Univerzita Pardubice
Fakulta chemicko-technologická
Katedra analytické chemie

Tvorba lineárních regresních modelů při analýze dat

Semestrální práce

Licenční studium GALILEO – Interaktivní statistická analýza dat

Brno, 2015

Mgr. Sylvie Pavloková
VFU Brno, Ústav technologie léků

Obsah

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

Zadání:

Úloha se týká stanovení glukózy v kalibračních vzorcích enzymatickou metodou GOD-POD (využívající reakční roztok s obsahem enzymů glukózaoxidázou GOD a peroxidázou POD, kdy se výsledný barevný produkt stanovuje v definovaném čase fotometricky). Záměrem je porovnat lineární regresní modely pro dva časy stanovení a zjistit tak, zda je nutné proměřovat vzorky na spektrofotometru vždy ve stejném časovém rozestupu, či nikoli.

Data:

Měření absorbance kalibračních roztoků o definovaných koncentracích v časech 30 a 60 minut od smíchání s reakčním roztokem (při vlnové délce $\lambda = 500$ nm):

c (mol/l)	*A* (t = 30 min)	*A* (t = 60 min)
1,22	0,183	0,193
2,44	0,308	0,322
4,87	0,564	0,607
7,31	0,852	0,908
9,74	1,097	1,166

Užitý program: QC.Expert 3.2

Řešení:

1. Návrh modelu: Pro stanovení glukózy ve vzorku pomocí enzymatické metody byl navržen přímkový regresní model pro závislost proměnných absorbance po 30 minutách, respektive po 60 minutách, na koncentraci glukózy ve vzorku (mol/l). Rovnice pro navržené modely je: $y = \beta_0 + \beta_1 x$. Po vyšetření regresního tripletu u obou modelů bude testována nulová hypotéza o statistické shodě modelů H_0: $\beta_A = \beta_B$, kde regresní parametry pro první model jsou $\beta_A = (\beta_{1,A}, \beta_{0,A})$ a regresní parametry pro druhý model jsou $\beta_B = (\beta_{1,B}, \beta_{0,B})$.

2. Odhadování parametrů: Pomocí klasické metody nejmenších čtverců (MNČ) byly stanoveny odhady regresních parametrů. Analýza dat probíhala na hladině významnosti $\alpha = 0,05$. Užitím Studentova t-testu bylo zjištěno, že úsek u modelu 1 lze považovat za statisticky nevýznamný, jelikož příslušný interval spolehlivosti zahrnuje i hodnotu 0, zatímco úsek u modelu 2 je statisticky významný. Směrnice lze u obou modelů považovat za statisticky významné.

Proměnná	Odhad	Směr.Odch.	Závěr	Pravděpodobnost	Spodní mez	Horní mez
model 1 (absorbance po 30 minutách A_30)						
Abs	0,0330	0,0164	Nevýznamný	0,1144	-0,0125	0,0785
c (mol/l)	0,1120	0,0022	Významný	9,1714E-007	0,1059	0,1182
model 2 (absorbance po 60 minutách A_60)						
Abs	0,0389	0,0120	Významný	0,0319	0,0055	0,0723
c (mol/l)	0,1179	0,0016	Významný	2,1776E-007	0,1134	0,1224

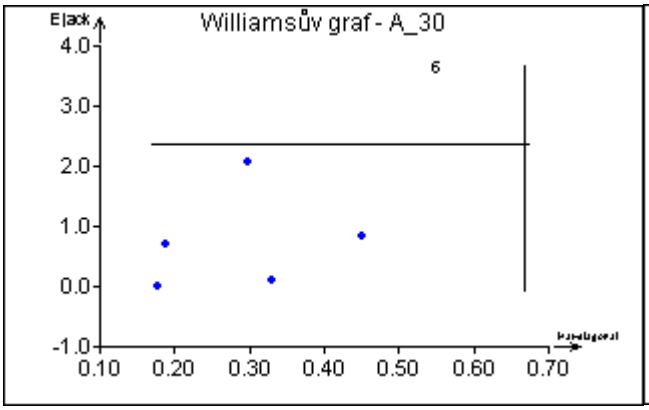
3. Základní statistické charakteristiky: U obou modelů byla zjištěna vysoká hodnota vícenásobného korelačního koeficientu, což svědčí pro významnost modelu. Hodnota koeficientu determinace u obou modelů je také poměrně vysoká, charakterizuje podíl experimentálních bodů vyhovujících navrženému modelu.

parametr	model 1 (A_30)	model 2 (A_60)
Vícenásobný korelační koeficient R	0,9992	0,9996
Koeficient determinace R^2	0,9984	0,9992
Predikovaný korelační koeficient Rp	0,9901	0,9956
Střední kvadratická chyba predikce MEP	0,0009386	0,0004563
Akaikeho informační kritérium	-44,76	-48,46

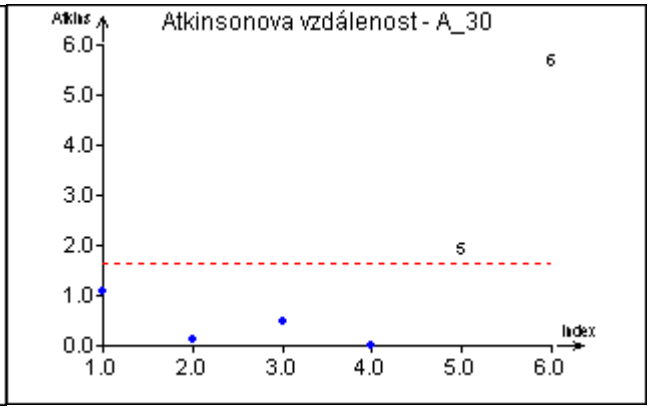
4. Regresní diagnostika – regresní triplet

Kritika dat: V tomto oddílu analýzy regresních dat je důležité identifikovat vlivné body a odlehlé body vyloučit, aby mohl být stanoven zpřesněný model. K tomuto účelu nejlépe slouží grafy znázorňující vlivné body, popřípadě indexové a rankitové grafy.

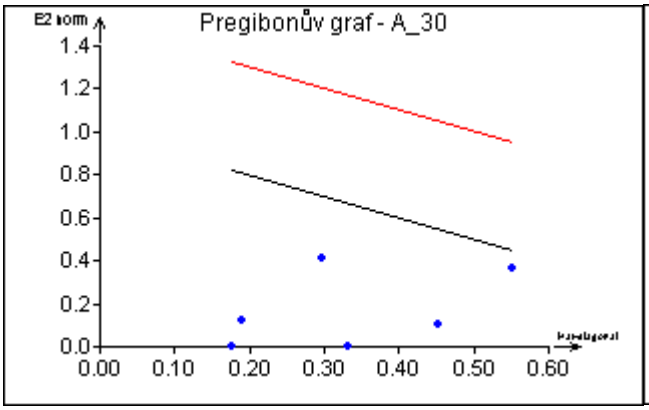
Model 1 (absorbance po 30 minutách)



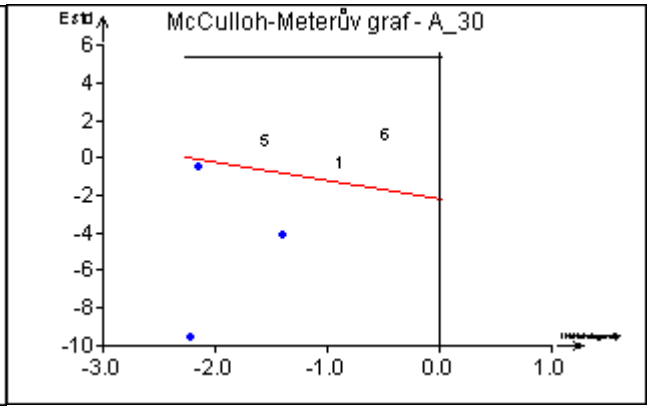
Obr. 1.1 – Williamsův graf (model 1)



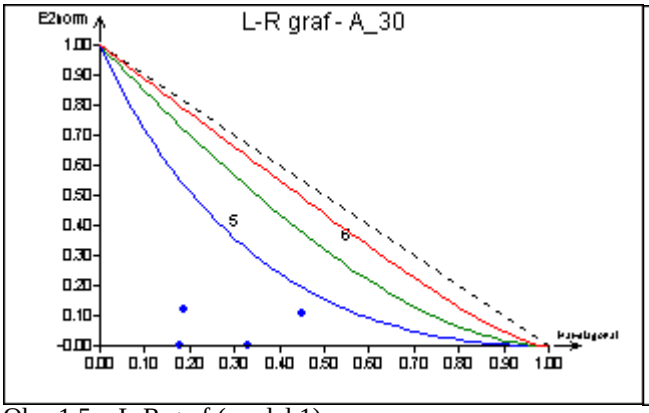
Obr. 1.2 – Atkinsova vzdálenost (model 1)



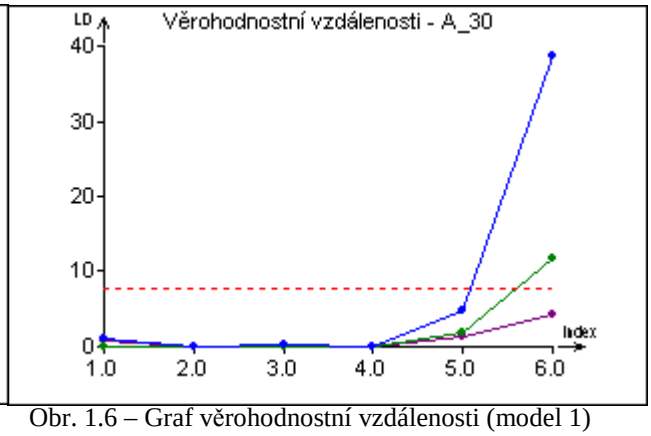
Obr. 1.3 – Pregibonův graf (model 1)



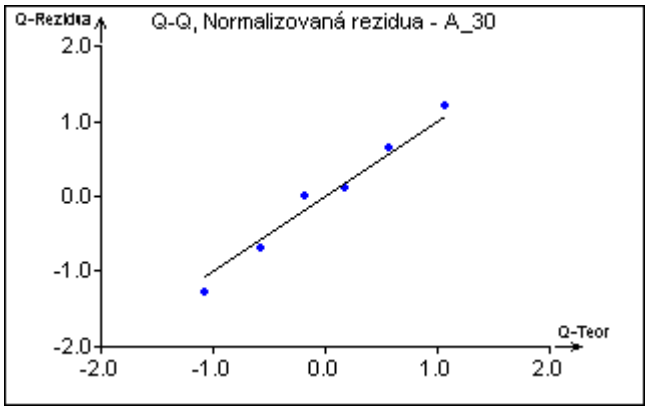
Obr. 1.4 – McCulloh-Meterův graf (model 1)



Obr. 1.5 – L-R graf (model 1)



Obr. 1.6 – Graf věrohodnostní vzdálenosti (model 1)



Obr. 1.7 – Q-Q graf normalizovaných reziduí (model 1)

Williamsův graf (Obr. 1.1) slouží k indikaci vlivných i vybočujících bodů. Zde můžeme vidět odlehlý bod 6, jelikož leží nad vodorovnou přímkou.

Atkinsonova vzdálenost (Obr. 1.2) znázorňuje jako odlehlý bod 6 a 5, jelikož tyto body leží nad vodorovnou přímkou. Bod 5 je zde těsně za hranicí pro odlehlost.

Pregibonův graf (Obr. 1.3) také slouží pro posouzení vlivných bodů. Podle tohoto zobrazení je vidět, že žádný bod nelze s jistotou označit jako vlivný.

McCulloh-Meterův graf (Obr. 1.4) identifikoval body podezřelé na odlehlost. Jsou to body 1, 5 a 6, jelikož leží nad šikmou (červenou) přímkou.

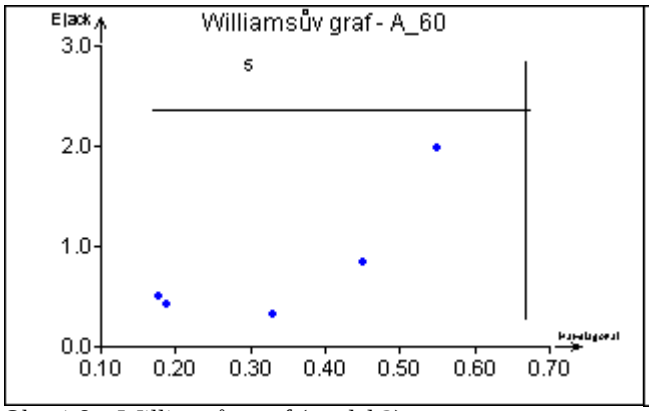
L-R graf (Obr. 1.5) naznačuje, že body 5 a 6 jsou vlivné.

V grafu věrohodnostní vzdálenosti (Obr. 1.6) je vidět, že bod 6 lze označit za silně odlehlý, zatímco bod 5 je spíše podezřelý na odlehlost.

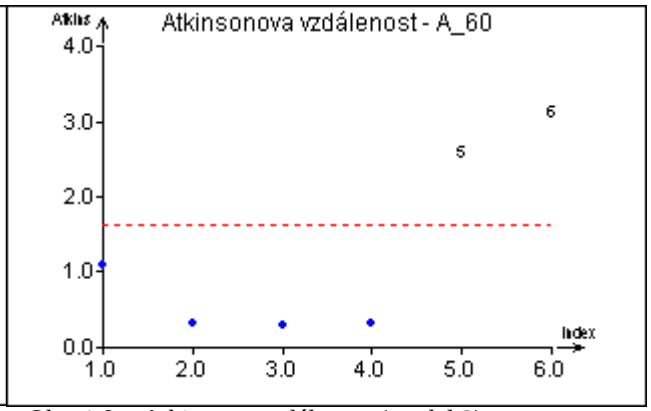
Q-Q graf normalizovaných reziduí (Obr. 1.7) slouží pro posouzení normality reziduí. Body poměrně dobře kopírují přímku, z tohoto grafu se nedá žádný z bodů s jistotou označit za odlehlý. Přímka odpovídající normálnímu rozdělení protíná bod o souřadnicích [0,0] a tento bod lze považovat za střed při proložení experimentálních bodů.

Kritika dat při tvorbě prvního modelu prokázala odlehlost bodu 6, jelikož se na tomto tvrzení shodlo minimálně pět grafických diagnostik. Tento bod bude tedy z datového souboru vyloučen a bude zkonstruován následný zpřesněný model.

