FSTA: Pokročilé statistické metody

Lineární modely – základy

Jiří Jarkovský, Simona Littnerová

FSTA: Pokročilé statistické metody

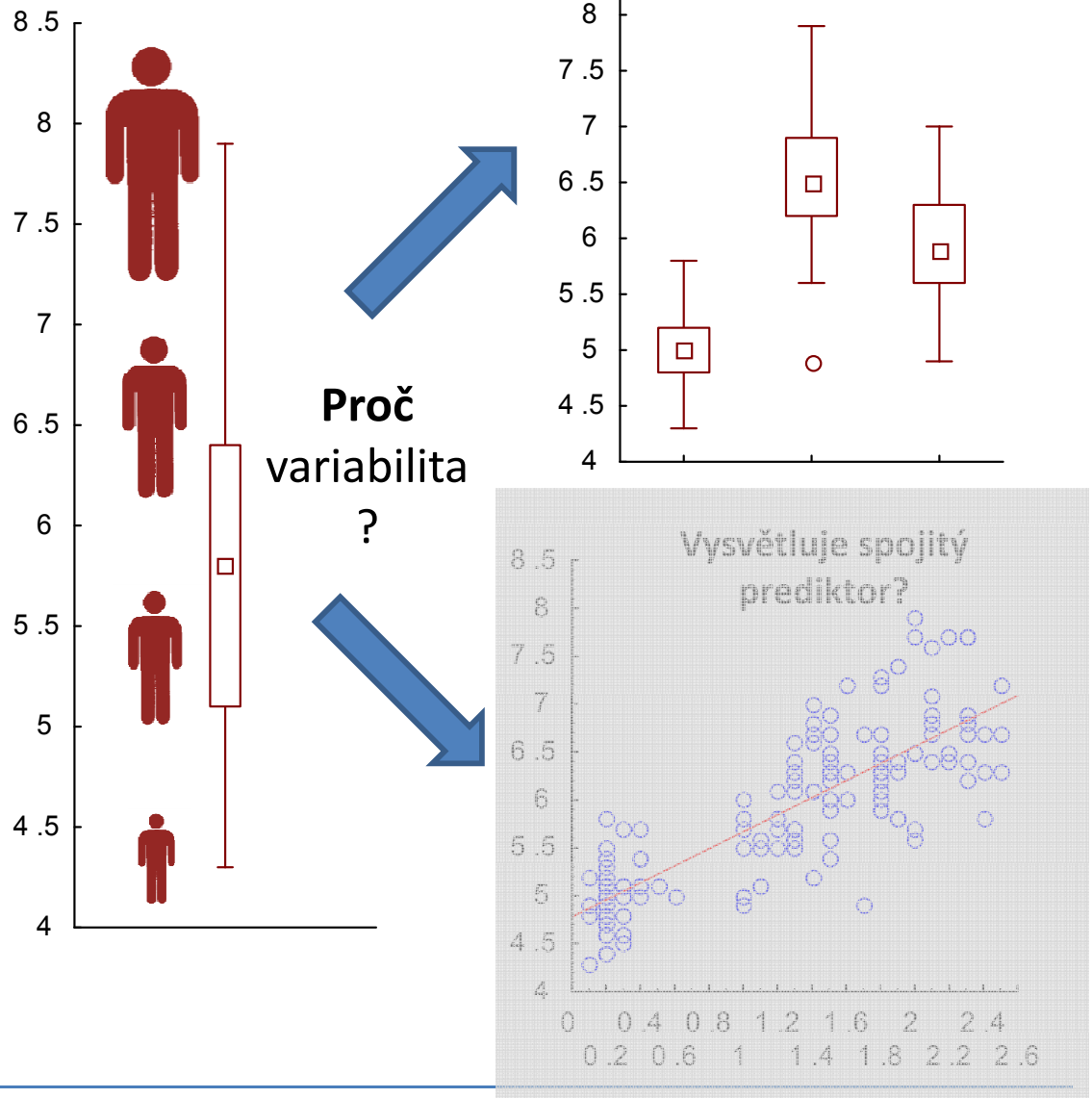
Stochastické modelování obecně - ANOVA

ANOVA

- Analýza rozptylu je základním nástrojem pro analýzu rozdílů mezi průměry v několika skupinách pacientů.
- Základní myšlenka, na níž je ANOVA založena, je rozdělení celkové variability v datech (neznámé, dané pouze náhodným rozložením) na část systematickou (spjatou s kategoriemi pacientů, vysvětlená variabilita) a část náhodnou. Pokud systematická, tedy nenáhodná a vysvětlitelná část variability převažuje, považujeme daný kategoriální faktor za významný pro vysvětlení variability dat.
- Analýza rozptylu vyhodnocuje pouze celkový vliv faktoru na variabilitu, v případě analýzy jednotlivých kategorií je třeba využít tzv. post-hoc testy

Cíl stochastického modelování

- Obecným cílem je snaha **vysvětlit variabilitu predikované proměnné** (endpoint, Y) pomocí **prediktorů** (vysvětlující proměnná, faktor, X)
- Jak predikovaná proměnná, tak prediktor mohou být různého typu
  - Binární
  - Kategoriální
  - Ordinální
  - Spojitá
  - Cenzorovaná (-> analýza přežití)
- Kombinace datového typu predikované proměnné a prediktoru určuje použitou metodu analýzy



ANOVA – předpoklady

- Symetrické rozložení hodnot a normalita odchylek od hodnoceného modelu ANOVA. Velkou část dat lze adekvátně normalizovat použitím logaritmické transformace. Předpoklad lognormální transformace může pochopitelně být teoreticky vyloučen u mnoha datových souborů obsahujících diskrétní parametry, kde je indikována vhodnost jiného typu transformace. U asymetricky rozložených a u diskrétních dat je nutné využít neparametrické alternativy analýzy rozptylu.
- Homogenita rozptylu je nutným předpokladem pro smysluplnost vzájemných srovnání pokusných variant. U testů toxicity by splnění tohoto předpokladu mělo být ověřováno (Bartlettův test), neboť vážné rozdíly (až řádové) v jednotkách testovaného parametru mohou nastat v důsledku inhibice dávkami látky. Nehomogenita rozptylu je často ve vztahu k nenormalitě (asymetrii) dat a lze ji odstranit vhodnou normalizující transformací.
- Statistická nezávislost reziduí vyhodnocovaného modelu ANOVA. Pokud odhad a posouzení korelačních vztahů mezi pokusnými variantami není přímo předmětem výzkumu, lze jejich vliv na vyhodnocení odstranit znáhodněním dat v rámci pokusných variant - tedy změnou pořadí v náhodné. Rozsah vlivu těchto autokorelačních vztahů musí být ovšem primárně omezen správností experimentálního uspořádání.
- Aditivita jako předpoklad týkající se složitějších experimentálních uspořádání. Exaktní otestování aditivity více pokusných faktorů je procedura poměrně náročná na experimentální design vyvážený co do počtu opakování. Je rovněž obtížné testovat interakci na nestandardních datech, neboť případná transformace může změnit charakter odchylek původních dat od hodnoceného modelu ANOVA.

