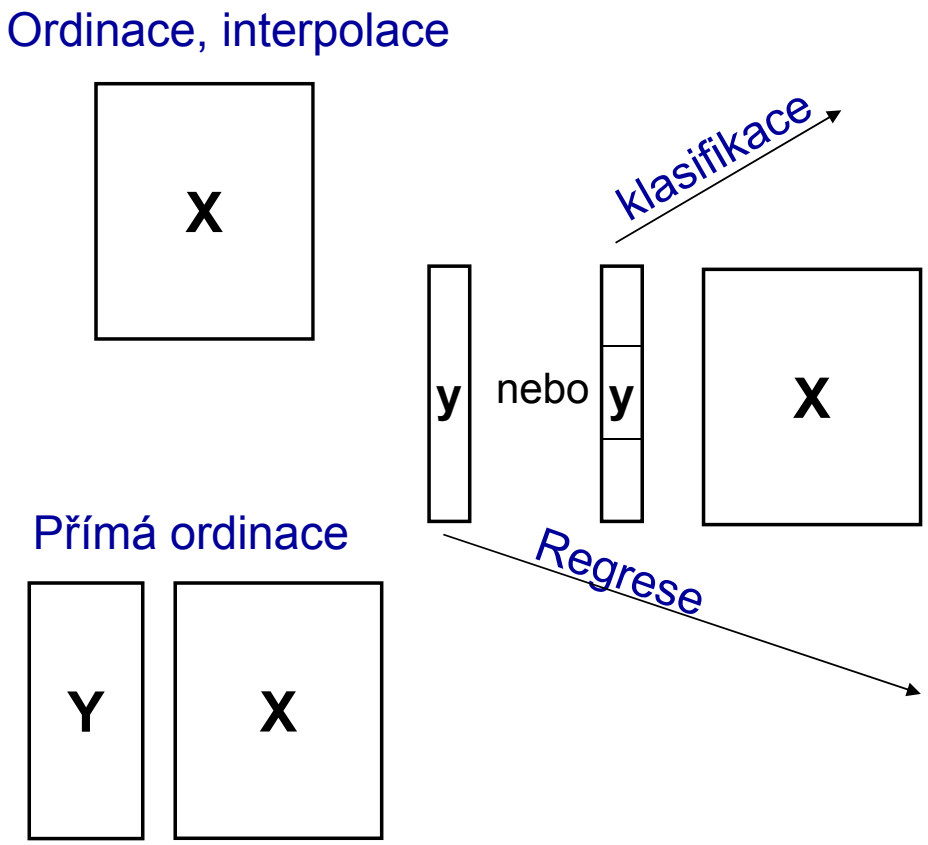
MUNI | RECETOX

Vícerozměrné metody pro predikci, identifikaci a klasifikaci znečištění

Pokročilejší modelovací přístupy



Klasifikace

- Metody založené na stromech
- Lineární dikriminační analýza
- Neuronové sítě
- Metoda podpůrných vektorů
- Logistická regrese
- Bayesovský klasifikátor

…

Regrese

- Klasický lineání model
- Lineární zobecněné a aditivní modely
- Nelineární regrese
- Na stromech založené techniky
- Neuronové sítě
- Metoda podpůrných vektorů
- Na stromech založené techniky

…

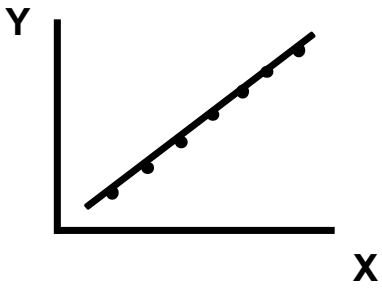
Regresní metody

Regrese

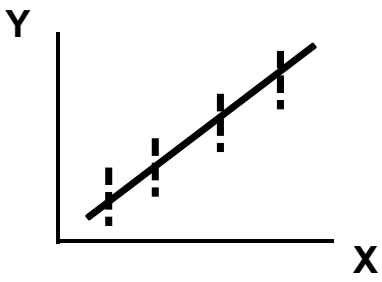
- Regrese - funkční vztah dvou nebo více proměnných závislost jedné veličiny na druhé

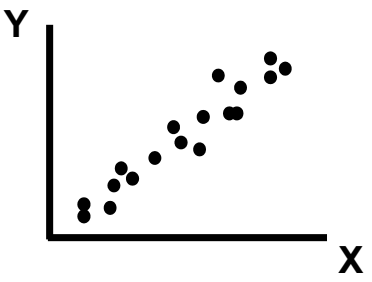
Vztah x, y

Jednorozměrná



Vícerozměrná
y = f(x1, x2, x3,





Pro každé x existuje pravděpodobnostní rozložení y

MUNI | RECETOX

Lineární regresní model

- Jedna a více nezávisle proměnných

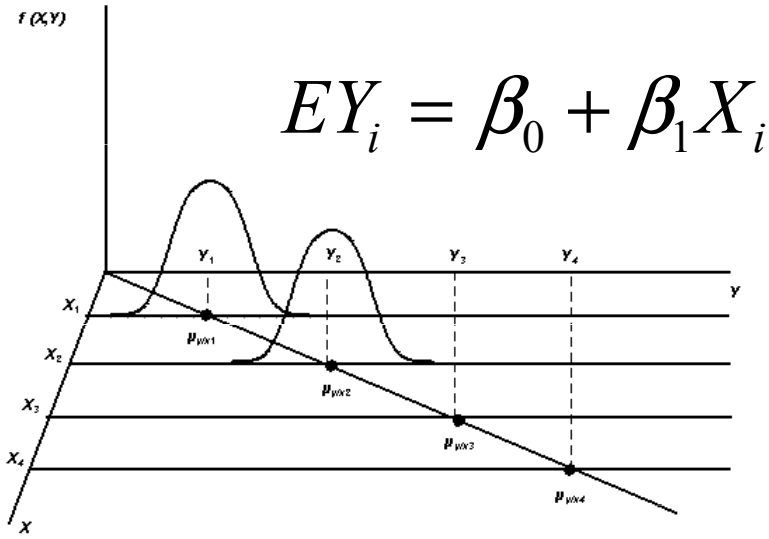
- *n* objektů
- Pro každý objekt: pozorované veličiny X a Y – spojité

- Pozorování, objekty – navzájem nezávislá

- Zajímá nás závislost veličiny Y na X – POZOR! – nutná podmínka je, že závislost je stejná pro všechny zkoumané objekty.

Lineární regresní model

- X,Y – náhodné veličiny (střední hodnota, rozptyl)
- Existuje souvislost mezi středními hodnotami X.?



