Lineární regrese

Komentované řešení pomocí MS Excel

Vstupní data

	A	B
1	**koncentrace NH3 (x)**	**napěťová odezva (y)**
2	mg/l	mV
3	12	62,42
4	17	85,16
5	33	165,84
6	41	205,25
7	56	280,18
8	64	320,2
9	78	392,74
10	89	447
11	92	496,8

- Tabulka se vstupními daty je umístěna v oblasti A1:B11 (viz. obrázek) na listu „ cela data“

Základní výpočty - regrese

Postup

- Výpočet základních ukazatelů regrese provedeme pomocí maticové funkce **LINREGRESE** (parametry této funkce i výsledné hodnoty viz. níže)
  - Práce s maticovými vzorci viz. nápověda k této funkci.

	A	B
15	**y = a+ bx**	
16	5,20	-5,72
17	0,13	7,93
18	1,00	11,04
19	1572,73	7
20	191826,62	853,79

	A	B
15	**y = a+ bx**	
16	=LINREGRESE($B$3:$B$11;$A$3:$A$11;1;1)	=LINREGRESE($B$3:$B$11;$A$3:$A$11;1;1)
17	=LINREGRESE($B$3:$B$11;$A$3:$A$11;1;1)	=LINREGRESE($B$3:$B$11;$A$3:$A$11;1;1)
18	=LINREGRESE($B$3:$B$11;$A$3:$A$11;1;1)	=LINREGRESE($B$3:$B$11;$A$3:$A$11;1;1)
19	=LINREGRESE($B$3:$B$11;$A$3:$A$11;1;1)	=LINREGRESE($B$3:$B$11;$A$3:$A$11;1;1)
20	=LINREGRESE($B$3:$B$11;$A$3:$A$11;1;1)	=LINREGRESE($B$3:$B$11;$A$3:$A$11;1;1)

Interpretace

- Význam jendotlivých čísel spočítaných pomocí funkce LINREGRESE v oblasti A16:B20 je popsán v tabulce níže.
- Jedná se zatím o pomocné výpočty, které budeme dále používat pro řešení jendotlivých úloh.

	A	B
15	**y = a+ bx**	
16	odhad *b*	odhad *a*
17	směrodatná chyba s_b	směrodatná chyba s_a
18	koeficient determinace r^2	směrodatná chyba s_Y
19	F	stupně volnosti
20	regresní součet čtverců	residuální součet čtverců

Základní výpočty - rezidua

Postup

- Na základě odhadů regresních parametrů určíme odhad střední (skutečné) odezvy. Rezidua jsou pak rozdílem naměřené odezvy a odhadu střední odezvy. Rezidua tudíž představují odhad (nepozorovatelných) náhodných chyb v měření.

	D	E
1	**střední odezva - odhad**	**rezidua**
2		
3	=+$B$16+$A$16*A3	=+B3-D3
4	=+$B$16+$A$16*A4	=+B4-D4
5	=+$B$16+$A$16*A5	=+B5-D5
6	=+$B$16+$A$16*A6	=+B6-D6
7	=+$B$16+$A$16*A7	=+B7-D7
8	=+$B$16+$A$16*A8	=+B8-D8
9	=+$B$16+$A$16*A9	=+B9-D9
10	=+$B$16+$A$16*A10	=+B10-D10
11	=+$B$16+$A$16*A11	=+B11-D11

