FSTA: Pokročilé statistické metody

Principy stochastického modelování

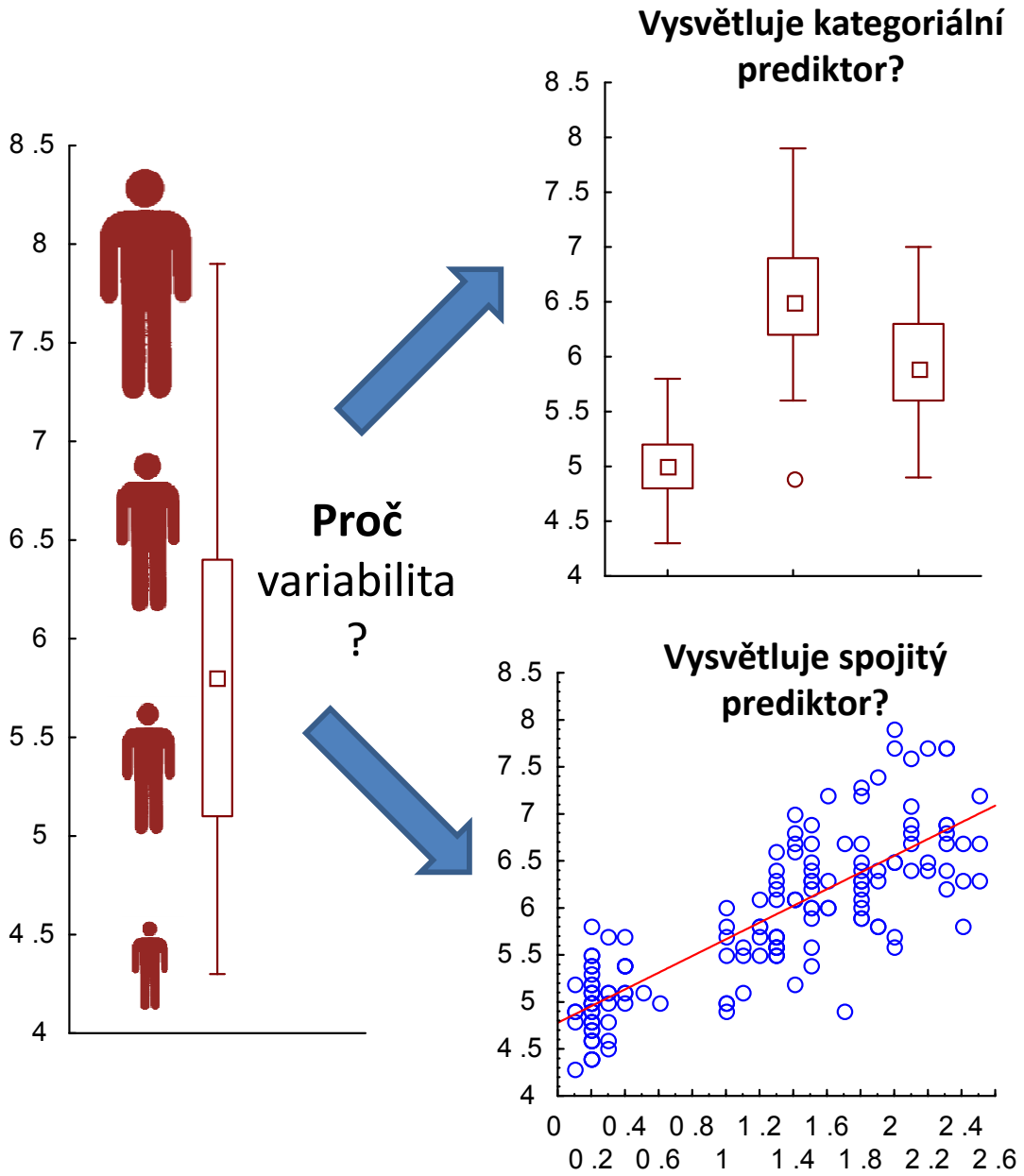
Jiří Jarkovský, Simona Littnerová

FSTA: Pokročilé statistické metody

Stochastické modelování - úvod

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



Kombinace typu predikované proměnné a prediktorů

Typ Y	Počet Y	Typ X	Metoda
Spojitá	1	Spojitá (binární)	Linární regrese
Spojitá	1	Binární, kategoriální	ANOVA
Spojitá	více	Spojitá (binární)	RDA, CCA, CC, co-inertia
Binární	1	Spojitá (binární)	Logistická regrese
Kategoriální	1	Spojitá (binární)	Diskriminační analýza

- Y – predikovaná proměnná
- X – prediktor

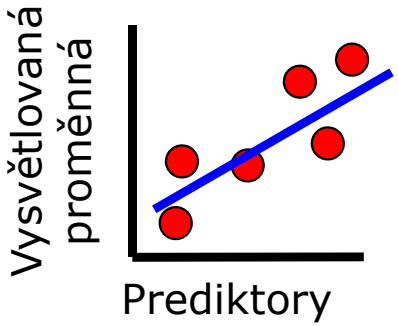
- Binární proměnné jsou častou používány jako prediktory v regresi nebo ordinační analýze
- Kategoriální proměnné jsou často překódovány do dummies, tedy do binárních proměnných
- Spojité proměnné nemusí být pouze normálně rozděleny a v lineárním vztahu, nicméně v takovém případě je nutné použít trasnformace nebo nelineární regrese/zobecněných lineárních modelů
- Existují i přístupy kombinující jako prediktory spojité i binární/kategoriální proměnné
- Častým přístupem je také konverze spojitých proměnných na binární s jasnou interpretací dělícího bodu

Obecné zásady tvorby predikčních modelů

- Požadavky na kvalitní predikční model
  - Maximální predikční síla
  - Maximální interpretovatelnost
  - Minimální složitost
- Tvorba modelů
  - Neobsahuje redundantní proměnné
  - Je otestován na nezávislých datech
- Výběr proměnných
  - Algoritmy typu dopředné a zpětné eliminace jsou pouze pomocným ukazatelem při výběru proměnných finálního modelu
  - Při výběru proměnných se uplatní jak klasické statistické metody (ANOVA), tak expertní znalost významu proměnných a jejich zastupitelnosti

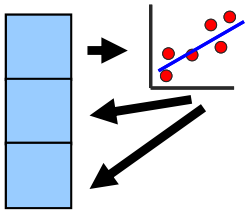
Vytváření modelů

1.Tvorba modelu



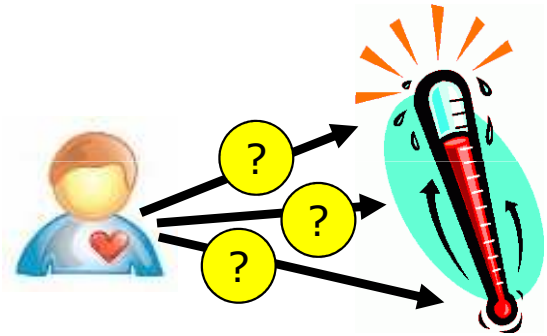
- Parametry ovlivňující vysvětlovanou charakteristiku pacienta
- Rovnice umožňující predikci
- Platnost modelu pouze v rozsahu prediktorů

2.Validace modelu



- Nebezpečí „přeučení" modelu
- Testování modelu na známých datech
- Krosvalidace

3. Aplikace modelu



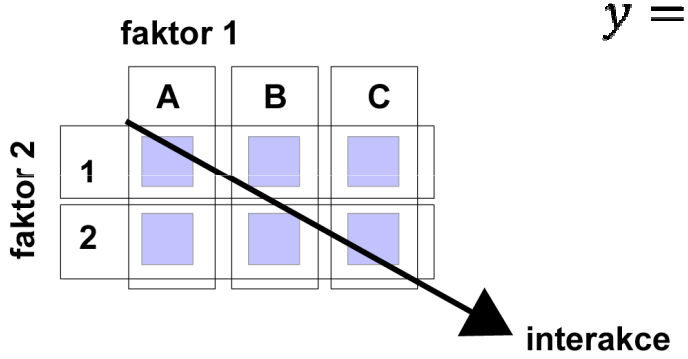
- Individuální predikce stavu nenámých pacientů
- Model musí být podložen korektní statistikou a rozsáhlými daty

Klíčové pojmy stochastického modelování

- Design modelu
  - Vhodně zvolená metodika a kombinace proměnných
- Výpočet modelu
  - Testování předpokladů zvolené metody
  - Redundance a kolinearita
  - Adjustace proměnných na vliv jiných proměnných
  - Výběr proměnných vícerozměrného modelu
- Kvalita modelu
  - Vyčerpaná variabilita a její statistická významnost
  - Testování výsledků modelu
- Interpretace modelu
  - Testování dílčích hypotéz
  - Hlavní efekty a interakce
  - Statistická významnost vs. praktické využití modelu
  - Rozsah aplikovatelnosti modelu

Design modelu

- Design modelu znamená jaké proměnné a v jakých kombinacích budou vysvětlovat hodnocenou proměnnou
- Obecně je vhodné ať již expertně nebo jako výsledek předběžné analýzy vytvořit a ověřit hypotézy o vzájemných vztazích proměnných a podle těchto předběžných výsledků vytvářet finální model
- Tvorba designu modelu úzce souvisí s pojmy:
  - Analýza pouze hlavních efektů proměnných
  - Analýza interakcí mezi proměnnými a složitost interakcí
- Design modelu lze vyjádřit graficky nebo v rovnici nebo pomocí maticoveho zápisu