Princip ANOVA

- Základním principem ANOVY je porovnání rozptylu připadajícího na:
  - Rozdělení dat do skupin (tzv. effect, variance between groups)
  - Variabilitu objektů uvnitř skupin (tzv. error, variance within groups), předpokládá se, že jde o náhodnou variabilitu (=error)

1. Variabilita mezi skupinami

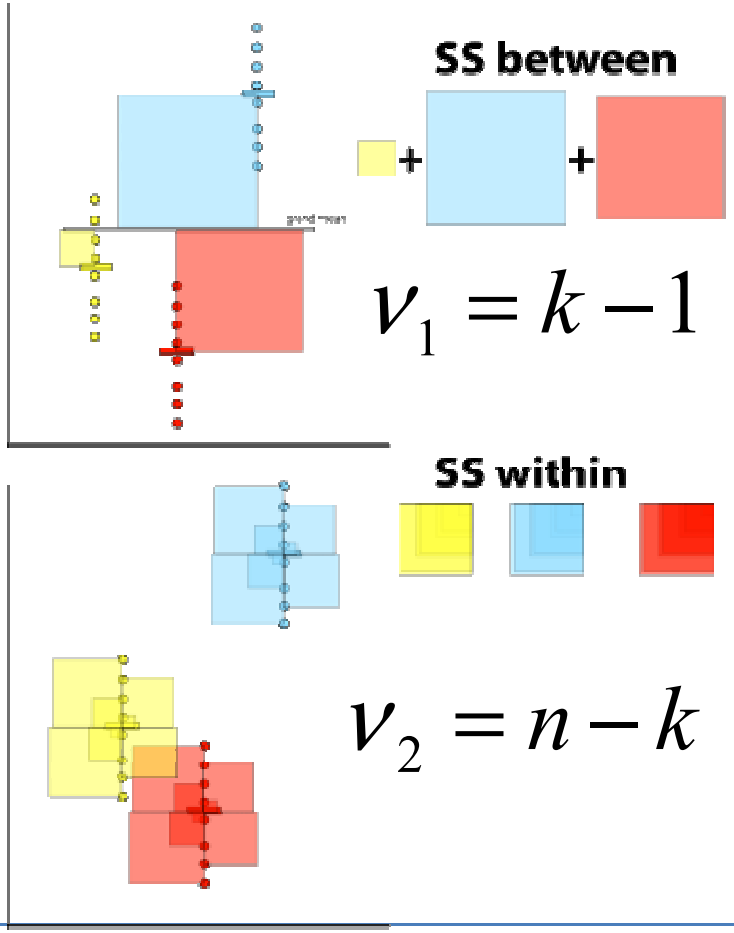
Rozptyl je počítán pro celkový průměr (tzv. grand mean) a průměry v jednotlivých skupinách dat

Stupně volnosti jsou odvozeny od počtu skupin (= počet skupin -1)

2. Variabilita uvnitř skupin

Rozptyl je počítán pro průměry jednotlivých skupin a objekty uvnitř příslušných, celková variabilita je pak sečtena pro všechny skupiny

Stupně volnosti jsou odvozeny od počtu hodnot (= počet hodnot - počet skupin)



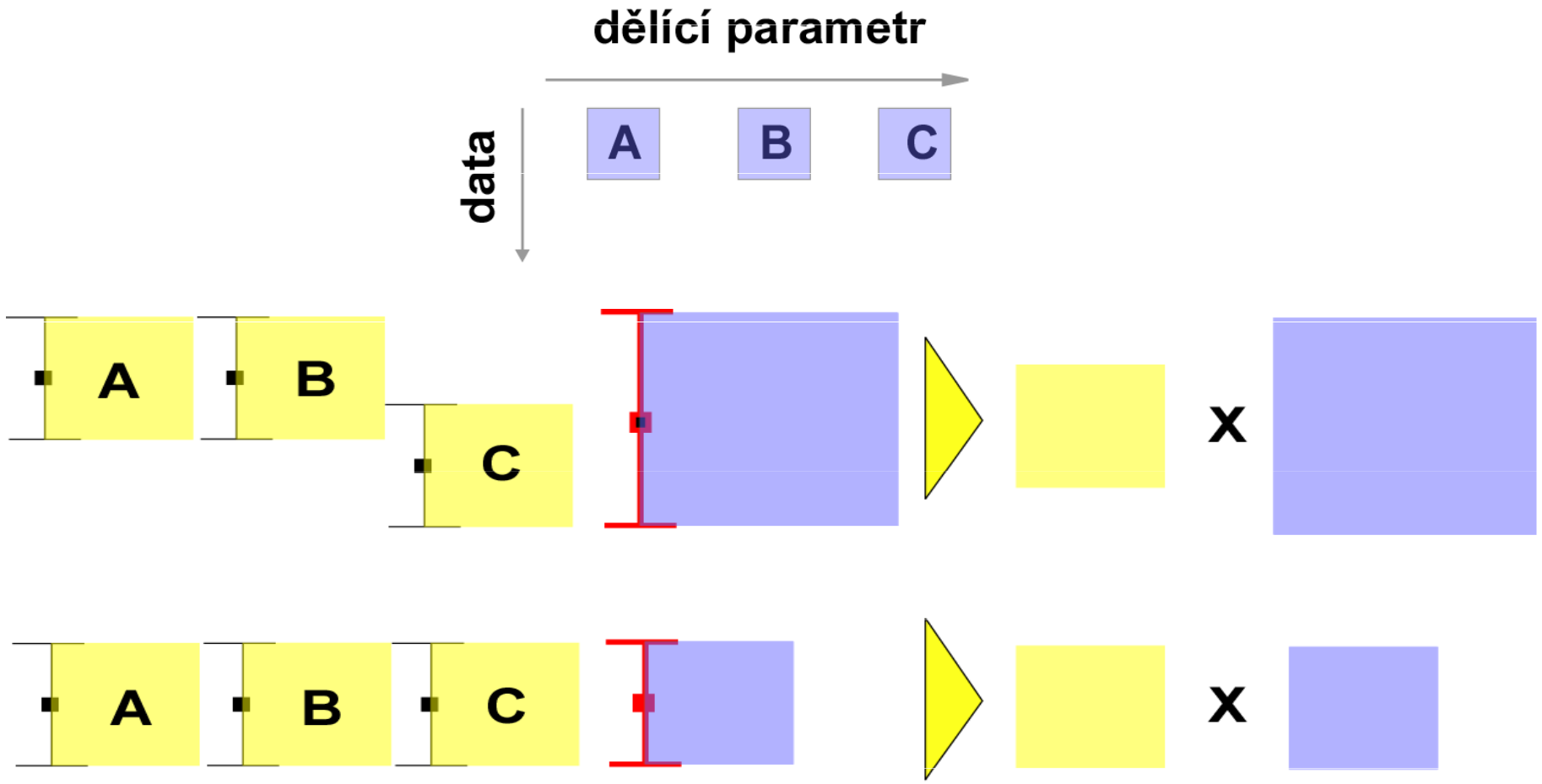
$$F = \frac{between\_groups}{within\_groups}$$

Výsledný poměr (F) porovnáme s tabulkami F rozložení pro v_1 a v_2 stupňů volnosti

SS=sum of squares

Jednoduchý ANOVA design

- Nejjednodušším případem ANOVA designu je rozdělení na skupiny podle jednoho parametru