Interpretace

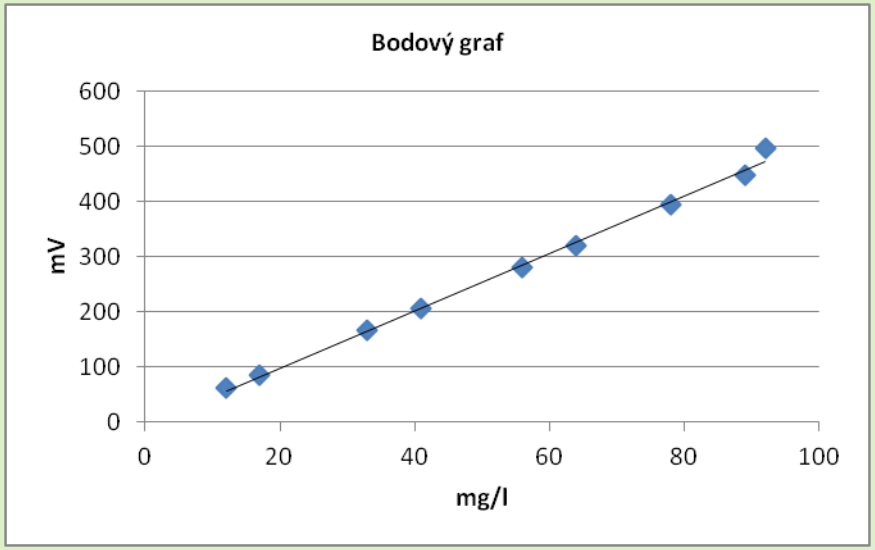
- Rezidua představují odhad (nepozorovatelných) náhodných chyb v měření.

	D	E
1	**střední odezva - odhad**	**rezidua**
2		
3	56,694	5,726
4	82,701	2,459
5	165,925	-0,085
6	207,536	-2,286
7	285,558	-5,378
8	327,170	-6,970
9	399,990	-7,250
10	457,206	-10,206
11	472,810	23,990

Ověření předpokladů – linearita

Postup:

- Zkonstruujeme bodový graf, kde každý bod bude reprezentovat jedno měření
  - x-ová souřadnice bude odpovídat koncentraci (nezávisle proměnná)
  - y-ová souřadnice bude odpovídat napěťové odezvě (závisle proměnná)
- Do bodového grafu přidáme ještě regresní přímku (označíme datovou řadu → pravé tlačítko → „přidat spojnici trendu“



Interpretace výsledků

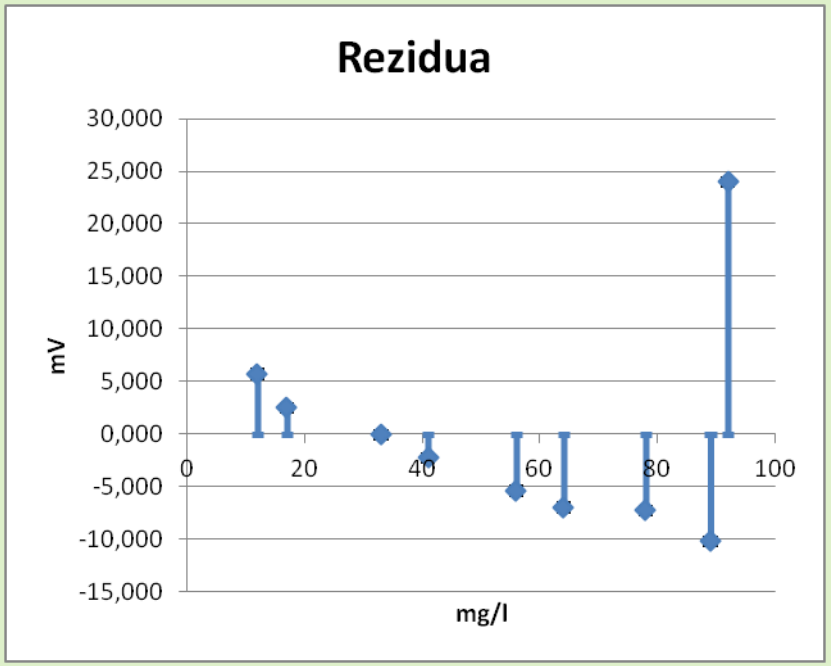
- Z grafu je patrné, že mezi měřenými veličinami skutečně existuje velmi silná lineární závislost, všechny body dosti těsně přiléhají k regresní přímce
- Poslední bod je poněkud dále od regresní přímky, než body ostatní. Mohlo by to signalizovat odlehlé pozorování. Detailnější pohled přinese analýza reziduí.