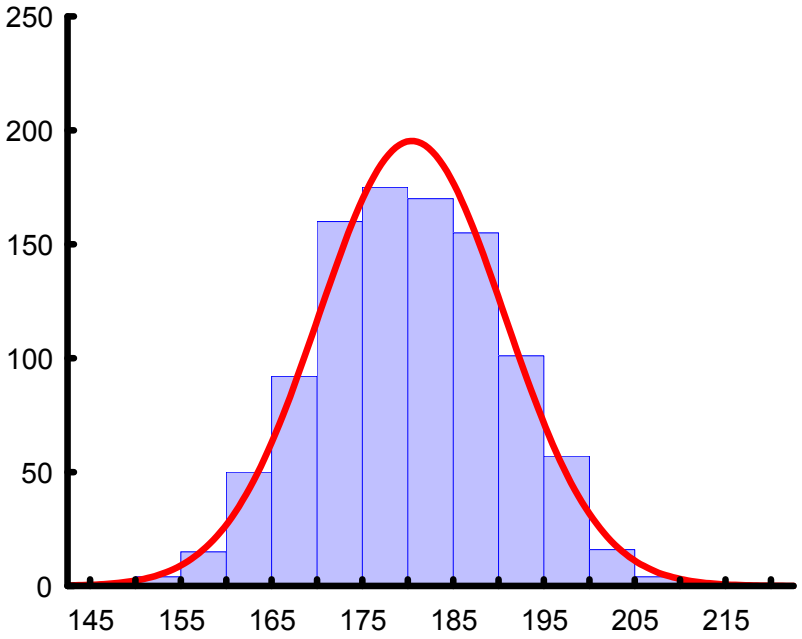


$$y = hmotnost * 1.5 + věk * 3.6 + hmotnost * věk * 1.8 + 9$$

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

Testování předpokladů

- Metody stochastického modelování jsou, stejně jako jiné statistické metody, závislé na dodržení předpokladů
- Nejčastějším předpokladem je normalita dat a linearita vztahu (ať již původních dat nebo po propojení linkovací funkcí)
- Testy normality pracují s nulovou hypotézou, že není rozdíl mezi zpracovávaným rozložením a normálním rozložením. Vždy je ovšem dobré prohlédnout si i histogram, protože některé odchylky od normality, např. bimodalitu některé testy neodhalí.



•Test dobré shody

V testu dobré shody jsou data rozdělena do kategorií (obdobně jako při tvorbě histogramu), tyto intervaly jsou normalizovány (převedeny na normální rozložení) a podle obecných vzorců normálního rozložení jsou k nim dopočítány očekávané hodnoty v intervalech, pokud by rozložení bylo normální. Pozorované normalizované četnosti jsou poté srovnány s očekávanými četnostmi pomocí $\chi 2$ testu dobré shody. Test dává dobré výsledky, ale je náročný na n, tedy množství dat, aby bylo možné vytvořit dostatečný počet tříd hodnot.

•Kolgomorov Smirnov test

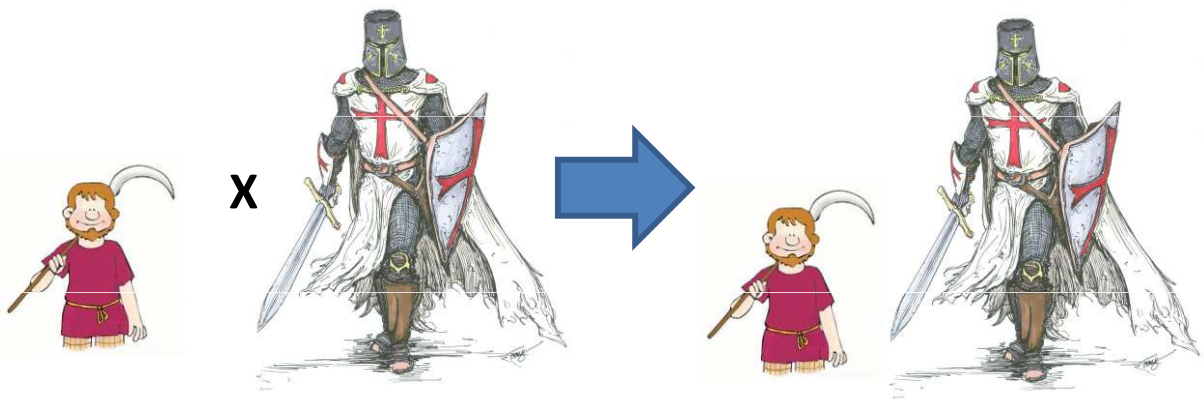
Tento test je často používán, dokáže dobře najít odlehlé hodnoty, ale počítá spíše se symetrií hodnot než přímo s normalitou. Jde o neparametrický test pro srovnání rozdílu dvou rozložení. Je založen na zjištění rozdílu mezi reálným kumulativním rozložením (vzorek) a teoretickým kumulativním rozložením. Měl by být počítán pouze v případě, že známe průměr a směrodatnou odchylku hypotetického rozložení, pokud tyto hodnoty neznáme, měla by být použita jeho modifikace – Lilieforsův test.

•Shapiro-Wilk`s test

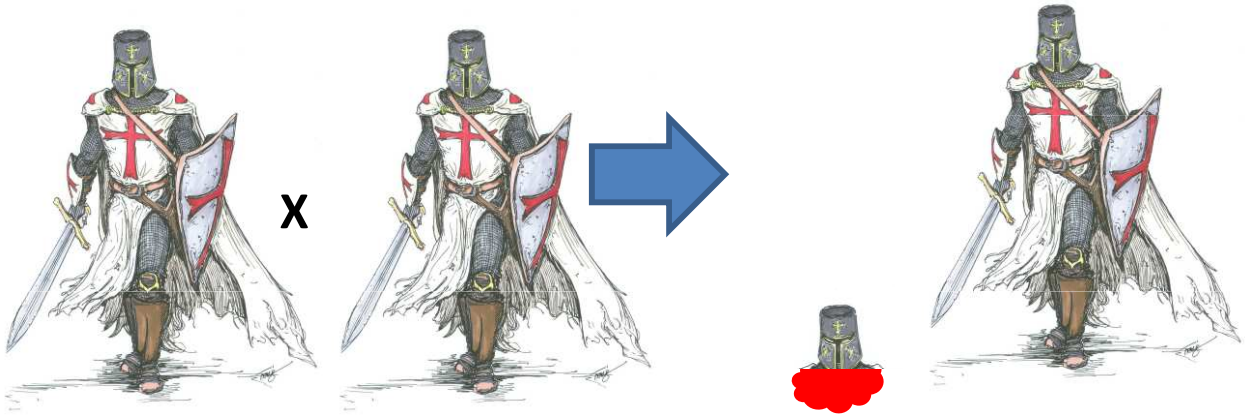
Jde o neparametrický test použitelný i při velmi malých n (10) s dobrou sílou testu, zvláště ve srovnání s alternativními typy testů, je zaměřen na testování symetrie.

Význam identifikace redundantních proměnných

- Redundantní proměnné snižují stabilitu modelu a mohou vést až k nesmyslným výsledkům



Proměnná se silnější diskriminační silou a nekorelovaná s druhou proměnnou snadno vyhrává zařazení do modelu, další proměnné následují dle jejich významu



V případě dvou korelovaných proměnných s obdobnou diskriminační silou pouze jedna vyhrává zařazení do modelu (výsledek dán nepatrnými náhodnými odlišnostmi), druhá je vyřazena nebo vstupuje s do modelu s minimálním významem -> problém s interpretací a stabilitou