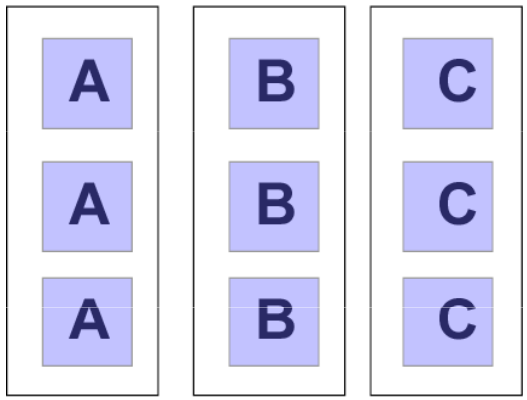


Nested ANOVA

- Rozdělení skupin na náhodné podskupiny (např. opakování experimentu)
- Cílem je zjistit, zda data v jedné skupině nejsou pouhou náhodou
- Nejprve je testována shoda podskupin v hlavních skupinách,
  - pokud jsou shodné, je vše v pořádku
  - pokud nejsou, stále lze zjišťovat, zda se variabilita uvnitř hlavních skupin liší od celkové variability

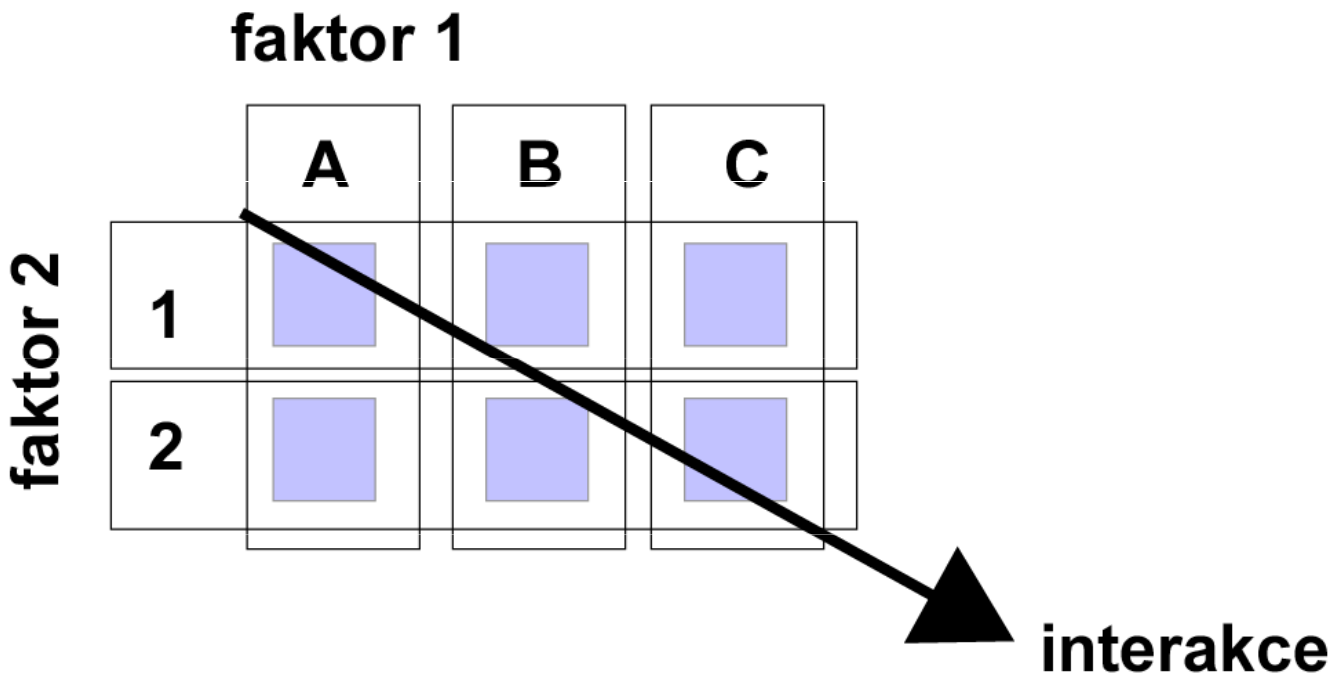
jednoduchá ANOVA

nested ANOVA



Two way ANOVA

- Pro rozdělení do kategorií je zde více parametrů
- Na rozdíl od nested ANOVY nejde o náhodná opakování experimentu, ale o řízené zásahy (např.vliv pH a koncentrace O2)
- Kromě vlivu hlavních faktorů se uplatňuje i jejich interakce



ANOVA – základní výstup

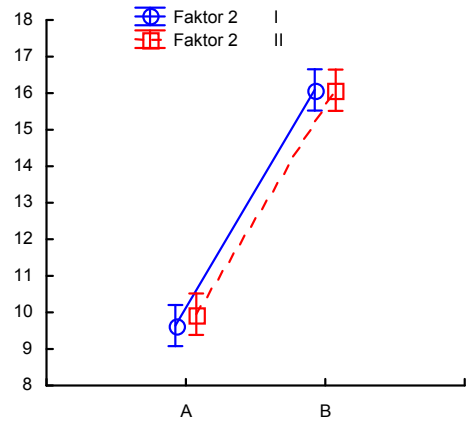
- Základním výstupem analýzy rozptylu je Tabulka ANOVA - frakcionace komponent rozptylu

Zdroj rozptylu	St. v.	SS	MS	F
Pok. zásah (mezi skupinami)	**a -1**	**SS_B**	**$SS_B/(a-1)$**	**MS_B/MS_E**
Uvnitř skupin	**N - a**	**SS_E**	**$SS_E/(N-a)$**	
Celkem	**N -1**	**SS_T**		

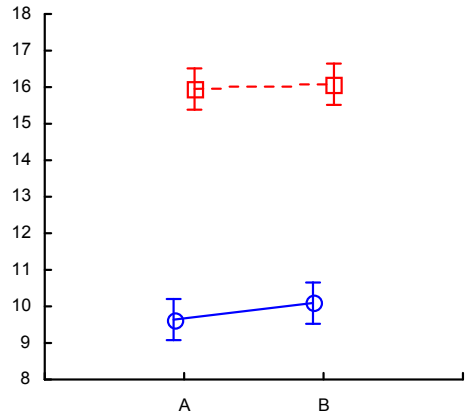
SS_B/SS_T → Kvantifikovaný podíl rozdílu mezi pokusnými zásahy na celkovém rozptylu

MS_B/MS_T → Statistická významnost rozdílu

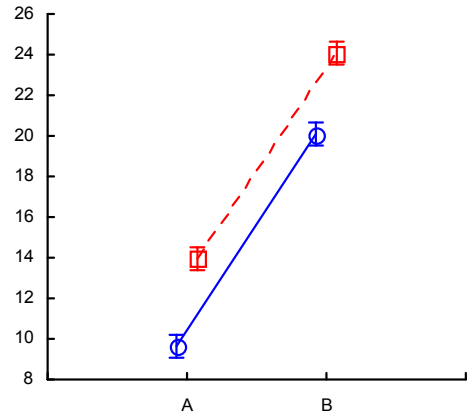
Hlavní efekty a interakce



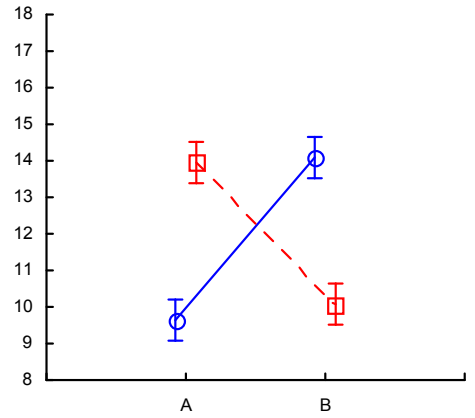
	SS	D.f.	MS	F	p
Intercept	33487	1	33487	8165.3	0.000
Faktor 1	**1978**	**1**	**1978**	**482.2**	**0.000**
Faktor 2	1	1	1	0.3	0.602
F1*F2	1	1	1	0.3	0.570
Error	804	196	4		