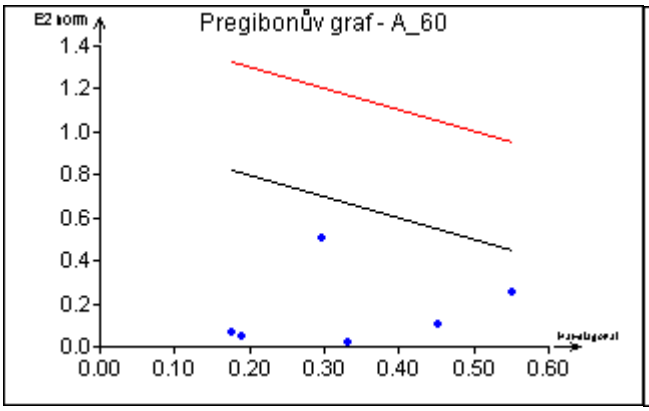
Model 2 (absorbance po 60 minutách)



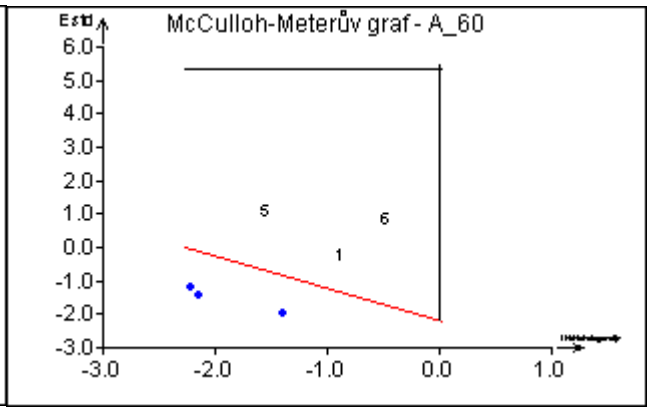
Obr. 1.8 – Williamsův graf (model 2)



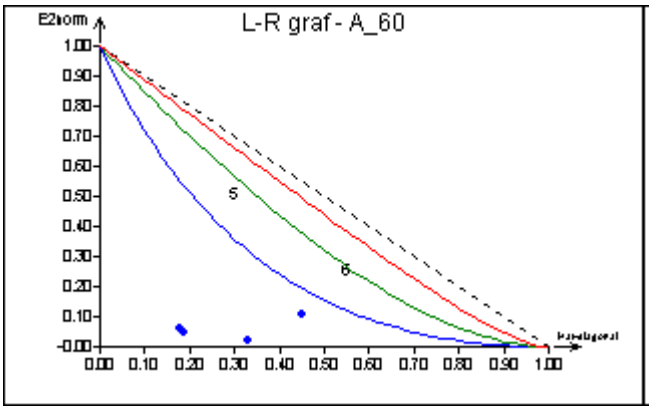
Obr. 1.9 – Atkinsova vzdálenost (model 2)



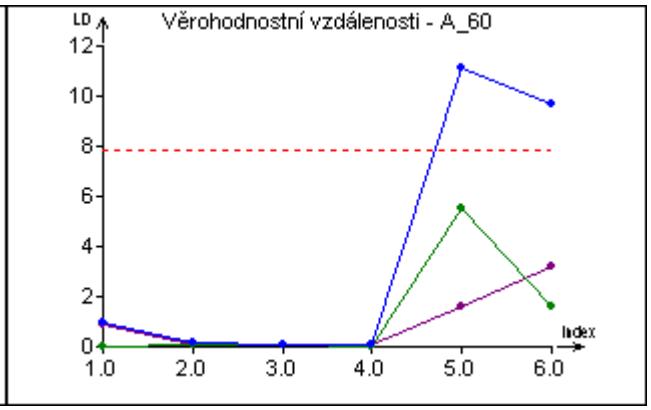
Obr. 1.10 – Pregibonův graf (model 2)



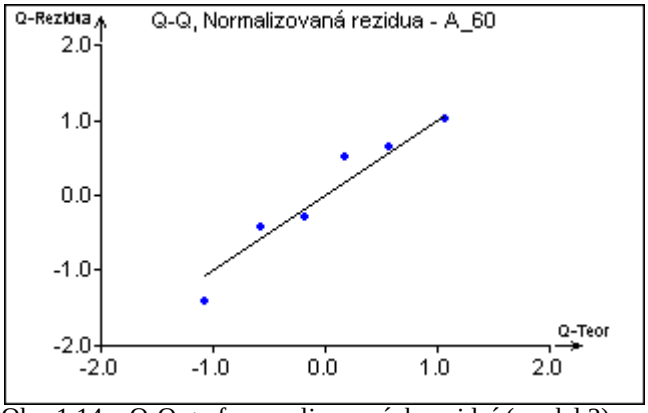
Obr. 1.11 – McCulloh-Meterův graf (model 2)



Obr. 1.12 – L-R graf (model 2)



Obr. 1.13 – Graf věrohodnost. vzdálenosti (model 2)



Obr. 1.14 – Q-Q graf normalizovaných reziduí (model 2)

Podle Williamsova grafu (Obr. 1.8) můžeme vidět odlehlý bod 5, jelikož leží nad vodorovnou přímkou.

Atkinsonova vzdálenost (Obr. 1.9) znázorňuje jako odlehlý bod 6 a 5, jelikož tyto body leží nad vodorovnou přímkou.

Pregibonův graf (Obr. 1.10) neindikuje žádné odlehlé body.

McCulloh-Meterův graf (Obr. 1.11) identifikoval body podezřelé na odlehlost. Jsou to body 1, 5 a 6, jelikož leží nad šikmou (červenou) přímkou.

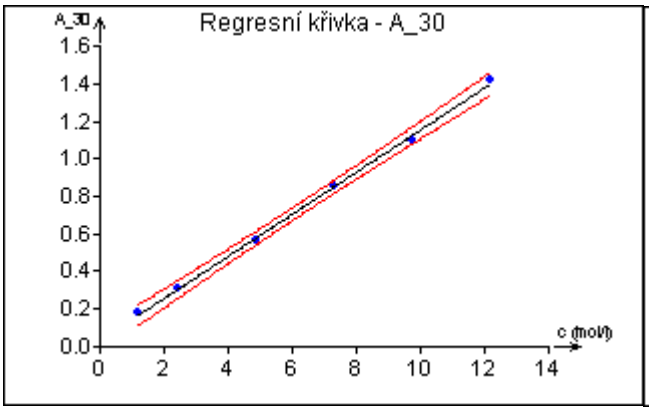
L-R graf (Obr. 1.12) naznačuje, že body 5 a 6 jsou vlivné, bod 5 odlehlý, bod 6 spíše extrém.

V grafu věrohodnostní vzdálenosti (Obr. 1.13) je vidět, že bod 5 a případně 6 lze označit za silně odlehlý.

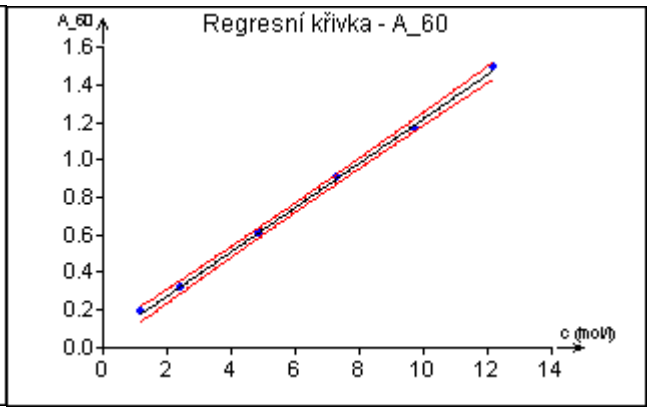
V Q-Q grafu normalizovaných reziduí (Obr. 1.14) body poměrně dobře kopírují přímku, z tohoto grafu se nedá žádný z bodů s jistotou označit za odlehlý. Přímka odpovídající normálnímu rozdělení protíná bod o souřadnicích [0,0] a tento bod lze považovat za střed při proložení experimentálních bodů.

Kritika dat při tvorbě prvního modelu prokázala odlehlost bodu 5, jelikož se na tomto tvrzení shodlo minimálně pět grafických diagnostik. Tento bod bude tedy z datového souboru vyloučen a bude zkonstruován následný zpřesněný model. Další bodem vhodným k vyloučení by byl bod 6.

Kritika modelu: Navržený lineární model pro stanovení glukózy po 30 i 60 minutách lze posoudit jako vhodný, jelikož je již z regresního grafu vidět (Obr. 1.15 a 1.16), že přímková závislost je dodržena.



Obr. 1.15 – Regresní přímka (model 1)



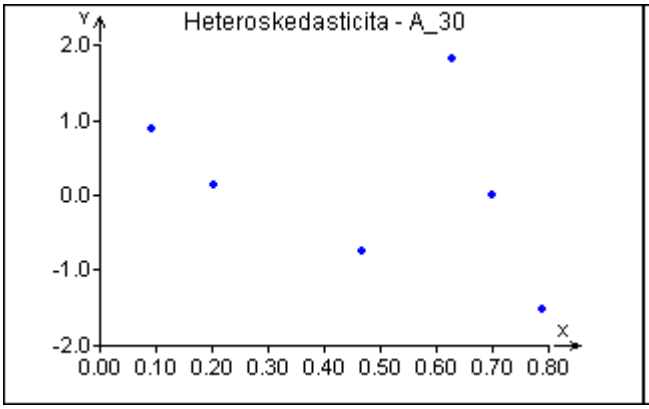
Obr. 1.16 – Regresní přímka (model 2)

Kritika metody: Součástí regresního tripletu je také posouzení splnění základních předpokladů MNČ. Hodnoty jsou vždy uvedeny ve dvou sloupcích (první sloupec pro model 1, druhý sloupec pro model 2).

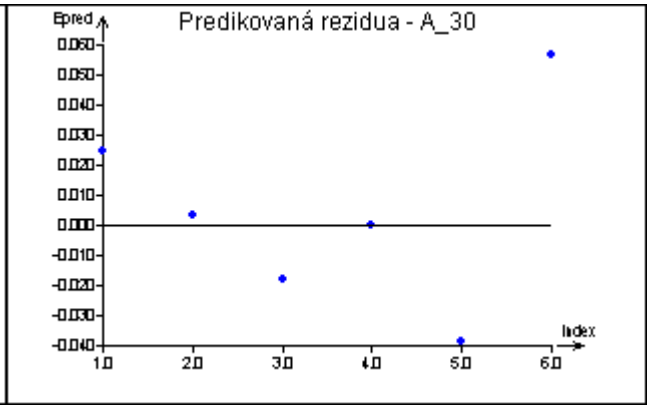
Fisher-Snedecorův test významnosti modelu		
Hodnota kritéria F:	2554,409306	5245,831772
Kvantil F (1-alfa, m-1, n-m):	7,708647422	7,708647422
Pravděpodobnost:	9,171443055E-007	2,17756341E-007
Závěr:	Model je významný	Model je významný
Cook-Weisbergův test heteroskedasticity		
Hodnota kritéria CW:	0,7276076583	0,6113491388
Kvantil Chi^2(1-alfa,1):	3,841458829	3,841458829
Pravděpodobnost:	0,3936597403	0,4342802873
Závěr:	Rezidua vykazují homosk.	Rezidua vykazují homosk.

Jarque-Berrův test normality		
Hodnota kritéria JB:	0,283697371	0,4606458531
Kvantil Chi^2(1-alfa,2):	5,991464547	5,991464547
Pravděpodobnost:	0,8677525501	0,7942770679
Závěr:	Rezidua mají N rozdělení.	Rezidua mají N rozdělení.
Waldův test autokorelace		
Hodnota kritéria WA:	0,6554899031	1,853004159
Kvantil Chi^2(1-alfa,1):	3,841458829	3,841458829
Pravděpodobnost:	0,4181567071	0,1734343968
Závěr:	Autokor. je nevýznamná	Autokor. je nevýznamná
Znaménkový test reziduí		
Hodnota kritéria Sg:	1,944543648	1,369306394
Kvantil N(1-alfa/2):	1,959963999	1,959963999
Pravděpodobnost:	0,05182992722	0,1709035202
Závěr:	V reziduích není trend.	V reziduích není trend.

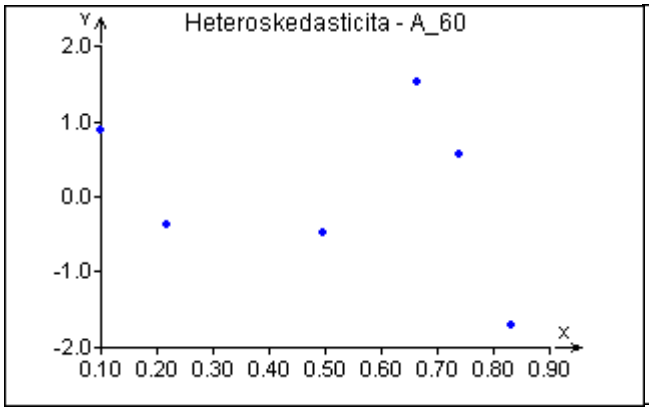
U obou modelů bylo potvrzeno následující: Fisher-Snedecorův test významnosti modelu prokázal, že model může být považován za významný, jelikož hodnota testovacího kritéria F přesáhla kritický kvantil. Redizua vykazují homoskedasticitu (jejich rozptyl je konstantní) podle Cook-Weisbergova testu heteroskadasticity a také podle grafických diagnostik heteroskedasticity (Obr. 1.17 a 1.19). Rovněž bylo potvrzeno normální rozdělení reziduí podle Jarque-Berrova testu normality. Waldův test zjistil statistickou nevýznamnost autokorelace. Znaménkový test potvrdil, že v reziduích není trend, což je patrné také z grafu predikovaných reziduí (Obr. 1.18 a 1.20)



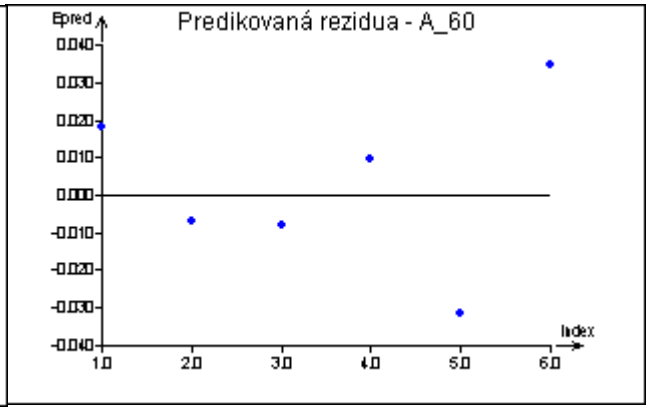
Obr. 1.17 – Graf heteroskedasticity (model 1)



Obr. 1.18 – Graf predikovaných reziduí (model 1)



Obr. 1.19 – Graf heteroskedasticity (model 2)



Obr. 1.20 – Graf predikovaných reziduí (model 2)

5. Konstrukce zpřesněného modelu: Z datového souboru pro model 1 (absorbance po 30 minutách) byl odstraněn bod 6, z datového souboru pro model 2 (absorbance po 60 minutách) byl odstraněn bod 5. Následující zpřesněný regresní model byl sestaven vždy bez tohoto bodu. Hodnota střední kvadratické chyby predikce klesla u obou modelů (pro model 1: MEP = 0,0009386 → 0,0001593; pro model 2: MEP = 0,0004563 → 0,0001671). Akaikeho informační kritérium se u obou modelů nepatrně zvýšilo (pro model 1: AIC = -44,76 → -44,12; pro model 2: AIC = -48,46 → -45,25). Snížení hodnot MEP znamená zkvalitnění modelu s odstraněnými odlehlými body oproti modelu původnímu. To vedlo u modelu 1 také k jinému závěru o významnosti úseku – u zpřesněného modelu je úsek již významný, zatímco u původního modelu byl označen za statisticky nevýznamný. Výhodnost odstranění odlehlých bodů však není jistá, jelikož již na počátku se regresní přímka skládala z nedostatečného počtu bodů (pro dva regresní parametry, směrnici a úsek, by jich mělo do modelu zahrnuto nejméně 10, zde však datový soubor obsahoval pouze 6 bodů). Proto by bylo nejspíše vhodnější v datových souborech pro regresi zachovat všechny body.

