

LINEÁRNÍ REGRESE

Komentované řešení pomocí programu *Statistica*

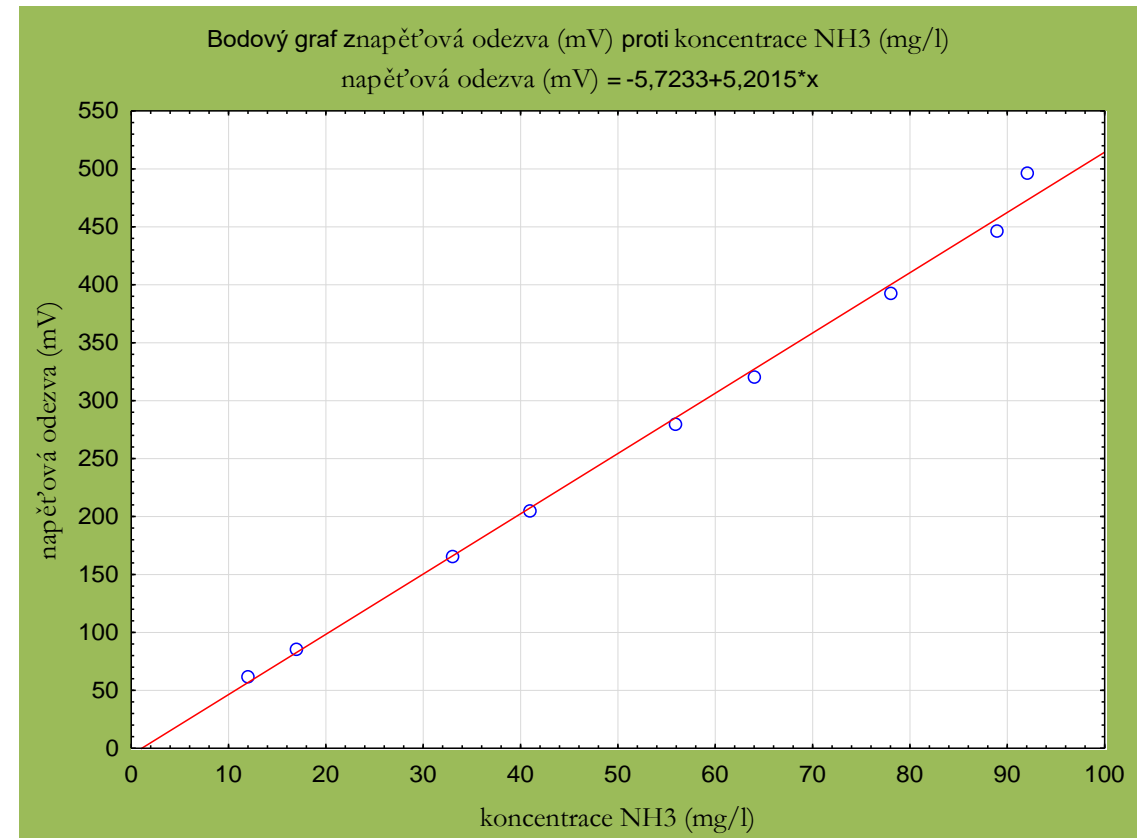
Vstupní data

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

- Data umístěná v *excelovském* souboru překopírujeme do tabulky ve *Statistice* a pojmenujeme proměnné, viz prezentace k tématu **Popisná statistika**.
- Úkolem je proložit našimi daty tzv. **kalibrační křivku** a udělat statistickou analýzu získaných dat.