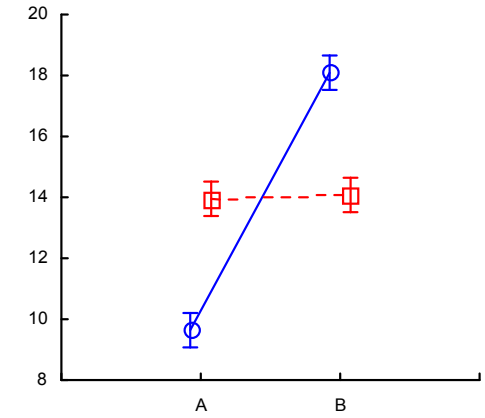
	SS	D.f.	MS	F	p
Intercept	33487	1	33487	8165.3	0.000
Faktor 1	4	1	4	1.0	0.314
Faktor 2	**1891**	**1**	**1891**	**461.1**	**0.000**
F1*F2	1	1	1	0.3	0.570
Error	804	196	4		



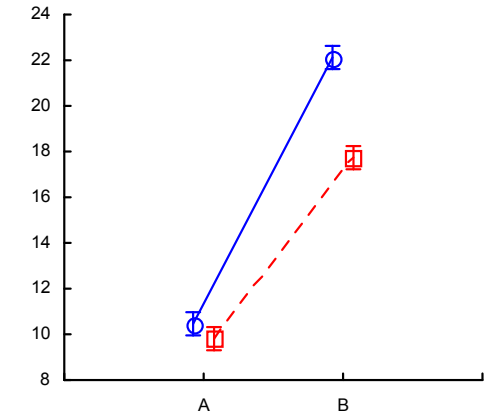
	SS	D.f.	MS	F	p
Intercept	57391	1	57391	13993	0.000
Faktor 1	5293	1	5293	1290.7	0.000
Faktor 2	861	1	861	209.9	0.000
F1*F2	1	1	1	0.3	0.570
Error	804	196	4		



	SS	D.f.	MS	F	p
Intercept	28511	1	28511	6952.0	0.000
Faktor 1	4	1	4	1.0	0.314
Faktor 2	1	1	1	0.3	0.602
F1*F2	**867**	**1**	**867**	**211.3**	**0.000**
Error	804	196	4		



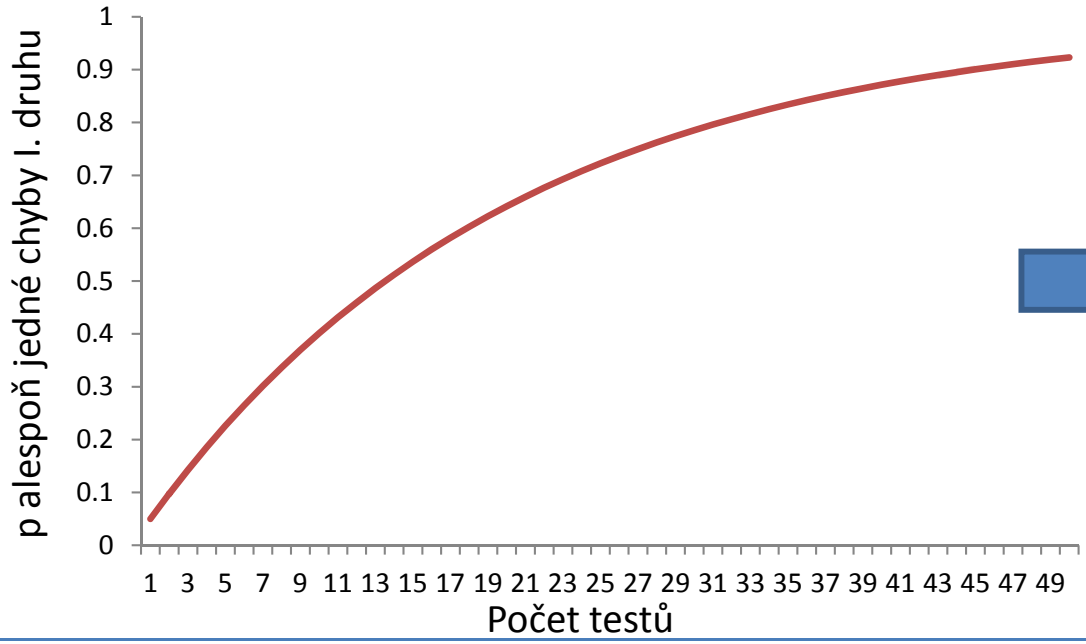
	SS	D.f.	MS	F	p
Intercept	38863	1	38863	9476.2	0.000
Faktor 1	**920**	**1**	**920**	**224.3**	**0.000**
Faktor 2	1	1	1	0.3	0.602
F1*F2	**867**	**1**	**867**	**211.3**	**0.000**
Error	804	196	4		



	SS	D.f.	MS	F	p
Intercept	45203	1	45203	13596	0.000
Faktor 1	**4799**	**1**	**4799**	**1443.4**	**0.000**
Faktor 2	**316**	**1**	**316**	**95.0**	**0.000**
F1*F2	**175**	**1**	**175**	**52.5**	**0.000**
Error	652	196	3		

