Téma 5: Analýza závislostí

Přednáška 15 – Závislost mezi jevy

Základní pojmy

Předmětem této kapitoly bude zkoumání závislostí (souvislostí) mezi dvěma a více jevy. Jedná se o proniknutí do vztahů mezi sledovanými jevy a tím i přiblížení k tzv. *příčinným*, tj. *kauzálním* souvislostem. Příčinnou souvislostí mezi jevy se rozumí situace, kdy změny jednoho jevu (příčina) podmiňují změny druhého jevu (důsledek, účinek). V praxi mohou nastat různě složité situace:

~ existence určitého jevu má za následek výskyt jiného jevu (párová závislost: $X \to Y$)

~ existence skupiny jevů má za následek výskyt jednoho jiného jevu (vícenásobná závislost: $X_1, X_2, \ldots, X_p \to Y$)

~ existence určitého jevu nebo skupiny jevů má za následek výskyt jiných jevů ($X \to Y_1, Y_2, \ldots, Y_r$ resp. $X_1, X_2, \ldots, X_p \to Y_1, Y_2, \ldots, Y_r$)

Závislost mezi jevy zkoumá statistika tak, že vyšetřuje souvislost mezi náhodnými veličinami (statistickými znaky), které dané jevy charakterizují. Při práci s nimi bude výhodné je nazývat proměnnými:

X ... ***nezávislá = vysvětlující proměnná***

Y ... ***závislá = vysvětlovaná proměnná***

Z hlediska metod zkoumání kauzálních souvislostí je vhodné rozlišovat tzv. *pevné* a *volné* (*stochastické*) závislosti:

~ ***pevnou*** závislostí se označuje případ, kdy změně např. jednoho jevu nutně odpovídá změna druhého jevu (a často i naopak) s pravděpodobností rovnou jedné; jedná se vlastně o funkci $y = f(x)$, $z = f(x,y)$, … , např. $P = a^2$, $V = \pi r^2 v$, $I = U/R$, …

~ ***volnou*** (***stochastickou***) závislostí se označuje případ, kdy změna např. jednoho jevu vyvolá změnu druhého jevu s určitou pravděpodobností

Statistická závislost je volnou závislostí v tom smyslu, že určité hodnotě jednoho znaku neodpovídá vždy stejná hodnota druhého znaku. Z tohoto pohledu nás bude nejčastěji zajímat, jak se při změně hodnot jedné veličiny mění podmíněné pravděpodobnostní rozdělení druhé veličiny, přesněji řečeno jak se při změně hodnot jedné veličiny mění podmíněné střední hodnoty druhé veličiny. V tomto případě budeme hovořit o ***korelační závislosti***. Např.: prospěch ve fyzice volně souvisí se znalostmi v matematice, …

K poznání a matematickému popisu korelačních závislostí slouží metody *regresní analýzy* a

korelační analýzy; pro potřeby těchto metod je vhodné rozlišovat ***jednostranné*** a ***vzájemné*** (***oboustranné***) závislosti. Např. jednostranné závislosti: příjmy - výdaje, rychlost auta – spotřeba benzinu, … , oboustranné závislosti: obrat - zásoby, prospěch v M a F, výkony ve sprintu a ve skoku do dálky, …

Při analýze korelační závislosti nás předně zajímá

a) její průběh, tj. tendence změn podmíněných průměrů → ***regresní analýza*** : zabývá se jednostrannými závislostmi a zkoumá *průběh = formu = tendenci* závislosti Y na X ; proti sobě stojí vysvětlující proměnná X v úloze „příčin“ a vysvětlovaná proměnná Y v úloze „důsledků“ → cílem regresní analýzy je co nejlepší přiblížení empirické (vypočítané) regresní funkce k teoretické regresní funkci; empirická regresní funkce bude za určitých podmínek odhadem teoretické regresní funkce

b) její intenzita → ***korelační analýza*** : zabývá se oboustrannými závislostmi a zkoumá *intenzitu = sílu = těsnost* vzájemného vztahu mezi X a Y

Otázka síly závislosti je souběžná s otázkou kvality regresní funkce. Z hlediska výpočtů a interpretací dochází však ke značnému prolínání obou přístupů. Speciální korelační charakteristiky informují o tom, do jaké míry mezi sebou jevy souvisí a jak spolu souvisí. Určitá síla závislosti však ještě neříká, že existuje příčinná závislost, existenci závislosti je nutné nejprve posoudit, např. vztah mezi IQ a obvodem hrudníku!