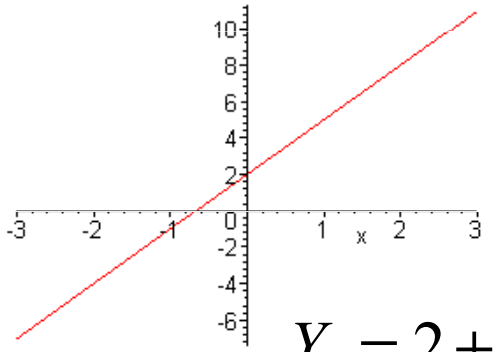
$$EY_i = \beta_0 + \beta_1 X_i, i = 1,...,12.$$

MUNI | RECETOX

Opakování z gymnázia – analytická geometrie

- Analytické vyjádření přímky, rovnice

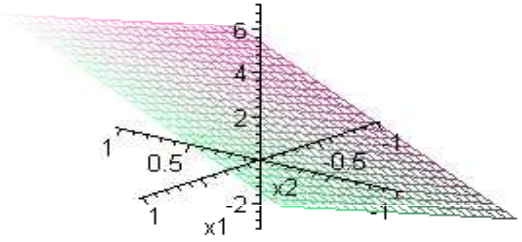
$$Y = \beta_0 + \beta_1 X$$



$$Y = 2 + 3X$$

- Analytické vyjádření roviny, rovnice

$$Y = \beta_0 + \beta_1 X1 + \beta 2 X2$$



$$Y = 2 + 3X1 + 2X2$$

MUNI | RECETOX

Nejjednodušší typ závislosti - lineární

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, i = 1,\ldots,n$$

Systematická část modelu

Náhodná část, složka modelu (náhodné chyby, *random error*)

Regresní rovnice - proměnné

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, i = 1,\ldots,n$$

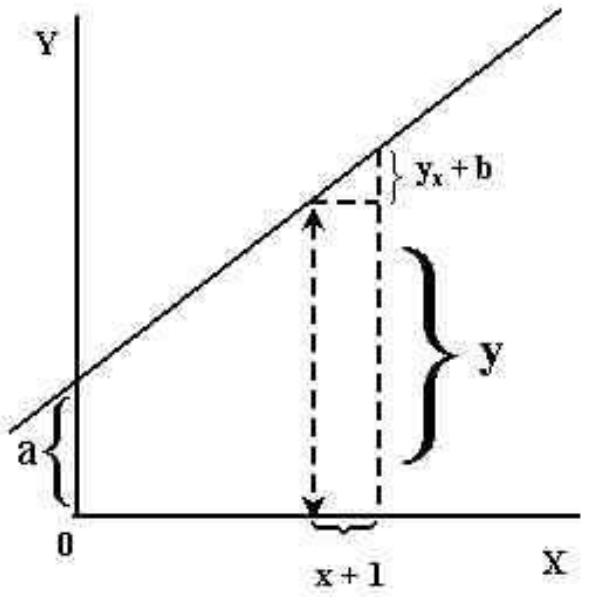
- Závisle proměnná
- *Dependent variable*

- Nezávisle proměnná
- *Independent variable*
- Kovariáta *(covariate)*
- Prediktor
- Regresor

Regresní rovnice, přímka - parametry

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, i = 1,\ldots,n$$

- Průsečík s osou Y
- Směrnice
- *Intercept*



Interpretace parametrů:

Směrnice: o kolik se změní hodnota závisle proměnné, jestliže hodnota nezávisle proměnné vzroste o 1 jednotku.

Průsečík: udává hodnotu závisle proměnné, jestliže hodnota nezávisle proměnné je rovna 0.

Tvorba lineárního regresního modelu

- Je-li závisle proměnná spojitá a nezávisle proměnné jsou spojité nebo diskrétní (podmínkou je, že alespoň jedna nezávisle proměnná je spojitá) a jsou-li splněny jisté předpoklady.....

- Při tvorbě modelu (obecně, nejen lineárního) postupujeme následujícím způsobem:

  1. Odhadneme parametry modelu
  2. Hledáme významné (signifikantní) prediktory
  3. Na závěr hodnotíme vhodnost námi vytvořeného modelu, jak dobře popisuje funkcionální závislost mezi závisle proměnnou a nezávisle proměnnými.

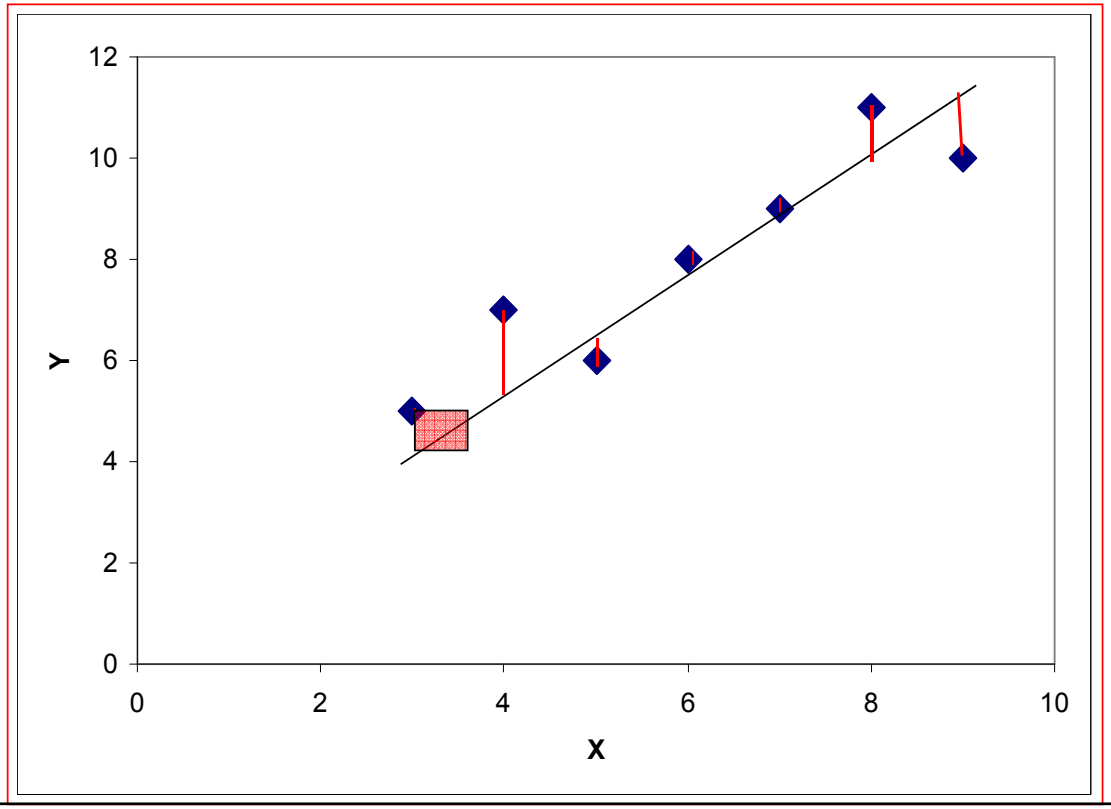
Residua

- Svislé odchylky naměřených hodnot od regresní přímky nazýváme **residua.**
- *i*-té residuum vypočteme jako rozdíl skutečně naměřené hodnoty Y a hodnoty predikované regresním modelem $\hat{Y}$

$$\text{Reziduum}_i = Y_i - \hat{Y}_i = Y_i - (\beta_0 + \beta_1 X_i) = Y_i - \beta_0 - \beta_1 X_i$$

Metoda nejmenších čtverců

- Výsledný minimální součet čtverců residuí (pro b0 a b1) nazýváme **residuální součet čtverců** (*residual sum of squares*)



Svislá - nikoliv kolmá vzdálenost k přímce!!!