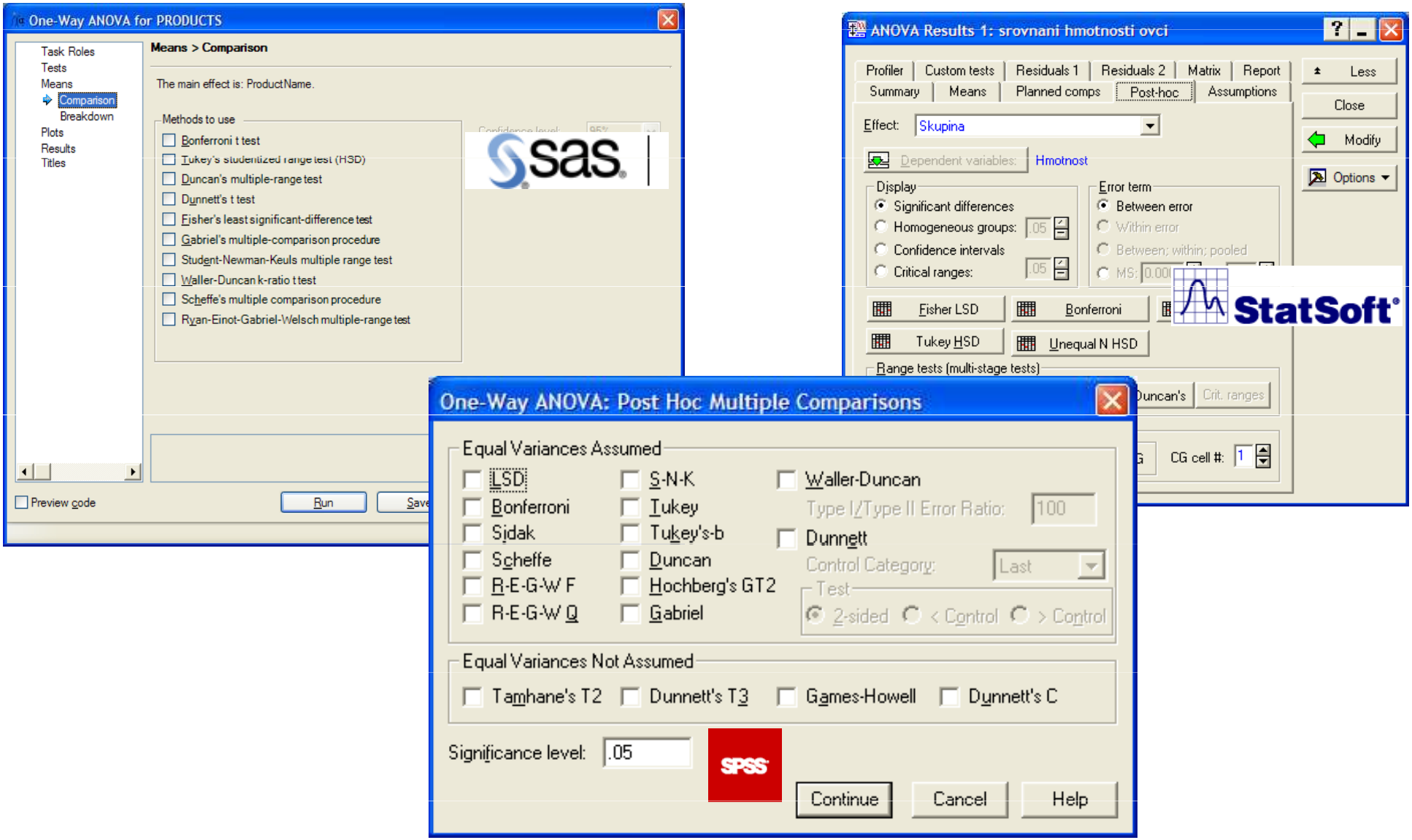
Testování dílčích hypotéz

- V řadě analýz je třeba pracovat se vzájemným testováním více skupin objektů stylem každý s každým
- Obecný postup analýzy je
  - Testování celkové významnosti – všechny skupiny navzájem (ENG: among groups)
  - Pokud je zjištěna celková významnost pokračuje testování analýzou již konkrétních kombinací dvojic skupin (ENG: between)
- Problémem je vliv mnohonásobného testování na statistickou významnost testů:
  - Každý jeden test má α=0.05 (chyba I. druhu)
  - Při mnohonásobném testování stoupá pravděpodobnost, že alespoň u jednoho testu dojde k chybnému zamítnutí nulové hypotézy (tedy k chybě I. druhu)



Řešením jsou různé procedury korigující hodnotu p (např. Bonferroniho korekce, FWR, FDR procedury apod.)

Řada různých post-hoc testů



Příklad: Anova - One way

Dávka rostlinného stimulátoru (0, 4, 8, 12 mg/l)

A = 4 ; n = 8

I. ANOVA

Bartlett's test: P = 0,9847

K-S test: P = 0,482 - 0,6525 pro jednotlivé kategorie

Source	D.f.	SS	MS	F	p
Between	3	305.8	101.9	8.56	<0.001
Within	28	322.2	11.9		
Total	31	638			

II. Multiple Range Test (NKS –test)

Level	Average	Homogeneous groups		
0	34.8	x		
4	41.4		x	
12	41.8		x	
8	52.6			x

FSTA: Pokročilé statistické metody

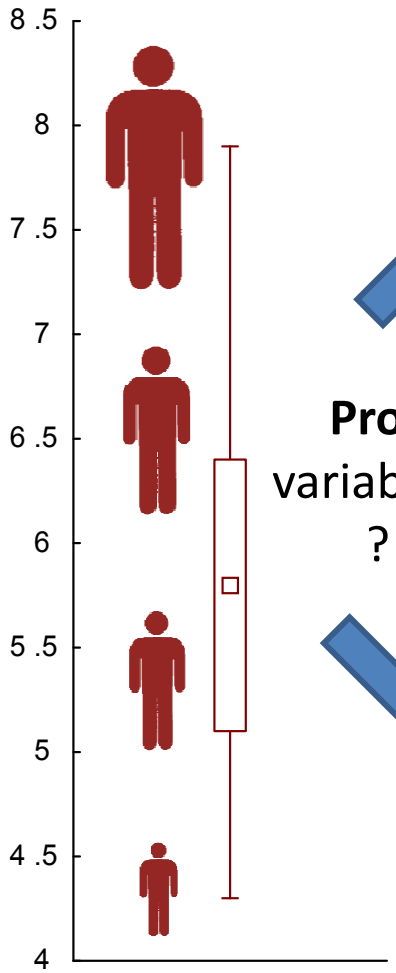
Stochastické modelování obecně – Lineární regrese

Lineární regrese

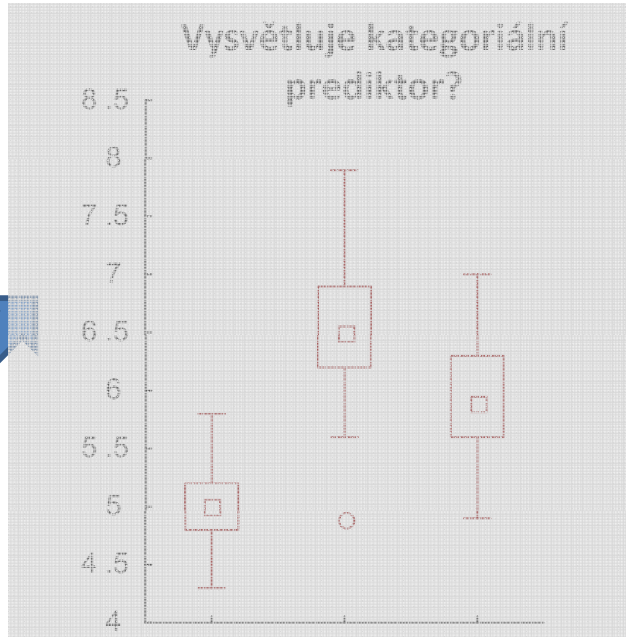
- Korelační analýza je využívána pro vyhodnocení míry vztahu dvou spojitých proměnných. Obdobně jako jiné statistické metody, i korelace mohou být parametrické nebo neparametrické
- Regresní analýza vytváří model vztahu dvou nebo více proměnných, tedy jakým způsobem jedna proměnná (vysvětlovaná) závisí na jiných proměnných (prediktorech). Regresní analýza je obdobně jako ANOVA nástrojem pro vysvětlení variability hodnocené proměnné