Odhadování parametrů:

Proměnná	Odhad	Směr.Odch.	Závěr	Pravděpodobnost	Spodní mez	Horní mez
model 1 (absorbance po 30 minutách A_30), vyloučení bodu 6						
Abs	0,0468	0,0090	Významný	0,0139	0,0181	0,0755
c (mol/l)	0,1083	0,0015	Významný	5,9109E-006	0,1035	0,1131
model 2 (absorbance po 60 minutách A_60), vyloučení bodu 5						
Abs	0,0366	0,0073	Významný	0,0155	0,0132	0,0600
c (mol/l)	0,1191	0,0011	Významný	1,6244E-006	0,1157	0,1226

Základní statistické charakteristiky:

parametr	model 1 (A_30)	model 2 (A_60)
Vícenásobný korelační koeficient R	0,9997	0,9999
Koeficient determinace R^2	0,9994	0,9998
Predikovaný korelační koeficient Rp	0,9972	0,9985
Střední kvadratická chyba predikce MEP	0,0001593	0,0001671
Akaikeho informační kritérium	-44,12	-45,25

Kritika metody: U obou zpřesněných modelů byly zachovány závěry testů ve stejném znění jako u modelů původních. Fisher-Snedecorův test významnosti modelu prokázal, že model je významný. Cook-Weisbergův test heteroskedasticity zjistil, že rezidua mají konstantní rozptyl. Jarque-Berrův test normality potvrdil normalitu reziduí. Waldův test autokorelace nezjistil statisticky významnou autokorelaci. Znaménkový test reziduí potvrdil, že v reziduích se nevyskytuje trend.

6. Společný model: Byl zkonstruován společný model pro všechny naměřené body (bez vyloučení odlehlých bodů). Byla prokázána významnost modelu; dále významnost směrnice a nevýznamnost úseku. Rezidua vykazují homoskedasticitu a normalitu. Autokorelace je nevýznamná. V reziduích byl odhalen trend.

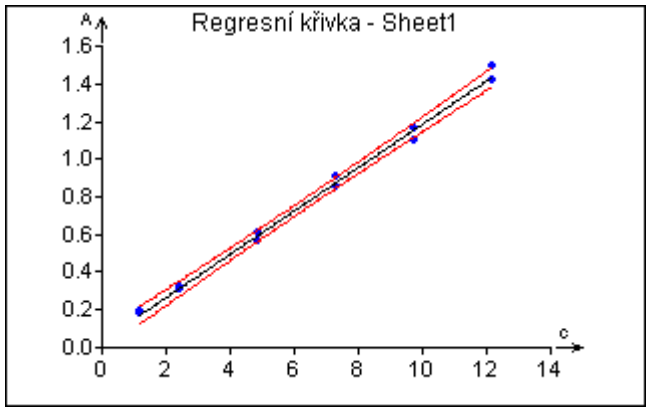
Odhadování parametrů:

Proměnná	Odhad	Směr.Odch.	Závěr	Pravděpodobnost	Spodní mez	Horní mez
Abs	0,0359	0,0173	Nevýznamný	0,0642	-0,0026	0,0745
c	0,1150	0,0023	Významný	2,9310E-013	0,1098	0,1202

Základní statistické charakteristiky:

parametr	společný model
Vícenásobný korelační koeficient R	0,9979
Koeficient determinace R^2	0,9959
Predikovaný korelační koeficient Rp	0,9876
Střední kvadratická chyba predikce MEP	0,0012431
Akaikeho informační kritérium	-81,24

Kritika modelu:



Obr. 1.21 – Regresní přímka

Kritika metody:

Fisher-Snedecorův test významnosti modelu
Hodnota kritéria F: 2416,59939
Kvantil F (1-alfa, m-1, n-m): 4,964602744
Pravděpodobnost: 2,929929842E-013
Závěr: Model je významný

Cook-Weisbergův test heteroskedasticity
Hodnota kritéria CW: 2,365792058
Kvantil Chi^2(1-alfa,1): 3,841458829
Pravděpodobnost: 0,1240209516
Závěr: Rezidua vykazují homoskedasticitu.

Jarque-Berrův test normality
Hodnota kritéria JB: 0,06791020093
Kvantil Chi^2(1-alfa,2): 5,991464547
Pravděpodobnost: 0,9666149042
Závěr: Rezidua mají normální rozdělení.

Waldův test autokorelace
Hodnota kritéria WA: 3,138545711
Kvantil Chi^2(1-alfa,1): 3,841458829
Pravděpodobnost: 0,07646195718
Závěr: Autokorelace je nevýznamná

Znaménkový test reziduí
Hodnota kritéria Sg: 2,914884513
Kvantil N(1-alfa/2): 1,959963999
Pravděpodobnost: 0,003558201855
Závěr: V reziduích je trend!

7. Test shody rozptylů: Pro testování shody rozptylů byl použit Fisher-Snedecorův F-test. Toto testování prokázalo, že data získaná proměřením kalibračních roztoků glukózy po 30 a 60 minutách mají shodné rozptyly.

Porovnání dvou výběrů		
Porovnávané sloupce:	A_30	A_60
Počet dat:	6	6
Průměr:	0,73785	0,7809333333
Směr. odchylka:	0,4762366418	0,5011358485
Rozptyl:	0,226801339	0,2511371387
Test shody rozptylů		
Poměr rozptylů:	1,107300071	
Kritická hodnota:	5,050329058	
Závěr:	Rozptyly jsou SHODNÉ	
Pravděpodobnost:	0,4568352382	

8. Porovnání lineárních modelů: Jelikož byla prokázána shoda rozptylů u dvou kalibračních měření, lze použít Chowův test pro porovnání regresních přímek s kritickou hodnotou testu F (m; n – 2m). Z důvodu nízkého počtu bodů zařazených do kalibračního měření nebyl pro toto porovnání vyloučen žádný bod. Grafické diagnostiky odlehlých bodů sice u obou souborů odhalily jeden potenciálně odlehlý bod, ale např. také podle grafů regresní přímky (Obr. 1.15 a 1.16) můžeme usuzovat, že tyto body nejsou značně odlehlé a pro zajištění stability testu bude lepší je z modelu nevyřazovat. Dále u modelu 1 a u společného modelu byla zjištěna statistická nevýznamnost úseku, proto byly tyto modely zkonstruovány znovu bez zahrnutí absolutního členu.

Proměnná	Odhad	Směr.Odch.	Závěr	Pravděpodobnost	Spodní mez	Horní mez
model 1 (absorbance po 30 minutách A_30)						
c (mol/l)	0,1158	0,0015	Významný	6,3649E-009	0,1120	0,1196
model 2 (absorbance po 60 minutách A_60)						
Abs	0,0389	0,0120	Významný	0,0319	0,0055	0,0723
c (mol/l)	0,1179	0,0016	Významný	2,1776E-007	0,1134	0,1224
společný model						
c	0,1191	0,0014	Významný	0	0,1160	0,1222

parametr	model 1 (A_30)	model 2 (A_60)	společný model
Vícenásobný korelační koeficient R	0,9984	0,9996	0,9970
Koeficient determinace R^2	0,9969	0,9992	0,9941
Predikovaný korelační koeficient Rp	0,9914	0,9956	0,9853
Střední kvadratická chyba predikce MEP	0,000819	0,0004563	0,001471
Akaikeho informační kritérium	-42,56	-48,46	-78,92

Stanovení testačního kritéria F_C:

$$F_C = \frac{(RSC - RSC_1 - RSC_2) \times (n - 2m)}{(RSC_1 + RSC_2) \times (m)},$$

kde *n* je celkový počet bodů ve společném modelu,

m je počet porovnávaných lineárních modelů,

RSC je reziduální součet čtverců (*RSC* pro společný model, RSC_1 pro první kalibraci, RSC_2 pro druhou kalibraci).

Reziduální součet čtverců pro jednotlivé modely:

RSC = 0,01413955082

RSC_1 = 0,003569621137

RSC_2 = 0,0009567435666

Vyčíslení testačního kritéria:

$$\boldsymbol{F_C} = \frac{(0{,}01414 - 0{,}00357 - 0{,}00096)\times(12-2\times2)}{(0{,}00357 + 0{,}00096)\times(2)} = \mathbf{8{,}49}$$

Vyčíslení kritické hodnoty:

$\boldsymbol{F_{0,95}(m;\ n-2m)} = F_{0,95}(2;8) = \mathbf{4{,}46}$

$F_C > F_{0,95}(m;\ n-2m)$ H_0 zamítnuta

Nalezené regresní modely mají následující tvar (v závorkách udány hodnoty příslušných směrodatných odchylek):

model 1: **y = 0,1158 (0,0015) x**

model 2: **y = 0,0389 (0,0120) + 0,1179 (0,0016) x**

Závěr: Bylo provedeno měření pro porovnání kalibračních křivek u stanovení glukózy pomocí enzymatické metody. Kalibrační roztoky byly měřeny 30 a 60 minut po přidání reakčního činidla a bylo snahou zjistit, zda má čas měření vliv na odezvu.

Nejprve byla u obou kalibračních měření provedena analýza regresního tripletu, které poukázala na fakt, že v každém modelu se vyskytuje jeden potenciálně odlehlý bod. Jelikož je však doporučováno na stanovení dvou regresních parametrů použít minimálně deset experimentálních bodů, žádný bod nebyl pro porovnání těchto dvou modelů vyloučen.

U obou modelů byla prokázána významnost, homoskedasticita a normalita reziduí, nepřítomnost autokorelace a trendu v reziduích. U modelu 1 byl zanedbán úseku z důvodu jeho nevýznamnosti.

Na porovnání dvou lineárních modelů byl použit Chowův test. Nejprve byla potvrzena shoda rozptylů obou modelů, podle toho bylo zvoleno kritérium Chowova testu. Testační statistika však nabývala hodnoty vyšší, než kritická hodnota (8,49 > 4,46), proto byla nulová hypotéza o nevýznamnosti rozdílu mezi regresními modely zamítnuta.

Enzymatická metoda stanovení glukózy ve vzorku tedy dává statisticky významně odlišnou odezvu vzhledem k času od smíchání vzorku s reakčním roztokem (30 a 60 minut). Je tedy vhodné zjišťovat absorbanci roztoků glukózy pro všechny vzorky vždy ve stejný čas, aby nedocházelo k zanášení chyb do měření a následného vyhodnocení.

Úloha 2. Určení stupně polynomu

Zadání: Záměrem této úlohy je naleznout vhodný lineární regresní model pro stanovení koncentrace železa v řece v závislosti na vzdálenosti místa odběru vzorku od ústí řeky. Předpokládá se polynomická regresní závislost. Analýzu dat proveďte pomocí MNČ, případně RH křivkové závislosti. (J. H. Zar, Biostatistical Analysis, 5th ed., p. 459)

Data:

Stanovení koncentrace železa v řece *c* v závislosti na vzdálenosti odběrového místa od ústí řeky *d*:

d (km)	1,22	1,34	1,51	1,66	1,72	1,93	2,14	2,39	2,51	2,78
c (µg/l)	40,9	41,8	42,4	43	43,4	43,9	44,3	44,7	45	45,1
d (km)	2,97	3,17	3,32	3,5	3,53	3,85	3,95	4,11	4,18	
c (µg/l)	45,4	46,2	47	48,6	49	49,7	50	50,8	51,5	

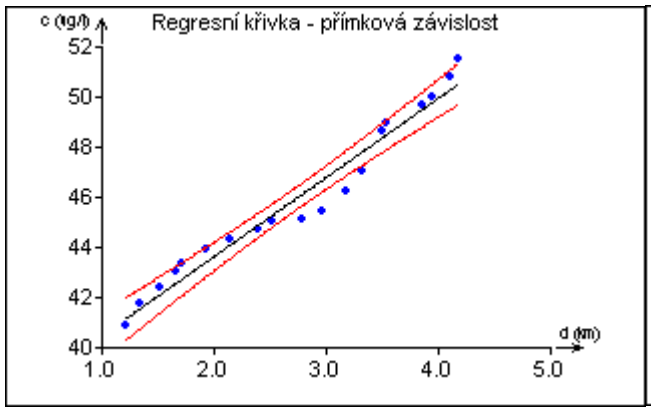
Užitý program: QC.Expert 3.2

Řešení:

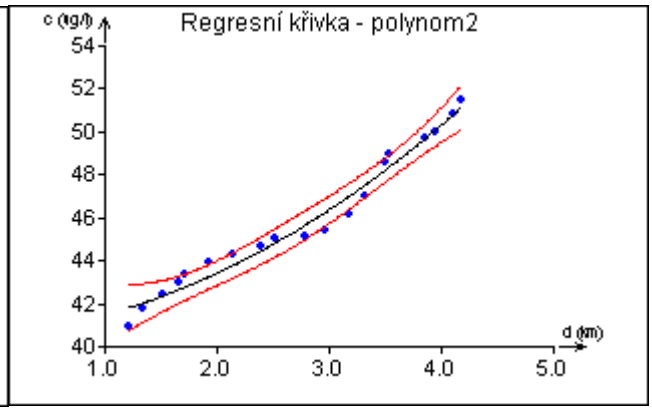
1. Návrh modelu: Pro závislost koncentrace železa v řece na místě odběru je nutné stanovit vhodný regresní model. Testována bude přímková závislost a potom několik stupňů polynomu. O vhodném stupni polynomu rozhodnou hodnoty střední kvadratické chyby predikce (MEP), Akaikova informačního kritéria (AIC) a reziduální směrodatné odchylky (reziduální SD).

model	stupeň pol.	MEP	AIC	reziduální SD
přímkový	-	0,4756	-13,75	0,6627
polynom	2	0,3768	-19,99	0,5499
polynom	3	0,2545	-26,64	0,4523
polynom	4	0,1680	-35,07	0,3558
polynom	5	0,2155	-34,15	0,3589
polynom	6	0,2778	-37,59	0,3237

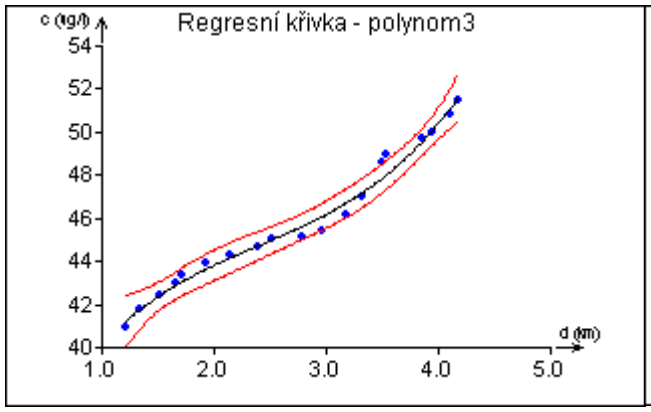
Jak je vidět z tabulky, hodnota MEP postupně klesala s rostoucí složitostí modelu, nejnižší je pro polynom čtvrtého stupně. S dalším zvyšováním stupně polynomu MEP opět začíná poměrně značně narůstat. Hodnota AIC se snižuje s rostoucí složitostí modelu také až po čtvrtý stupeň polynomu, pro pátý stupeň je mírně vyšší a pro šestý stupeň nejnižší ze všech zkoumaných modelů. Reziduální směrodatná odchylka v podstatě klesá přes všechny sledované modely s rostoucím stupněm polynomu.
Regresní závislosti pro všechny sledované modely jsou znázorněny na grafech (Obr. 2.1 – 2.6). Jak je vidět, polynom čtvrtého stupně poskytuje dostatečné proložení, u modelů pro polynomy pátého a šestého stupně již lze pozorovat rozšíření pásů spolehlivosti na obou koncích regresní závislosti, což je nežádoucí.