$$\sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2$$

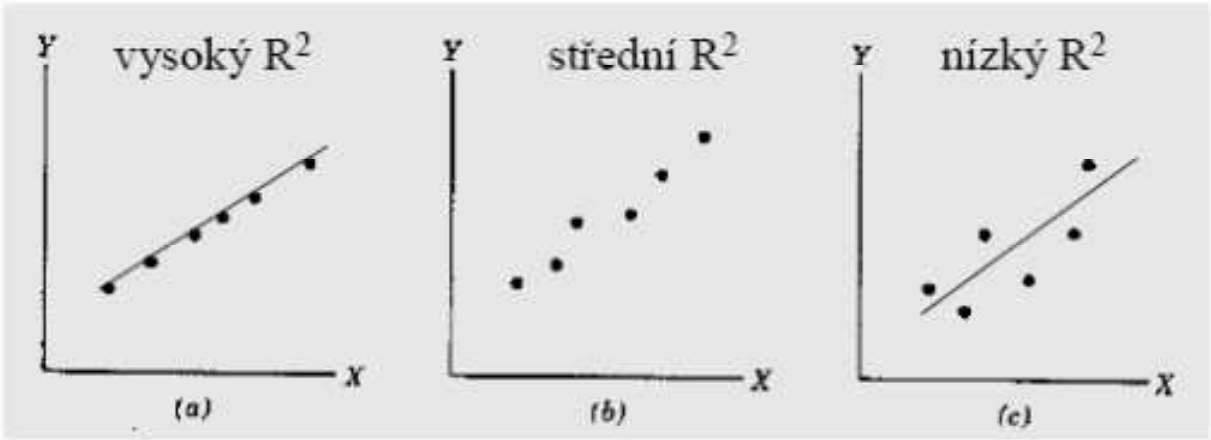
$$S_e = \sum_{i'=1}^{n} (Y - b_0 - b_1 X_i)^2$$

MUNI | RECETOX

Koeficient determinace - procento vysvětlené variability

$$R^2 = \frac{variabilita\_vysvetlena\_modelem}{celkova\_variabilita\_Y} = 1 - \frac{residualni\_variabilita}{celkova\_variabilita\_Y} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2},$$

$$R^2 = \frac{SS_{REG}}{SS_{TOT}} = 1 - \frac{SS_e}{SS_{TOT}}$$



(Lepš 1996)

MUNI | RECETOX

Koeficient determinace - vlastnosti

- Koeficient determinace udává relativní velikost variability závisle proměnné, kterou se uvažovanou závislostí podařilo vysvětlit.
- Koeficient determinace nabývá hodnot od 0 do 1.
- Čím vyšší je hodnota koeficientu determinace, tím je náš regresní model lepší.
- V případě regrese s jedinou nezávisle proměnnou je hodnota koeficientu determinace rovna kvadrátu Pearsonova korelačního koeficientu mezi veličinami X a Y.

$$R^2 = corr(X,Y)^2$$

Pearsonův korelační koeficient

- **Pearsonův korelační koeficient**

  postižení lineárního vztahu mezi veličinami

$$r = \frac{Cov(x,y)}{S_x . S_y} = \frac{\sum x_i y_i - \frac{1}{n}\sum x_i \sum y_i}{\sqrt{\left[\sum x_i^2 - \frac{1}{n}\left(\sum x_i\right)^2\right]\left[\sum y_i^2 - \frac{1}{n}\left(\sum y_i\right)^2\right]}}$$

- **R=1 … přímá úměra, kladná korelace**
- **R=-1… záporná korelace**
- **R=0… mezi veličinami není žádná spojitost, žádná korelace, není lineární vztah mezi proměnnými**

- **Předpoklady: dvourozměrné normální rozdělení**

Vzorce pro odhad parametrů regresní přímky – metoda nejmenších čtverců

Odhad b je zatížený chybou:

$$S_b^2 \sim \sigma_\beta^2: \quad \frac{1}{\sum (X_i - \bar{X})^2} \cdot S_{y \cdot x}^2$$

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

I. $$b \sim \beta: \quad b = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}$$

II. intercept

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

III. $\hat{Y}$: modelová hodnota

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

MUNI | RECETOX

Hledáme významné (signifikantní) prediktory

Při konstrukci regresního modelu bychom chtěli prokázat, že závislá veličina skutečně závisí na nezávisle proměnné.Tuto závislost na *X* prokazujeme testováním nulové hypotézy