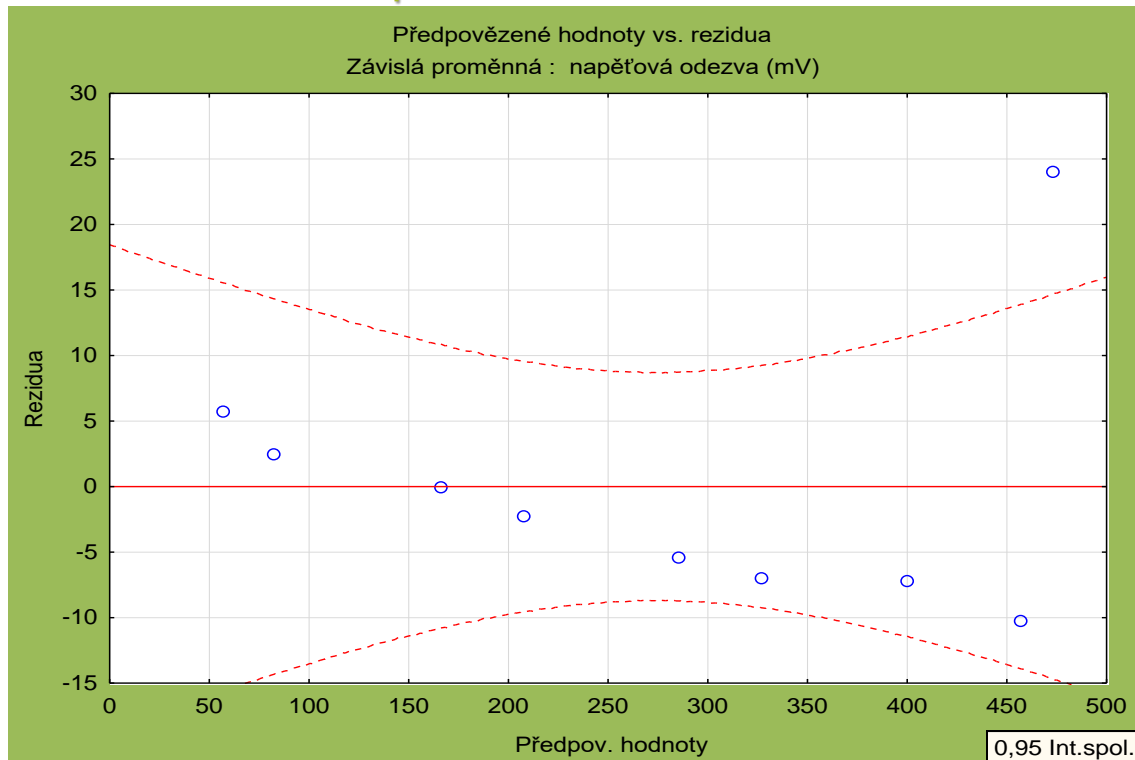
Ověření předpokladů – linearita

- Vykreslíme si graf dat proložených regresní přímkou:
 - **Grafy** → **Bodový graf** → **Proměnné** – na osu x zvolíme Koncentraci a na osu y Napěťovou odezvu → OK → OK
- 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

- Abychom posoudili vzájemnou nekorelovanost i neměnný rozptyl náhodných odchylek, vykreslíme bodový graf **reziduí**:
 - **Statistiky** → **Vícenásobná regrese** → **Proměnné** – nezáv. prom. je Koncentrace a záv. prom. je Napěťová odezva → OK → OK → **Reziduální analýza** → **Bodové grafy** → **Předpovědi vs. rezidua**