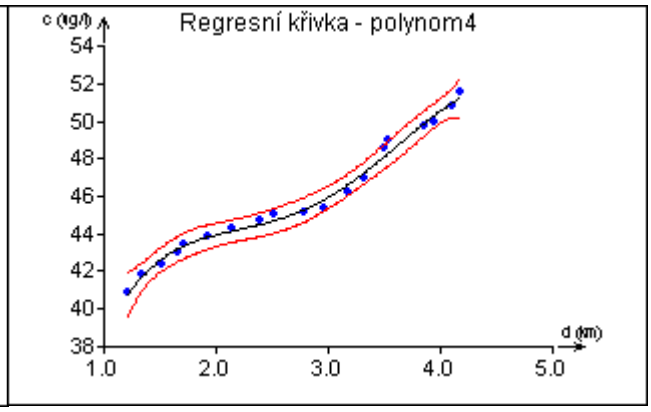
Obr. 2.1 – Regresní závislost (přímka)



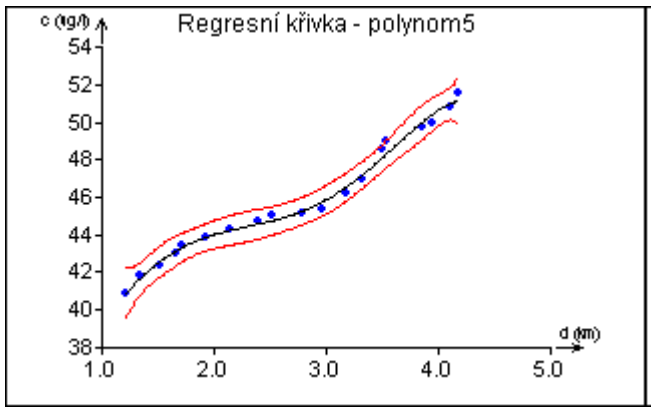
Obr. 2.2 – Regresní závislost (polynom stupně 2)



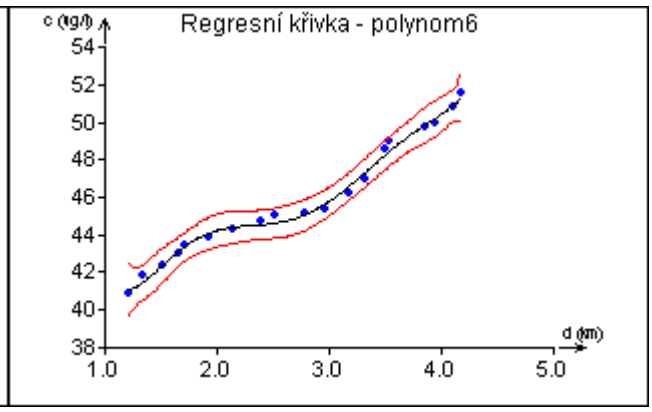
Obr. 2.3 – Regresní závislost (polynom stupně 3)



Obr. 2.4 – Regresní závislost (polynom stupně 4)



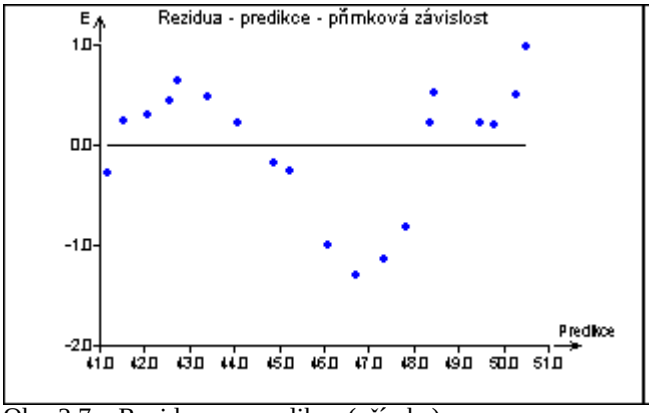
Obr. 2.5 – Regresní závislost (polynom stupně 5)



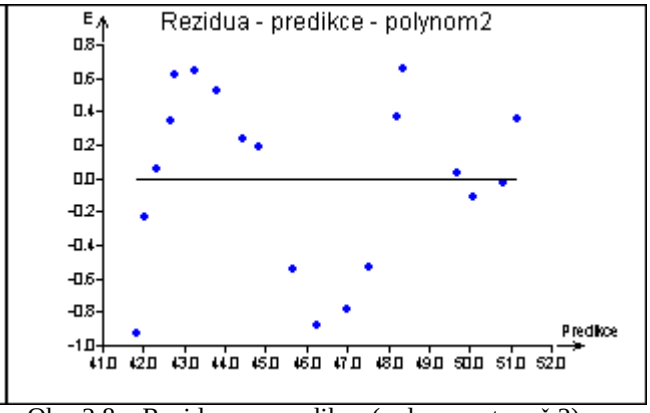
Obr. 2.6 – Regresní závislost (polynom stupně 6)

Také byla posuzována grafická znázornění rezidua versus predikce. Je vidět, že se zvyšujícím se stupněm polynomu tvoří body v tomto grafu více náhodný mrak. U přímkové lineární regrese je patrná jasná závislost zobrazených bodů, méně potom u polynomu druhého a polynomu třetího stupně. U polynomu čtvrtého stupně jsou již body rozmístěny na první pohled náhodně, což značí správnost modelu.

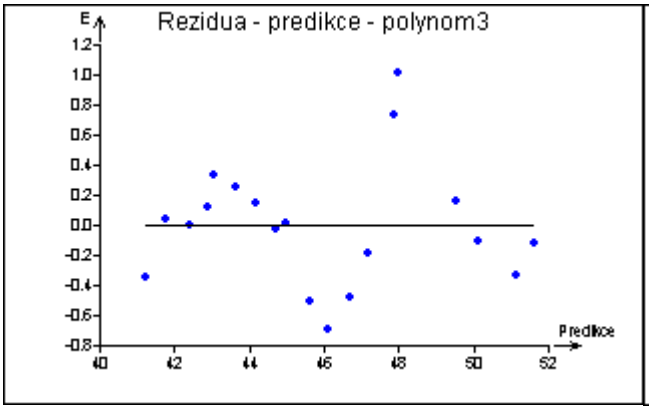
Z těchto důvodů byl vybrán polynom čtvrtého stupně jako vhodný model pro stanovení železa v řece v závislosti na odběrovém místě.



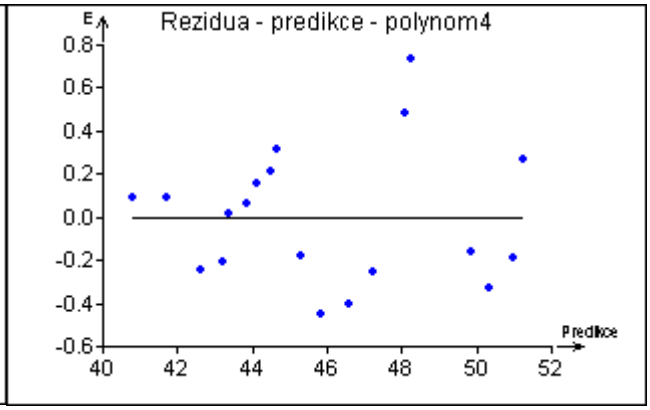
Obr. 2.7 – Rezidua vs. predikce (přímka)



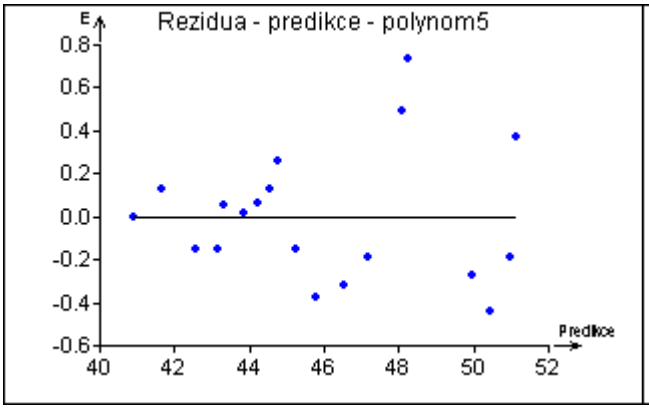
Obr. 2.8 – Rezidua vs. predikce (polynom stupně 2)



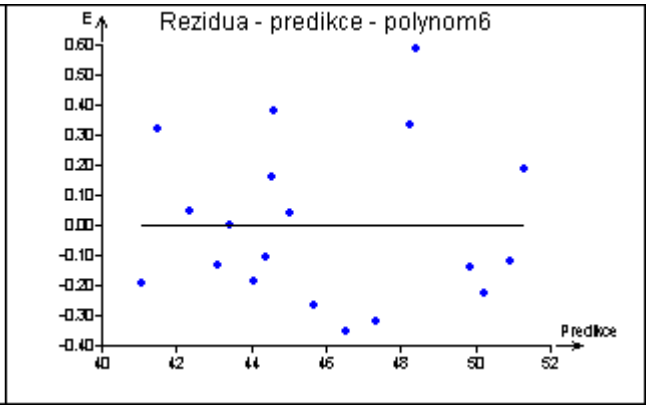
Obr. 2.9 – Rezidua vs. predikce (polynom stupně 3)



Obr. 2.10 – Rezidua vs. predikce (polynom stupně 4)



Obr. 2.11 – Rezidua vs. predikce (polynom stupně 5)

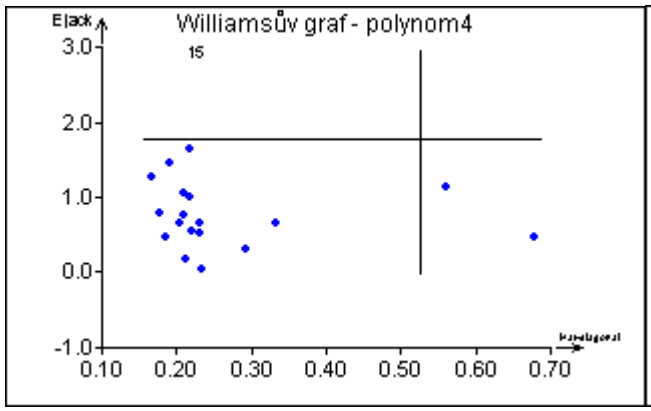


Obr. 2.12 – Rezidua vs. predikce (polynom stupně 6)

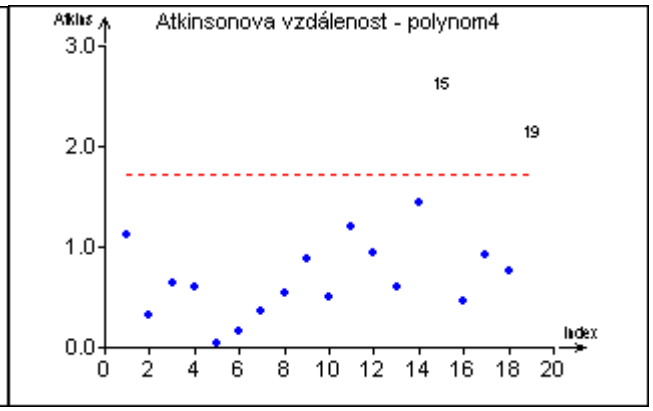
2. Odhadování parametrů: Pomocí klasické metody nejmenších čtverců (MNČ) byly stanoveny odhady regresních parametrů. Analýza dat probíhala na hladině významnosti $\alpha = 0{,}05$. Užitím Studentova t-testu bylo zjištěno, že úsek lze považovat za statisticky nevýznamný, zatímco ostatní regresní parametry jsou statisticky významné.

Proměnná	Odhad	Směr.Odch.	Závěr	Pravděpodobnost	Spodní mez	Horní mez
Abs	9,0072	7,5955	Nevýznamný	0,2554	-7,2835	25,2978
d (km)	51,945	13,0261	Významný	0,0013	24,0071	79,884
d (km)^2	-28,9085	7,9290	Významný	0,0026	-45,9146	-11,9025
d (km)^3	7,0424	2,0406	Významný	0,0039	2,6658	11,4191
d (km)^4	-0,6032	0,1885	Významný	0,0064	-1,0074	-0,1990

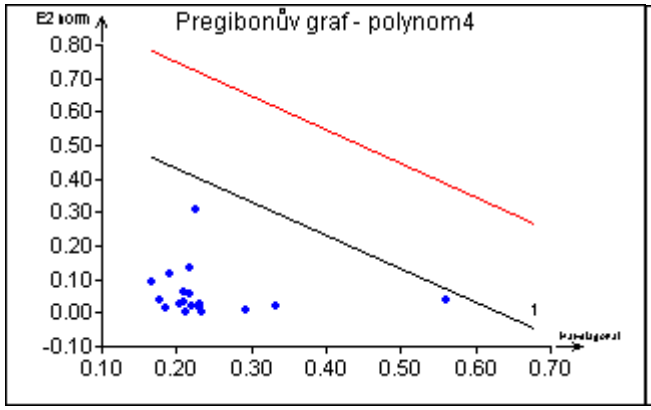
3. Kritika dat: V této části analýzy regresních dat je především důležité identifikovat vlivné body a odlehlé body vyloučit, aby mohl být stanoven zpřesněný model.