Úkoly regresní a korelační analýzy:

1. Posouzení existence závislosti, pokud není věcně zřejmá.
2. Konstrukce regresní funkce – spočívá v ověření známého modelu nebo odhadu neznámého modelu pomocí náhodného výběru (k odhadu slouží výsledky elementárního zpracování dat, tj. tabulky, grafy, charakteristiky a volba regresní funkce).
3. Posouzení kvality zvolené regresní funkce – představuje testování hypotéz o parametrech regresní funkce a o mírách těsnosti korelační závislosti.

Elementární zpracování dvourozměrného statistického souboru

Hodnoty proměnných X a Y získané na základě měření nebo zjišťování na jednotlivých statistických jednotkách vyjadřujeme ve tvaru uspořádaných dvojic $[x, y]$ → elementární popis závislosti :

a) Tabulkové vyjádření

- Při malém počtu měření lze datové dvojice $[x_i, y_i]$, kde $i = 1, 2, \ldots, n$, zapsat do *jednoduché tabulky* (je analogií neroztříděného jednorozměrného souboru, kdy každá hodnota se

vyskytuje pouze jednou):

$\boldsymbol{x}$	x_1	x_2	….	x_n
$\boldsymbol{y}$	y_1	y_2	….	y_n

- Při velkém počtu měření vyjádříme tzv. podmíněné rozdělení četností → sestrojíme dvourozměrnou tzv. ***korelační tabulku*** :

1. ***úplná*** : pro bodové nebo intervalové rozdělení četností hodnot znaků *X*, *Y* označíme

   x_i jako varianty znaku *X* pro $i = 1, 2, \ldots, k$

   y_j jako varianty znaku *Y* pro $j = 1, 2, \ldots, s$

   n_{ij} ... *sdružené četnosti*, n ... *celková četnost*,

   $n_{i\bullet}$ resp. $n_{\bullet j}$ ... *marginální* (*okrajové*) *četnosti*

   zde platí : $n_{i\bullet} = \sum_{j=1}^{s} n_{ij}$; $n_{\bullet j} = \sum_{i=1}^{k} n_{ij}$; $\sum_{i=1}^{k} n_{i\bullet} = \sum_{j=1}^{s} n_{\bullet j} = \sum_{i=1}^{k}\sum_{j=1}^{s} n_{ij} = n$

$\boldsymbol{x_i}$	$\boldsymbol{y_j}$						$\boldsymbol{n_{i\bullet}}$
	y_1	y_2	…	y_j	…	y_s	
x_1	n_{11}	n_{12}	…	n_{1j}	…	n_{1s}	$n_{1\bullet}$
x_2	n_{21}	n_{22}	…	n_{2j}	…	n_{2s}	$n_{2\bullet}$
…	…	…	…	…	…	…	…
x_i	n_{i1}	n_{i2}	…	n_{ij}	…	n_{is}	$n_{i\bullet}$
…	…	…	…	…	…	…	…
x_k	n_{k1}	n_{k2}	…	n_{kj}	…	n_{ks}	$n_{k\bullet}$
$\boldsymbol{n_{\bullet j}}$	$n_{\bullet 1}$	$n_{\bullet 2}$	…	$n_{\bullet j}$	…	$n_{\bullet s}$	$\boldsymbol{n}$

2. ***neúplná*** : pro bodové nebo intervalové rozdělení četností hodnot znaku *X* (hodnoty y_{ij} znaku *Y* nejsou roztříděné) označíme

