PSY117/454
Statistická analýza dat v psychologii
Přednáška 13

Vícerozměrné metody

Schematický úvod

Co je na slově *statistika* tak divného, že jeho vyslovení tak často způsobuje napjaté ticho?

William Kruskal

Přehled vícerozměrných metod

Analýza závislostí

...externální, strukturní

Modelujeme vliv nezávislých proměnných na závislé

- Vícerozměrná lineární regrese a strukturní modelování
- Faktoriální ANOVA a MANOVA
- Diskriminační analýza (Logistická regrese)

Klasifikace a struktura dat

...internální

Hledáme strukturu vzájemných vztahů mezi proměnnými či jedinci s cílem je klasifikovat popř. redukovat složitost

- Explorační faktorová analýza
- Shluková analýza

ZÁVISLOSTI 1 – Vícerozměrná lineární regrese

Jak dobře lze předpovědět inteligenci dítěte z inteligence otce, matky, vzdělání otce a vzdělání matky?

Který z uvedených prediktorů má nejvyšší predikční sílu?

$$Y = a + b_1X_1 + b_2X_2 + \ldots + b_nX_n$$

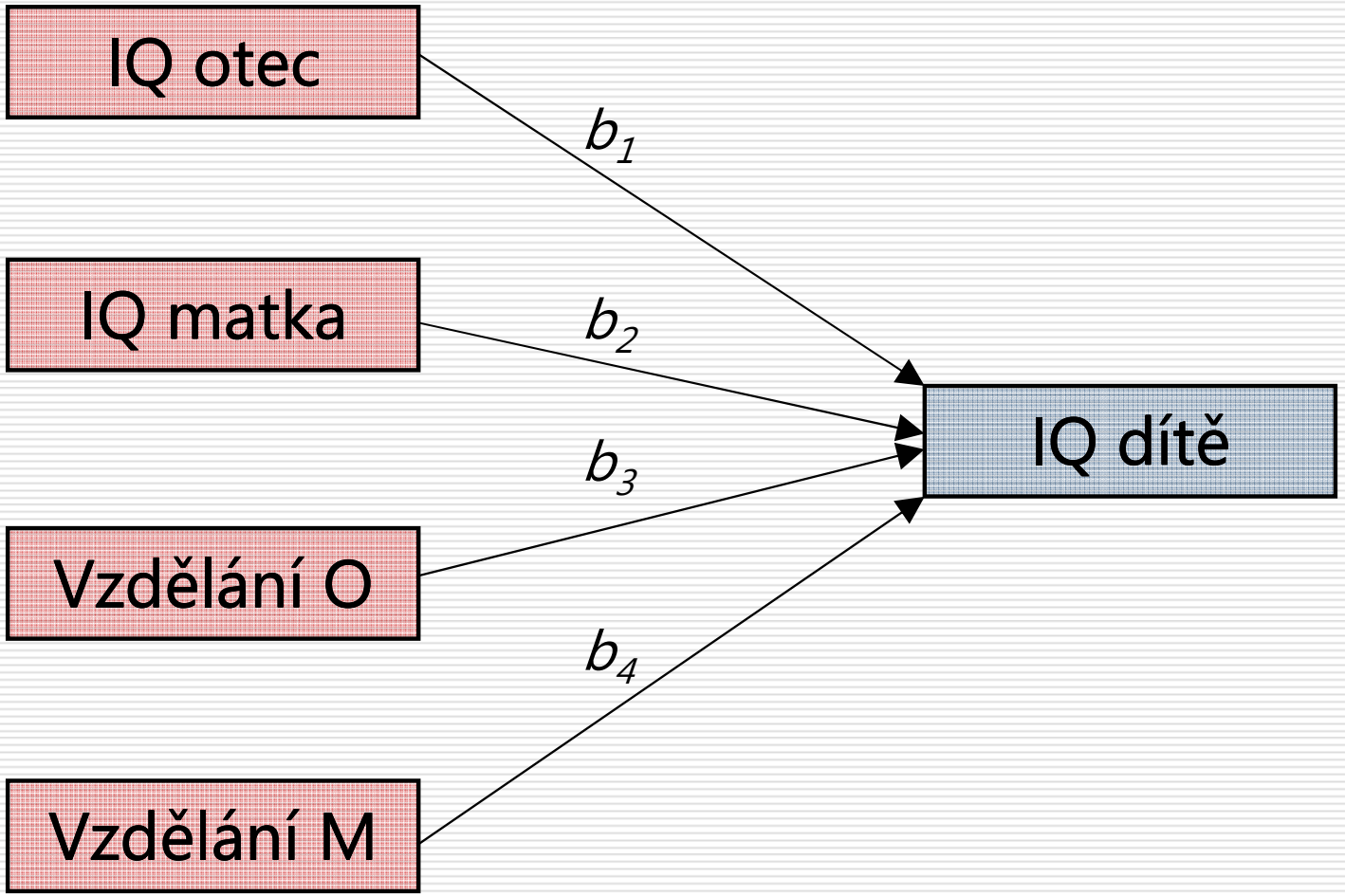
1 intervalová závislá

n intervalových nezávislých – prediktorů

- Oproti jednoduché lineární regresi je zde novinkou nutnost vypořádat se se vztahy mezi prediktory (čím menší tím lépe).
- Pro možnost srovnávání predikční síly prediktorů zavedena standardizovaná verze koreficientu $b_{n:}$ β_n (beta)

AJ: multiple regression

ZÁVISLOSTI 1 – Vícerozměrná lineární regrese



ZÁVISLOSTI 2 – Strukturní modelování SEM, LISREL

Velmi obecné rozšíření regresního modelu o

- více závislých včetně vztahů mezi nimi
- zohlednění vztahů (korelací) mezi prediktory
- latentní (neměřené) proměnné