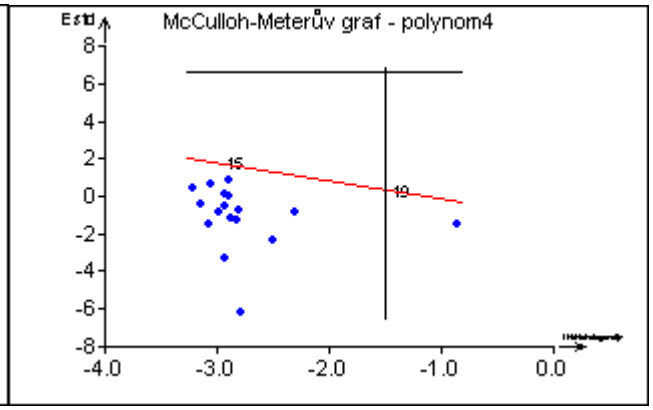
Obr. 2.13 – Williamsův graf



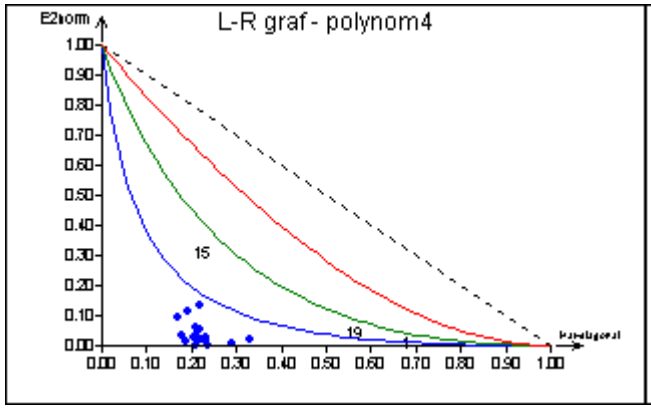
Obr. 2.14 – Atkinsova vzdálenost



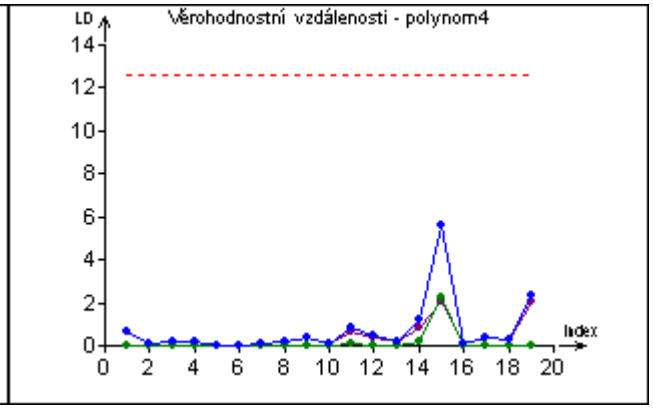
Obr. 2.15 – Pregibonův graf



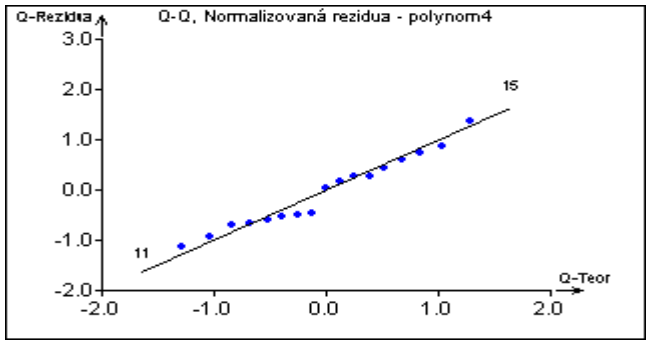
Obr. 2.16 – McCulloh-Meterův graf



Obr. 2.17 – L-R graf



Obr. 2.18 – Graf věrohodnost. vzdálenosti



Obr. 2.19 – Q-Q graf normalizovaných reziduí

Podle Williamsova grafu (Obr. 2.13) můžeme vidět odlehlý bod 15, jelikož leží nad vodorovnou přímkou.

Atkinsonova vzdálenost (Obr. 2.14) znázorňuje jako odlehlý bod 15 a 19, jelikož tyto body leží nad vodorovnou přímkou.

Pregibonův graf (Obr. 1.10) ukázal jako středně vlivný bod 1.

McCulloh-Meterův graf (Obr. 2.15) identifikoval body podezřelé na odlehlost. Jsou to body 15 a 19, jelikož leží přesně na šikmé (červené) přímce.

L-R graf (Obr. 2.16) naznačuje, že bod 15 je odlehlý, zatímco body 19 a 1 jsou spíše extrémy.

V grafu věrohodnostní vzdálenosti (Obr. 2.17) je vidět, že bod 15 je podezřelý na odlehlost.

V Q-Q grafu normalizovaných reziduí (Obr. 2.18) body poměrně dobře kopírují přímku, na koncích závislosti se mírně odchylují pouze body 11 a 15.

Kritika dat při tvorbě prvního modelu prokázala odlehlost bodu 15, jelikož se na tomto tvrzení shodlo minimálně pět grafických diagnostik. Tento bod bude tedy z datového souboru vyloučen a bude zkonstruován následný zpřesněný model.

4. Konstrukce zpřesněného modelu:

Odhadování parametrů (model po vyloučení odlehlého bodu 15):

Proměnná	Odhad	Směr.Odch.	Závěr	Pravděpodobnost	Spodní mez	Horní mez
Abs	12,2234	6,2335	Nevýznamný	0,0717	-1,2432	25,6900
d (km)	45,7540	10,7340	Významný	0,0009	22,5647	68,9433
d (km)^2	-24,6912	6,5661	Významný	0,0024	-38,8763	-10,5060
d (km)^3	5,8412	1,6991	Významný	0,0044	2,1705	9,5118
d (km)^4	-0,4827	0,1577	Významný	0,0091	-0,8235	-0,1419

Odhadování parametrů (model po vyloučení odlehlého bodu 15 a zanedbání úseku):

Proměnná	Odhad	Směr.Odch.	Závěr	Pravděpodobnost	Spodní mez	Horní mez
d (km)	66,7083	1,1129	Významný	0	64,3214	69,0953
d (km)^2	-37,3566	1,2963	Významný	7,2609E-014	-40,1368	-34,5763
d (km)^3	9,0636	0,4736	Významný	1,9506E-011	8,0479	10,0793
d (km)^4	-0,7760	0,0548	Významný	1,0974E-009	-0,8937	-0,6584

Základní statistické charakteristiky:

parametr	model s úsekem	model s úsekem (vyloučen bod 15)	model bez úseku (vyloučen bod 15)
Vícenásobný korelační koeficient R	0,9952	0,9969	0,9960
Koeficient determinace R^2	0,9903	0,9938	0,9920
Predikovaný korelační koeficient Rp	0,9655	0,9763	0,9700
Střední kvadratická chyba predikce MEP	0,1680	0,1150	0,1453
Akaikeho informační kritérium	-35,07	-40,75	-38,09

Bylo prokázáno, že po vyloučení bodu 15 vykazuje model snížení hodnoty MEP i AIC. Jelikož byla i po této proceduře zjištěna statistická nevýznamnost úseku, byl stanoven model bez úseku. U tohoto výsledného modelu se sice nepatrně zvýšila hodnota jak MEP, tak i AIC, přesto byl však tento model vybrán jako nejlepší možný.

Kritika metody: Součástí regresního tripletu je také posouzení splnění základních předpokladů MNČ.

Fisher-Snedecorův test významnosti modelu	
Hodnota kritéria F:	576,7161602
Kvantil F (1-alfa, m-1, n-m):	3,343888678
Pravděpodobnost:	6,722024896E-015
Závěr:	Model je významný
Scottovo kritérium multikolinearity	
Hodnota kritéria SC:	0,1065218334
Závěr:	Model je korektní.
Cook-Weisbergův test heteroskedasticity	
Hodnota kritéria CW:	0,01255089064
Kvantil Chi^2(1-alfa,1):	3,841458829
Pravděpodobnost:	0,9107990191
Závěr:	Rezidua vykazují homoskedasticitu.
Jarque-Berrův test normality	
Hodnota kritéria JB :	1,53517616
Kvantil Chi^2(1-alfa,2) :	5,991464547
Pravděpodobnost:	0,4641311665
Závěr:	Rezidua mají normální rozdělení.
Waldův test autokorelace	
Hodnota kritéria WA:	0,5043904664
Kvantil Chi^2(1-alfa,1):	3,841458829
Pravděpodobnost:	0,4775773156
Závěr:	Autokorelace je nevýznamná.
Znaménkový test reziduí	
Hodnota kritéria Sg :	1,054135332
Kvantil N(1-alfa/2):	1,959963999
Pravděpodobnost:	0,2918209619
Závěr:	V reziduích není trend.

Fisher-Snedecorův test prokázal významnost modelu, protože hodnota testovacího kritéria F přesáhla kritický kvantil. Redizua vykazují homoskedasticitu podle Cook-Weisbergova testu heteroskadasticity. Rovněž bylo potvrzeno normální rozdělení reziduí podle Jarque-Berrova testu normality a nepřítomnost trendu v reziduích podle znaménkového testu. Podle Waldova testu byla autokorelace označena jako nevýznamná. Především bylo potvrzeno, že model je korektní, tudíž nevykazuje multikolinearitu podle Scottova kritéria, a není proto nutné použít metodu racionálních hodností (RH), která jinak slouží ke korekci multikolinearity.

5. Zhodnocení kvality modelu

Nalezený regresní model polynomické závislosti má tvar (odhad směrodatné odchylky parametru uveden v závorce):

$$\mathbf{c = 66{,}7\ (1{,}1)\ d - 37{,}4\ (1{,}3)\ d^2 + 9{,}1\ (0{,}5)\ d^3 - 0{,}8\ (0{,}1)\ d^4}$$

Závěr: Byl stanoven **polynomický regresní model pro závislost koncentrace železa v řece na místě odběru**. Pomocí hodnot MEP, AIC a reziduální směrodatné odchylky byl vybrán vhodný stupeň polynomu (**polynom čtvrtého stupně**) a pomocí kritiky dat byl z datového souboru vyloučen jeden bod. Dále byla zjištěna statistická nevýznamnost úseku. Výsledný model splňoval všechny předpoklady pro použití MNČ, proto nebylo třeba použití metody RH.

Úloha 3. Validace nové analytické metody

Zadání: Cílem této úlohy je zjistit vhodnost nově využívané metody pro stanovení glukózy v peletách. Nová metoda GOD-POD využívá enzymatické stanovení pomocí glukózaoxidázy GOD a peroxidázy POD, kdy se výsledný barevný produkt stanovuje fotometricky. Záměrem je provést validaci této nově používané analytické metody pomocí stanovení standardů roztoků glukózy.

Data:

Známé hodnoty koncentrací roztoků standardů v porovnání s hodnotami stanovenými pomocí nově využívané enzymatické metody:

standardy (mg/l)	stanoveno (mg/l)
56,1	29,5
168,3	158,2
280,5	254,8
392,7	381,9
504,9	529,3
617,1	613,7
729,3	793,2
841,5	924,2
953,7	970,1
1065,9	1057,5
1178,0	1171,9
1290,2	1199,5
1402,4	1267,5

Užitý program: QC.Expert 3.2

Řešení:

1. Návrh modelu: Za účelem validace metody je navržen přímkový regresní model pro závislost proměnných „stanoveno (mg/l)“ na proměnné „standardy (mg/l)“. Rovnice navrženého modelu je: $y = \beta_0 + \beta_1 x$, u tohoto modelu bude testována nulová hypotéza o jednotkovosti směrnice a nulovosti úseku H_0: $\beta_0 = 0$, $\beta_1 = 1$.

2. Odhadování parametrů: Pomocí klasické metody nejmenších čtverců (MNČ) byly stanoveny odhady regresních parametrů. Analýza dat probíhala na hladině významnosti $\alpha = 0,05$. Užitím Studentova t-testu bylo zjištěno, že úsek lze považovat za statisticky nevýznamný, zatímco směrnice je statisticky významná. Navíc má úsek velmi vysokou směrodatnou odchylku. Z intervalu spolehlivosti (spodní a horní mez) můžeme usuzovat, že nová metoda dává požadovanou odezvu, jelikož interval spolehlivosti úseku obsahuje hodnotu 0 a interval spolehlivosti směrnice zahrnuje hodnotu 1.

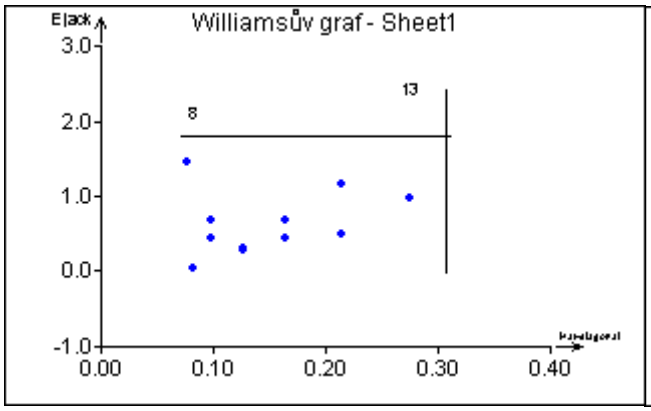
Proměnná	Odhad	Směr.Odch.	Závěr	Pravděpodobnost	Spodní mez	Horní mez
Abs	22,1344	31,0115	Nevýznamný	0,4902	-46,1213	90,3901
standardy (mg/l)	0,9560	0,0369	Významný	3,2321E-011	0,8749	1,0371

3. Základní statistické charakteristiky: Vysoká hodnota vícenásobného korelačního koeficientu naznačuje významnost modelu. Hodnota koeficientu determinace je rovněž číslo poměrně vysoké, charakterizuje podíl experimentálních bodů vyhovujících navrženému modelu.

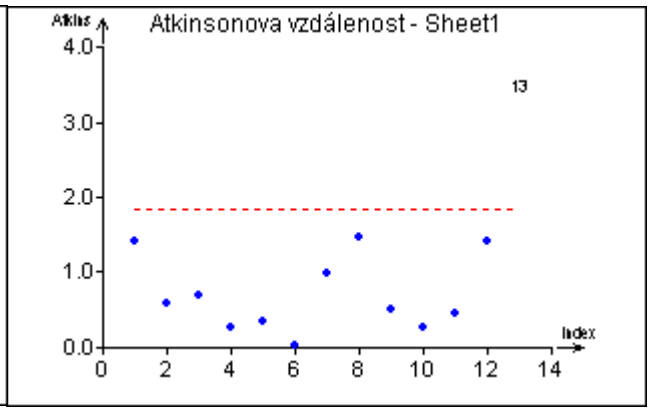
Vícenásobný korelační koeficient R	0,9919
Koeficient determinace R^2	0,9839
Predikovaný korelační koeficient Rp	0,9528
Střední kvadratická chyba predikce MEP	3908,8
Akaikeho informační kritérium	106,4

4. Regresní diagnostika – regresní triplet

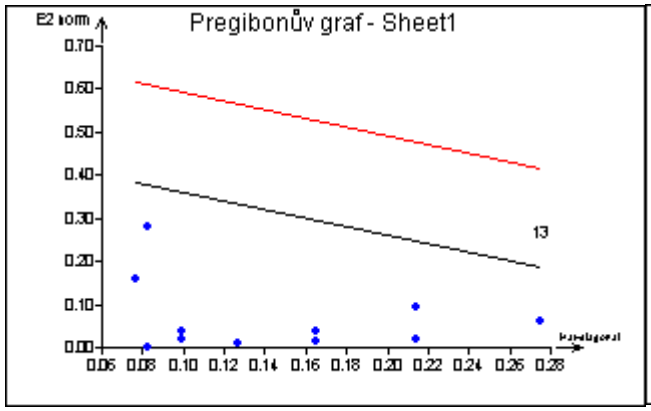
Kritika dat: V této části analýzy regresních dat je především důležité identifikovat vlivné body a odlehlé body vyloučit, aby mohl být stanoven zpřesněný model. K tomuto nejlépe slouží grafy znázorňující vlivné body, popřípadě indexové a rankitové grafy.