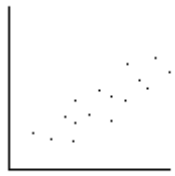
   x_i jako varianty znaku *X* pro $i = 1, 2, \ldots, k$

   n_i ... *třídní četnosti* znaku *X*, $n = \sum_{i=1}^{k} n_i$ … *celková četnost*

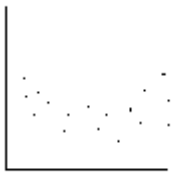
$\boldsymbol{x_i}$	$\boldsymbol{y_{ij}}$					$\boldsymbol{n_i}$
x_1	y_{11}	y_{12}	y_{13}	…	y_{1n_1}	n_1
x_2	y_{21}	y_{22}	y_{23}	…	y_{2n_2}	n_2
…	…	…	…	…	…	…
x_k	y_{k1}	y_{k2}	y_{k3}	…	y_{kn_k}	n_k
$\boldsymbol{\Sigma}$						$\boldsymbol{n}$

b) Grafické vyjádření

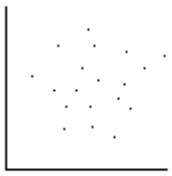
- ***bodový diagram*** - zobrazení datových dvojic $[x_i, y_i]$ z jednoduché tabulky nebo $[x_i, y_{ij}]$ z neúplné korelační tabulky $\rightarrow$ informace o formě i síle závislosti



přímková závislost

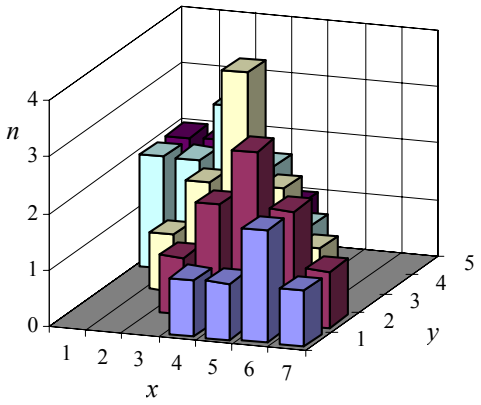


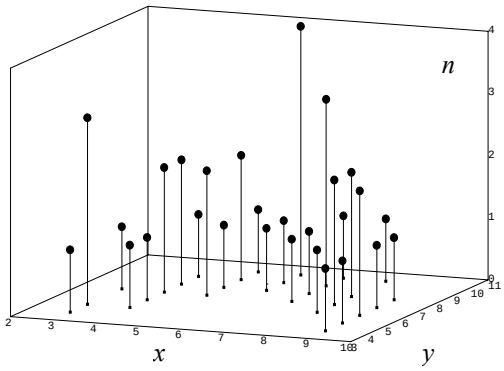
parabolická závislost



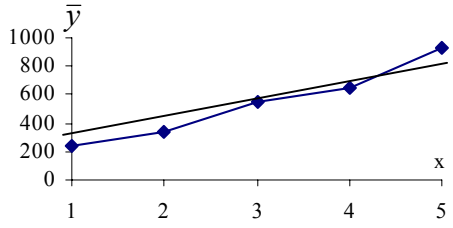
korelační nezávislost

- ***trojrozměrný (prostorový) histogram*** resp. ***trojrozměrný bodový graf*** - zobrazení dvojic $[x_i, y_j]$ z úplné korelační tabulky: osa $x \rightarrow x_i$, osa $y \rightarrow y_j$, osa $z \rightarrow n_{ij}$

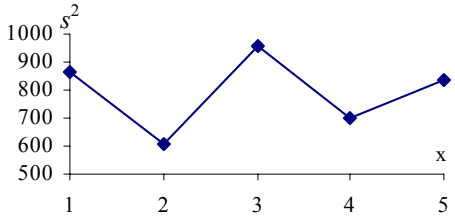




- ***graf podmíněných průměrů*** - zobrazení dvojic $[x_i , \bar{y}_i]$ v rovině $\rightarrow$ vyjadřuje tendenci změn podmíněných průměrů závisle proměnné Y při změnách hodnot nezáv. proměnné X
- ***graf podmíněných rozptylů*** - zobrazení dvojic $[x_i , s_i^2(y)]$ v rovině $\rightarrow$ vyjadřuje tendenci změn podmíněných rozptylů závisle proměnné Y při změnách hodnot nezáv. proměnné X