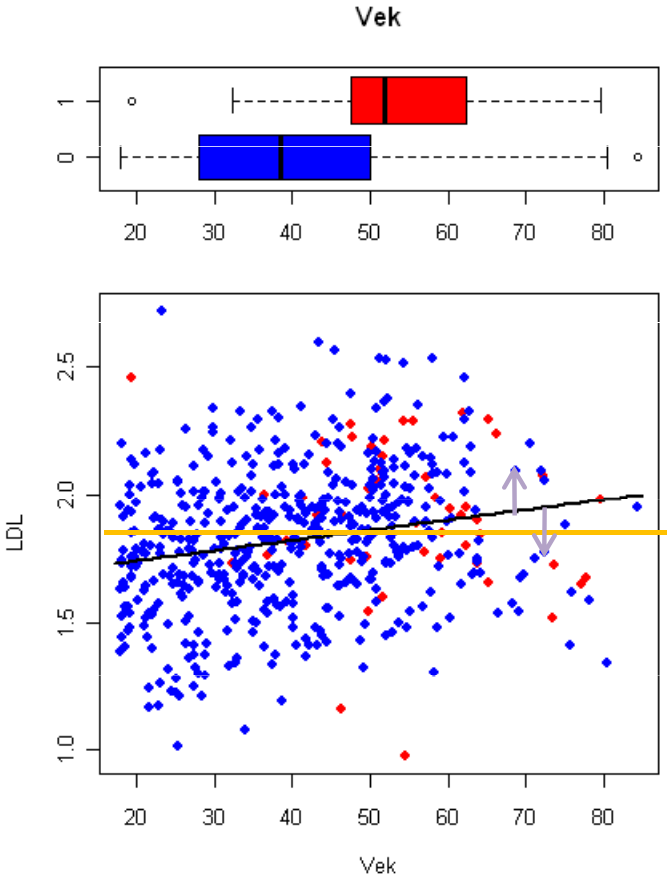
Identifikace redundantních proměnných

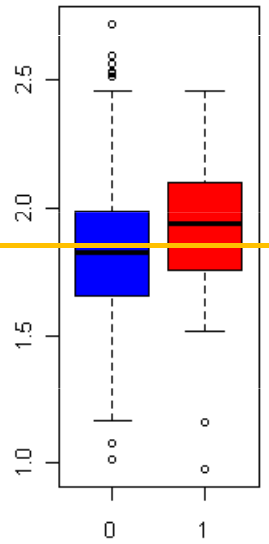
- Korelační analýza a XY grafy
  - Jednoduchý výpočet
  - Analyzuje vztahy pouze dvojic proměnných
- Analýza hlavních komponent nebo faktorová analýza
  - Analyzuje vzájemné vztahy sady proměnných
  - Usnadňuje výběr neredundantních proměnných nebo nahrazení proměnných faktorovými osami
- Analýza vzájemného vysvětlení proměnných (analýza redundance)
  - Ve statistických software často součást regresní analýzy nebo diskriminační analýzy
  - R^2 a Tolerance – R^2 popisuje kolik variability dané proměnné je vysvětleno ostatními proměnnými v modelu? Tolerance je $1-R^2$, tedy kolik unikátní variability na proměnnou připadá (principem je vícerozměrná regrese, ta determinuje i předpoklady výpočtu)
  - VIF (Variance Inflation Factor) je počítán jako 1/Tolerance, při VIF>10 je kolinearita považována za velmi závažnou (nicméně nejsou dány žádné závazné hranice VIF)
- Expertní znalost proměnných
  - Vyřazovány jsou korelované proměnné s obtížným měřením, zatížené chybami, nízkou vyplněností apod.

Adjustace proměnných na vliv jiných proměnných

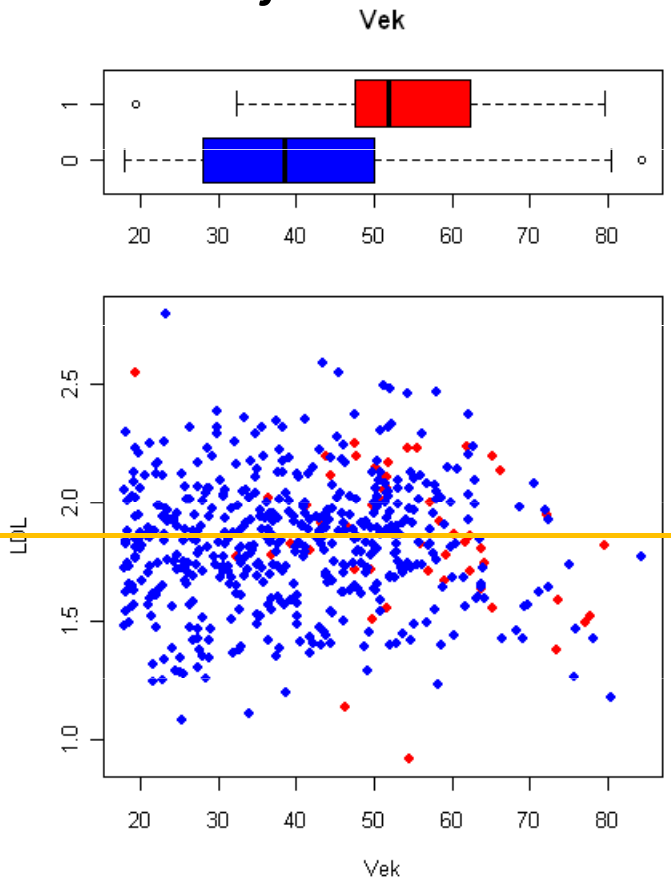
1. V prvním kroku definujeme regresní model vztahu věku a adjustovaného parametru
2. Pro každého pacienta je vypočteno jeho reziduum od regresní přímky ↑↓
3. Reziduum (představující hodnotu parametru po odečtení vlivu věku, jeho průměr je 0) je přičteno k průměrné hodnotě parametru
4. Výsledná adjustovaná hodnota má odečten vliv věku, ale zároveň není změněna číselná hodnota parametru

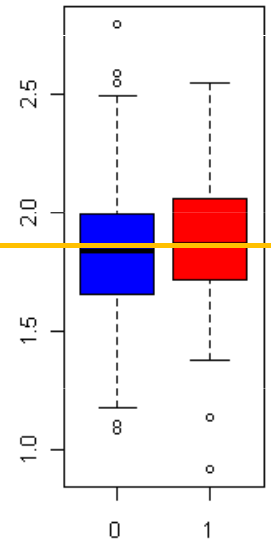
Původní data





Adjustovaná data





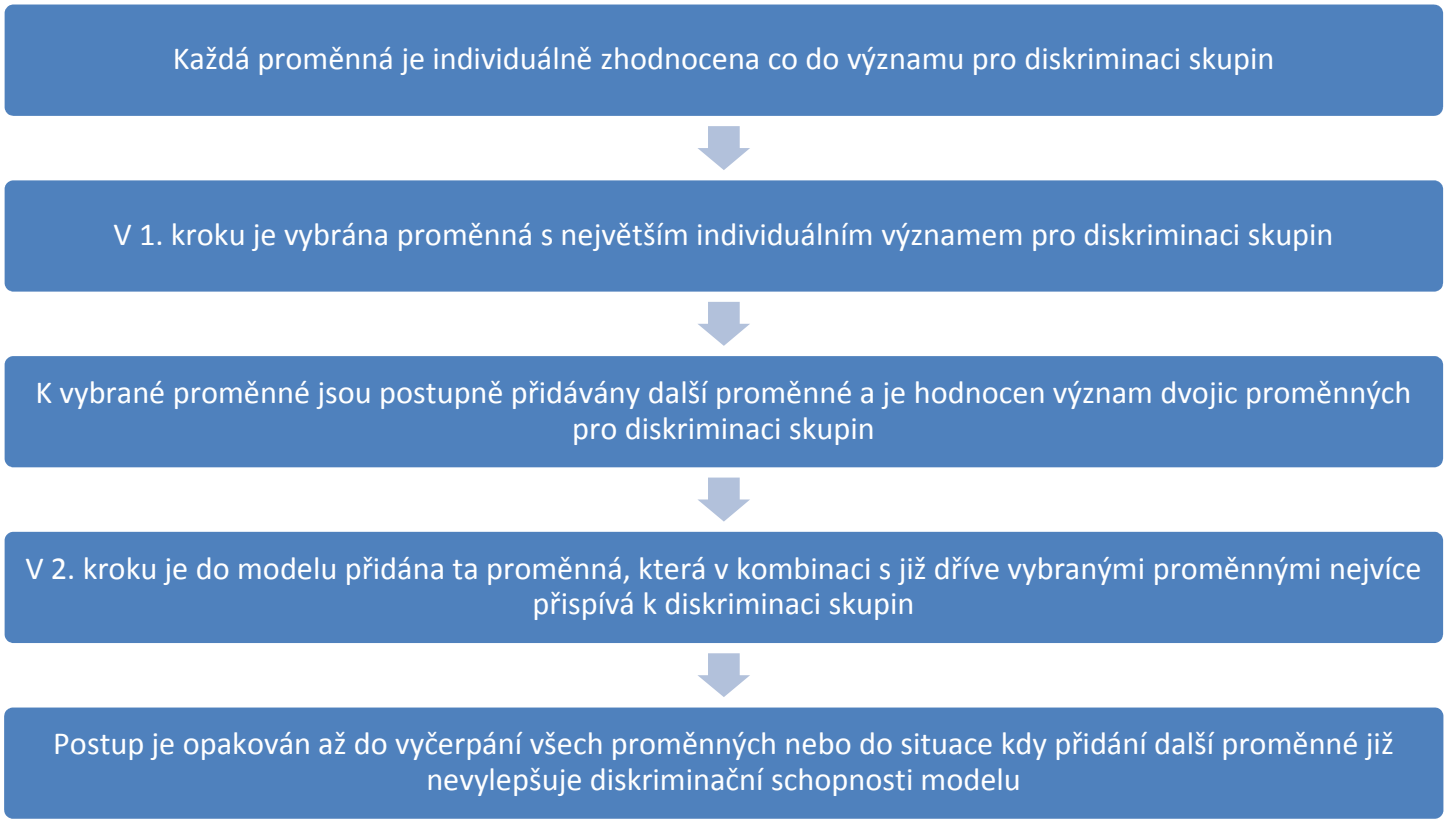
Dopředná a zpětná eliminace

- Dopředná a zpětná eliminace proměnných z modelu (forward, backward stepwise) je obecná technika používaná při tvorbě regresních, diskriminačních a jiných modelů
- Proměnné jsou do modelu postupně přidávány (ubírány) podle jejich významu v modelu