$Y_1 = a + b_{11}X_1 + b_{12}X_2 + \ldots + b_{1m}X_n + cY_2$

$Y_2 = a + b_{21}X_1 + b_{22}X_2 + \ldots + b_{2m}X_n + cY_1$ … až Y_m

m intervalových závislých

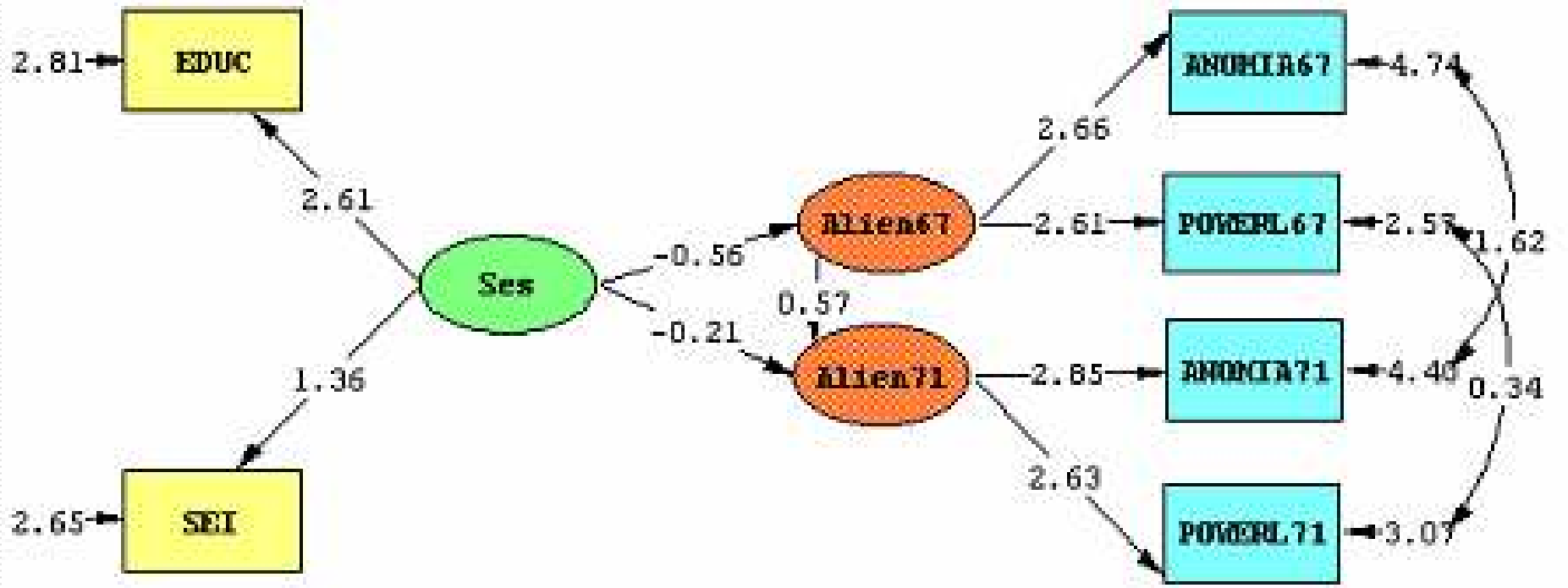
n intervalových prediktorů

Ověřují se jím složité hypotézy, které mají obecný tvar:

„Odpovídají vztahy mezi daty specifikovanému modelu?“

AJ: structural equation modeling, latent variables

ZÁVISLOSTI 2 – Strukturní modelování SEM, LISREL



Vzdělání a socioekonomický index predikují stabilitu anomie a bezmoci mezi lety 1967 a 1971.

Vážený součet vzdělání a socioekonomického indexu dává **SES** – latentní proměnnou. Podobně anomie a bezmoc jsou složkami latentní proměnné **pocit odcizení**.

ZÁVISLOSTI 3 – Faktoriální ANOVA

Jak ovlivňují inteligenci dítěte jeho pohlaví, etnická příslušnost otce, matky, vzdělání otce a vzdělání matky?

Který z uvedených faktorů má největší vliv na inteligenci dítěte?

$$Y = X_1 + X_2 + \ldots + X_n + interakce$$

1 intervalová závislá

n kategoriálních nezávislých – faktorů

- Lze zde uvažovat o kombinovaném vlivu 2(či více) faktorů - **interakce**
- Pro možnost srovnávání velikosti vlivu faktorů používáme ukazatel velikosti účinku – η^2, ω^2.

ZÁVISLOSTI 4 – Vícerozm. ANOVA: MANOVA

Jak ovlivňují inteligenci a školní výkon dítěte jeho pohlaví, etnická příslušnost otce, matky, vzdělání otce a vzdělání matky?

Který z uvedených faktorů má největší vliv na inteligenci dítěte?

$$Y_1 + Y_2 + \ldots + Y_m = X_1 + X_2 + \ldots + X_n + interakce$$