Ověření předpokladů – rezidua

Postup:

- Abychom posoudili vzájemnou nekorelovanost i neměnný rozptyl náhodných odchylek, vykreslíme si bodový graf reziduí.Zkonstruujeme bodový graf, kde každý bod bude reprezentovat jedno měření
- Do bodového grafu přidáme ještě svislé čáry (chybové úsečky):
  - Nejdříve si rezidua rozdělíme na kladná a záporná
  - Označíme datovou řadu → „Rozložení“ → „Chybové úsečky“
  - Do kladných chybový hodnot zadáme kladná rezidua a do záporný zase záporná rezidua

	A	B
34	**kladná rezidua**	**záporná rezidua**
35		
36	5,725851877	0
37	2,458596972	0
38	0	0,084618724
39	0	2,286226572
40	0	5,377991287
41	0	6,969599135
42	0	7,249912869
43	0	10,20587366
44	23,9897734	0



Interpretace výsledků

- Rezidua nevypadají náhodně a nesystematicky. Naopak, s výjimkou posledního měření jsou silně pozitivně korelovaná.
- Pro tato data tedy nelze považovat předpoklad o nekorelovanosti náhodných složek za splněný.
- Poslední měření se zjevně vymyká těm ostatním a způsobilo vychýlení regresní přímky, které se následně projevilo korelovaností reziduí.
- V dalším tedy budeme pracovat s opravenými daty – tj. v Excelu se přesuneme na nový list a vymažeme poslední měření

Vstupní data - opravená

	A	B
1	**koncentrace NH3 (x)**	**napěťová odezva (y)**
2	mg/l	mV
3	12	62,42
4	17	85,16
5	33	165,84
6	41	205,25
7	56	280,18
8	64	320,2
9	78	392,74
10	89	447

- Tabulka se vstupními daty je tedy po opravě umístěna v oblasti A1:B11 na listu „opravena data“

Základní výpočty – regrese (opravená data)

Postup a interpretace

- Je analogický jako v případě původních výpočtů se všemi daty
- F statistika je mnohem vyšší než v případě původních výpočtů, což znamená těsnější proložení naměřených bodů regresní přímkou.

	A	B
19	**y = a+ bx**	
20	5,010	0,597
21	0,016	0,866
22	1,000	1,155
23	102209,612	6
24	136246,249	7,998

	A	B
15	**y = a+ bx**	
16	odhad *b*	odhad *a*
17	směrodatná chyba s_b	směrodatná chyba s_a
18	koeficient determinace r^2	směrodatná chyba s_Y
19	F	stupně volnosti
20	regresní součet čtverců	residuální součet čtverců

Základní výpočty – rezidua (opravená data)

Postup a interpretace

- Je analogický jako v případě původních výpočtů se všemi daty

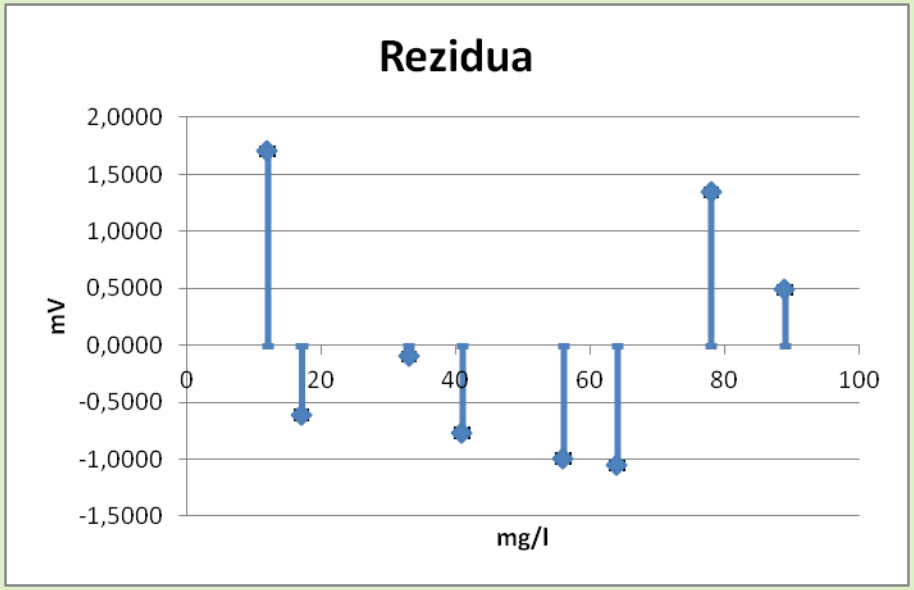
	D	E
1	**střední odezva - odhad**	**rezidua**
2		
3	60,7208	1,6992
4	85,7722	-0,6122
5	165,9368	-0,0968
6	206,0190	-0,7690
7	281,1733	-0,9933
8	321,2556	-1,0556
9	391,3996	1,3404
10	446,5127	0,4873