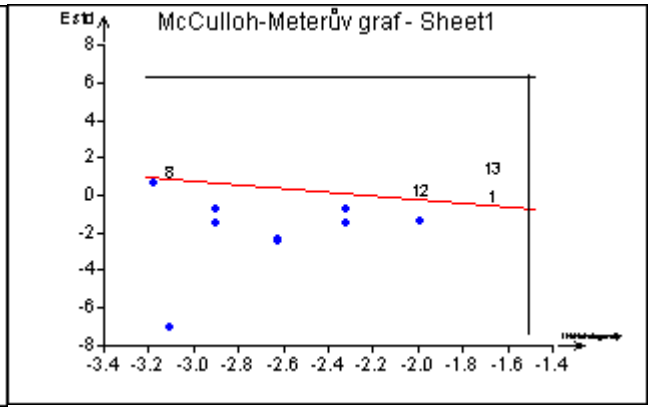
Obr. 3.1 – Williamsův graf



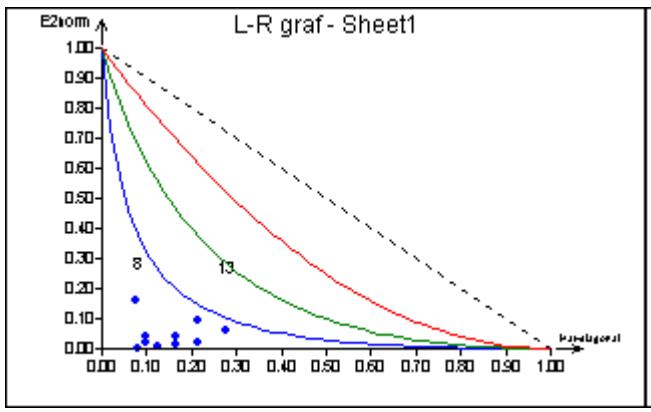
Obr. 3.2 – Atkinsova vzdálenost



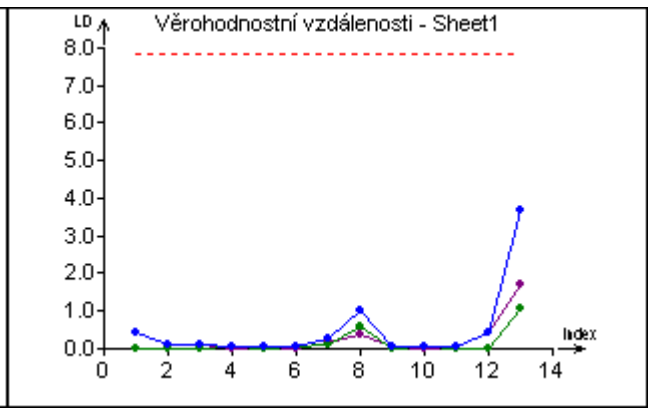
Obr. 3.3 – Pregibonův graf



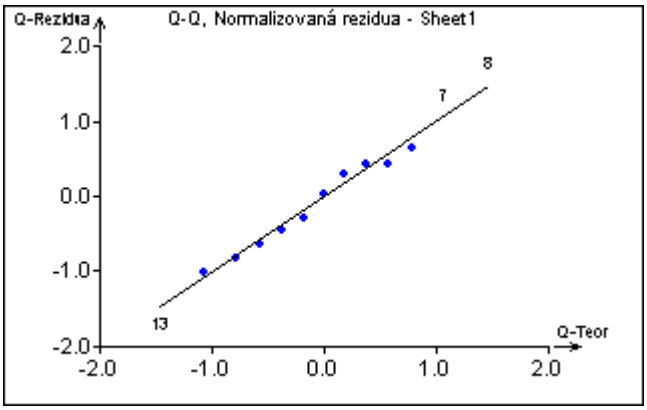
Obr. 3.4 – McCulloh-Meterův graf



Obr. 3.5 – L-R graf



Obr. 3.6 – Graf věrohodnostní vzdálenosti



Obr. 3.7 – Q-Q graf normalizovaných reziduí

Williamsův graf (Obr. 3.1) slouží k indikaci vlivných i vybočujících bodů. Zde můžeme vidět body 8 a 13 podezřelé na odlehlost, jelikož leží nad vodorovnou přímkou.

Atkinsonova vzdálenost (Obr. 3.2) znázorňuje jako odlehlý bod 13 (leží nad vodorovnou přímkou).

Pregibonův graf (Obr. 3.3) slouží pro společné posouzení vybočujících bodů a vlivných bodů. Podle tohoto zobrazení můžeme označit bod 13 jako středně vlivný.

McCulloh-Meterův graf (Obr. 3.4) prokázal body podezřelé z označení jako vybočující, a to body 1, 8, 12 a 13. Všechny tyto body totiž leží nad šikmou (červenou) přímkou.

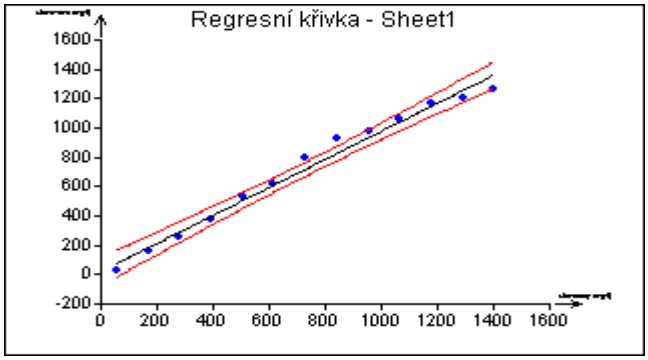
L-R graf (Obr. 3.5) naznačuje, že body 8 a 13 budou nejspíše vybočujícími body.

V grafu věrohodnostní vzdálenosti (Obr. 3.6) je vidět, že bod 13 lze označit za silně odlehlý, zatímco bod 8 je spíše středně odlehlý. Dále by mohly být podezřelé body 1 a 12.

Q-Q graf (Obr. 3.7) normalizovaných reziduí slouží pro posouzení normality reziduí. Body poměrně dobře kopírují přímku, pouze body 13 a 8 na obou koncích se zdají být odlehlé, možná také bod 7. Přímka odpovídající normálnímu rozdělení protíná bod o souřadnicích [0,0] a tento bod lze považovat za střed při proložení experimentálních bodů.

Kritika dat při tvorbě prvního modelu prokázala odlehlost bodů 8 a 13, jelikož se na tomto tvrzení shodlo minimálně pět grafických diagnostik. Tyto body budou tedy z datového souboru vyloučeny a bude zkonstruován následný zpřesněný model.

Kritika modelu: Navržený lineární model lze posoudit jako vhodný, jelikož je již z regresního grafu vidět, že přímková závislost je dodržena.

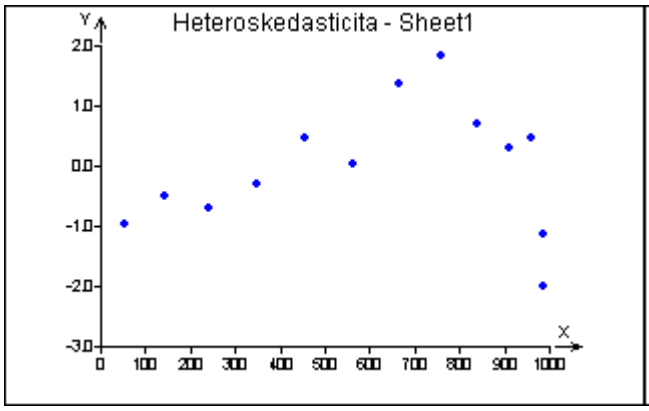


Obr. 3.8 – Regresní přímka

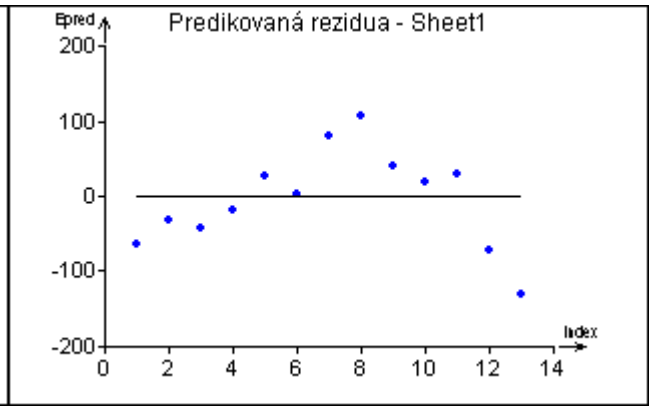
Kritika metody: Součástí regresního tripletu je také posouzení splnění základních předpokladů MNČ.

Fisher-Snedecorův test významnosti modelu	
Hodnota kritéria F:	672,9233436
Kvantil F (1-alfa, m-1, n-m):	4,844335675
Pravděpodobnost:	3,23207946E-011
Závěr:	Model je významný
Cook-Weisbergův test heteroskedasticity	
Hodnota kritéria CW:	1,114024055
Kvantil Chi^2(1-alfa,1):	3,841458829
Pravděpodobnost:	0,2912088782
Závěr:	Rezidua vykazují homoskedasticitu.
Jarque-Berrův test normality	
Hodnota kritéria JB:	0,1837010223
Kvantil Chi^2(1-alfa,2):	5,991464547
Pravděpodobnost:	0,9122415093
Závěr:	Rezidua mají normální rozdělení.
Waldův test autokorelace	
Hodnota kritéria WA:	3,998449985
Kvantil Chi^2(1-alfa,1):	3,841458829
Pravděpodobnost:	0,04554212759
Závěr:	Autokorelace je významná.
Znaménkový test reziduí	
Hodnota kritéria Sg:	2,310090716
Kvantil N(1-alfa/2):	1,959963999
Pravděpodobnost:	0,02088313235
Závěr:	V reziduích je trend!

Fisher-Snedecorův test prokázal významnost modelu, jelikož hodnota testovacího kritéria F přesáhla kritický kvantil. Redizua vykazují homoskedasticitu podle Cook-Weisbergova testu heteroskadasticity a také podle grafických diagnostik heteroskedasticity (Obr. 3.9). Rovněž bylo potvrzeno normální rozdělení reziduí podle Jarque-Berrova testu normality. Waldův test odhalil významnou autokorelaci. Znaménkový test zjistil trend v reziduích, což je patrné také z grafu predikovaných reziduí (Obr. 3.10).



Obr. 3.9 – Graf heteroskedasticity



Obr. 3.10 – Graf predikovaných reziduí

5. Konstrukce zpřesněného modelu: Z datového souboru byly odstraněny body 8 a 13 a následující zpřesněný regresní model byl sestaven bez těchto bodů. Hodnota střední kvadratické chyby predikce klesla (MEP = 3908,8 → 2112,8) a také se snížilo Akaikeho informační kritérium (AIC = 106,6 → 82,6), což značí zkvalitnění modelu s odstraněnými odlehlými body oproti modelu původnímu. Odstranění odlehlých bodů také vedlo ke zlepšení charakteristik modelu, co se týče autokorelace a trendu v reziduích. Autokorelace byla prokázána jako nevýznamná a v reziduích nebyl zjištěn trend.

Odhadování parametrů:

Proměnná	Odhad	Směr.Odch.	Závěr	Pravděpodobnost	Spodní mez	Horní mez
Abs	2,1861	22,8569	Nevýznamný	0,9259	-49,5199	53,8920
standardy (mg/l)	0,9860	0,0297	Významný	9,9478E-011	0,9189	1,0532

Základní statistické charakteristiky:

Vícenásobný korelační koeficient R	0,9960
Koeficient determinace R^2	0,9919
Predikovaný korelační koeficient Rp	0,9734
Střední kvadratická chyba predikce MEP	2112,8
Akaikeho informační kritérium	82,6

Kritika metody (vypsány pouze testy, kde došlo ke změně závěru vůči původnímu modelu):

Waldův test autokorelace
Hodnota kritéria WA: 0,2460499196
Kvantil Chi^2(1-alfa,1): 3,841458829
Pravděpodobnost: 0,6198702875
Závěr: Autokorelace je nevýznamná
Znaménkový test reziduí
Hodnota kritéria Sg: 1,895438491
Kvantil N(1-alfa/2): 1,959963999
Pravděpodobnost: 0,05803433331
Závěr: V reziduích není trend.

6. Zhodnocení kvality modelu

Nalezený regresní model má následující tvar (v závorkách udány hodnoty příslušných směrodatných odchylek):

$$\mathbf{y = 2{,}1861\ (22{,}8569) + 0{,}9860\ (0{,}0297)\ x}$$

Intervalový odhad směrnice:

$$\mathbf{0{,}9189 < \beta_1 < 1{,}0532}$$

Intervalový odhad úseku:

$$\mathbf{-49{,}5199 < \beta_0 < 53{,}8920}$$

Závěr: Bylo provedeno měření pro validaci nové analytické metody. Enzymatickou reakcí a následným fotometrickým stanovením byla zjištěna koncentrace glukózy ve standardních roztocích glukózy. Původní model regresní přímky zahrnovat třináct bodů. Body 8 a 13 byly pomocí grafických diagnostik označeny jako odlehlé a z modelu vyloučeny. Poté byla provedena konstrukce zpřesněného modelu, který vykazoval kvalitnější vlastnosti (snížení MEP, AIC a označení autokorelace a trendu v reziduích jako nevýznamné).

Jelikož intervalový odhad směrnice pro výsledný model obsahoval hodnotu 1, **lze směrnici označit za jednotkovou**. Intervalový odhad úseku zahrnoval hodnotu 0, proto **lze úsek považovat za statisticky nevýznamný**. Tímto byly **splněny předpoklady pro validaci nové analytické metody**, tato metoda dává požadovanou odezvu na hladině významnosti $\alpha = 0{,}05$.