Schéma dopředné eliminace proměnných v modelu

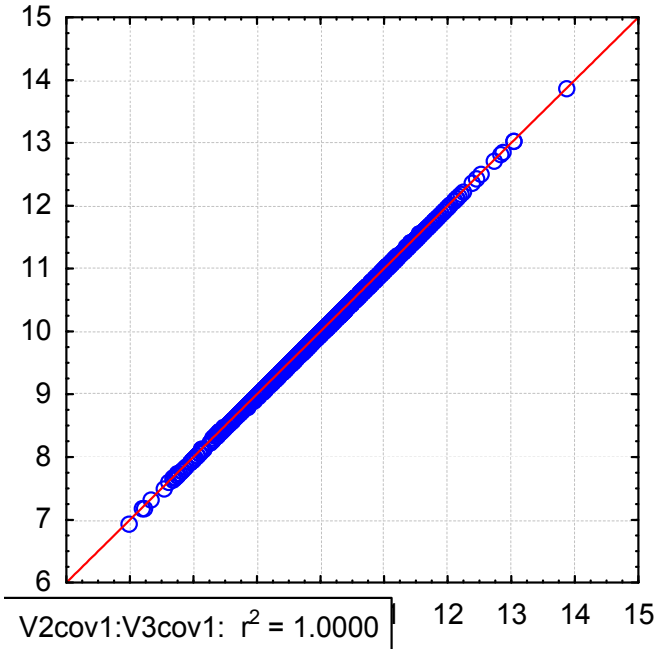
V případě zpětné eliminace začíná proces od modelu se všemi proměnnými a postupně jsou vyřazovány proměnné s nejmenším příspěvkem k diskriminační síle modelu

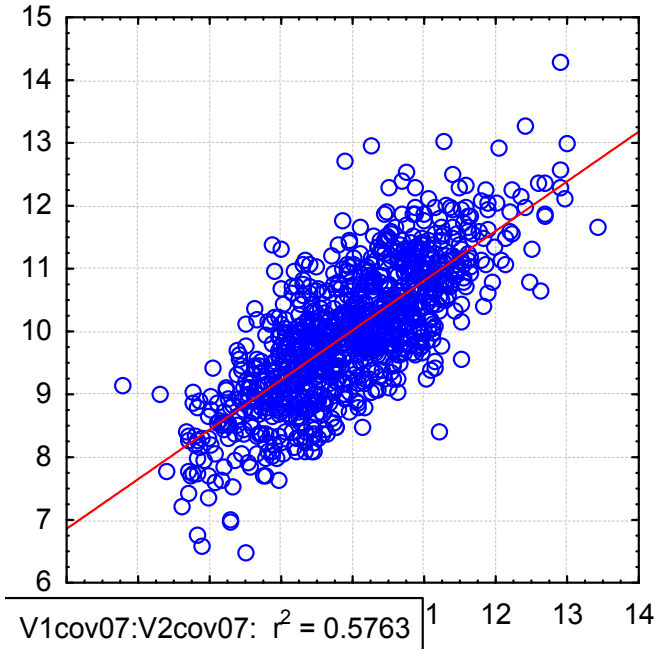
Proces je třeba expertně kontrolovat, riziková je např. přítomnost redundantních proměnných

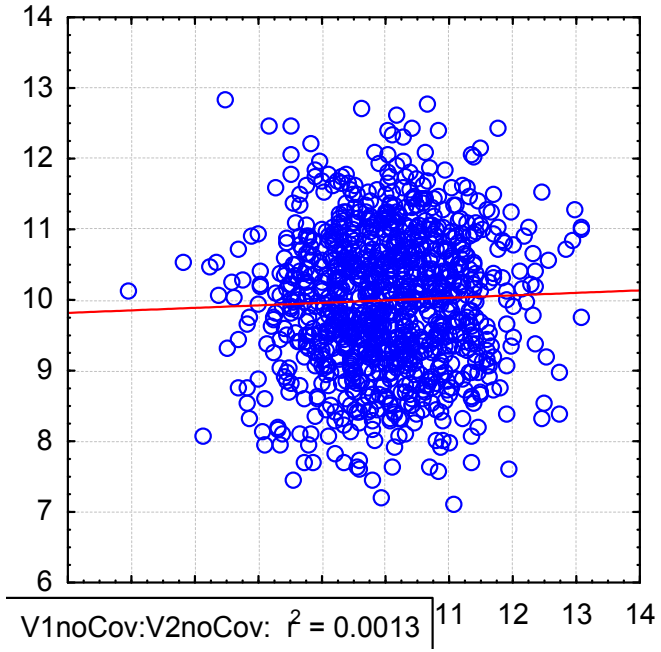


Vyčerpaná variabilita a její statistická významnost

- Základním ukazatelem kvality modelu je množství varibility, které je modelem vysvětleno
- Obecně se značí R^2 a uvádí se v procentech nebo podílu celkové variability (v případe lineární regrese jde o Pearsonovu korelaci na druhou)
- Statisckou významnost vyčepané variability je možné testovat pomocí analýzy rozptylu

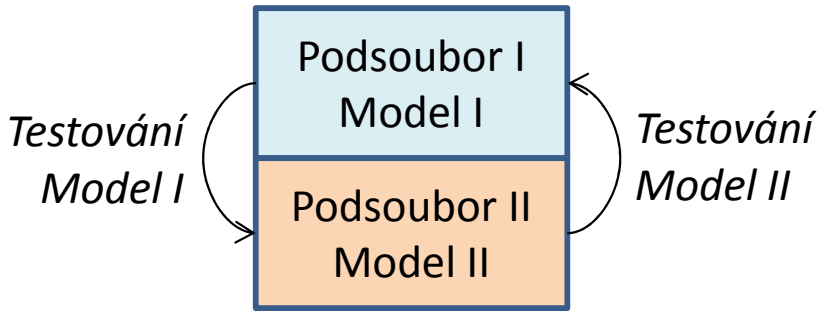






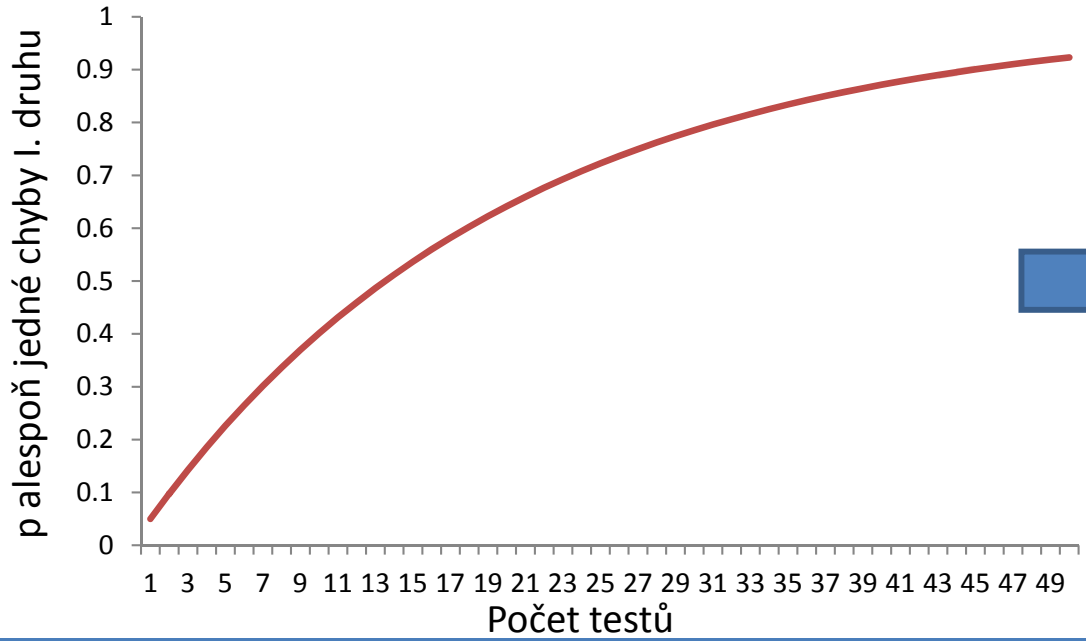
Ověření modelu na nezávislém souboru

- Při tvorbě modelů může dojít k problému, kdy vytvořený model je perfektně „vycvičen“ řešit danou úlohu na datovém soubor na němž byla vytvořena
- Z tohoto důvodu je problematické testovat výsledky modelu na stejném souboru, na němž byla vytvořena -> jde o důkaz kruhem
- Řešením je testování výsledků modelu na souboru se známým výsledkem (zde známým zařazením objektů do skupin), který se nepodílel na definici modelu
  - Krosvalidace
    - datový soubor je náhodně rozdělen na několik podsouborů (2 nebo více)
    - Na jednom podsouboru je vytvořen model a jeho výsledky testovány na zbývajících podsouborech
    - Výpočet je proveden postupně na všech podsouborech
  - One out leave out
    - Model je vytvořen na celém souboru bez jednoho objektu
    - na tomto objektu je model testován
    - postup je zopakován pro všechny objekty
  - Permutační metody
    - Jackknife, bootstrap – model je postupně vytvářen na náhodných podvýběrech souboru a testován na zbytku dat



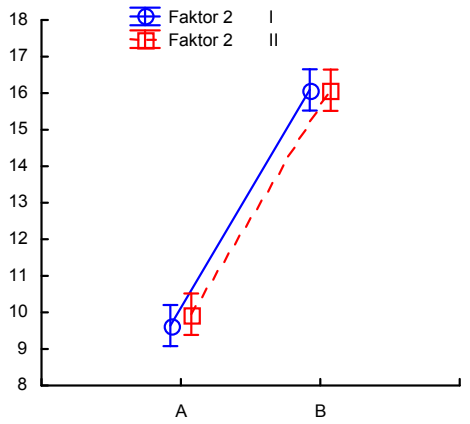
Testování dílčích hypotéz

- 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)

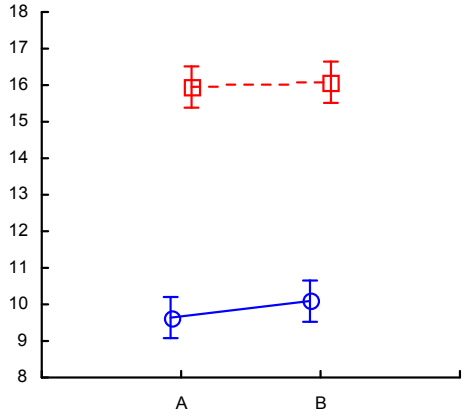


Řešením jsou různé procedury korigující hodnotu p (např. Bonferroniho korekce, FWR, FDR procedury apod.)