Ověření předpokladů – rezidua (opravená data)

Postup a interpretace

- Je analogický jako v případě původních výpočtů se všemi daty
- Rezidua už vypadají náhodně, nesystematicky. Nevykazují ani zjevnou autokorelaci ani heteroskedasticitu.

	A	B
36	kladná rezidua	záporná rezidua
37		
38	1,69919438	0
39	0	0,612226624
40	0	0,096773837
41	0	0,769047444
42	0	0,993310456
43	0	1,055584063
44	1,340437126	0
45	0,487310917	0

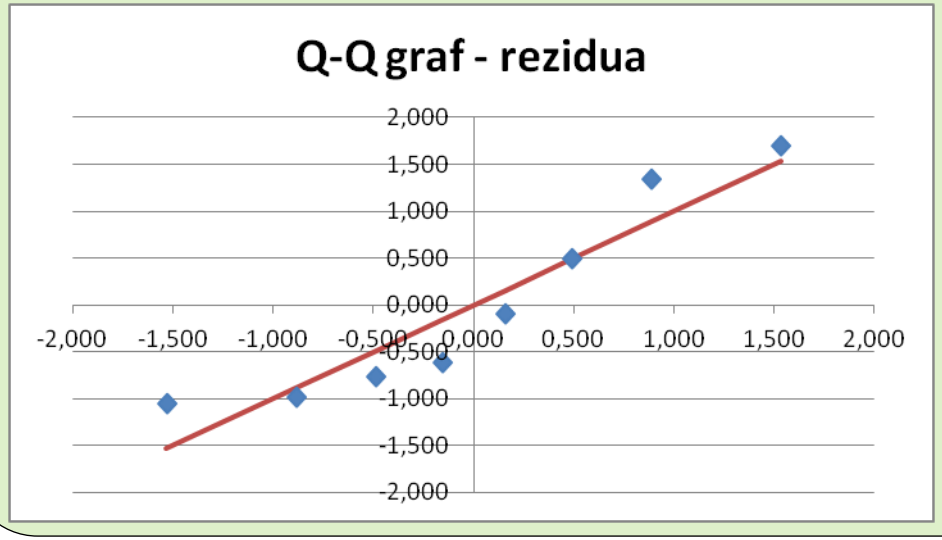


Ověření předpokladů – normalita (opravená data)

Postup

- Pro posouzení normality použijeme **Q-Q graf**. Na histogram máme příliš málo pozorování.
- Postup tvorby Q-Q grafu v MS Excel je podrobně popsán v prezentaci k jednovýběrovým testům.

	I	J	K	L	M
36	**Pořadí hodnoty**	**Uspořádaná rezidua**	**Kumulativní pravděp.**	**Kvantil N(0,1)**	**Standardizovaná rezidua**
37					
38	**1**	-1,05558406	=(I38-0,5)/POČET($I$38:$I$45)	=NORMSINV(K38)	=(J38-PRŮMĚR($J$38:$J$45))/SMODCH($J$38:$J$45)
39	**2**	-0,99331046	0,188	-0,887	-0,993
40	**3**	-0,76904744	0,313	-0,489	-0,769
41	**4**	-0,61222662	0,438	-0,157	-0,612
42	**5**	-0,09677384	0,563	0,157	-0,097
43	**6**	0,48731092	0,688	0,489	0,487
44	**7**	1,34043713	0,813	0,887	1,341
45	**8**	1,69919438	0,938	1,534	1,699