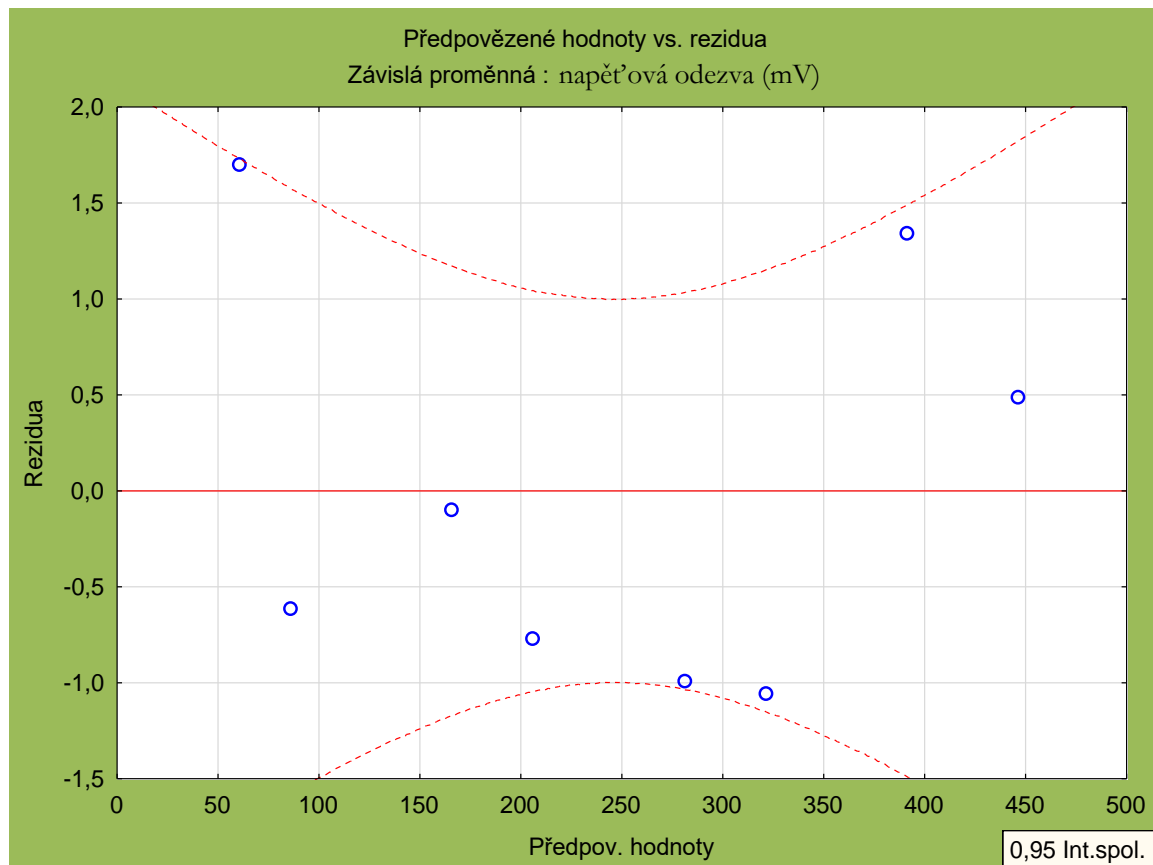
- Rezidua nevypadají náhodně a nesystematicky. Naopak, s výjimkou posledního měření jsou silně pozitivně korelovaná.
- Pro tato data tedy nelze považovat předpoklad o nekorelovanosti náhodných složek za splněný.
- Poslední měření se zjevně vymyká těm ostatním a způsobilo vychýlení regresní přímky, které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

Opravená vstupní data

	1 koncentrace NH ₃ (mg/l)	2 napět'ová odezva (mV)
1	12	62,42
2	17	85,16
3	33	165,84
4	41	205,25
5	56	280,18
6	64	320,2
7	78	392,74
8	89	447

- Data se jednoduše opraví tak, že se vymaže poslední řádek z tabulky.
- Provedeme opět analýzu reziduí zcela stejně, jako s původními daty.