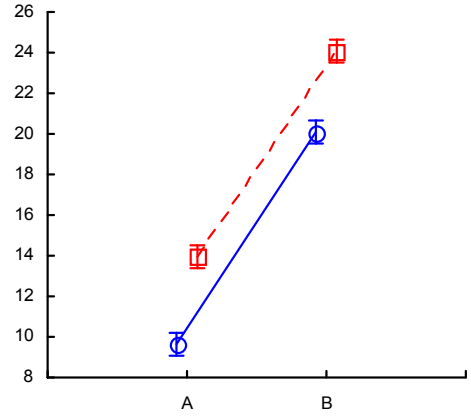
Hlavní efekty a interakce



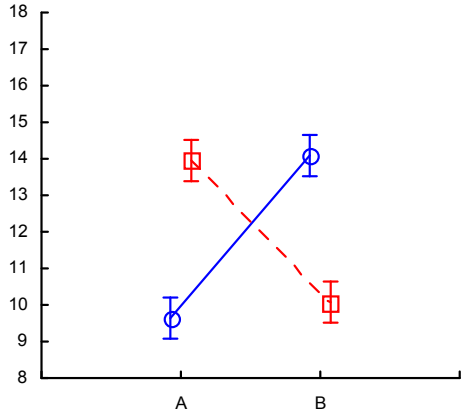
	SS	D.f.	MS	F	p
Intercept	33487	1	33487	8165.3	0.000
Faktor 1	**1978**	**1**	**1978**	**482.2**	**0.000**
Faktor 2	1	1	1	0.3	0.602
F1*F2	1	1	1	0.3	0.570
Error	804	196	4		



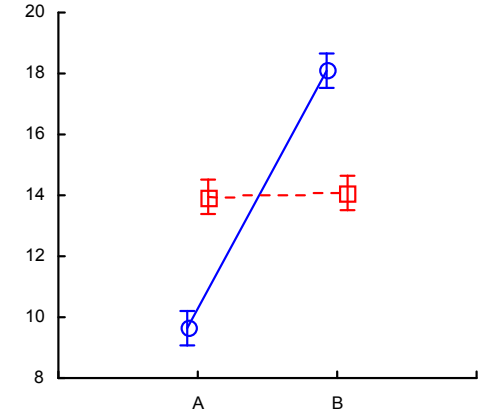
	SS	D.f.	MS	F	p
Intercept	33487	1	33487	8165.3	0.000
Faktor 1	4	1	4	1.0	0.314
Faktor 2	**1891**	**1**	**1891**	**461.1**	**0.000**
F1*F2	1	1	1	0.3	0.570
Error	804	196	4		



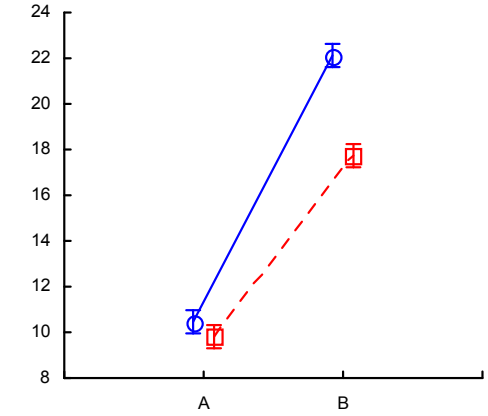
	SS	D.f.	MS	F	p
Intercept	57391	1	57391	13993	0.000
Faktor 1	5293	1	5293	1290.7	0.000
Faktor 2	861	1	861	209.9	0.000
F1*F2	1	1	1	0.3	0.570
Error	804	196	4		



	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		

Statistická významnost vs. praktické využití modelu

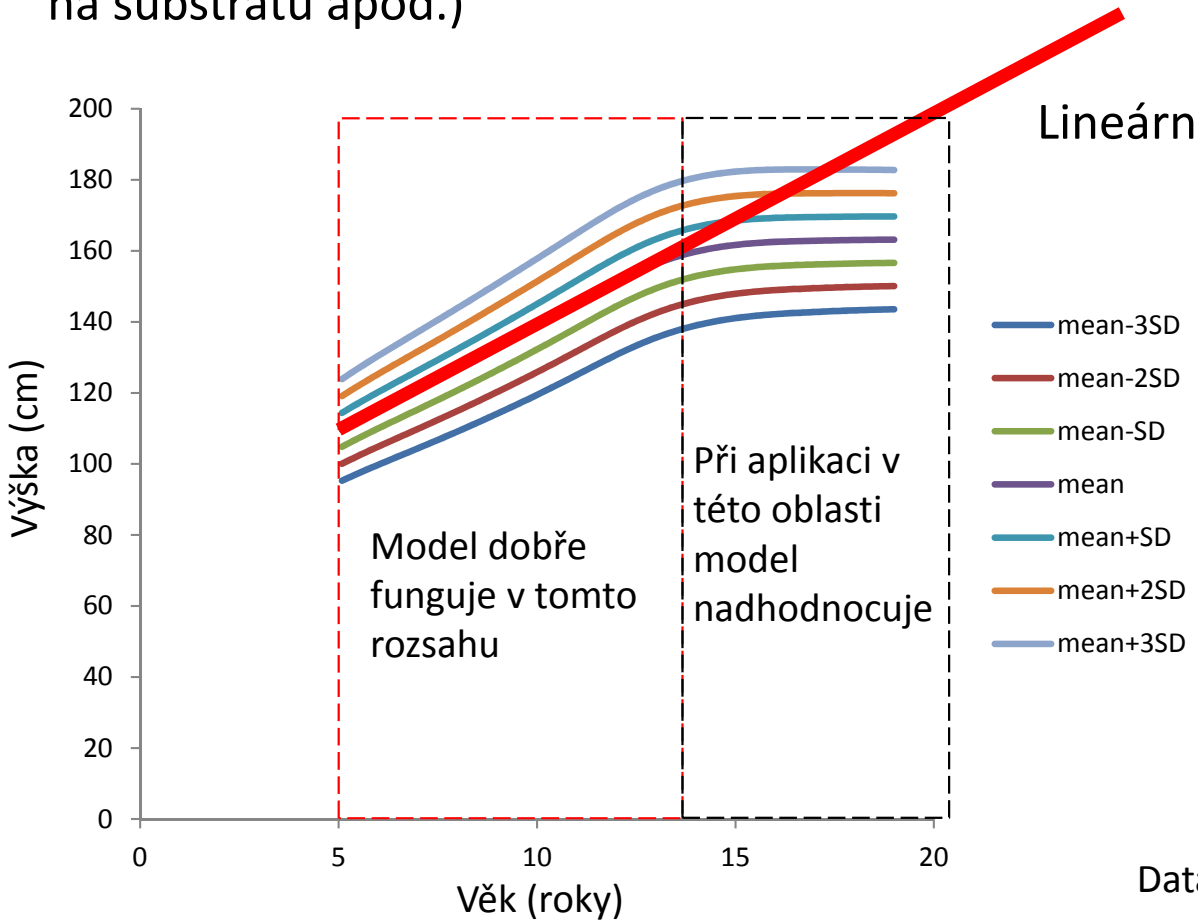
- Při aplikaci modelu v praxi je třeba zohlednit jak zjištěné statistické významnosti, tak praktický význam výstupů modelu
- Jde o analogii k statistické vs. praktické významnosti rozdílů např. v t –testu

- Statistická významnost = vztah mezi proměnnými, rozdíl mezi skupinami není pouhá náhoda (respektivě je dostatečně nízká pravděpodobnost, že nejde o náhodu)

- Praktický význam modelu
  - Z hlediska prediktorů: změna predikované hodnoty při změně prediktoru je prakticky významná (např. velikost nárůstu krevního tlaku při změně věku o 10 let)
  - Z hlediska objektů: Individuální predikce pacienta je dostatečně přesná aby byla prakticky využitelná (predikce různých událostí – hospitalizace, úmrtí, vznik komplikací, výsledek léčby atd.)