graf podmíněných průměrů



graf podmíněných rozptylů

Přednáška 16 – Jednofaktorová analýza rozptylu

Podmíněné charakteristiky

Každý řádek korelační tabulky (u úplné tabulky i sloupec) obsahuje rozdělení četností hodnot znaku Y (X) za podmínky, že znak X (Y) nabyl určité obměny, tj. obsahuje tzv. ***podmíněné rozdělení četností*** znaku Y (X), které lze popsat pomocí ***podmíněných charakteristik***:

- Podobně jako u jednorozměrného rozdělení četností počítáme i z korelační tabulky nejdůležitější charakteristiky :

  $\bar{y}_i$ … podmíněné průměry, $s_i^2(y)$ … podmíněné rozptyly ,

  $\bar{y}$ … celkový průměr, $s^2(y)$ … celkový rozptyl ,

  $s^2(\bar{y}_i)$ … rozptyl podmíněných průměrů, $\overline{s_i^2(y)}$ … průměr podmíněných rozptylů

- Pro celkový rozptyl platí důležitá vlastnost → ***rozklad rozptylu*** :

$$s_n^2(y) = s_n^2(\bar{y}_i) + \overline{s_{n,i}^2(y)} \quad \rightarrow \quad S_c(y) = S_m(y) + S_v(y)$$

  kde $S_m(y)$ představuje *meziskupinovou* variabilitu znaku Y (vnější variabilitu)

  $S_v(y)$ představuje *vnitroskupinovou* variabilitu znaku Y (vnitřní variabilitu)

  Praktické určení veličin $S_m(y)$, $S_v(y)$ *a* $S_c(y)$

  ~ meziskupinová variabilita vychází z rozptylu podmíněných průměrů:

$$s^2(\bar{y}_i) = \frac{1}{n}\sum_{i=1}^{k}(\bar{y}_i - \bar{y})^2 = \frac{1}{n}S_m(y) \quad \Rightarrow \quad S_m(y) = n \cdot s^2(\bar{y}_i)$$ s k–1 stupni volnosti

  ~ vnitroskupinová variabilita vychází z podmíněných rozptylů :

$$s_i^2(y) = \frac{1}{n_i}\sum_{i=1}^{k}(y_{ij} - \bar{y}_i)^2 = \frac{1}{n_i}S_i(y) \quad \Rightarrow \quad S_i(y) = n_i \cdot s_i^2(y)$$

  potom $S_v(y) = \sum_{i=1}^{k} S_i(y)$ s n–k stupni volnosti

  ~ celková variabilita vychází ze všech hodnot:

$$s^2(y) = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{s}(y_{ij} - \bar{y})^2 = \frac{1}{n}S_c(y) \quad \Rightarrow \quad S_c(y) = n \cdot s^2(y)$$ s n–1 stupni volnosti

- Potom charakteristika $p_{yx}^2 = \dfrac{s^2(\bar{y})}{s^2(y)} = \dfrac{S_m(y)}{S_c(y)}$ je tzv. ***poměr determinace*** ; $p_{yx}^2 \in \langle 0,1 \rangle$

  ~ udává, jaké % rozptylu závisle proměnné Y lze vysvětlit vlivem nezávisle proměnné X, doplněk do 100 % udává vliv blíže nespecifikovaných činitelů

  ~ závislost znaku Y na X se považuje za tím silnější, čím více se p_{yx}^2 blíží k 1 a naopak

~ $p_{yx} = \sqrt{p_{yx}^2}$ … *korelační poměr*

Analýza rozptylu

Cílem analýzy rozptylu je rozhodnout, zda pozorovaná data hovoří ve prospěch hypotézy o nezávislosti, či zda lze s vysokou spolehlivostí tvrdit, že znak *X* ovlivňuje znak *Y*. Základem analýzy rozptylu je rozklad celkové variability $S_c(y) = S_m(y) + S_v(y)$. Základním předpokladem analýzy rozptylu je, že každý z *k* nezávislých výběrů (v korelačních tabulkách odpovídají nezávislé výběry řádkům) pochází z normálního rozdělení s konstantním rozptylem.