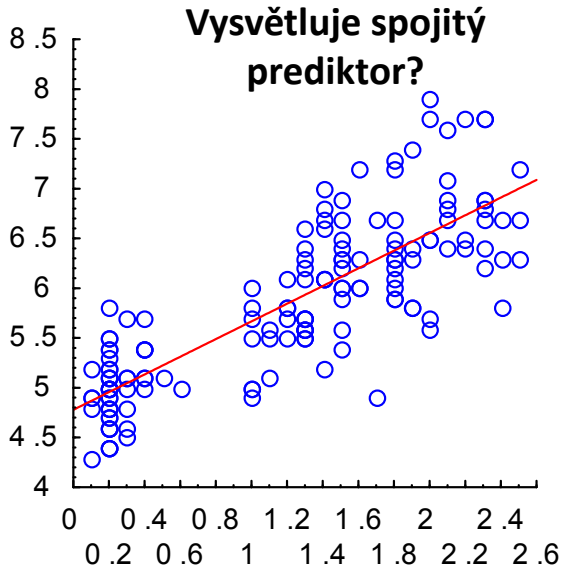
Cíl stochastického modelování

- Obecným cílem je snaha **vysvětlit variabilitu predikované proměnné** (endpoint, Y) pomocí **prediktorů** (vysvětlující proměnná, faktor, X)
- Jak predikovaná proměnná, tak prediktor mohou být různého typu
  - Binární
  - Kategoriální
  - Ordinální
  - Spojitá
  - Cenzorovaná (-> analýza přežití)
- Kombinace datového typu predikované proměnné a prediktoru určuje použitou metodu analýzy



Proč variabilita ?





Základy regresní analýzy

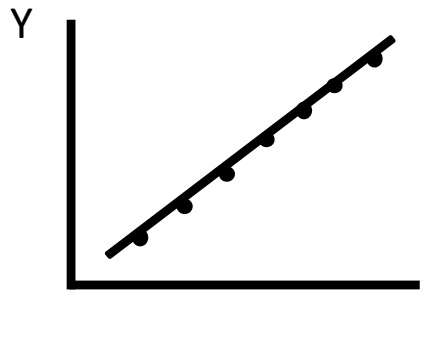
- Regrese - funkční vztah dvou nebo více proměnných

Jednorozměrná
y = f(x)

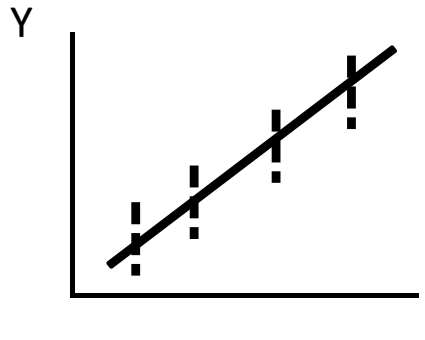
Vícerozměrná
y = f(x1, x2, x3, ……xp)

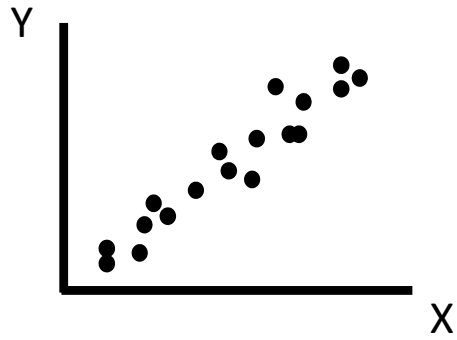
Vztah x, y

Deterministický



Regresní, stochastický





Pro každé x existuje pravděpodobnostní rozložení y

Lineární regrese I

$$Y = a + b \cdot x + e \quad \approx \quad \alpha + \beta \cdot X + \varepsilon$$

y —

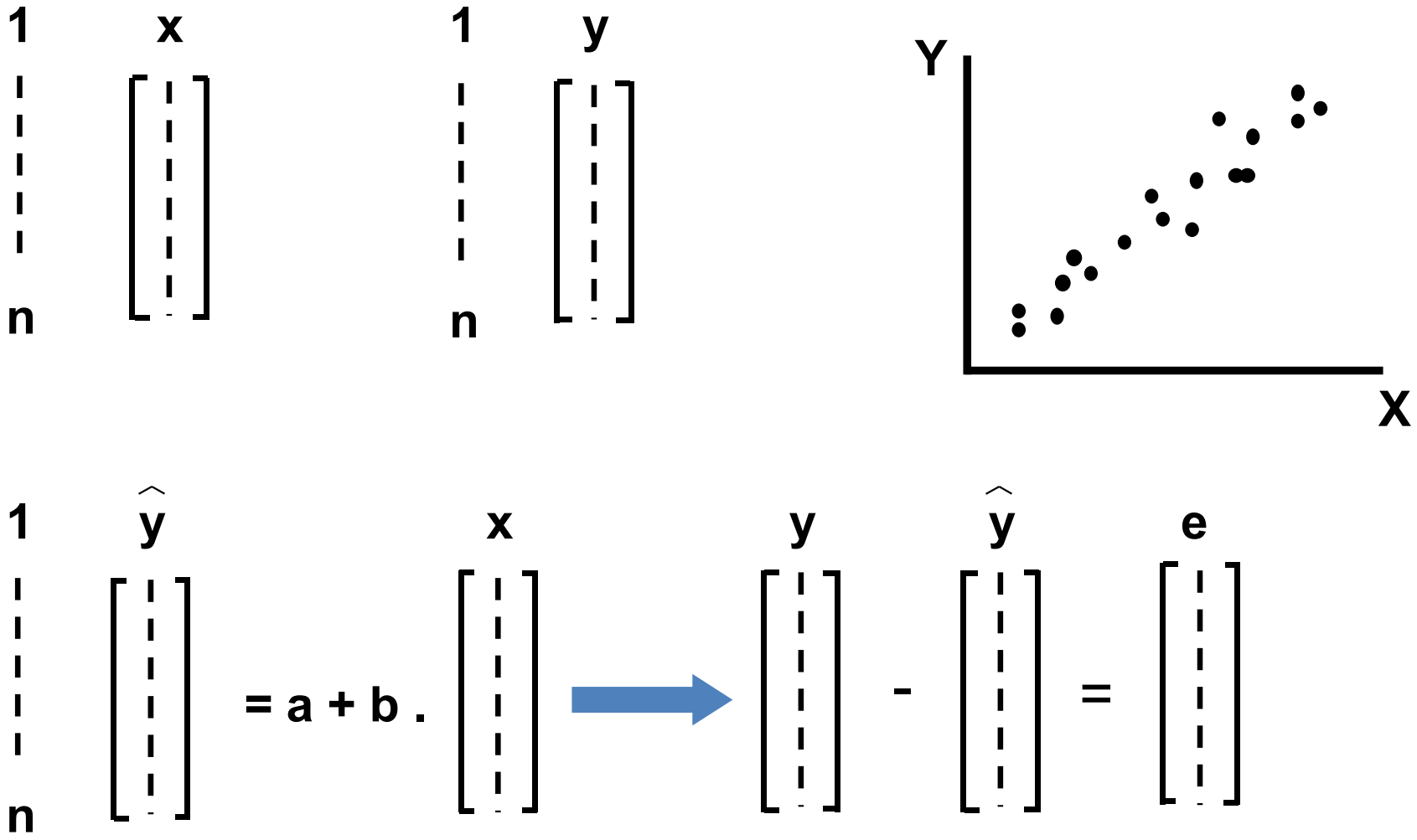
- $\alpha \approx a$ **(intercept):** $a = \bar{y} - b \cdot \bar{x}$
- $\beta \cdot X \approx b \cdot x$ **(sklon; slope)**
- $\varepsilon \approx e$ **- náhodná složka :** $\mathrm{N}\left(0; \sigma_e^2\right) = N\left(0; \sigma_y^2 x\right)$