Rozsah aplikovatelnosti modelu

- Modely je možné aplikovat pouze v rozsahu prediktorů, na nichž byly vyvinuty
- Důvodem je naše neznalost chování vztahů mezi prediktory a predikovanou proměnnou mimo hranice v nichž byl model definován (typickými příklady jsou např. křivky dávka-odpověď, růst dětí v závislosti na věku, růst baktérií v závislosti na substrátu apod.)



Lineární model odvozený z části dat

Data: WHO Growth reference 5-19 years

FSTA: Pokročilé statistické metody

Stochastické modelování - 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 převažujeme, 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

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

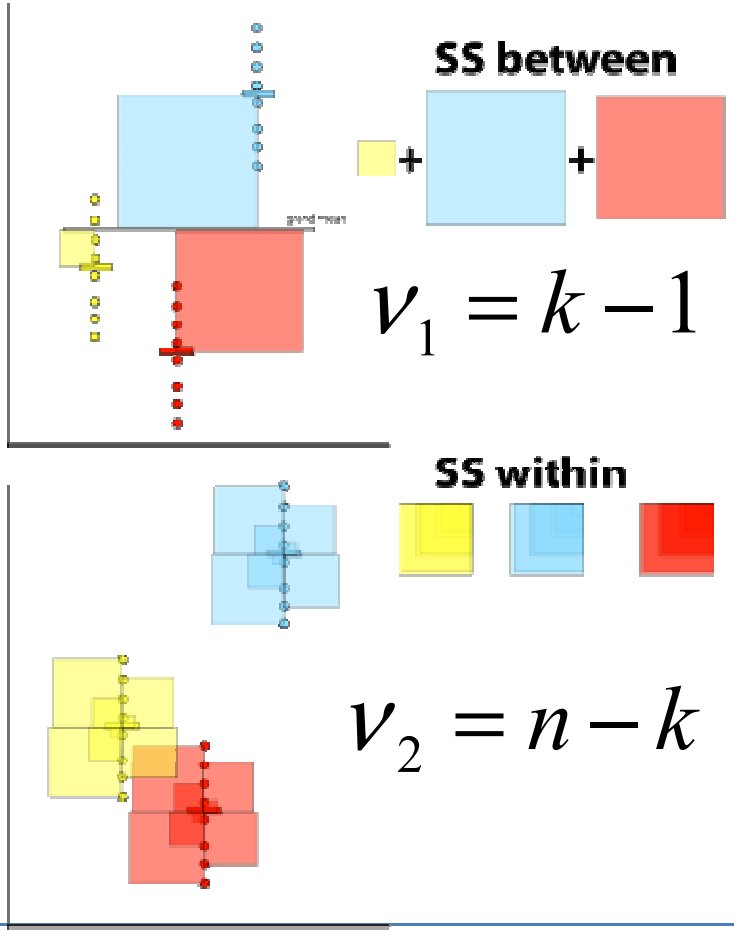
Rozptyl je počítán pro celkový průměr (tzv. grand mean) a průměry v jednotlivých skupinách dat

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)



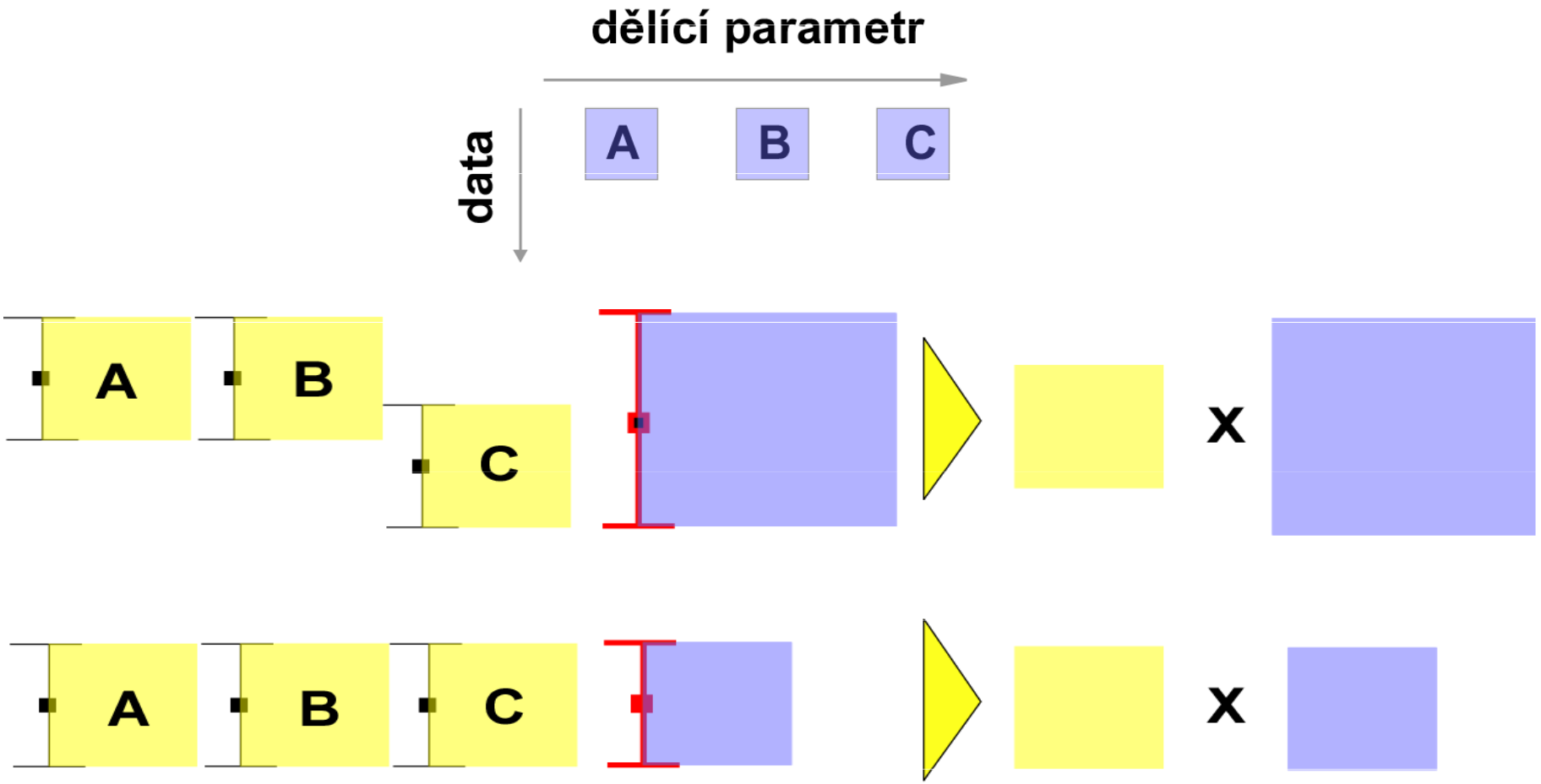
$$F = \frac{between\_groups}{within\_groups}$$

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

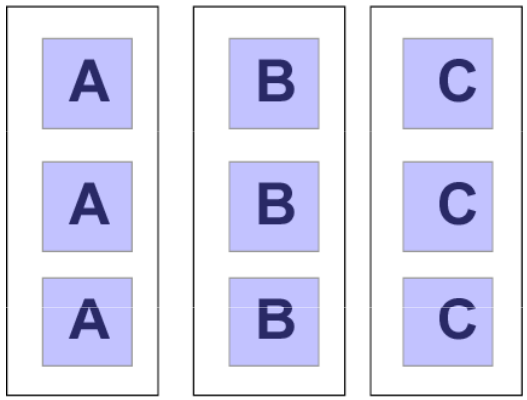


Nested ANOVA

- 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

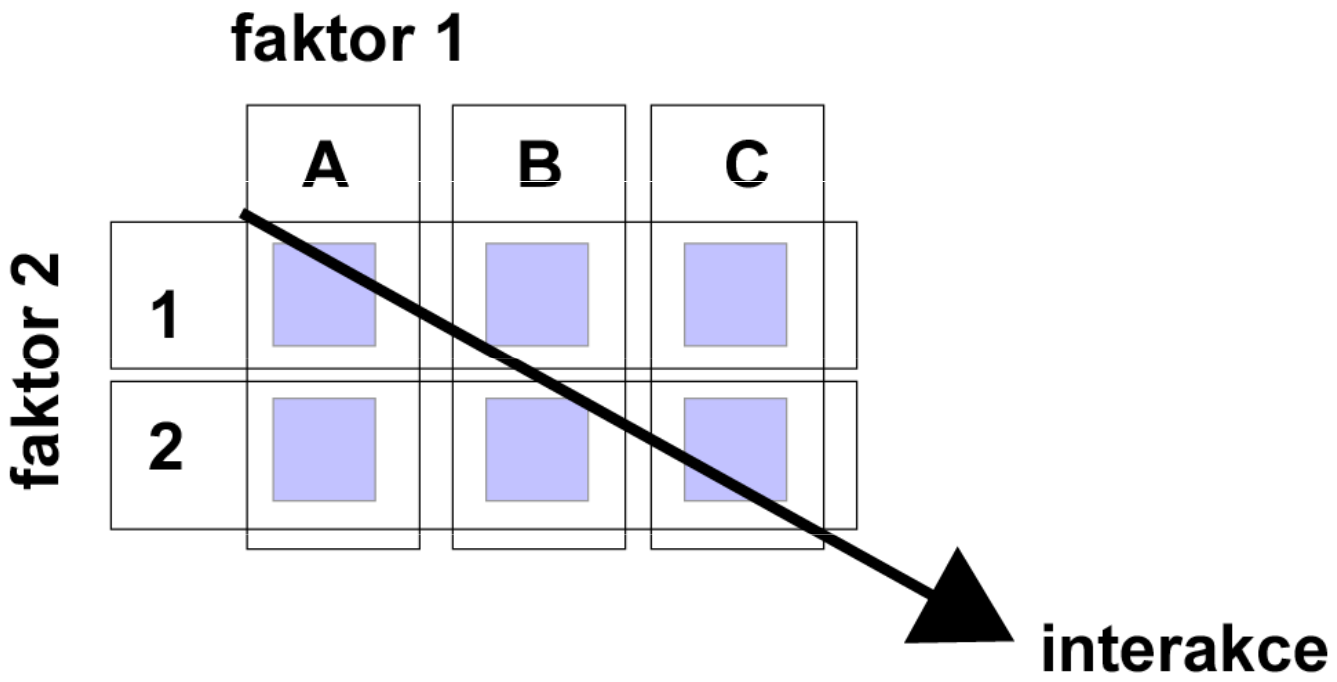
jednoduchá ANOVA

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

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 – 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