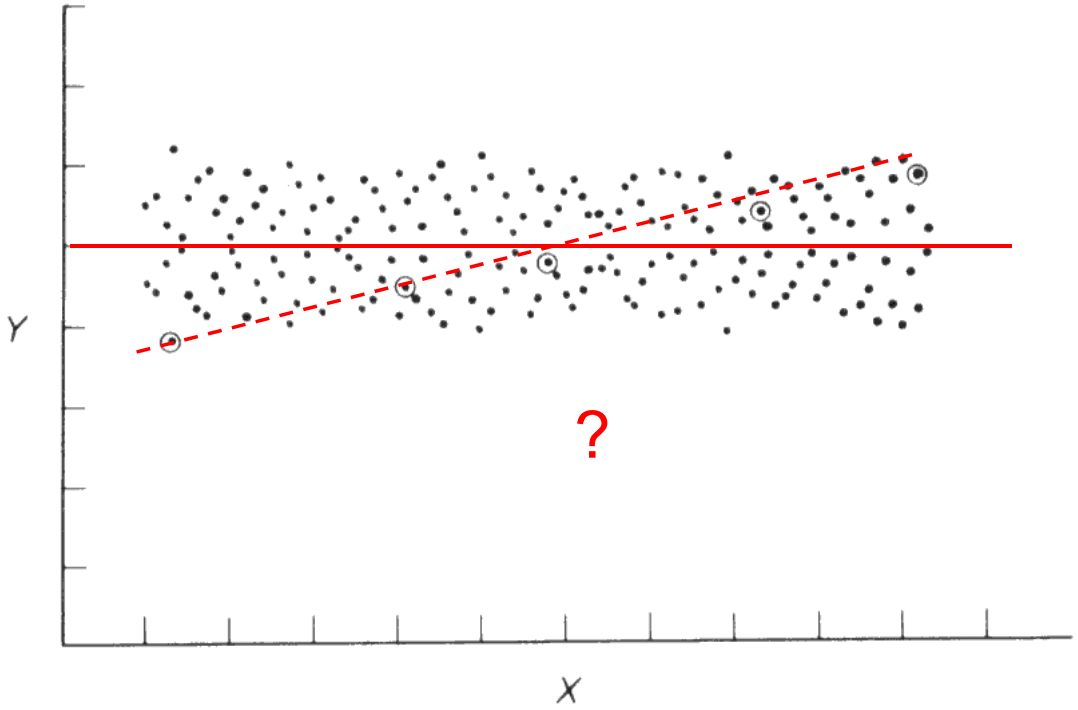
$$H_0 : \beta_1 = 0$$

proti alternativní hypotéze

$$H_A : \beta_1 \neq 0.$$

Testujeme T-testem

b je výběrovým odhadem skutečné hodnoty *β*

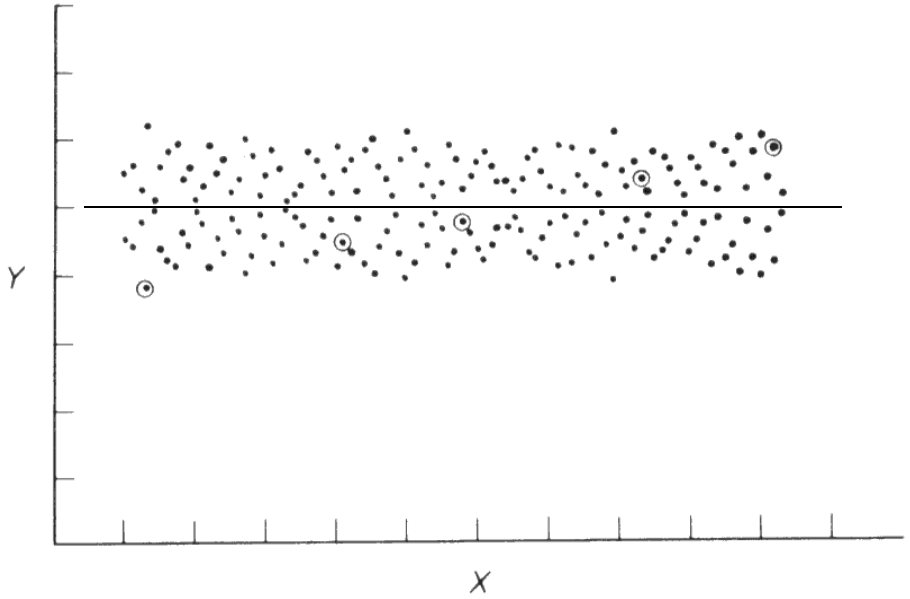


Každý odhad je zatížen nějakou chybou - z variability dat můžeme spočítat střední chybu odhadu *b*

Hypotetický základní soubor dat, s regresním koeficientem β rovným nule. Zakroužkované body mohou být možným výběrem pěti pozorování.

MUNI | RECETOX

V případě nezávislosti β=0



Dosažená hladina významnosti pro test

H_0: β=0

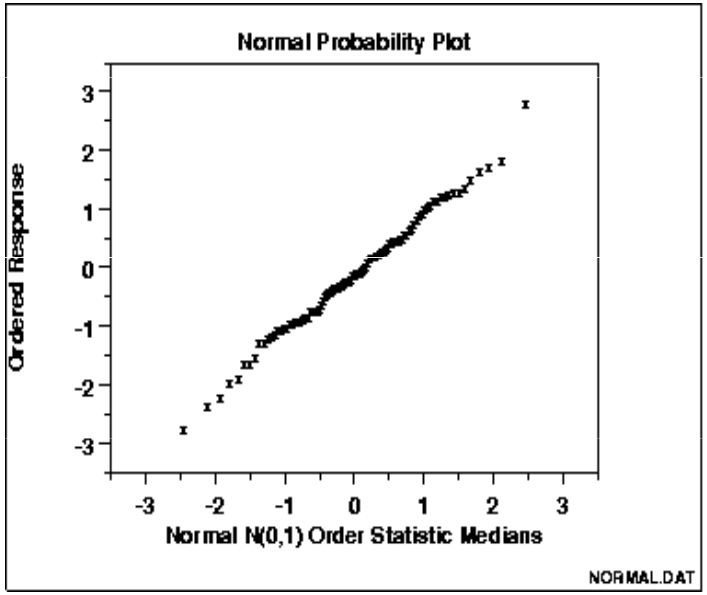
je pravděpodobnost, že takhle dobrou závislost dostaneme čistě náhodou, pokud jsou proměnné nezávislé

Předpoklady

- Nutný předpoklad potřebný ke všem testům spojeným s regresním modelem je **normalita residuí.**
- Residua mají mít normální rozdělení s **nulovou střední hodnotou a konstantním rozptylem** .
- Dále předpokládáme, že všechna **pozorování** jsou **navzájem nezávislá.**

Normalita residuí – graficky Q-Q plot *(Quantile-Quantile plot)*

- Grafická metoda pro srovnání rozdělení dvou výběrů.
- Vodorovná osa – empirické kvantily rozdělení 1. výběru. (jestliže vynášíme teoretické kvantily normovaného normálního rozdělení – **normal probability plot**)
- Svislá osa – empirické kvantily rozdělení 2. výběru (např. reziduí).
- Jsou-li obě rozdělení totožná, leží body (odpovídající si kvantily) na diagonální přímce



MUNI | RECETOX

Normalita residuí - testy

- Testy normality:

1. Kolmogorov-Smirnov
2. Shapiro-Wilks

Není-li splněn předpoklad normality – mohou pomoci **transformace**

Diagnostika residuí

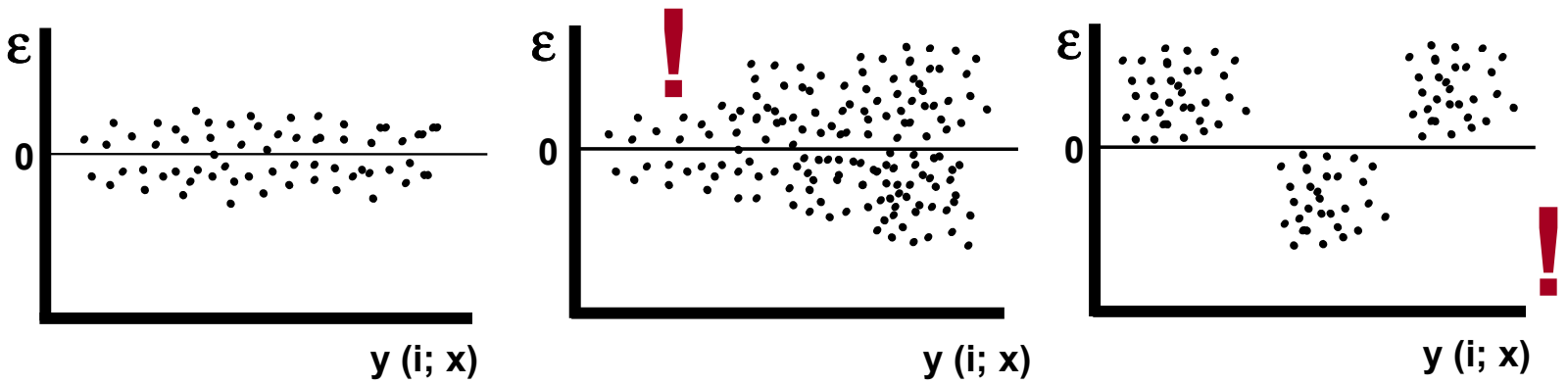
- Je námi zvolená závislost (lineární) vhodná?
- Pomoc grafické znázornění – **grafy závislosti hodnot residuí na hodnotách Y nebo X.**

- V případě, že zvolený tvar závislosti byl vhodný, jsou residua

1. umístěna náhodně kolem nulové střední hodnoty
2. nevykazují žádný systematický trend
3. jejich rozptyl je homogenní