} Komponenty tvořící y se sčítají

ε - náhodná složka modelu přímky = rezidua přímky

$$\sigma_e^2 \left(\sigma_{y \cdot x}^2\right) \Rightarrow \textbf{ rozptyl reziduí}$$

Lineární regrese II

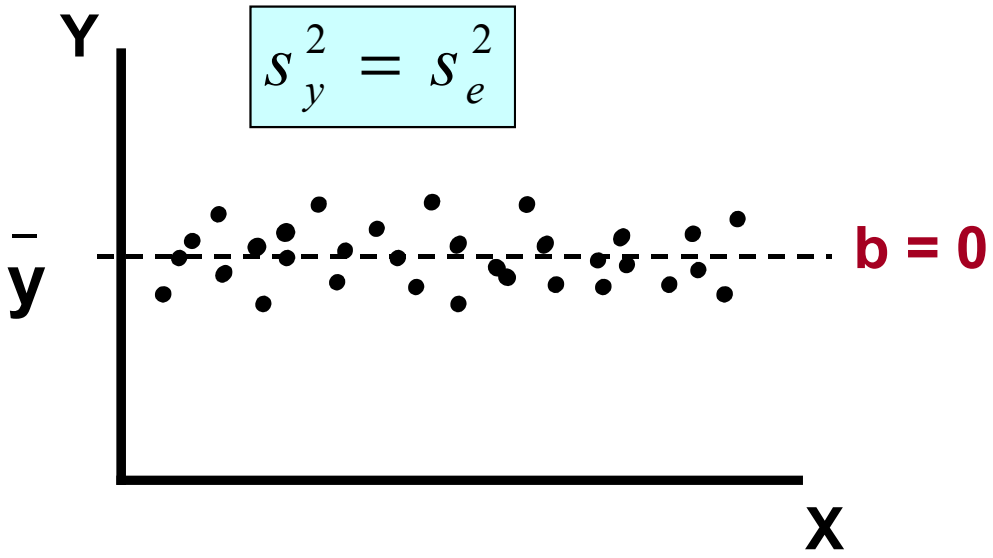


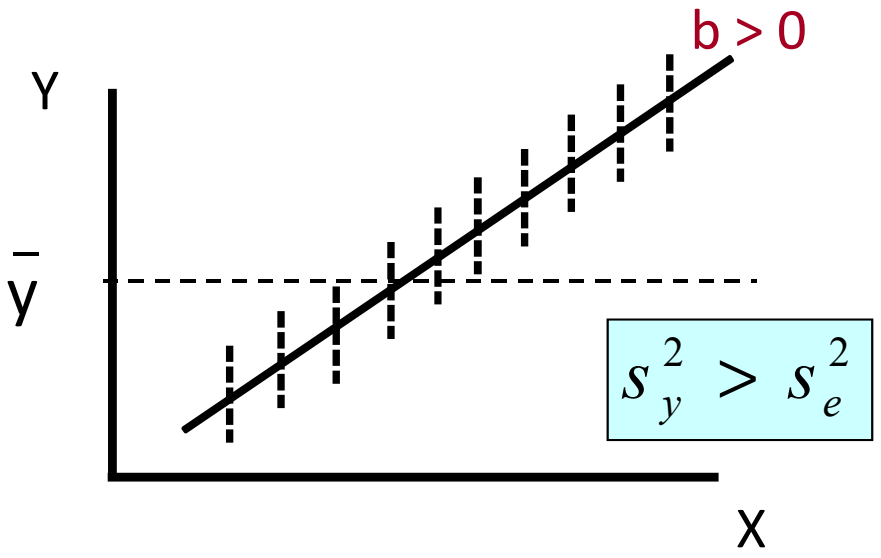
Lineární regrese III

$$\mathbf{x} \quad \mathbf{y} \quad \mathbf{y} \quad \mathbf{e}$$

$$\begin{bmatrix} \\ \\ \\ \end{bmatrix} \quad \begin{bmatrix} \\ \\ \\ \end{bmatrix} s_y^2 \quad \begin{bmatrix} \\ \\ \\ \end{bmatrix} \quad \begin{bmatrix} \\ \\ \\ \end{bmatrix} s_e^2$$

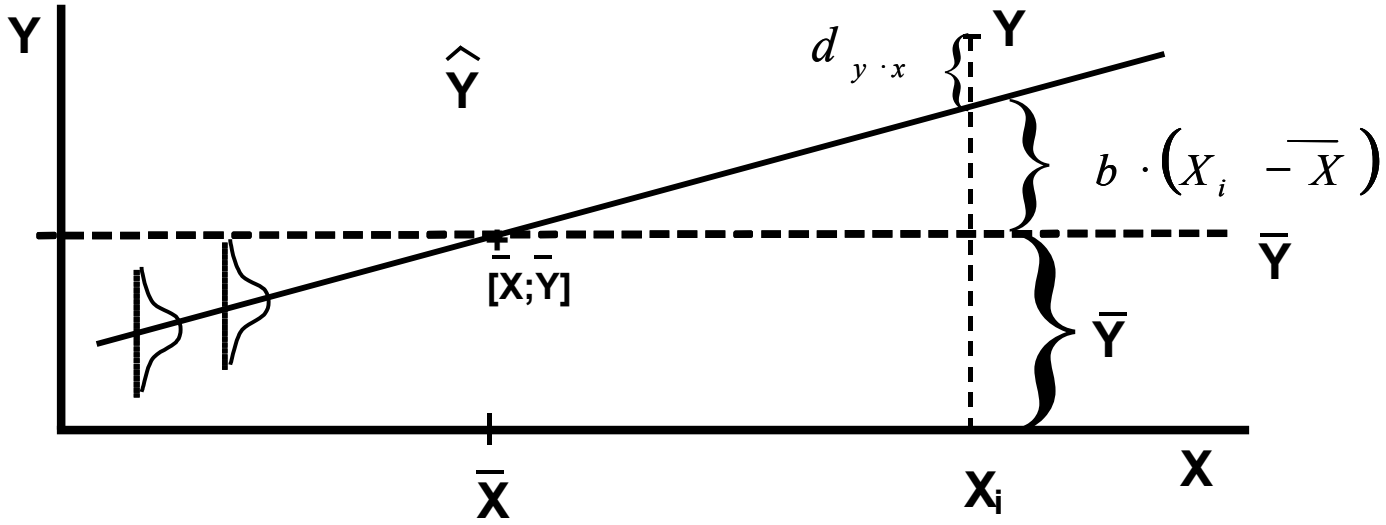
$$\bar{x} \quad \bar{y} \quad \bar{\hat{y}} \quad \bar{e} = 0$$