Úloha 4. Vícerozměrný lineární regresní model

Zadání: Při plnění nádrže benzínem jsou do ovzduší uvolňovány uhlovodíky. Pro vyhodnocení efektivity kontroly tohoto znečištění byl proveden experiment, při němž byly měřeny následující parametry: teplota nádrže a benzínu, počáteční tlak v nádrži a tlak benzínu. V závislosti na těchto proměnných bylo stanoveno množství uvolněných uhlovodíků. Úkolem je zjistit vliv jednotlivých parametrů na výsledné znečištění ovzduší uhlovodíky a vytvořit pro tento experiment vícerozměrný lineární regresní model. (http://www.filewatcher.com/m/sniffer.txt.2355-0.html; S. Weisberg, Applied Linear Regression, p. 182)

Data: Stanovení hmotnosti uhlovodíků v závislosti na čtyřech parametrech při plnění nádrže:

teplota nádrže (°F)	33	31	33	37	36	35	59	60	59	60	34
teplota benzínu (°F)	53	36	51	51	54	35	56	60	60	60	35
počáteční tlak v nádrži (psi)	3,32	3,10	3,18	3,39	3,20	3,03	4,78	4,72	4,60	4,53	2,90
tlak benzínu (psi)	3,42	3,26	3,18	3,08	3,41	3,03	4,57	4,72	4,41	4,53	2,95
hmotnost uhlovodíků (g)	29	24	26	22	27	21	33	34	32	34	20
teplota nádrže (°F)	60	60	60	62	62	90	90	92	91	61	59
teplota benzínu (°F)	59	62	36	38	61	64	60	92	92	62	42
počáteční tlak v nádrži (psi)	4,40	4,31	4,27	4,41	4,39	7,32	7,32	7,45	7,27	3,91	3,75
tlak benzínu (psi)	4,36	4,42	3,94	3,49	4,39	6,70	7,20	7,45	7,26	4,08	3,45
hmotnost uhlovodíků (g)	36	34	23	24	32	40	46	55	52	29	22
teplota nádrže (°F)	88	91	63	60	60	59	59	37	35	37	
teplota benzínu (°F)	65	89	62	61	62	62	62	35	35	37	
počáteční tlak v nádrži (psi)	6,48	6,70	4,30	4,02	4,02	3,98	4,39	2,75	2,59	2,73	
tlak benzínu (psi)	5,80	6,60	4,30	4,10	3,89	4,02	4,53	2,64	2,59	2,59	
hmotnost uhlovodíků (g)	31	45	37	37	33	27	34	19	16	22	

Užitý program: QC.Expert 3.2

Řešení:

1. Návrh modelu: Pro řešení závislosti hmotnosti uhlovodíků uvolněných při čerpání benzínu (y) na parametrech: teplota nádrže (x_1), teplota benzínu (x_2), počáteční tlak v nádrži (x_3) a tlak benzínu (x_4) je navržen regresní model o tvaru: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$. U tohoto modelu bude testována významnost jednotlivých regresních parametrů.

2. Předběžná analýza dat: Pro první náhled na data byly studovány průměrné hodnoty a směrodatné odchylky nezávislých proměnných. Také byly vypočteny korelační koeficienty vůči závislé proměnné; extrémně silná korelace nebyla odhalena u žádné proměnné, nicméně je zřejmé, že všechny parametry se závislou proměnnou korelují. Hodnoty korelačních koeficientů se pohybovaly v rozmezí 0,82 až 0,92.

Proměnná	Průměr	Směr.Odch.	Kor.vs.Y	Významnost
teplota nádrže	57,9063	19,5107	0,8261	5,8227E-009
teplota benzínu	55,9063	15,7324	0,9094	5,9153E-013
počáteční tlak nádrže	4,4222	1,4522	0,8699	1,0190E-010
tlak benzínu	4,3238	1,3917	0,9213	7,6605E-014

Dále byly studovány párové korelace podle Pearsona pro všechny dvojice nezávislých proměnných. Hodnoty korelačních koeficientů se pohybovaly v rozmezí 0,77 až 0,98. Nejsilnější korelace byla zjištěna u dvojice „počáteční tlak nádrže - tlak benzínu“ a dále „teplota nádrže - počáteční tlak nádrže“.

párové korelace	r	Pravděpodobnost
teplota nádrže - teplota benzínu	0,7743	1,9990E-007
teplota nádrže - počáteční tlak nádrže	0,9554	0
teplota nádrže - tlak benzínu	0,9338	6,2172E-015
teplota benzínu - počáteční tlak nádrže	0,7815	1,2923E-007
teplota benzínu - tlak benzínu	0,8375	2,2841E-009
počáteční tlak nádrže - tlak benzínu	0,9851	0

3. Odhadování parametrů: Pomocí klasické metody nejmenších čtverců (MNČ) byly stanoveny odhady regresních parametrů. Analýza dat probíhala na hladině významnosti $\alpha = 0,05$. Užitím Studentova t-testu bylo zjištěno, že úsek lze považovat za statisticky nevýznamný, stejně tak parametry „teplota nádrže“ a „počáteční tlak nádrže“.

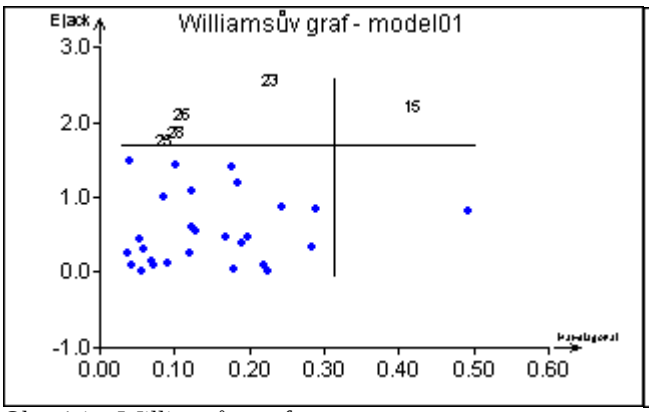
Proměnná	Odhad	Směr.Odch	Závěr	p	Spodní mez	Horní mez
Abs	1,0150	1,8613	Nevýznamný	0,5900	-2,8041	4,8341
teplota nádrže	-0,0286	0,0906	Nevýznamný	0,7546	-0,2145	0,1573
teplota benzínu	0,21584	0,0677	Významný	0,0036	0,0769	0,3548
počáteční tlak nádrže	-4,3201	2,8510	Nevýznamný	0,1413	-10,1698	1,5297
tlak benzínu	8,9749	2,7726	Významný	0,0032	3,2859	14,6639

4. Základní statistické charakteristiky: Vysoká hodnota vícenásobného korelačního koeficientu indikuje významnost modelu. Hodnota koeficientu determinace, charakterizující podíl experimentálních bodů vyhovujících navrženému modelu, je přibližně 0,92. Je zřejmé, že vyloučením úseku, případně některých proměnných z modelu bude možná dosaženo vyšší hodnoty a naopak snížení MEP a AIC.

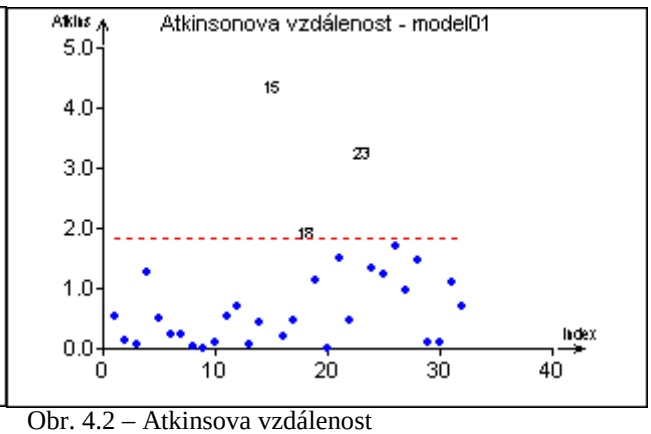
Vícenásobný korelační koeficient R	0,9623
Koeficient determinace R^2	0,9261
Predikovaný korelační koeficient Rp	0,7849
Střední kvadratická chyba predikce MEP	9,70
Akaikeho informační kritérium	68,84

5. Regresní diagnostika – regresní triplet

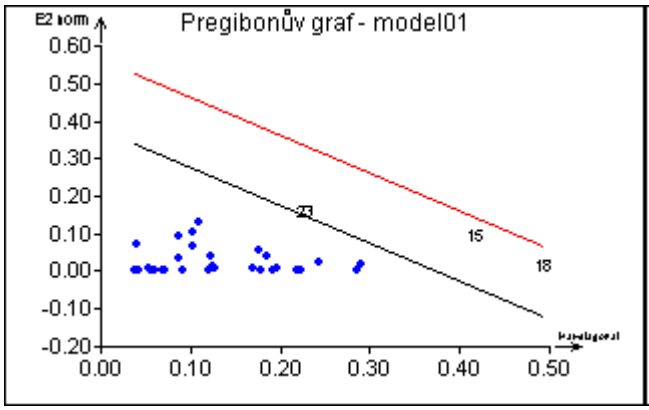
Kritika dat: V této části analýzy regresních dat je především důležité identifikovat vlivné body a odlehlé body vyloučit, aby mohl být stanoven zpřesněný model. K tomuto nejlépe slouží grafy znázorňující vlivné body, popřípadě indexové a rankitové grafy.



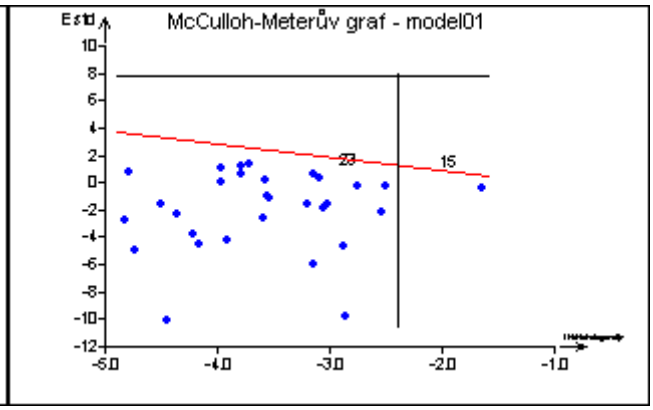
Obr. 4.1 – Williamsův graf



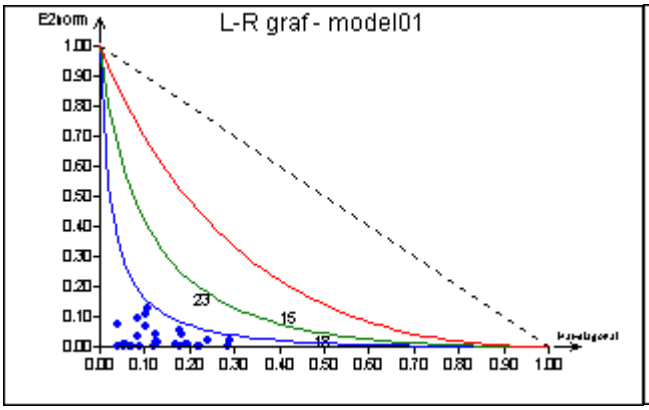
Obr. 4.2 – Atkinsova vzdálenost



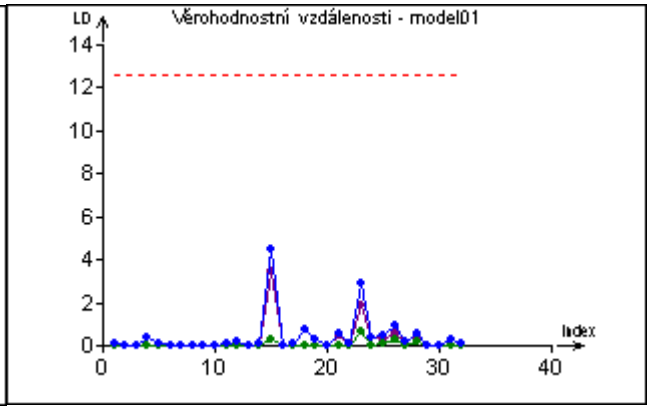
Obr. 4.3 – Pregibonův graf



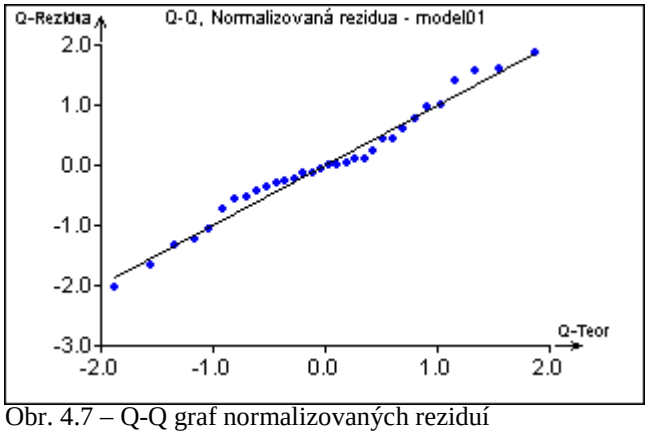
Obr. 4.4 – McCulloh-Meterův graf



Obr. 4.5 – L-R graf



Obr. 4.6 – Graf věrohodnostní vzdálenosti



Obr. 4.7 – Q-Q graf normalizovaných reziduí

Williamsův graf (Obr. 3.1) slouží k indikaci vlivných i vybočujících bodů. Zde můžeme vidět body 15, 23, 25, 26 a 28 podezřelé na odlehlost, jelikož leží nad vodorovnou přímkou.

Atkinsonova vzdálenost (Obr. 3.2) znázorňuje jako odlehlý bod 15 a 23 (leží nad vodorovnou přímkou), případně bod 18 ležící těsně nad hranicí.

Pregibonův graf (Obr. 3.3) slouží pro společné posouzení vybočujících bodů a vlivných bodů. Podle tohoto zobrazení můžeme označit jako středně vlivné body 15, 18 a 23.

McCulloh-Meterův graf (Obr. 3.4) odhalil jako vlivné body 15, případně 23, jelikož leží těsně nad šikmou (červenou) přímkou.

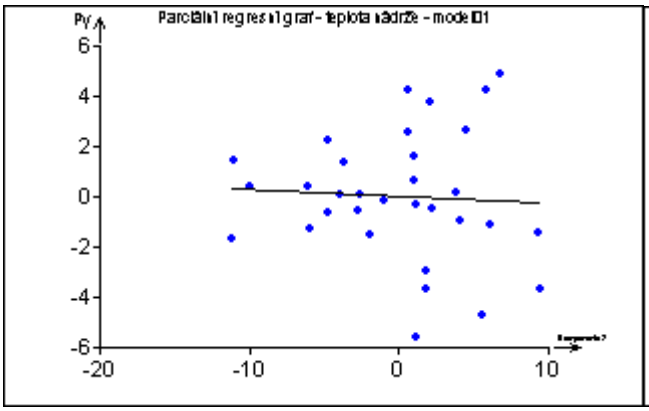
L-R graf (Obr. 3.5) naznačuje, že bod 23 je nejspíše vybočující, zatímco body 15 a 18 se jeví spíše jako extrémy.

V grafu věrohodnostní vzdálenosti (Obr. 3.6) je vidět, že body 15 a 23 lze označit za potenciálně odlehlé.