Základy regresní analýzy

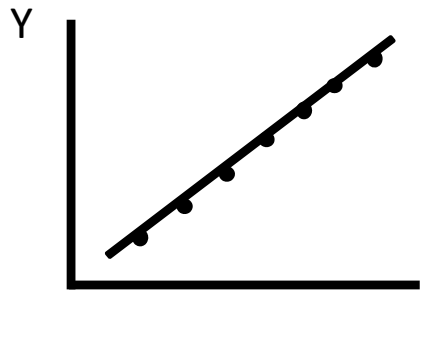
- Regrese - funkční vztah dvou nebo více proměnných

Jednorozměrná
y = f(x)

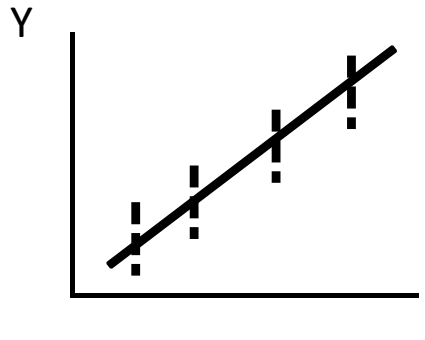
Vícerozměrná
y = f(x1, x2, x3, ……xp)

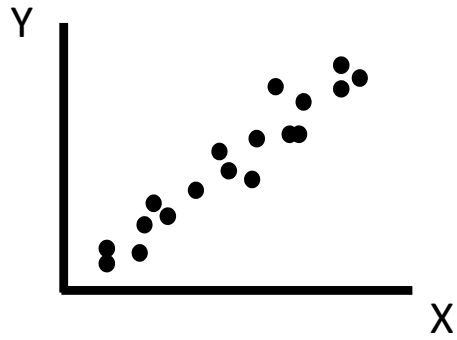
Vztah x, y

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 —

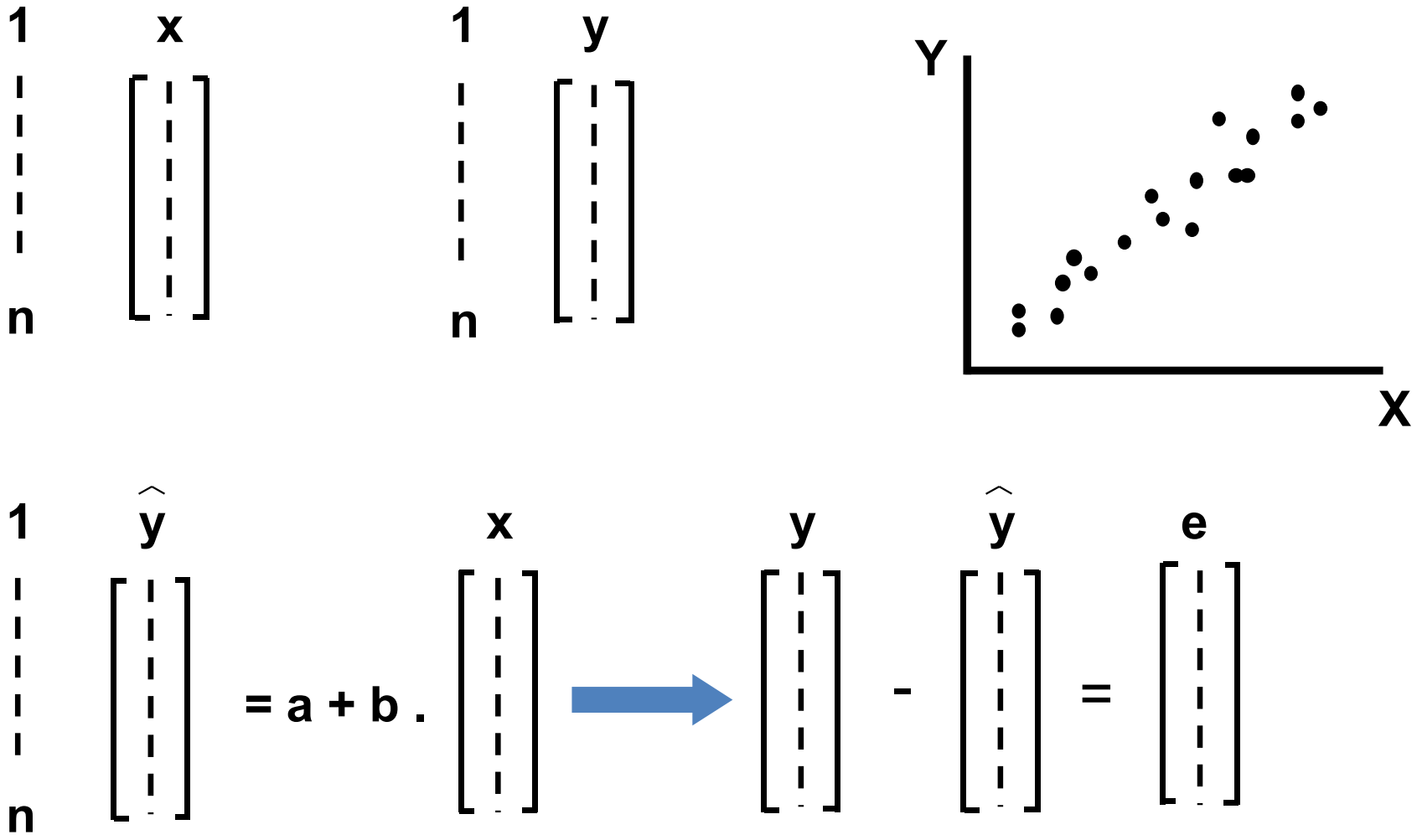
- $\alpha \approx a$ **(intercept):** $a = \bar{y} - b \cdot \bar{x}$
- $\beta \cdot X \approx b \cdot x$ **(sklon; slope)**
- $\varepsilon \approx e$ **- náhodná složka :** $\mathrm{N}\left(0; \sigma_e^2\right) = N\left(0; \sigma_y^2 x\right)$

} Komponenty tvořící y se sčítají

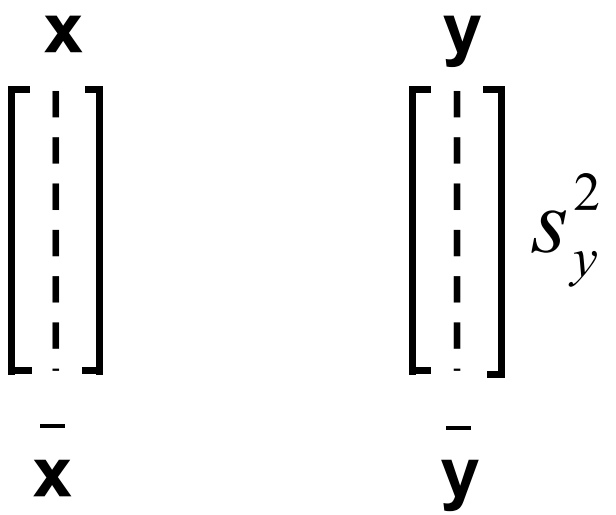
ε - náhodná složka modelu přímky = rezidua přímky

