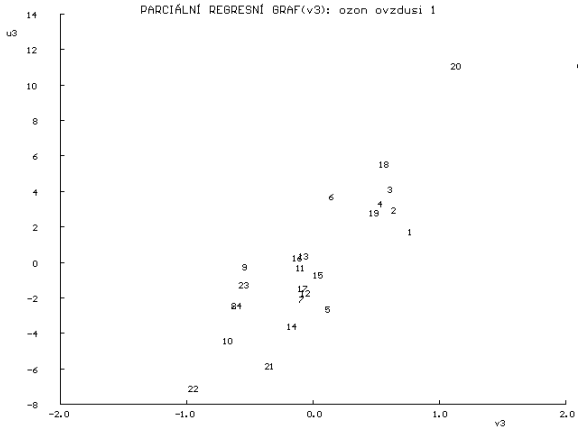
Hodnoty korelačních koeficientů poukazují na to, že model je statisticky významný.
Oproti původnímu modelu se MEP a AIC se snížily a Rp^2 zvýšil, model je proto lepší než původní.



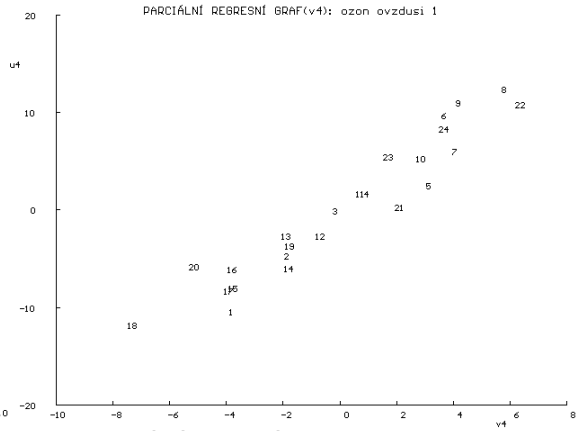
Obr.č.1 Parciální regresní graf pro X_1(Prach PM_{10})



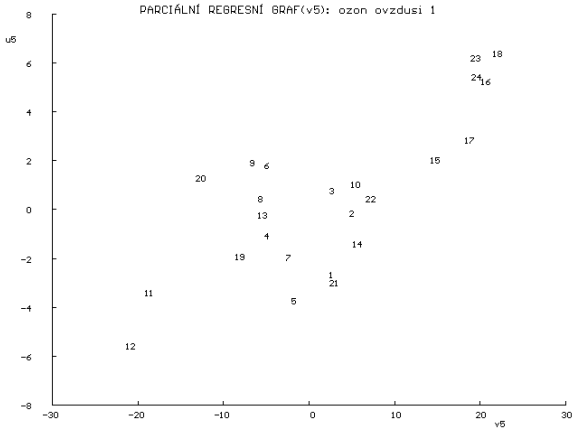
Obr.č.2 Parciální regresní graf pro X_2(Směr větru)



Obr.č.3 Parciální regresní graf pro X_3(Rychlost větru)



Obr.č.4 Parciální regresní graf pro X_4(Teplota)



Obr.č.5 Parciální regresní graf pro X_5(Relativní vlhkost)

TESTOVÁNÍ REGRESNÍHO TRIPLETU (DATA + MODEL + METODA):

Fisher-Snedocor–v test významnosti regrese,F	: 2.7759E+02
Tabulkový kvantil, F(1-alpha,m-1,n-m)	: 2.8951E+00
Závěr	: Navržený model je přijat jako významný.
Spočtená hladina významnosti	: 0.000
Scottovo kriterium multikolinearity, M	: 6.2859E-01
Závěr	: Navržený model není korektní.
Cook-Weisberg–v test heteroskedasticity, Sf	: 2.8918E+01
Tabulkový kvantil, Chi^2(1-alpha,1)	: 3.8415E+00
Závěr:	: Rezidua vykazují heteroskedasticitu.
Spočtená hladina významnost	: 0.000
Jarque-Berraův test normality reziduí, L(e)	: 5.8877E-01
Tabulkový kvantil, Chi^2(1-alpha,2)	: 3.8415E+00
Závěr	: Normalita je přijata.
Spočtená hladina významnosti	: 0.745
Waldův test autokorelace, Wa	: 1.5479E+00
Tabulkový kvantil, Chi^2(1-alpha,1)	: 3.8415E+00
Závěr	: Rezidua nejsou autokorelována.
Spočtená hladina významnosti	: 0.213
Znaménkový test, Dt	: 1.9272E+00
Tabulkový kvantil, N(1-alpha/2)	: 1.6449E+00
Závěr	: Rezidua vykazují trend.
Spočtená hladina významnosti	: 0.027

Závěr:

Byl nalezen lineární regresní model ve tvaru:

$$y = -0{,}09933(\pm 0{,}02897)x_1 - 0{,}03028(\pm 0{,}00933)x_2 + 6{,}416(\pm 0{,}895)x_3 + 1{,}873(\pm 0{,}1279)x_4 + 0{,}2015(\pm 0{,}03792)x_5$$

pro

- y ozon
- X_1 prach PM_{10}
- X_2 směr větru
- X_3 rychlost větru
- X_4 teplota
- X_5 relativní vlhkost

Tento regresní model platí pouze statisticky hodnocená data v této úloze (místo,čas,způsob měření škodlivin automatickými analyzátory).