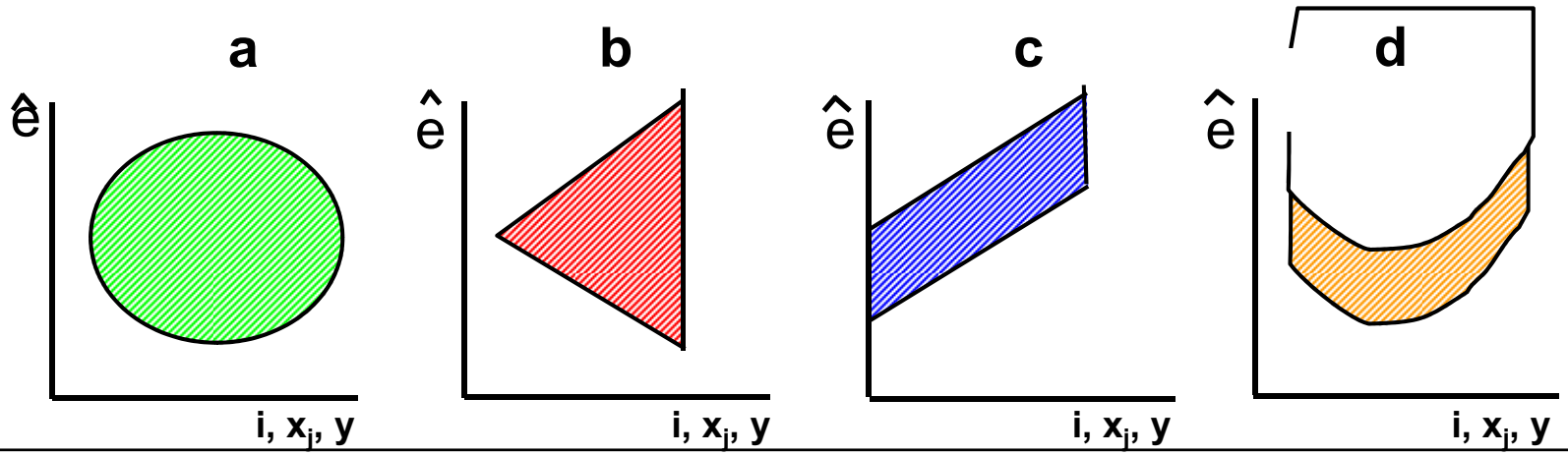
Diagnostika residuí

Grafy residuí modelů (příklady)



Pozor na odlehlé hodnoty!

Obecné tvary residuí modelů (schéma)



MUNI | RECETOX

Interakce

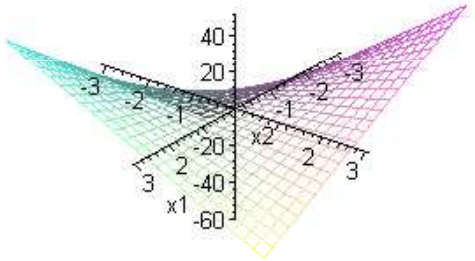
- Při zkoumání závislosti veličiny **y** na regresoru **x** třeba vzít v úvahu také další veličiny, **z**.
- **Interakce (effect modifcation)** – skutečná na hodnota veličiny z ovlivňuje závislost y na x.
- Vyjadřujeme pomocí součinu x·z.
- **Příklad:** závislost platu na délce praxe, když se zjistí, že směrnice příslušné přímky je jiná u mužů a u žen.
- Kdyby byly přímky rovnoběžné, byl by vliv veličin délka praxe a pohlaví aditivní. Každý rok praxe by v průměru přidal stejnou částku k platu mužům i ženám.
- Vliv délky praxe by naopak byl modifikován proměnnou pohlavi, kdyby tyto průměrné přírůstky byly u mužů a u žen různé.
- Model s interakcemi:

$$Y_i = \beta_0 + \beta_1 X1_i + \beta_2 X2_i + \beta_3 X1 \cdot X2 + \varepsilon_i$$

Regresní plocha*(Response surface, regression surface)*

- Model s interakcemi

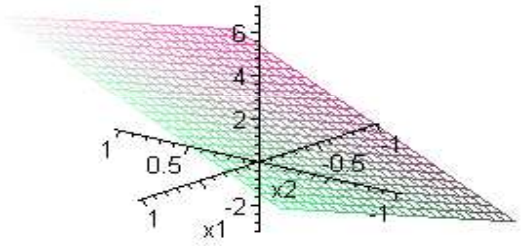
$$Y = \beta_0 + \beta_1 X1 + \beta_2 X2 + \beta_3 X1 \cdot X2$$



$$Y = 2 + 3X1 + 2X2 - 5X1X2$$

- Model bez interakcí – regresní rovina (*plane*)

$$Y = \beta_0 + \beta_1 X1 + \beta 2 X2$$



$$Y = 2 + 3X1 + 2X2$$

Snaha o co nejjednodušší model, obsahující jenom významné prediktory (nezávisle proměnné)

MUNI | RECETOX

T-test, F-test

- t-test: $H_0 : \beta_1 = 0$ nebo $H_0 : \beta_2 = 0$
- F test: $H_0 : \beta_1 = 0$ nebo $H_0 : \beta_2 = 0$ nebo $H_0 : \beta_1 = \beta_2 = 0$
- Upozornění: opakovaný t-test a F-test mohou dávat nekonzistentní výsledky

- **Podmodel** = jednodušší model obsahující pouze některé nezávisle proměnné (signifikantní) původního regresního modelu.
- S každou mocninou veličiny musí být v modelu všechny mocniny nižšího stupně, se součinem veličin musí být v modelu také všechny složky tohoto součinu.

Strategie hledání vhodného podmodelu

Sekvenční postupy

- **Sestupný výběr** - Nejprve se spočítá nejbohatší model, pak se jednotlivé regresory postupně z modelu vylučují. V každém kroku se vylučuje takový regresor, který v daném modelu nejméně přispívá k vysvětlení.

- **Vzestupný výběr** – opak sestupného výběru. Vyjde se z prázdné množiny regresorů, do níž se pak v každém kroku přidá vždy ten z ještě nezařazených regresorů, který v daném kroku co možná nejlépe zlepší vysvětlení závisle proměnné.

- **Kroková (stepwise) regrese** - kombinuje oba předešlé postupy. Vzestupný výběr je v každém kroku kombinován s pokusem o zjednodušení pomoci sestupného výběru.
- Každá z popsaných metod může dát jiný výsledný model, kromě jiného závisí také na volbě hladin testů.
- Zejména u krokové regrese se doporučuje najít několik téměř optimálních modelů a pokusit se najit mezi nimi ten,který má nejlepší interpretaci.

- Všechny modely!

Multikolinearita

- **Multikolinearita** - Existují-li závislosti mezi jednotlivými nezávisle proměnnými modelu. Koeficienty determinace lineárních modelů (jedné nezávisle proměnné na ostatních nezávisle proměnných) jsou vysoké (větší než 0,5). Nezávisle proměnné jsou navzájem korelované.
- Odhad regresních parametrů – velký rozptyl.
- I významné nezávisle proměnné se jeví jako nevýznamné, popř. parametry mohou mít opačné znaménko…
- Obtížná interpretace parametrů beta. (Obvykle: Koeficient beta1 lze interpretovat jako střední změnu Y při jednotkové změně X1 a nezměněné hodnotě X2. Nyní však X1 a X2 vzájemně korelované, proto nelze předpokládat, že při změně X1 zůstane X2 nezměněna.)
- Příklad 1: obvod pasu a váha významně korelované
- Příklad 2: Výška platu a daně úzce korelované
- **Řešení:** méně proměnných v modelu, vyloučení korelovaných nezávislých proměnných.