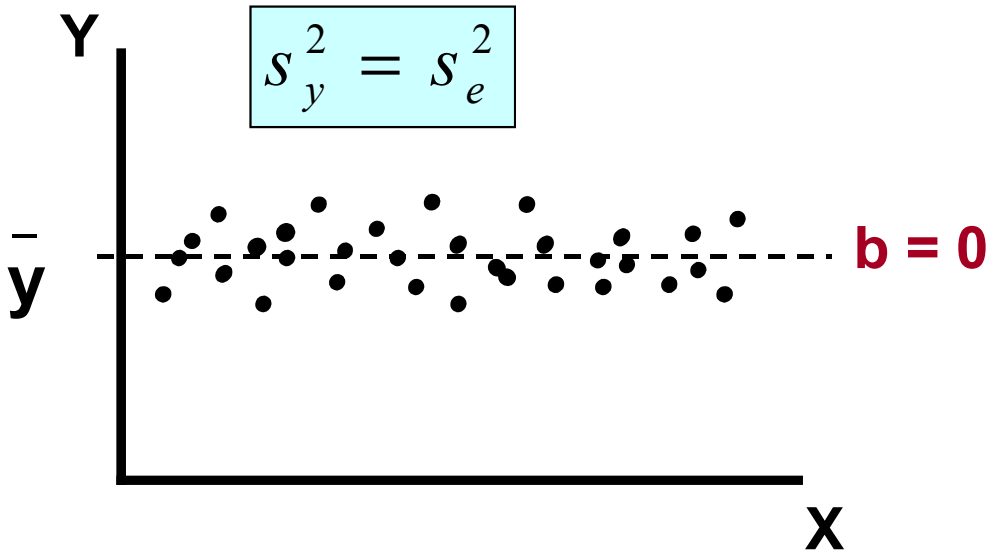
$$\sigma_e^2 \left(\sigma_{y \cdot x}^2\right) \Rightarrow \textbf{ rozptyl reziduí}$$

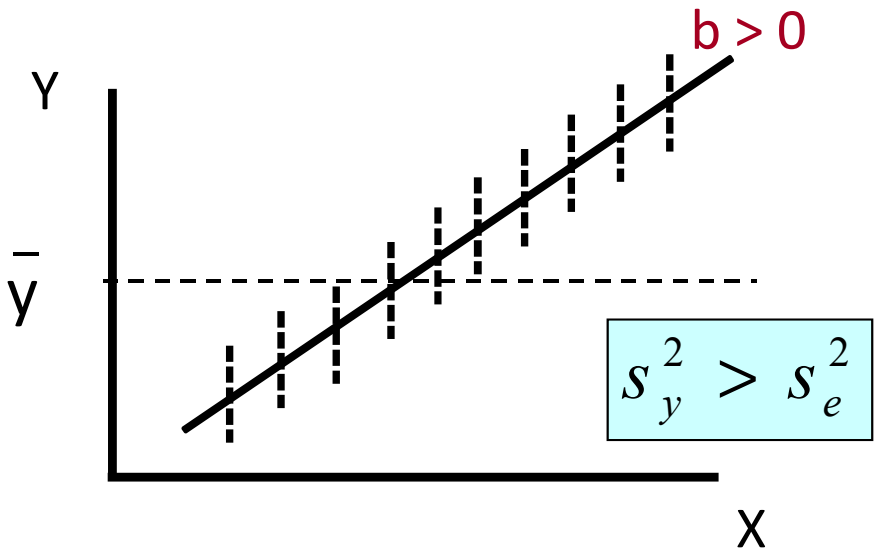
Lineární regrese II



Lineární regrese III

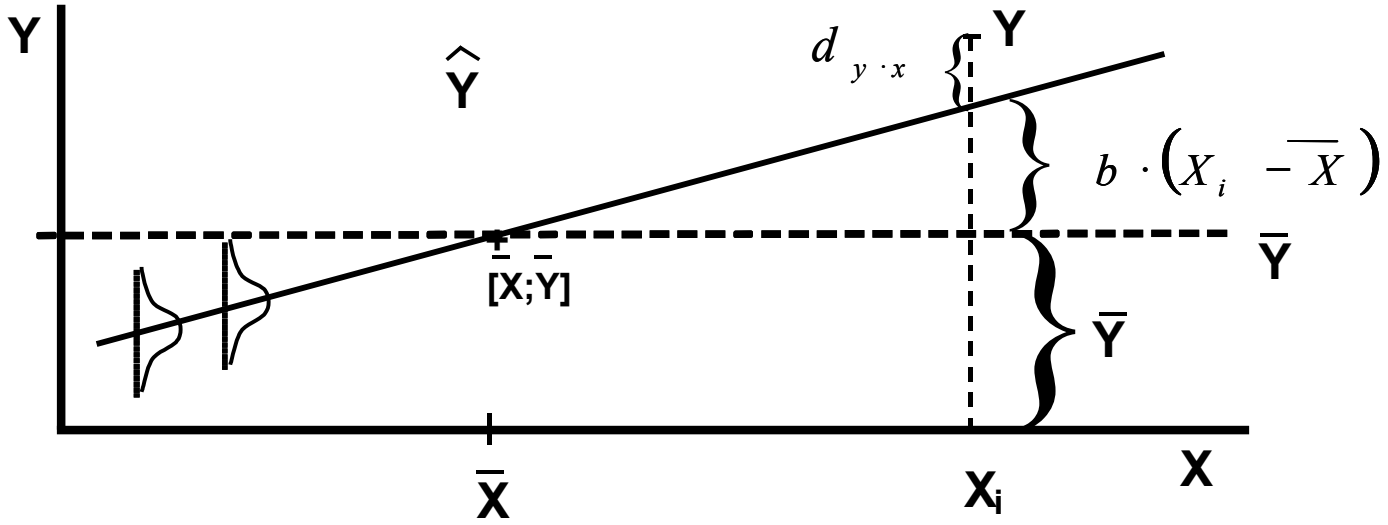






Lineární regrese III

- 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

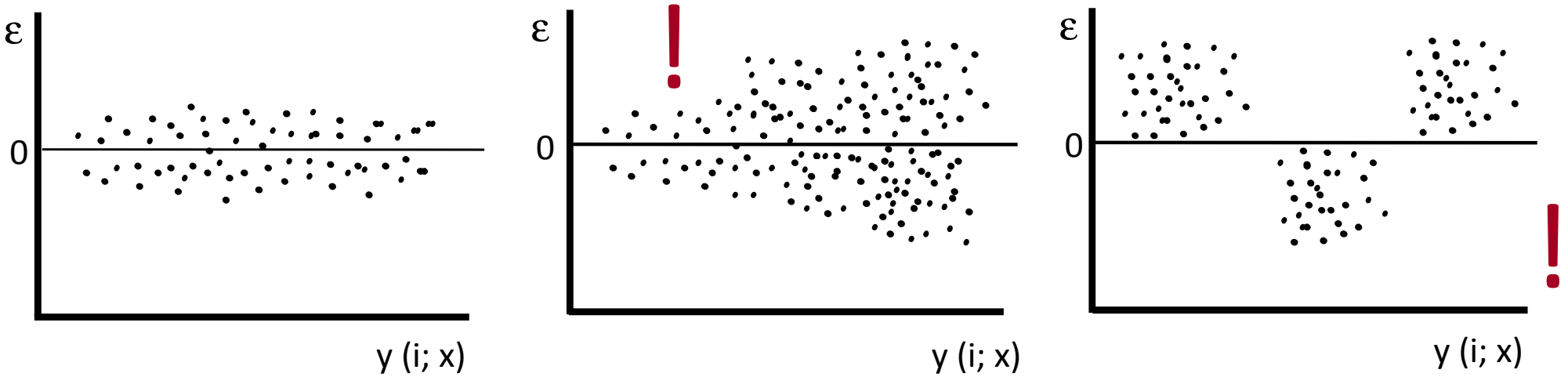
III. $\hat{Y}$: **modelová hodnota**

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

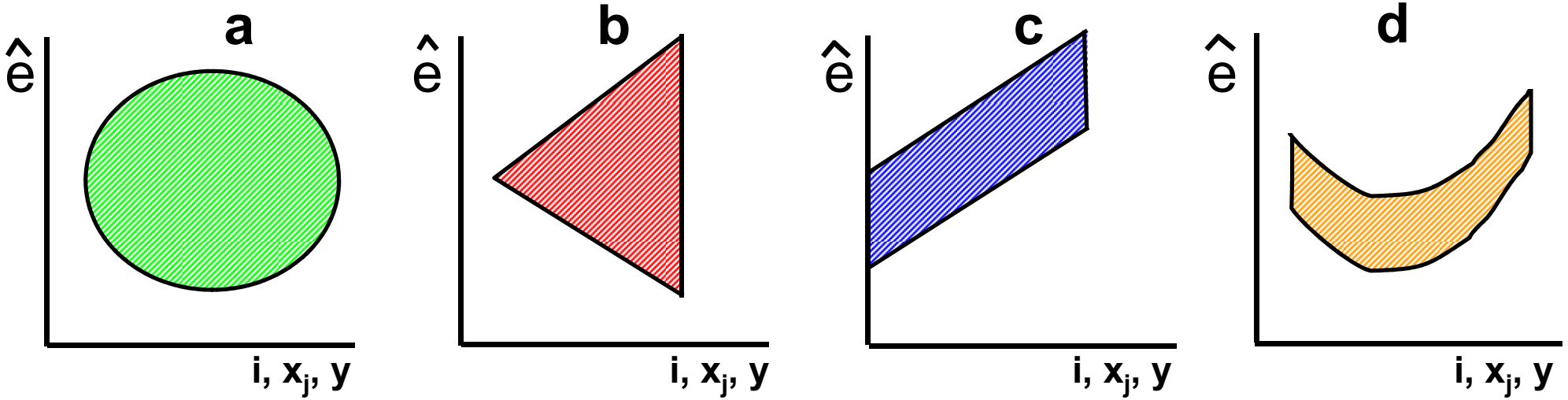
$$S_{\hat{y}_i} = (S_{y \cdot x}) \cdot \sqrt{\frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum X^2}}$$

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