Nejjednodušší formou analýzy rozptylu je tzv. *jednofaktorová analýza rozptylu*, ve které se předpokládá, že obměny faktoru *X* jsou dané a jeho změny mohou vést pouze ke změně střední hodnoty, ale ne ke změně rozptylu normálního rozdělení, z něhož výběr hodnot znaku *Y* pochází. Pokud je tedy oprávněné předpokládat shodu podmíněných středních hodnot znaku *Y*, vyjádříme to v podobě hypotézy H; ve prospěch tvrzení o vlivu znaku X na znak *Y* vypovídá alternativa A, pro jejíž přijetí stačí neshoda dvou různých středních hodnot:

$$\text{H: } \mu_1 = \mu_2 = \ldots = \mu_k \quad \rightarrow \quad \text{A: } \mu_i \neq \mu_{i'}$$

Testové kritérium je konstruované tak, aby ve prospěch alternativy hovořily vysoké hodnoty $S_m(y)$ na úkor $S_v(y)$, tj. $F = \dfrac{\dfrac{S_m(y)}{k-1}}{\dfrac{S_v(y)}{n-k}} = \dfrac{(n-k)\cdot S_m(y)}{(k-1)\cdot S_v(y)}$.

Toto testové kritérium má při platnosti H rozdělení $F(k-1,\ n-k)$, proto hypotézu H zamítneme na hladině α, když $F \geq F_{1-\alpha}(k-1,\ n-k)$.

Přednáška 17 – Regresní analýza

Obecný regresní model

V analýze rozptylu nám šlo o zjištění, zda mezi proměnnými *X* a *Y* existuje nějaká závislost, a případně jakou těsnost tato závislost vykazuje. Cílem regresní analýzy je stanovení formy (trendu, tvaru, průběhu) této závislosti pomocí vhodné funkce $\rightarrow$ hovoříme o stanovení vhodného teoretického regresního modelu, resp. jeho odhadu pomocí empirického regresního modelu.

Obecně lze závislost proměnné *Y* na *X* vyjádřit jako závislost

~ funkční … $y = f(x)$,

~ stochastickou … $y = f(X) + \varepsilon = Y + \varepsilon$, kde

$Y = f(X)$ je tzv. *regresní funkce* (teoret. regresní model), ε … *náhodná složka*.

Volba tvaru regresní funkce

- se zpravidla opírá o odhad z grafů (graf podmíněných průměrů a bodový diagram),
- může však vycházet i z věcné povahy závislosti.

Regresní model = regresní funkce … je definována jako podmíněná střední hodnota

$$Y = E(Y/X) = f(X, \beta_0, \beta_1, \dots, \beta_p), \text{ kde } \beta_j \dots \textit{parametry}, \; j = 0, 1, 2, \dots, p.$$

Velmi často se uvažují regresní funkce, jejichž rovnice se obecně vyjadřují ve tvaru

$Y = \beta_0 f_0(X) + \beta_1 f_1(X) + \dots + \beta_p f_p(X)$ … regresní funkce *lineární vzhledem k parametrům*

kde $f_j(X)$ … *regresory*.

Pokud regresor $f_0(X) = 1$, potom (viz dále lineární modely)

$Y = \beta_0 + \beta_1 f_1(X) + \dots + \beta_p f_p(X)$.

$Y_i = \beta_0 + \beta_1 f_1(x_i) + \dots + \beta_p f_p(x_i)$ … je potom hodnota teoretické regresní funkce pro i-té měření $\rightarrow$ v důsledku působení mnoha vlivů na proměnnou Y se budou teoretické hodnoty Y_i a empirické hodnoty y_i lišit !!! Obecně tedy platí

$$y_i = Y_i + \varepsilon_i ,$$

kde ε_i je náhodná složka pro i-té měření, pro kterou platí $\varepsilon_i \sim N(0, \sigma^2)$.