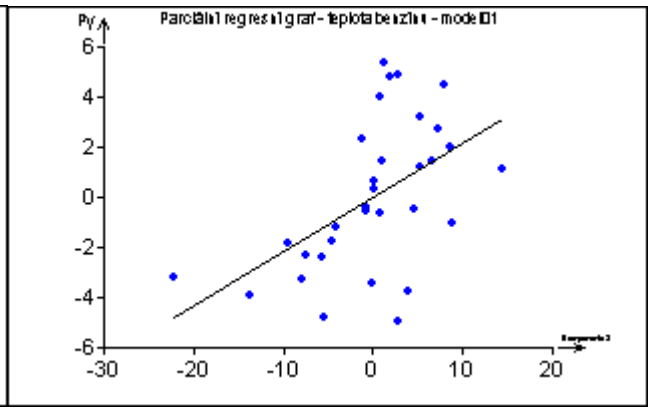
Q-Q graf (Obr. 3.7) normalizovaných reziduí slouží pro posouzení normality reziduí. Body poměrně dobře kopírují přímku, ačkoli je pozorovatelné jisté prohnutí. Podle tohoto zobrazení nelze žádný bod označit jako odlehlý.

Kritika dat při tvorbě prvního modelu prokázala odlehlost bodů 15 a 23, jelikož se na tomto tvrzení shodlo minimálně pět grafických diagnostik. Tyto body budou tedy z datového souboru vyloučeny a bude zkonstruován následný zpřesněný model.

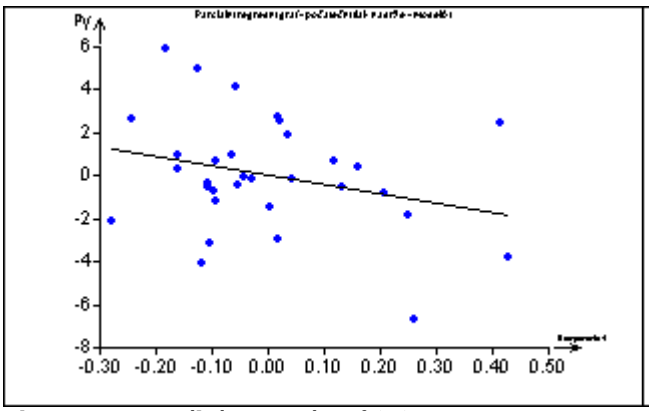
Kritika modelu: Navržený lineární model nelze posoudit jako vhodný, jelikož byla prokázána statistická nevýznamnost některých parametrů, proto je nejprve potřeba zjistit, které regresní parametry jsou i po vyloučení odlehlých bodů významné a ty potom zahrnout do výsledného modelu. Pro posouzení významnosti regresních parametrů lze z grafických diagnostik využít parciálních regresních a reziduálních grafů. Parciální regresní graf zobrazuje závislost závislé proměnné na vybrané nezávislé proměnné za eliminace vlivu ostatních nezávislých proměnných, přičemž směrnice přímky odpovídá danému regresnímu koeficientu a těsnost proložení souvisí s významností koeficientu. Parciální reziduální graf je modifikací parciálního regresního grafu.



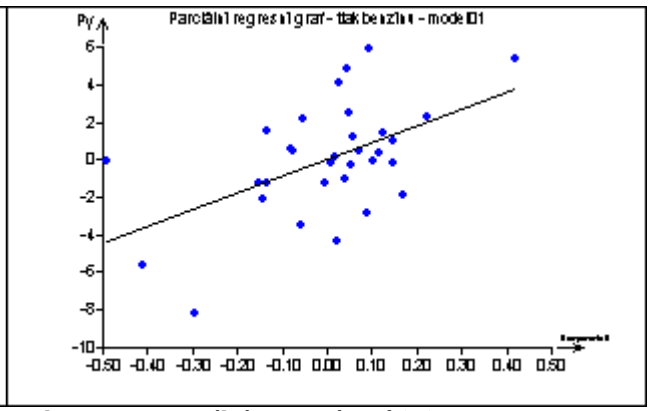
Obr. 4.8 – Parciální regresní graf (x_1)



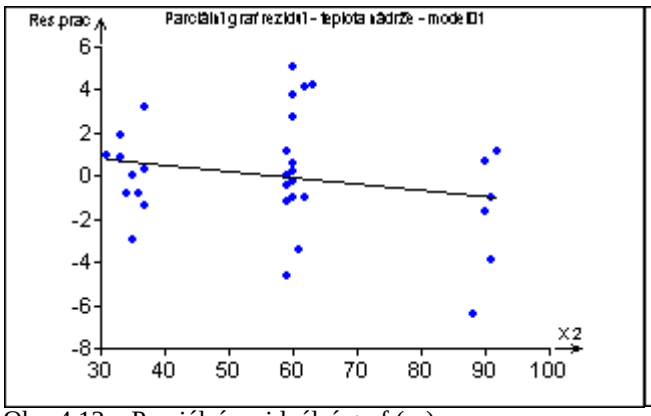
Obr. 4.9 – Parciální regresní graf (x_2)



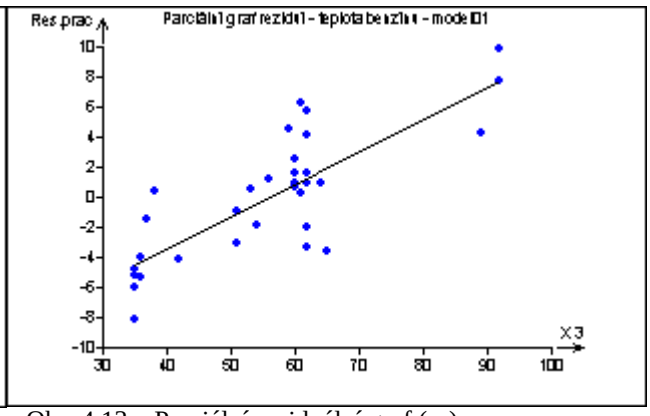
Obr. 4.10 – Parciální regresní graf (x_3)



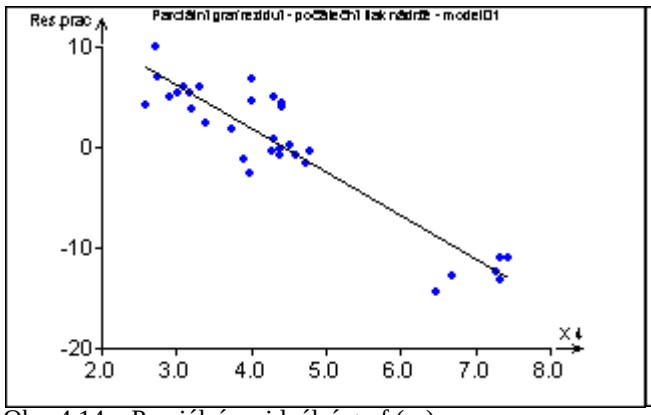
Obr. 4.11 – Parciální regresní graf (x_4)



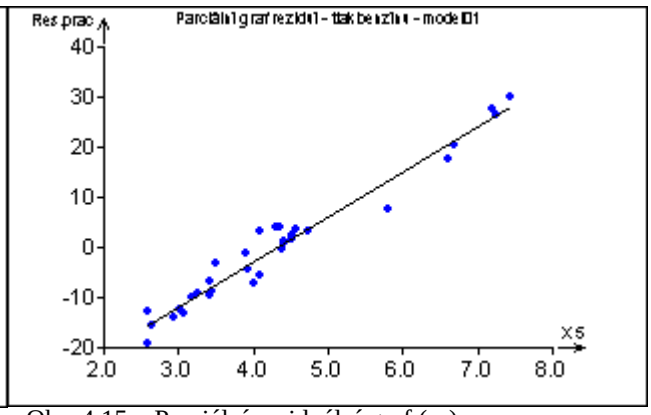
Obr. 4.12 – Parciální reziduální graf (x_1)



Obr. 4.13 – Parciální reziduální graf (x_2)



Obr. 4.14 – Parciální reziduální graf (x_3)



Obr. 4.15 – Parciální reziduální graf (x_4)

Z parciálních regresních a reziduálních grafů můžeme usuzovat na statistickou nevýznamnost proměnné x_1 „teplota nádrže“, jelikož je v příslušném grafu vysoký rozptyl bodů s téměř nulovou směrnicí. Taktéž u proměnné x_3 „počáteční tlak nádrže“ je patrná nízká hodnota směrnice. Nejvyšší významnost vykazuje proměnné x_2 „teplota benzínu“, jelikož jí přísluší nejvyšší směrnice a x_4 „tlak benzínu“.

Kritika metody: Součástí regresního tripletu je také posouzení splnění základních předpokladů MNČ.

Fisher-Snedecorův test významnosti modelu

Hodnota kritéria F:	84,54028518
Kvantil F (1-alfa, m-1, n-m):	2,727765306
Pravděpodobnost:	7,248720131E-015
Závěr:	Model je významný

Scottovo kritérium multikolinearity

Hodnota kritéria SC:	0,8953969161
Závěr:	Model je nekorektní!

Cook-Weisbergův test heteroskedasticity

Hodnota kritéria CW:	0,0002455819305
Kvantil Chi^2(1-alfa,1 :	3,841458829
Pravděpodobnost:	0,9874968198
Závěr:	Rezidua vykazují homoskedasticitu.

Jarque-Berrův test normality

Hodnota kritéria JB:	0,04669756543
Kvantil Chi^2(1-alfa,2):	5,991464547
Pravděpodobnost:	0,9769216909
Závěr:	Rezidua mají normální rozdělení.

Waldův test autokorelace

Hodnota kritéria WA:	0,3166680771
Kvantil Chi^2(1-alfa,1):	3,841458829
Pravděpodobnost:	0,5736169226
Závěr:	Autokorelace je nevýznamná

Znaménkový test reziduí

Hodnota kritéria Sg:	0,5187279417
Kvantil N(1-alfa/2):	1,959963999
Pravděpodobnost:	0,6039504737
Závěr:	V reziduích není trend.

6. Konstrukce zpřesněného modelu: Pro nalezení nejlepšího modelu byly studovány rozdíly následujících modelů:

model01 – celkový

model02 – po vyloučení úseku

model03 – po vyloučení úseku a odstranění odlehlých bodů

model04 – navíc vyloučení proměnné x_1

model05 – navíc vyloučení proměnných x_1 a x_3

Sledován byl vliv parametrů zahrnutých do modelu na výsledné statistiky jako MEP, AIC, R^2 a dále regresní charakteristiky pro určení významnosti proměnných.

Základní statistické charakteristiky:

parametr	model01	model02	model03	model04	model05
Vícenásobný korelační koeficient R	0,9623	0,9619	0,9717	0,9716	0,9677
Koeficient determinace R^2	0,9261	0,9252	0,9442	0,9440	0,9364
Predikovaný korelační koeficient Rp	0,7849	0,7901	0,8617	0,8724	0,8528
Střední kvadratická chyba predikce MEP	9,70	9,45	6,38	5,87	6,81
Akaikeho informační kritérium	68,84	67,19	56,05	54,16	56,01

Odhadování parametrů (model03):

Proměnná	Odhad	Směr.Odch	Závěr	p	Spodní mez	Horní mez
teplota nádrže	-0,0329	0,0891	Nevýznamný	0,7151	-0,2154	0,1497
teplota benzínu	0,2345	0,0577	Významný	0,0004	0,1164	0,3526
počáteční tlak nádrže	-3,9800	2,7467	Nevýznamný	0,1585	-9,6059	1,6466
tlak benzínu	8,6607	2,6779	Významný	0,0031	3,1753	14,1461

Odhadování parametrů (model04):

Proměnná	Odhad	Směr.Odch	Závěr	p	Spodní mez	Horní mez
teplota benzínu	0,2141	0,0490	Významný	0,0002	0,1136	0,3146
počáteční tlak nádrže	-5,3579	2,7856	Nevýznamný	0,0650	-11,0735	0,3579
tlak benzínu	9,9140	3,1443	Významný	0,0039	3,4625	16,3655

Odhadování parametrů (model05):

Proměnná	Odhad	Směr.Odch.	Závěr	p	Spodní mez	Horní mez
teplota benzínu	0,2548	0,0463	Významný	6,9190E-006	0,1601	0,3496
tlak benzínu	3,9666	0,5970	Významný	3,2913E-007	2,7438	5,1895

Bylo zjištěno, že po vyloučení úseku a odstranění odlehlých bodů dosahuje model snížení hodnot MEP a AIC a také zvýšení koeficientu determinace. Proměnné x_1 a x_3 však byly stále detekovány jako statisticky nevýznamné. Nejprve byl vytvořen model s vyloučením proměnné x_1, jelikož příslušná p hodnota pro tuto proměnnou byla nejvyšší. Takto bylo dosaženo ještě dalšího snížení MEP i AIC, proměnná x_3 však byla opět zjištěna jako nevýznamná, proto byla z modelu vyloučena. Výsledný model tedy zahrnuje pouze proměnné x_2 a x_4, což jsou veličiny „teplota benzínu“ a „tlak benzínu“. Podle tohoto modelu nemají parametry pro nádrž („teplota nádrže“ a „počáteční tlak nádrže“) statisticky významný vliv na množství uvolněných uhlovodíků při plnění nádrže benzínem.

Kritika metody: U zpřesněného modelu bylo dosaženo stejných závěrů, co se týče kritiky metody, jako u modelu původního. Čili data splňují všechna požadovaná kritéria, vyjma korektnosti modelu. Scottovo kritérium (M_T = 0,806) potvrdilo multikolinearitu mezi zbývajícími dvěma proměnnými. Multikolinearita by mohla být odstraněna vyloučením některé proměnné z modelu, tím by však mohlo dojít ke ztrátě informace a proto byly obě proměnné v modelu zachovány. Navíc leží hodnota Scottova kritéria přesně na hranici, kdy je úprava modelu pouze doporučována ($0,33 < M_T < 0,8$), nikoli nutná ($M_T > 0,8$).

7. Zhodnocení kvality modelu

Nalezený regresní model má následující tvar (v závorkách udány hodnoty příslušných směrodatných odchylek):

$$\mathbf{y = 0,2548\ (0,0463)\ x_2 + 3,9666\ (0,5970)\ x_4}$$

Čili ve fyzikálním smyslu:

hmotnost uhlovodíků (g) = 0,2548 (0,0463) teplota benzínu (°F) + 3,9666 (0,5970) tlak benzínu (psi)

Závěr: Byl stanoven **vícerozměrný lineární regresní model pro závislost množství uvolněných uhlovodíků** při plnění nádrže benzínem na parametrech nádrže a benzínu. Bylo zjištěno, že počáteční tlak v nádrži a teplota nádrže nemají statisticky významný vliv na hmotnost uvolněných uhlovodíků, zatímco **teplota a tlak benzínu jsou statisticky významné**.

Vybraný model poskytuje nejlepší výsledky, co se týče studovaných statistik i zhodnocení významnosti regresních parametrů.