m intervalových závislých

n kategoriálních nezávislých – faktorů

- Jde o rozšíření faktoriální ANOVY, testuje, zda se skupiny dané nezávislými proměnnými liší u alespoň jedné závislé (H_0).

ZÁVISLOSTI 5 – Diskriminační analýza

Známe-li schopnost rodičů intonovat, vzdělání rodičů a příjem rodičů, dokážeme predikovat, zda je jejich 15letý syn diskant, hoper, technař, nebo goth?

Který z uvedených prediktorů má největší predikční sílu?

$$Y = b_1X_1 + b_2X_2 + \ldots + b_nX_n$$

1 kategoriální závislá s 2 a více hodnotami

n intervalových nezávislých – prediktorů

Má-li závislá pouze 2 hodnoty, jde o logistickou regresi.

AJ: discriminant analysis, logistic regression

KLASIFIKACE 1 – Faktorová analýza I. - Použití

- Účelem FA je redukce většího množství proměnných na menší množství proměnných – **faktorů** nesoucích podstatné množství informace (varibility).
- Umožňuje tedy zredukovat počet proměnných v analýze.
- Typickým nasazením je analýza dotazníkových položek s cílem zjistit, které lze sečíst do jednoho skóru.
- Kromě **explorační FA** existuje i **konfirmační FA** (součást SEM)
- Předpoklady FA
  - Proměnné měřené minimálně na intervalové škále
  - Rozložení proměnných symetrické, blízké normálnímu
  - Velikost vzorku N > 20 * počet položek
  - Věcně významné korelace v matici

AJ: factor analysis, factor, exploratory FA (=FA), confirmatory FA (=CFA)

KLASIFIKACE 1 –

Faktorová analýza II. - Princip

Jde o analýzu **korelační** matice.