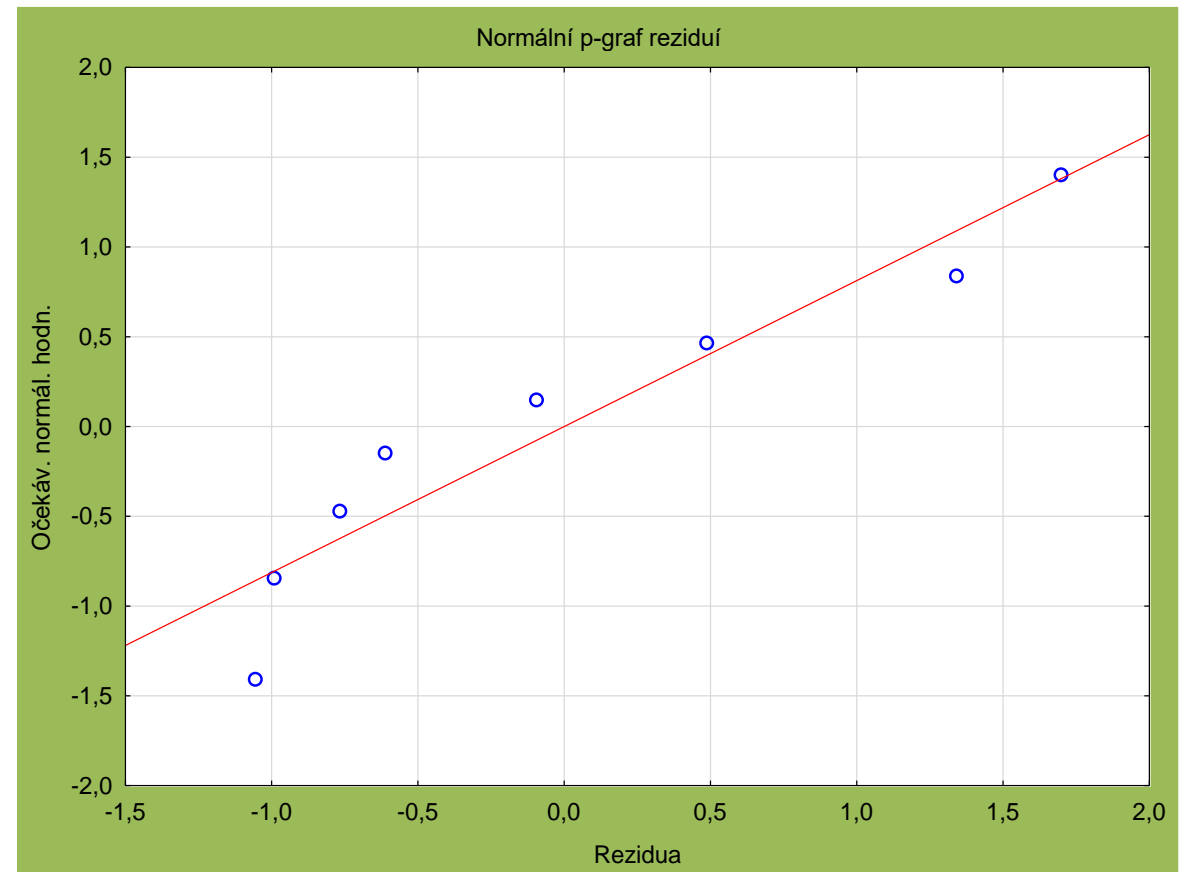
Rezidua – opravená data



Rezidua už vypadají náhodně,
nesystematicky.