Interpretace

- Je patrný rozdíl mezi skutečnými kvantily (modré body) a těmi gaussovskými (červená čára).
- Vzhledem k malému počtu pozorování však ještě tento rozdíl nemusí být zcela statisticky významný.
- Statistickou významnost rozdílu bychom mohli otestovat nějakým testem normality, to však přesahuje rámec této prezentace
- V dalším tedy budeme předpokládat normalitu (bez ní bychom nedokázali úlohy vyřešit) a to i vzhledem k tomu, že chyby v měření mívají typicky normální rozdělení. K výsledkům (a jejich použití) však musíme přistupovat opatrně, protože normalitu se nepodařilo jednoznačně prokázat.

Úloha (A) - opravená data

Postup

- Hypotézu o nulovosti parametru posunu otestujeme pomocí oboustranného intervalu spolehlivosti
- Z regresních výpočtů si vybereme potřebné vstupy a z nich spočítáme 95% interval spolehlivosti .

	A	B
51	**úloha a)**	
52	odhad *a*	0,597
53	směrodatná chyba s_a	0,866
54	stupně volnosti	6
55	α	0,05
56	**95% interval spolehlivosti pro koef. *a***	
57	**dolni mez**	**horni mez**
58	**-1,522**	**2,717**

	A	B
51	**úloha a)**	
52	odhad *a*	=B20
53	směrodatná chyba s_a	=B21
54	stupně volnosti	=B23
55	α	0,05
56	**95% interval spolehlivosti pro koef. *a***	
57	**dolni mez**	**horni mez**
58	**=+$B$52-$B$53*TINV($B$55;$B$54)**	**=+$B$52+$B$53*TINV($B$55;$B$54)**

Interpretace

- 95% interval spolehlivosti pro parametr posunu obsahuje nulu. Na hladině významnosti 5 % tedy nemůžeme zamítnout hypotézu o nulovosti tohoto koeficientu. Jinými slovy, nelze vyloučit, že kalibrační křivka prochází počátkem (ovšem nemůžeme to ani potvrdit).

Úloha (C) - opravená data 1

Postup

- Nejdříve si provedeme několik pomocných výpočtů

	A	B
64	**úloha c)**	
65		
66	s^2	1,333
67	x prumer	48,75
68	sum x^2	24440
69	n	8
70	stupne volnosti	6
71	α	0,05

	A	B
64	**úloha c)**	
65		
66	s^2	=$B$24/$B$23
67	x prumer	=+PRŮMĚR($A$3:$A$10)
68	sum x^2	=SUMA.ČTVERCŮ($A$3:$A$10)
69	n	=+POČET($A$3:$A$10)
70	stupne volnosti	=+$B$23
71	α	0,05

- Dále pro každou hodnotu *x* dopočítáme rozptyl odhadu regresní funkce, z něj pak rozptyl predikce a nakonec predikční pás

	I	J	K	L
1	$s^2_{\eta(x)}$	$s^2_{Y(x)}$	**Predičkní pás**	
2			**dolní mez**	**horní mez**
3	0,498	1,831	57,409	64,032
4	0,414	1,747	82,538	89,007
5	0,228	1,561	162,880	168,994
6	0,181	1,514	203,008	209,030
7	0,180	1,513	278,164	284,183
8	0,224	1,557	318,203	324,309
9	0,377	1,710	388,200	394,599
10	0,565	1,898	443,142	449,883

	L
3	=+$D3+TINV($B$71;$B$70)*ODMOCNINA($J3)

	K
3	=+$D3-TINV($B$71;$B$70)*ODMOCNINA($J3)