- Proměnné, které spolu vysoce korelují, pravděpodobně měří totéž.
- Mohou tedy být nahrazeny jedinou proměnnou, která je jejich lineární kombinací (váženým součtem) – **faktorem**.
- Váhy v tom váženém součtu jsou faktorové **náboje**.
- **„Dobrá struktura"** je požadavek na jasnost faktorové matice. Každá položka by měla vysoko (více než -+0,7) skórovat v právě jednom faktoru, každý faktor by měl obsahovat dva nebo více vysokých faktorových nábojů.

Korelační matice	P1	P2	P3	P4
P1	1	-,14	,74	,08
P2	-,14	1	,19	,59
P3	,74	,19	1	,17
P4	,08	,59	,17	1

Faktorová matice	F1	F2	h^2
P1	,73	-,60	,89
P2	,47	,77	,81
P3	,87	-,35	,87
P4	,58	,66	,77

AJ: correlation matrix, factor loading, good structure

KLASIFIKACE 2

Shluková analýza I. - Použití

- Účelem shlukové analýzy je kategorizovat objekty
  - typicky respondenty (za objekty lze považovat po transpozici datové matice i proměnné)
  - výsledkem je empirická typologie, nejistá zobecnitelnost
- Objekty kategorizujeme podle jejich vlastností - hodnot ve vstupních proměnných
  - např. kategorizujeme lidi podle *věku* a *pohlaví*
    - v takovém případě bychom měli získat 4 kategorie (shluky) – chlapce, muže, dívky a ženy
- Objekty jsou kategorizovány na základě **podobnosti**
  - existují různé **ukazatele podobnosti** (např. vzdálenost v *n*-rozměrném prostoru, kde *n* = počet vstupních proměnných)
  - maximem podobnosti je „stejnost“ - identita

AJ: cluster analysis, similarity indices (measures),

Shluková analýza II. - Princip

Analýza se skládá z neustálého opakování následujícího kroku:

- Najdi 2 nejpodobnější objekty a vytvoř z nich shluk
  - „vytvoř shluk" = utvoř z nich skupinu, kt. bude dále vystupovat jako pomyslný průměrný objekt vytvořený ze 2 původních objektů
  - po několika opakováních již budou shlukovány shluky vytvořené v předchozích krocích (do větších shluků)
- Postup končí, když jsou všechny případy v jednom velkém shluku
- Takto vzniká hierarchická struktura připomínající strom
  - na jedné straně *n* objektů, které se postupným slučováním nakonec slučují až do jednoho velkého shluku
  - grafickou podobou této struktury je **dendrogram**
- Na základě výsledků se rozhodujeme, které shluky jsou smysluplné a které ne
  - pro posouzení používáme běžná kategorizační pravidla, tj. chceme aby objekty ve shluku si byly velmi (kvalitativně) podobné a co nejvíce se odlišovaly od objektů v ostatních shlucích

KLASIFIKACE 2

Shluková analýza III. - Praktické

1. Jaký ukazatel podobnosti objektů využít?
   - musí odpovídat úrovni měření dat
2. Jak definovat podobnost shluků (cluster method)
   - např. vzdálenost „průměrů"(středů) shluků (centroid), nebo vzdálenost nejbližších prvků…
   - často se používá tzv. Wardova metoda
3. Kolik shluků chceme?
   - jen málo formálních pravidel a i ta jsou přibližná
   - rozhodnutí je dáno „smysluplností" shluků
4. Jaká data potřebujeme
   - na velikosti vzorku příliš nezáleží (záleží na účelu klasifikace)
   - počet vstupních proměnných (vlastností) držíme na minimu (<10)
   - vstupní proměnné by spolu ideálně neměly moc korelovat

Shrnutí

Vícerozměrné analýzy jsou realističtější

- můžeme zařadit do analýzy vše, co je relevantní ☺
- realističnost = složitost ☹
  - vztahy mezi nezávislými – co má vlastně vliv?
  - mnoho možností při specifikování modelu

Velmi obecné hypotézy.

Více proměnných vyžaduje **větší vzorky** a obvykle i **lepší měření** (více prostoru pro to, aby se projevil každý defekt)

Je dobré vyhledat pomoc zkušenějších.