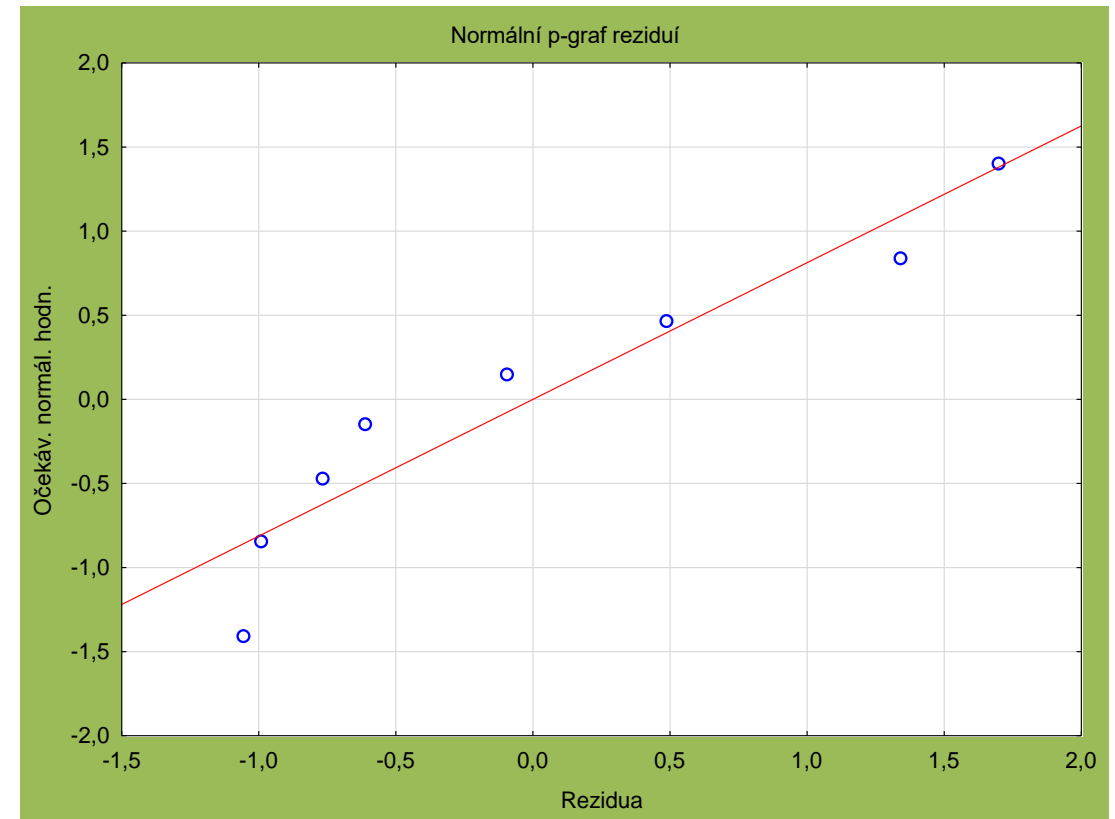
Ověření předpokladů – normalita I

- Pro posouzení normality použijeme Q – Q plot. Na histogram máme příliš málo pozorování.
- Statistiky → Vícenásobná regrese → Proměnné – nezáv. prom. je Koncentrace a záv. prom. je Napěťová odezva → OK → OK → Residuální analýza → Normální p – graf reziduí



Ověření předpokladů – normalita II

- Je patrný rozdíl mezi skutečnými kvantily (modré body) a těmi gaussovskými (červená čára).
- Vzhledem k malému počtu pozorování však tento rozdíl nemusí být statisticky významný.
- V dalším tedy budeme předpokládat normalitu (bez ní bychom nedokázali úlohu 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.



Rovnice kalibrační křivky

- Rovnici kalibrační křivky $y = 0,597 + 5,01 x$ dostaneme následovně:
 - Statistiky → Vícenásobná regrese → Proměnné – nezáv. prom. je Koncentrace a záv. prom. je Napěťová odezva → OK → OK → Výpočet: Výsledky regrese

		Výsledky regrese se závislou proměnnou : napěťová odezva (mV)					
		R= ,99997065 R2= ,99994130 Upravené R2= ,99993152					
		F(1,6)=1022E2 p<,00000 Směrod. chyba odhadu : 1,1546					
N=8		b*	Sm.chyba z b*	b	Sm.chyba z b	t(6)	p-hodn.
Abs.člen				0,597395	0,866208	0,6897	0,516176
koncentrace NH3 (mg/l)		0,999971	0,003128	5,010284	0,015672	319,7024	0,000000

- Hypotézu $H_0: \beta_1 = 0$ nemůžeme na hladině 5 % zamítnout ve prospěch alternativy $H_1: \beta_1 \neq 0$, neboť p – hodnota testu je **0,516**. Jinými slovy, nelze vyloučit, že kalibrační křivka prochází počátkem (ovšem nemůžeme to ani potvrdit). **Tím jsme vyřešili úlohu (A).**
- Hypotézu $H_0: \beta_2 = 0$ zamítáme na hladině 5 % ve prospěch alternativy $H_1: \beta_2 \neq 0$, neboť p – hodnota testu je blízká nule. Tvrzení o nulové směrnici můžeme tedy zamítnout. **Tím jsme částečně vyřešili úlohu (E).**

Úlohy (A) a (E) – intervaly spolehlivosti

- Interval spolehlivosti pro β_1 lze získat přímo:
 - zvolíme si **Rezidua/předpoklady/předpovědi** → **Předpověď závislé proměnné** → **Koncentrace** – necháme **0** (interval spolehlivosti pro regresní funkci je pro $x = 0$ roven intervalu spolehlivosti pro parameter β_1) → **OK**