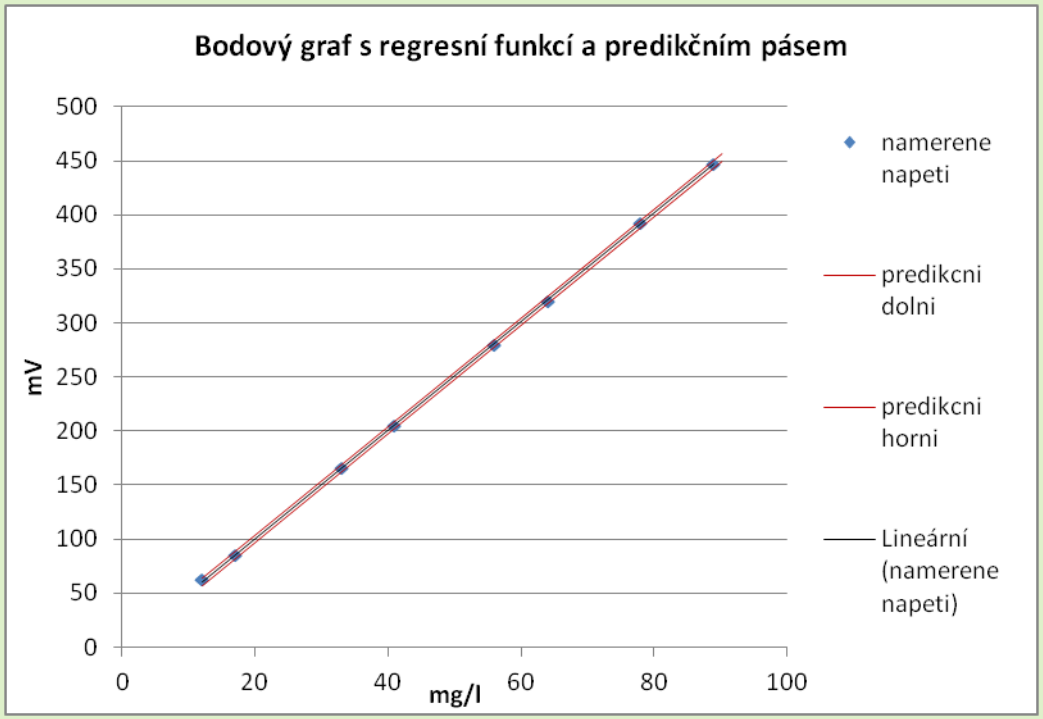
	J
3	=$I3+$B$66

	I
3	=$B$66*((1/$B$69)+(($A3-$B$67)^(2)/($B$68-$B$69*$B$67^2)))

Úloha (C) - opravená data 2

Interpretace

- Přibližně 95 % měření by mělo ležet v predikčním pásu. V našem případě je všech 8 měření uvnitř tohoto pásu, což není v rozporu s očekáváním.
- Pro lepší představu bývá dobré doplnit dolní a horní mez predikčního pásu do bodového grafu.
  - Dolní i horní mez se přidají jako nové řady do bodového grafu a upraví se typ grafu pro tyto řady na „Bodový s vyhlazenými spojnicemi“
  - V našem příkladu jsou předpovědi natolik přesné, že predikční pás v grafu prakticky splývá s regresní funkcí (predikční pás by byl zřetelný při velkém přiblížení). Predikce pomocí regresní funkce je tedy velmi přesná.



Úloha (B) - opravená data

Interpretace

- Vzhledem k předchozí analýze reziduí lze usuzovat, že poslední měření z původních dat je systematicky chybné.
- Po odebrání tohoto měření jsou už všechna pozorování uvnitř 95% predikčního pásu z úlohy (C) a rezidua nevykazují žádné výrazně odlehlé hodnoty, takže další systematicky chybné měření už nepředpokládáme.

	B	K	L
1	napěťová odezva (y)	Predičkní pás	
2	mV	dolní mez	horní mez
3	62,42	57,409	64,032
4	85,16	82,538	89,007
5	165,84	162,880	168,994
6	205,25	203,008	209,030
7	280,18	278,164	284,183
8	320,2	318,203	324,309
9	392,74	388,200	394,599
10	447	443,142	449,883