Odhad regresního modelu = odhad regresní funkce

$\hat{y} = f(x, b_0, b_1, \dots, b_p)$ … je *výběrová* (empirická) *regresní funkce*,

$b_j = \hat{\beta}_j$ … odhady regresních parametrů .

Je zřejmé, že vztah mezi hodnotami náhodné veličiny Y a výběrovou regresní funkcí lze potom vyjádřit ve tvaru

$$y = \hat{y} + e , \text{ kde } e = y - \hat{y} \text{ je tzv. } \textbf{\textit{reziduum}}$$

resp. ve tvaru

$$y_i = \hat{y}_i + e_i , \text{ kde } e_i = y_i - \hat{y}_i \text{ je reziduum pro i-té měření}$$

a $\hat{y}_i$ je tzv. *vyrovnaná hodnota* pro i-té měření.

Odhad lineární regresní funkce potom vyjádříme ve tvaru

$$\hat{y} = b_0 + b_1 f_1(x) + \dots + b_p f_p(x) \text{ resp. } \hat{y}_i = b_0 + b_1 f_1(x_i) + \dots + b_p f_p(x_i).$$

Regresní funkce, které jsou lineární vzhledem k parametrům, se nazývají *lineární regresní funkce*. Jsou to:

~ přímková regrese $Y = \beta_0 + \beta_1 X \quad \rightarrow \quad \hat{y} = b_0 + b_1 x$

~ hyperbolická regrese $Y = \beta_0 + \beta_1 \frac{1}{X} \quad \rightarrow \quad \hat{y} = b_0 + b_1 \frac{1}{x}$

~ logaritmická regrese $Y = \beta_0 + \beta_1 \ln X \rightarrow \quad \hat{y} = b_0 + b_1 \ln x$

~ parabolická regrese $Y = \beta_0 + \beta_1 X + \beta_2 X^2 \rightarrow \hat{y} = b_0 + b_1 x + b_2 x^2$

V praxi se používají i funkce, které nejsou lineární vzhledem k parametrům. Jsou to např.

~ exponenciální regrese $Y = \beta_0 \beta_1^x \quad \rightarrow \quad \hat{y} = b_0 b_1^x$

~ mocninná regrese $Y = \beta_0 x^{\beta_1} \quad \rightarrow \quad \hat{y} = b_0 x^{b_1}$

~ a řada dalších jako např. $Y = \beta_0 \beta_1^x + \beta_2$, …

Tyto funkce lze někdy vhodnou transformací převést na funkce lineární vzhledem k parametrům (např. u uvedené exponenciální a mocninné regrese).

V této kapitole nás budou zajímat jen lineární regresní modely. Nejjednodušší z lineárních regresních modelů je tzv. *klasický regresní model*. Ten předpokládá:

1. hodnoty vysvětlující proměnné X se volí (X není náhodná veličina),
2. regresní funkce je lineární vzhledem k parametrům,
3. matice hodnot regresorů $f_j(x)$, $j = 0, 1, \ldots, p$, má hodnost $p+1$,
4. náhodné složky ε_i jsou nezávislé a mají normální rozdělení $N(0, \sigma^2)$.

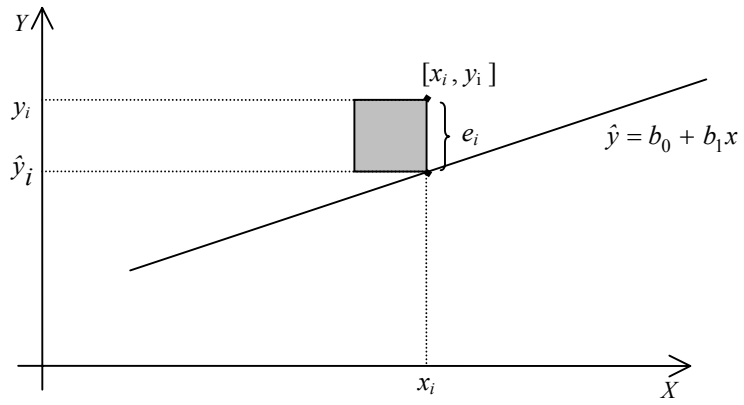
Z předpokladu o rozdělení náhodných složek ε_i vyplývá, že v klasickém modelu mají pozorované hodnoty y_i vysvětlované proměnné Y normální rozdělení