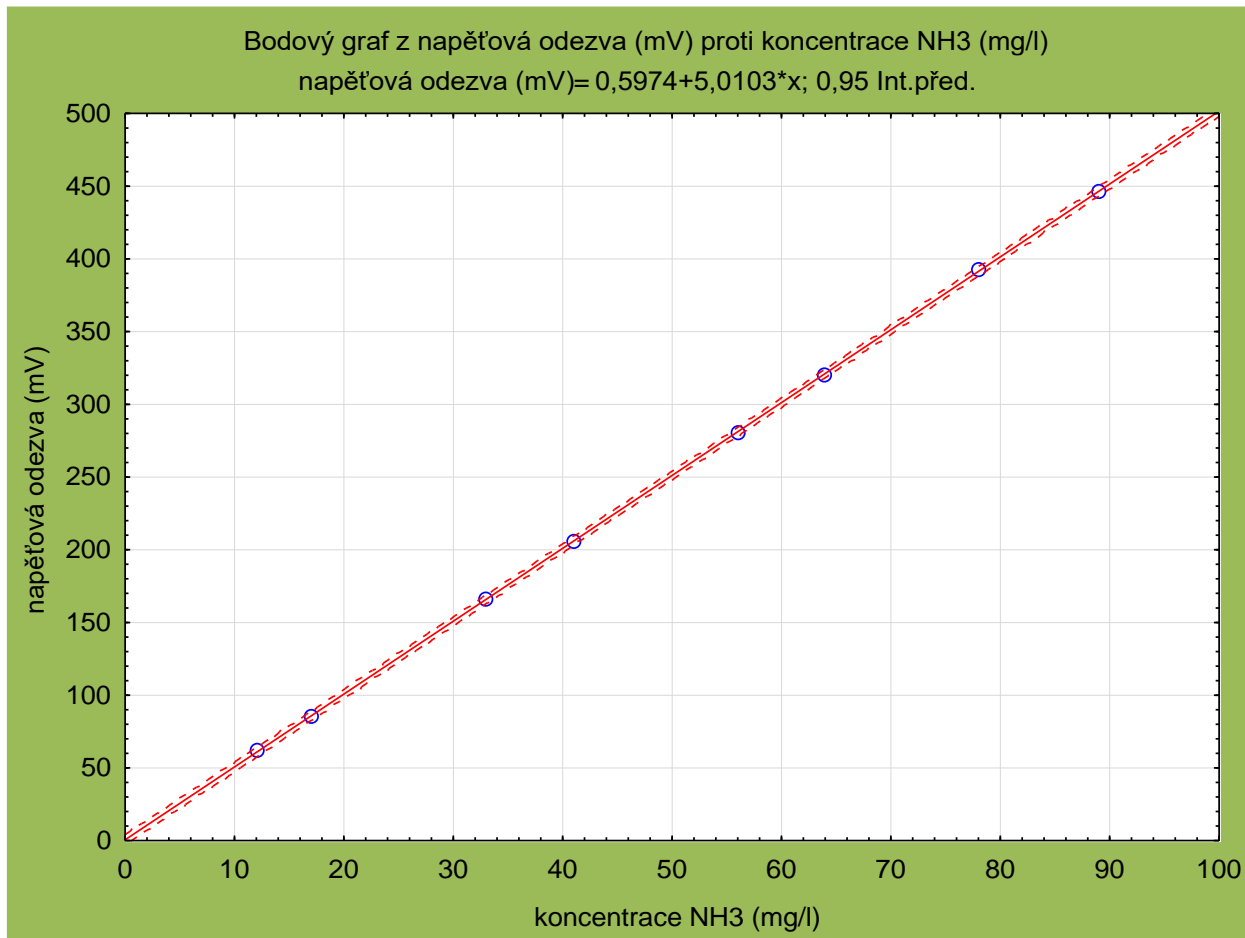
Proměnná	Předpovězené hodnoty proměnné: napěťová odezva (mV)		
	b-váha	Hodnota	b-váha * Hodnot
koncentrace NH3 (mg/l)	5,010284	0,00	0,00000
Abs. člen			0,59740
Předpověď			0,59740
-95,0%LS			-1,52214
+95,0%LS			2,71693

- Vidíme, že $0 \in [-1,522; 2,717]$, proto nulovost β_1 nemůžeme zamítnout, ale ani potvrdit.

- Interval spolehlivosti pro β_2 ovšem podobným způsobem získat nelze, proto se musí spočítat přímo (viz např. prezentace v *Excelu*) – $[4,972; 5,049] \not\ni 0$, takže nulovost směrnice β_2 zamítáme na hladině 5 %.

Úloha (C) – opravená data

- Přibližně 95 % měření by mělo ležet v predikčním pásu.