Umělé proměnné *(Dummy variables, dummies)*

- Vyjádření nominální veličiny s více než 2 hodnotami
- j úrovní faktoru -> j-1 umělých proměnných (v modelu buďto všech j-1 umělých proměnných nebo žádná)

Proměnná	Umělé proměnné (stačí 3)			
Rodinný příslušník (4 úrovně)	Otec (0/1)	Matka (0/1)	Strýc (0/1)	Dědeček (0/1) (zbytečná)
„otec“	1	0	0	0
„matka“	0	1	0	0
„strýc“	0	0	1	0
„dědeček“	0	0	0	1

Ozón cvičení

- V tomto příkladu budeme sledovat závislost denního měření koncentrace ozónu (ppb) na rychlosti větru (míle/h), teplotě vzduchu (denní maximum ve stupních Fahrenheita) a intenzitě slunečního záření (cal/cm^2) v New Yorku. Soubor obsahuje celkem 111 měření, která proběhla od května do září v roce 1973.

- Přízemní ozón je součástí tzv. fotochemického smogu, který se vyskytuje v místech s intenzivní automobilovou dopravou. Jeho původcem jsou oxidy dusíku emitované jako součást spalin ze spalovacích motorů. Působením slunečního záření se tyto oxidy štěpí a vzniklé radikály reagují s kyslíkem za vzniku ozónu. Jeho zvýšené koncentrace můžeme tedy očekávat v letních měsících při vyšších teplotách. Určitý nárůst koncentrací ozónu lze ale očekávat i za slunečného počasí v chladnějších měsících, pokud jsou zhoršené rozptylové podmínky. Podíváme se, zdali jsou tato očekávání ověřitelná pomocí výše zmíněných měření.

Vícerozměrné metody

Vícerozměrné metody

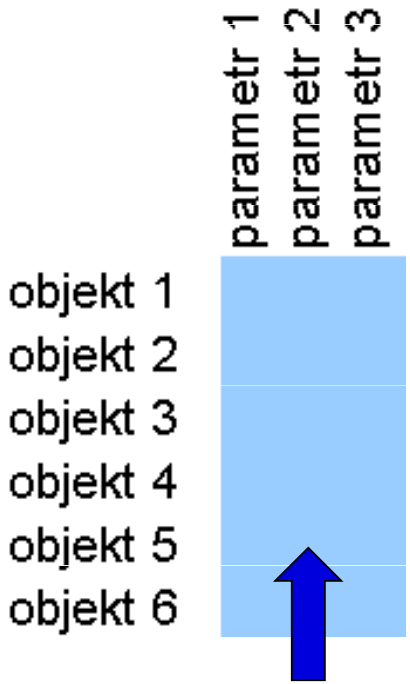
1. Cluster Analysis
2. Principal component analysis
3. Correspondence analysis
4. Canonical analysis
5. Discriminant analysis
6. Factor analysis
7. Multidimensional scaling

Úvod do vícerozměrných metod I.

- parametry (věk, příjem atd.) a každý z těchto parametrů můžeme považovat za jeden rozměr objektu.
- **Maticová algebra**: Základem práce s daty a výpočtů vícerozměrných metod je maticová algebra, matice tvoří jak vstupní, tak výstupní data a probíhají na nich výpočty.
- **NxP matice**: N objektů s p parametry pak vytváří tzv. NxP matici, která je prvním typem vstupu dat do vícerozměrných analýz.
- **Asociační matice**: Na základě těchto matic jsou počítány matice asociační na nichž pak probíhají další výpočty, jde o čtvercové matice obsahující informace o podobnosti nebo rozdílnosti (tzv. **metriky**) buď objektů (Q mode analýza) nebo parametrů (R mode analýza).Měřítko podobnosti se liší podle použité metody a typu dat, některé metody umožňují použití uživatelských metrik.

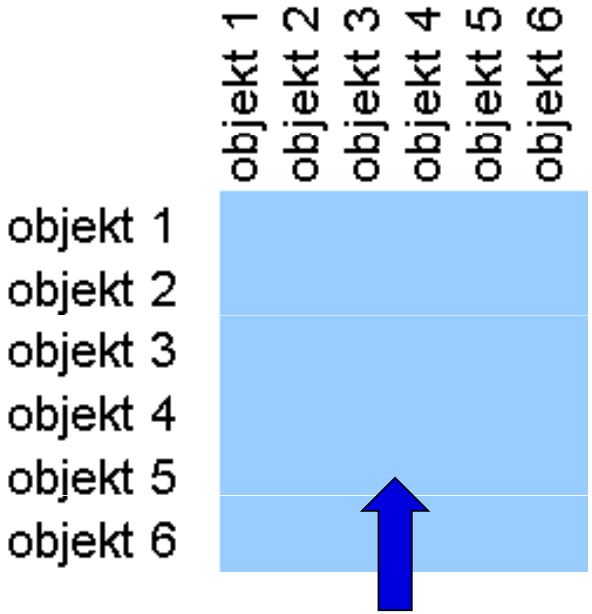
Vstupní matice vícerozměrných analýz

NxP MATICE



Hodnoty parametrů pro jednotlivé objekty

ASOCIAČNÍ MATICE



Korelace, kovariance, vzdálenost, podobnost

MUNI | RECETOX

Úvod do vícerozměrných metod II.

SHLUKOVÁ ANALÝZA

- vytváření shluků objektů na základě jejich podobnosti
- identifikace typů objektů

ORDINAČNÍ METODY

- zjednodušení vícerozměrného problému do menšího počtu rozměrů
- principem je tvorba nových rozměrů, které lépe vyčerpávají variabilitu dat

Shluková analýza

- Existuje několik typů shlukové analýzy, které se liší postupem shlukování. Shlukování může být ***hierarchické*** nebo ***nehierarchické***.
- ***Hierarchická shluková analýza*** vytváří systém skupin a podskupin tak, že každá skupina může obsahovat několik podskupin nižšího řádu a sama může být součástí skupiny vyšší-ho řádu. Výsledek se dá graficky znázornit stromem – dendrogramem.
- ***Nehierarchická shluková analýza*** (*partitioning methods*) rozdělí objekty do několika shluků stejného řádu.

Shluková analýza

- Vstupní data:
  - Tabulka spojitých nebo kategoriálních dat popisujících respondenty nebo jejich skupiny
- Výstupy analýzy
  - Tzv. dendrogram popisující vazby mezi vzorky nebo parametry
  - Rozdělení respondentů nebo parametrů do daného počtu skupin
- Kritické problémy analýzy
  - Velké množství parametrů nebo respondentů v dendrogramu je obtížně interpretovatelné
  - Analýza je silně závislá na zvolení vhodné metriky vzdáleností
  - Analýza je silně závislá na shlukovacím algoritmu
  - Korelace proměnných
  - Převážení informace

Měření vzdálenosti objektů

Spojitá data

Euklidovská vzdálenost

$$d_{ij} = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2}$$

Vážená euklidovská vzdálenost

$$d_{ij} = \sqrt{\sum_{k=1}^{p} w_k^2 (x_{ik} - x_{jk})^2}$$

i,j – označení objektů
d_{ij} – vzdálenost objektů i a j
p – počet parametrů
k – k-tý parametr
w_k – váha parametru k

Minkowski (power distance)

$$d_{ij} = \sqrt[\lambda]{\sum_{k=1}^{p} \left| x_{ik} - x_{jk} \right|^{\lambda}}$$