Úloha (D) - opravená data

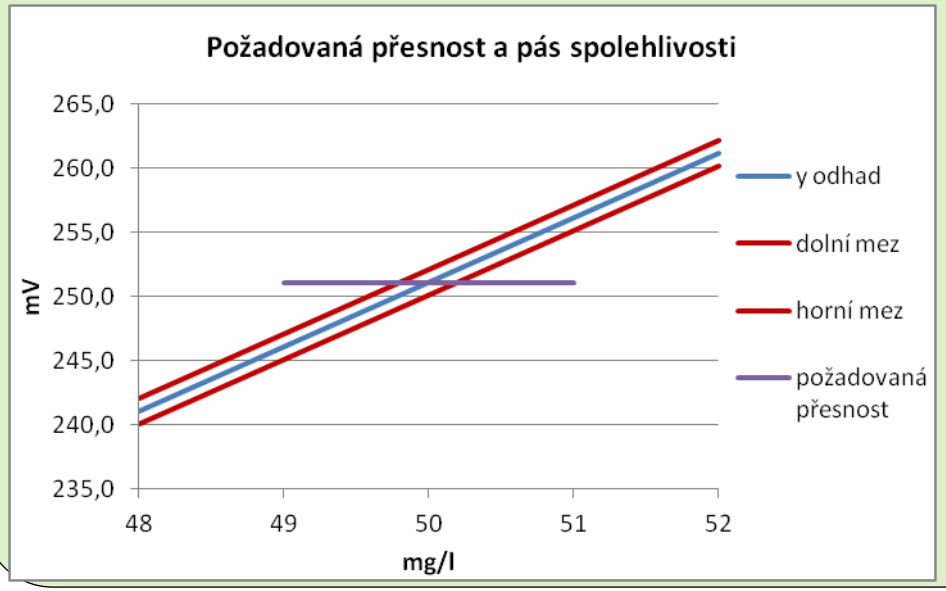
Postup

- Vzhledem k tomu, že na celkovém grafu horní i dolní mez predikčního pásu prakticky splývají, splývala by i horní a dolní mez pásu spolehlivosti. Vykreslíme si tedy zvětšenou část grafu s pásem spolehlivosti jen na okolí bodu $x = 50$
  - Pás spolehlivosti se spočítá podobně jako predikční pás, akorát místo směrodatné chyby odhadu Y ($s^2_{Y(x)}$) se použije směrodatná chyba odhadu regresní funkce ($s^2_{\eta(x)}$) .

	A	B	D	I	M	N
1	koncentrace NH3 (x)	napěťová odezva (y)	střední odezva - odhad	$s^2_{\eta(x)}$	Pás spolehlivosti	
2	mg/l	mV			dolní mez	horní mez
11	48		241,0910	0,167	240,092	242,090
12	50		251,1116	0,167	250,112	252,112
13	52		261,1322	0,169	260,126	262,139

- Dále do grafu nakreslíme úsečku představující požadovanou přesnost odhadu koncentrace, neboli $y = a + b*50 = 251{,}1$ a $50-0{,}02*50 < x < 50+0{,}02*50$, neboli $49 < x < 51$

úsečka s přesností	
x	y
49	251,1
50	251,1
51	251,1



Interpretace

- 95% pás spolehlivosti obsahuje skutečnou (neznámou) regresní funkci s pravděpodobností 95 % a tento pás je výrazně užší než požadovaná přesnost 2 %. Lze tedy tvrdit, že na hladině významnosti 95 % lze určit koncentraci s přesností lepší než 2 %.

Úloha (E) - opravená data

Postup

- Je velmi podobný postupu v úloze (A)

	A	B
108	**úloha e)**	
109	odhad *b*	5,0
110	směrodatná chyba s_b	0,016
111	stupně volnosti	6
112	α	0,05
113	**95% interval spolehlivosti pro koef. b**	
114	**dolní mez**	**horní mez**
115	**4,972**	**5,049**

	A	B
108	**úloha e)**	
109	odhad *b*	=$A$20
110	směrodatná chyba s_b	=$A$21
111	stupně volnosti	=$B$23
112	α	0,05
113	**95% interval spolehlivosti pro koef. b**	
114	**dolní mez**	**horní mez**
115	**=+B109-B110*TINV(B112;B111)**	**=+B109+B110*TINV(B112;B111)**

Interpretace

- 95% interval spolehlivosti pro směrnici regresní přímky je velmi úzký a neobsahuje nulu. Tvrzení o nulové směrnici tedy můžeme zamítnout.