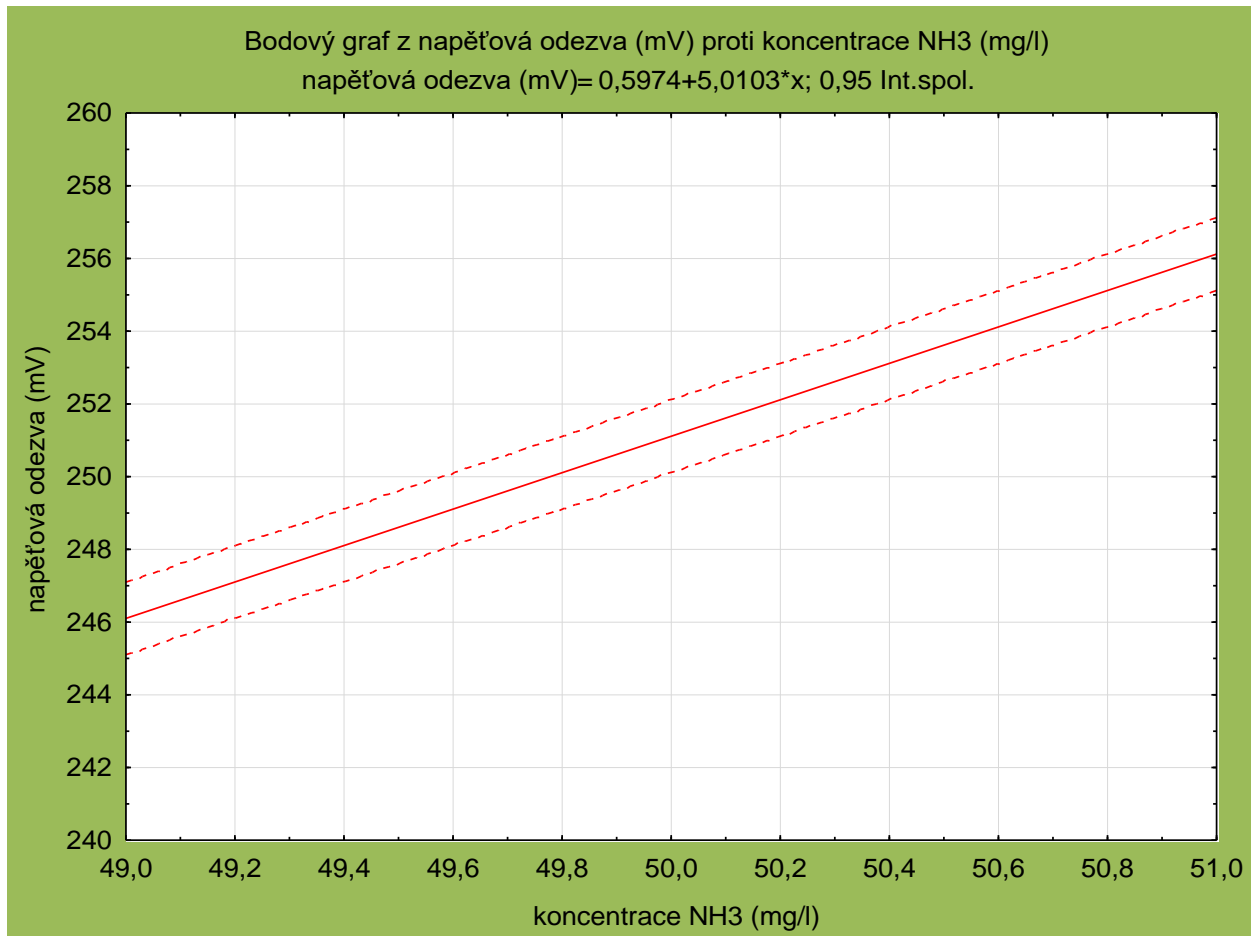
- Grafy → Bodový graf → Proměnné – osa X je koncentrace, osa Y je napěťová odezva → Regresní pásy – Predikce → OK
- V našem případě je všech 8 měření uvnitř tohoto pásu, což není v rozporu s očekáváním.

Úloha (B) – opravená data

- 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.

Úloha (D) – opravená data

- V úloze (C) zvolíme v **Regresních pásech** místo **Predikce Spolehlivost**, a aby bylo něco vidět, v **Měřítku** zvolíme na ose



- X ručně **Minimum 49** a **Maximum 51**, což jsou meze intervalu $50 \pm 0,02 \times 50$;
- Y ručně **Minimum 240** a **Maximum 260**.
- Jelikož úsečka o rovnici
$$y = 0,597 + 5,01x50, x \in [49,51],$$
překrývá celý 95% pás spolehlivosti, lze určit koncentraci amoniaku 50 mg/l s přesností větší než 2 % (pás spolehlivosti je výrazně užší než požadovaná přesnost).