- λ - celé číslo
- λ =1 Manhattan (city block)
- λ= 2 Euklidovská vzdálenost

Chebychev

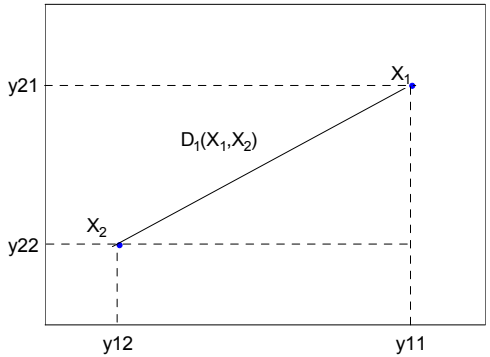
$$d_{ij} = \max \left| x_{ik} - x_{jk} \right|$$

Euklidovská vzdálenost jako základní vícerozměrná metrika

- Jde o základní metrické měřítko vzdálenosti a počítá vzdálenost objektů obdobně jako Pythagorova věta počítá přeponu pravoúhlého trojúhelníku.
- Metoda je citlivá na rozdílný rozsah hodnot vstupujících proměnných (vhodným řešením může být standardizace) a double zero problém.
- Nemá horní hranici hodnot.

$$D_1(x_1, x_2) = \sqrt{\sum_{j=1}^{p} (y_{1j} - y_{2j})^2}$$

- Jako další měřítko se používá také čtverec této vzdálenosti. . Jeho nevýhodou jsou semimetrické vlastnosti.



$$D_1^2(x_1, x_2) = \sum_{j=1}^{p} (y_{1j} - y_{2j})^2$$

MUNI | RECETOX

Měření podobnosti objektů

Binární koeficienty podobnosti

Objekt 2 \ Objekt 1	1	0
1	a	b
0	c	d

a, b, c, d = počet případů, kdy souhlasí binární charakteristika objektu 1 a 2
a+b+c+d=p

Symetrické binární koficienty - není rozdíl mezi případem 1-1 a 0-0

Simple matching coefficient

$$S(x_1 x_2) = \frac{a+d}{p}$$

Měření podobnosti objektů

Asymetrické binární koeficienty – odstranění double zero

Jaccard`s coefficient

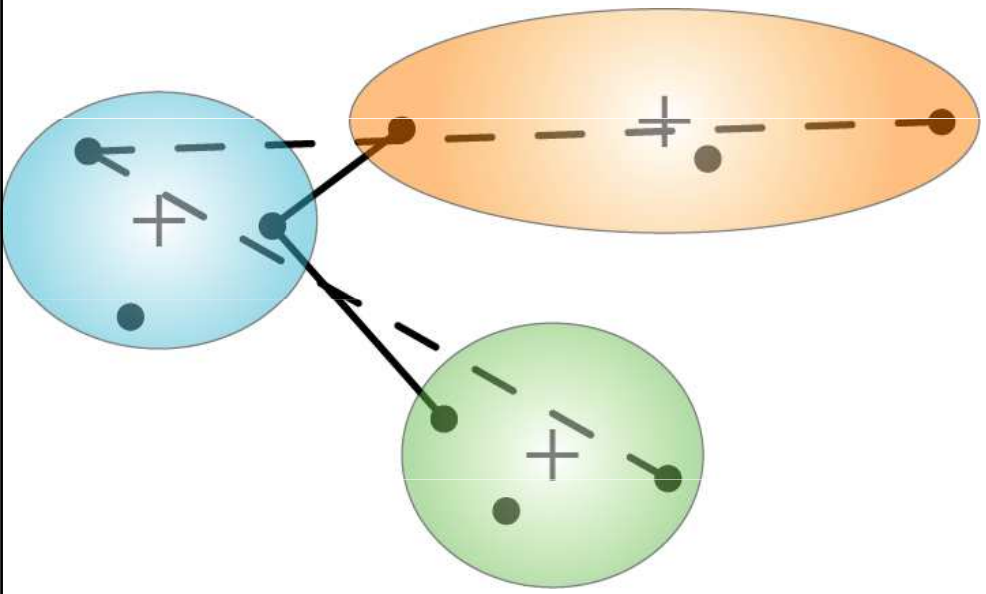
$$S(x_1 x_2) = \frac{a}{a+b+c}$$

Sorensen`s coefficient

$$S(x_1 x_2) = \frac{2a}{2a+b+c}$$

Řada dalších koeficientů dávajících různou váhu jednotlivým kombinacím parametrů

Joining (Tree Clustering) – shlukovací algoritmy



"**Klasická" shluková analýza** hierarchicky spojující objekty do skupin podle vzdálenosti v asociační matici

\+ centroid

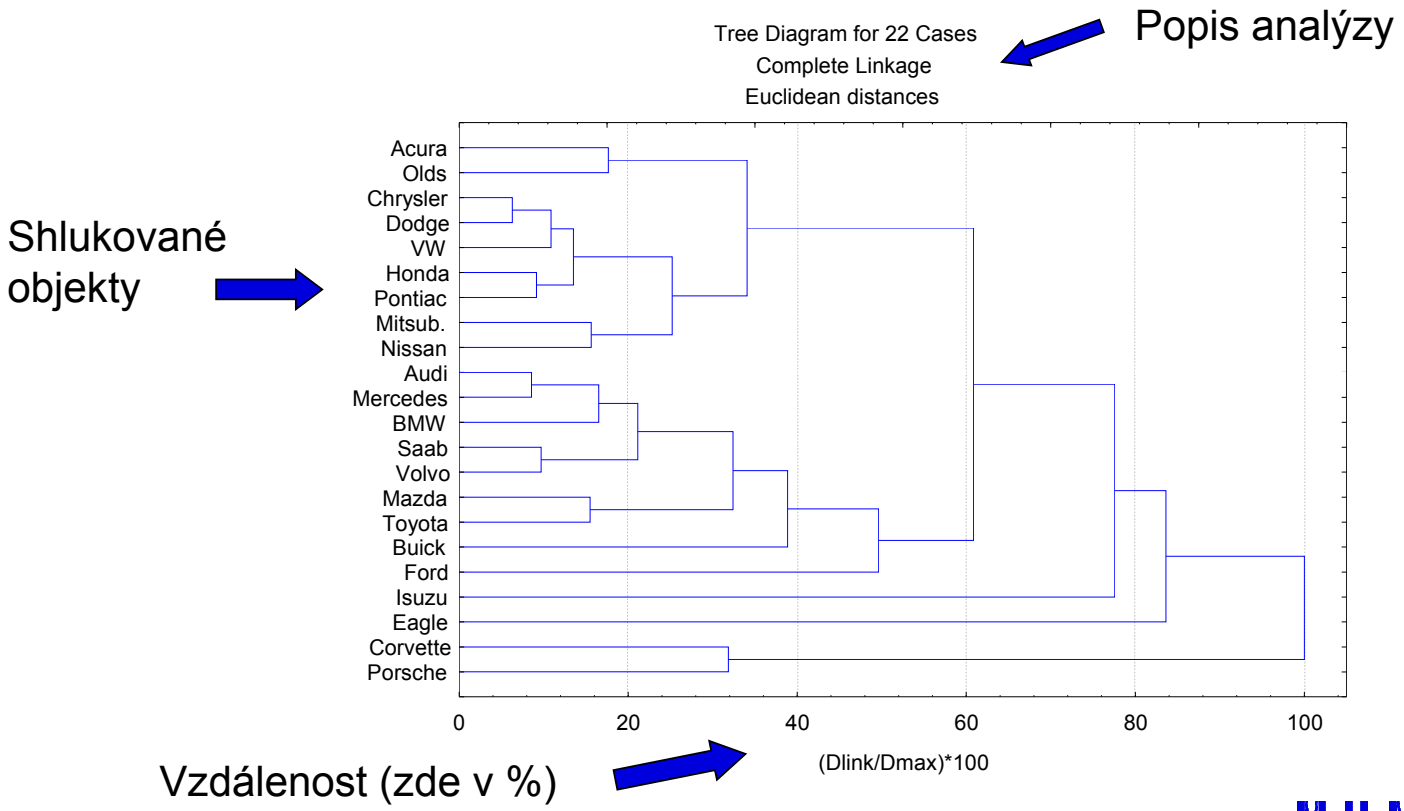
Na tuto vzdálenost se ptá **single linkage**
Na tuto vzdálenost se ptá **complete linkage**

Další metody počítají s **průměrnou vzdáleností** všech objektů shluků nebo vzdáleností **centroidů** (vzdálenost může být **vážena** velikostí shluků).
Wardova metoda se snaží minimalizovat variabilitu uvnitř shluků.