Lineární regrese III

- Metoda nejmenších čtverců
  - X: Pevná, nestochastická proměnná
  - Rozložení hodnot y pro každé x je normální
  - Rozložení hodnot y pro každé x má stejný rozptyl
  - Rezidua jsou navzájem nezávislá a mají normální rozložení



$$d_{y\cdot x} = y - \hat{y} \qquad \boxed{d_{y\cdot x} = y - \bar{y} - b(X_i - \bar{X})} \qquad \hat{y} = \bar{y} + b(X_i - \bar{X})$$

Smysl proložení přímky
minimalizace odchylek

$$d^2_{y\cdot x} \rightarrow \sum \left[y - \hat{\alpha} - \hat{\beta}(X_i - \bar{X}) \right]$$

Lineární regrese IV

I. $b \sim \beta$: $b = \dfrac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}$ $\qquad S_b^2 \sim \sigma_\beta^2$: $\dfrac{1}{\sum (X_i - \bar{X})^2} \cdot S_{y \cdot x}^2$

$S_{y \cdot x}^2$ = **mean squared deviation from regression**

$S_{y \cdot x}$ = **sample standard deviation from regression**

$$S_{y \cdot x}^2 = \frac{\sum d_{y \cdot x}^2}{n-2} = \frac{\sum Y_i^2 - \dfrac{\sum Y_i^2}{n} - b^2 \cdot \sum (X_i - \bar{X})^2}{n-2}$$

II. $a \sim \alpha$: $a = \bar{Y} - b \cdot \bar{X}$ $\qquad S_a^2 \sim \sigma_\alpha^2$ $\qquad S_\alpha^2 = \left[\dfrac{1}{n} + \dfrac{\bar{X}^2}{\sum X^2} \right] \cdot S_{y \cdot x}^2$

intercept

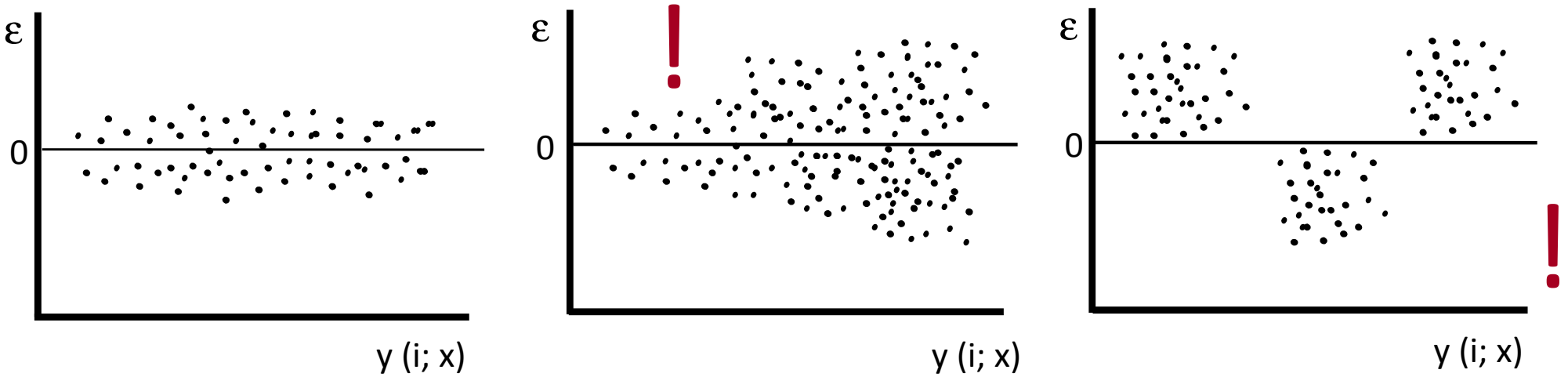
III. $\hat{Y}$: **modelová hodnota**

$$\hat{Y}_i = a - b \cdot X_i$$

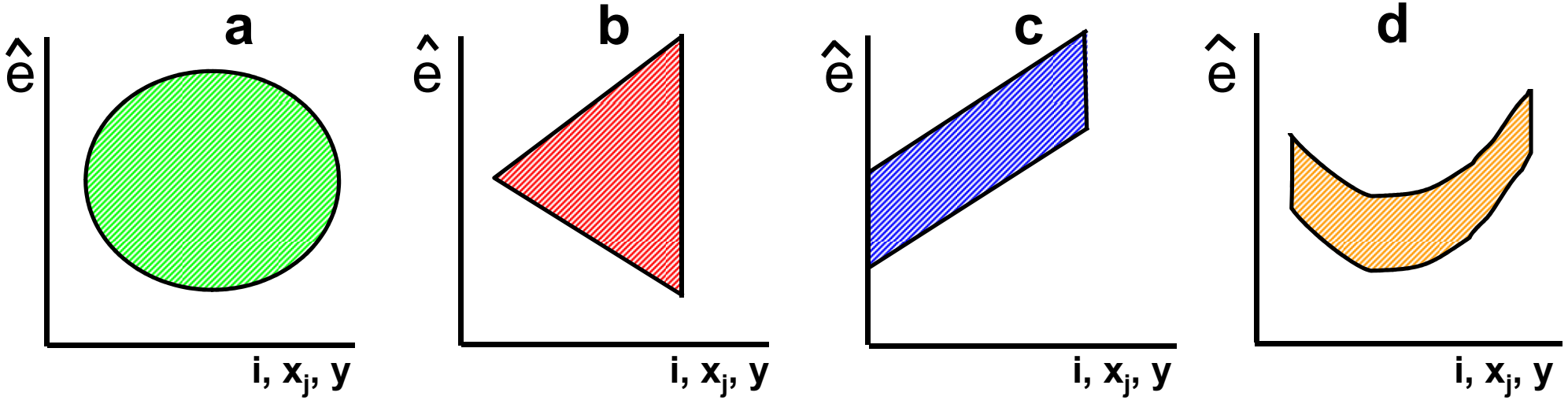
$$S_{\hat{y}_i} = (S_{y \cdot x}) \cdot \sqrt{\frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum X^2}}$$

Lineární regrese: analýza reziduí

Grafy residuí modelů (příklady)



Obecné tvary residuí modelů (schéma)



Analýza rozptylu v regresi

- Výpočet statistické významnosti rozptylu vyčerpaného regresním modelem

Celková ANOVA → SS_B/SS_T (variance ratio)

→ $MS_B/MS_E = F$

Analýza rozptylu regresního modelu (zde přímky)

Zdroj rozptylu	st.v.	SS	MS	F
Model (přímka)	1	SS_{MOD}	MS_{MOD}	MS_{MOD} / MS_R
Residuum	na - 2	SS_R	MS_R	
celkem	na - 1	SS_T		

(SS_{MOD}/SS_T) . 100 = % rozptylu Y "vyčerpaného" přímkou = koeficient determinace (R^2)

Kroky regresní analýzy

- Regresní analýza (a obecně i jiné stochastické modely) by měla probíhat v následujících krocích
  1. Ověření obecných předpokladů – normalita dat, linearita vztahu
  2. Výpočet modelu
  3. Analýza reziduí modelu umožňující ověřit vhodnost aplikace lineárního nebo jiného modelu
  4. Analýza vyčepané variability testující, zda model variabilitu dat významně vysvětluje
  5. Testování regresních koeficientů
     1. Posouzení významnosti komponent modelu
     2. Praktická smysluplnost modelu
  6. Závěr o využitelnosti a smysluplnosti modelu