MUNI | RECETOX

Dendrogram

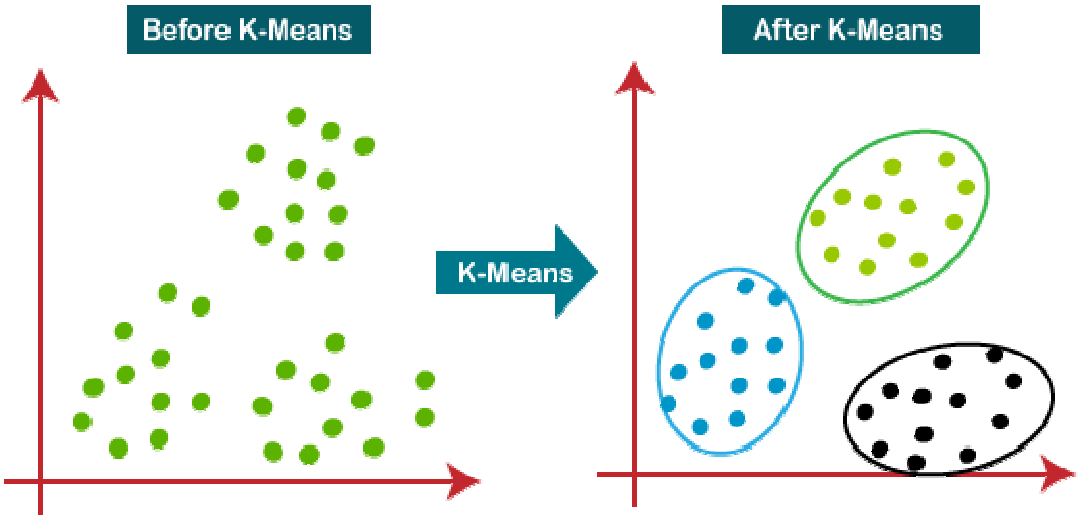
Dendrogram představuje grafický výstup shlukové analýzy, kde jsou objekty propojeny tak, jak postupovalo jejich shlukování



MUNI | RECETOX

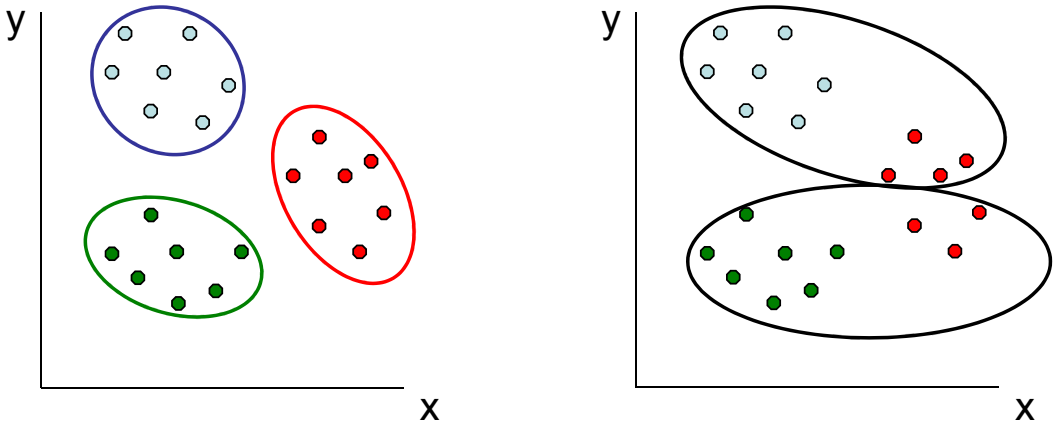
Shluková analýza K-means clustering

- **K-means clustering** se snaží rozdělit objekty do zadaného počtu shluků tak, aby byla minimalizována variabilita uvnitř shluků a maximalizována mezi shluky



Shluková analýza K-means clustering

- Vzorky jsou na základě zadaného počtu shluků rozděleni podle kritéria maximální homogenity shluků
- Rizika analýzy
  - Při špatném odhadu počtu shluků dává metoda chybné výsledky
  - Výpočet je možný pouze na Euklidovských vzdálenostech se všemi jejich omezeními



Ordinační analýzy

- Analýza hlavních komponent, faktorová analýza, korespondenční analýza a diskriminační analýza se snaží zjednodušit vícerozměrnou strukturu dat výpočtem souhrnných os
- Metody se liší v logice tvorby těchto os
  - Maximální variabilita (analýza hlavních komponent, korespondenční analýza)
  - Maximální interpretovatelnost os (faktorová analýza)
  - Maximální diskriminace skupin (diskriminační analýza)

PCA

- nové (latentní) proměnné (hlavní komponenty, principal components) vysvětlují maximum celkového rozptylu původních proměnných, případně maximálně reprodukují celkovou kovarianční (nebo korelační) matici výchozích proměnných
- Matice kovariancí – data jsou standardizována na průměr, ale je zohledněn rozptyl primárních dat-proměnné mají srovnatelný význam a absolutní hodnota rozptylu zohledňuje vzájemné váhy proměnných.
- Matice korelačních koeficientů – data jsou standardizována jak na průměr, tak na rozptyl, analýza pracuje s jednotkovým rozptylem proměnných a zohledňuje pouze sílu jejich vazby v rozsahu -1 až 1.

PCA

- Proces hledání hlavních komponent je postupný
- Výsledkem jsou ortogonální (nekorelované) faktory
- Hlavní komponenty jsou uspořádány podle jejich klesajícího rozptylu.
- Algebraicky PCA hledá vlastní hodnoty (eigenvalues) a vlastní vektory (eigenvectors) asociační matice.
- Prvky vlastních vektorů jsou váhy původních proměnných, udávají pozici objektů vzhledem k novému systému vytvořenému hlavními komponentami

Předpoklady PCA

- mnohorozměrné normální rozdělení proměnných
  - na menší odchylky od mnohorozměrného normálního rozdělení je PCA dostatečně robustní.
- kvantitativní proměnné- je možné pro ně vypočítat kovarianci nebo korelaci.
  - částečně robustní i pro zpracování semikvantitativních a binárních proměnných
  - není vhodná pro vícestavové kvalitativní proměnné, na které nelze použít euklidovskou metriku.
- nezávislost pozorování (objektů)

Předpoklady a omezení PCA

- nevhodná pro data obsahují mnoho nul (double zero problem)
- korelace větších skupiny proměnných
- počet proměnných by měl být menší, než je počet objektů n
- odlehlé hodnoty