- se středními hodnotami $E(y_i \mid x_i) = \mu_i$,
- s rozptylem $D^2(y_i \mid x_i) = D^2(\varepsilon_i) = \sigma^2$ a
- hodnoty y_i jsou vzájemně nezávislé.

Lineární regresní modely

Odhady regresních parametrů klasickou metodou nejmenších čtverců (MNČ) → klasická MNČ vychází z požadavku, aby součet čtverců odchylek empirických hodnot y_i a vyrovnaných hodnot $\hat{y}_i$, tzv. *reziduální součet čtverců* S_R , byl minimální, tj.

$$S_R = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2 = \min.$$



K určení odhadů regresních koeficientů potom slouží *soustava normálních rovnic*.

Např. pro regresní přímku $\hat{y} = b_0 + b_1 x$ dostaneme:

$$S_R = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2 = \text{ min., kde } b_0 \text{ a } b_1 \text{ jsou proměnné, potom}$$

$$\frac{\partial S_R}{\partial b_0} = 2\sum_{i=1}^{n} (y_i - b_0 - b_1 x_i) \cdot (-1) = 0 ,$$

$$\frac{\partial S_R}{\partial b_1} = 2\sum_{i=1}^{n} (y_i - b_0 - b_1 x_i) \cdot (-x_i) = 0 ,$$

a odtud dostaneme soustavu normálních rovnic

$$b_0 n \quad + \quad b_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i ,$$

$$b_0 \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i .$$

Řešením této soustavy dostaneme odhady $b_0 = \hat{\beta}_0$ *a* $b_1 = \hat{\beta}_1$ ve tvaru

$$b_0 = \frac{\sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} \quad \text{a} \quad b_1 = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} .$$

Pro ostatní lineární regresní modely se postupuje analogicky $\rightarrow$ soustava normálních rovnic a vztahy pro určení b_0 a b_1 viz přehled vzorců.

Reziduální rozptyl

Pro posouzení vhodnosti několika regresních modelů se použije rozptyl náhodné složky ε_i, tj. $D^2(\varepsilon_i) = \sigma_R^2 \rightarrow$ za vhodnější model se považuje model s menším rozptylem σ_R^2.

Nestranným odhadem rozptylu σ_R^2 je *reziduální rozptyl*, tj. $\hat{\sigma}_R^2 = s_R^2$, kde

$s_R^2 = \frac{S_R}{n-c}$, tj. $s_R^2 = \frac{1}{n-c}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$, kde c je počet neznámých regresních parametrů.

Např. pro regresní přímku $\hat{y} = b_0 + b_1 x$ dostaneme:

$$S_R = \sum_{i=1}^{n}(y_i - b_0 - b_1 x_i)^2 = \ \ldots \ = \sum_{i=1}^{n} y_i^2 - b_0\sum_{i=1}^{n} y_i - b_1\sum_{i=1}^{n} x_i y_i \ , \text{potom}$$

$$s_R^2 = \frac{1}{n-2}\left[\sum_{i=1}^{n} y_i^2 - b_0\sum_{i=1}^{n} y_i - b_1\sum_{i=1}^{n} x_i y_i\right].$$

Pro ostatní lineární regresní modely se postupuje analogicky $\rightarrow$ viz přehled vzorců.

Těsnost závislosti

Vztah mezi proměnnými X a Y může mít různou intenzitu, od úplné nezávislosti až po pevnou (funkční) závislost. Těsností závislosti se rozumí stupeň, s jakým se zkoumaná závislost blíží k funkční závislosti.

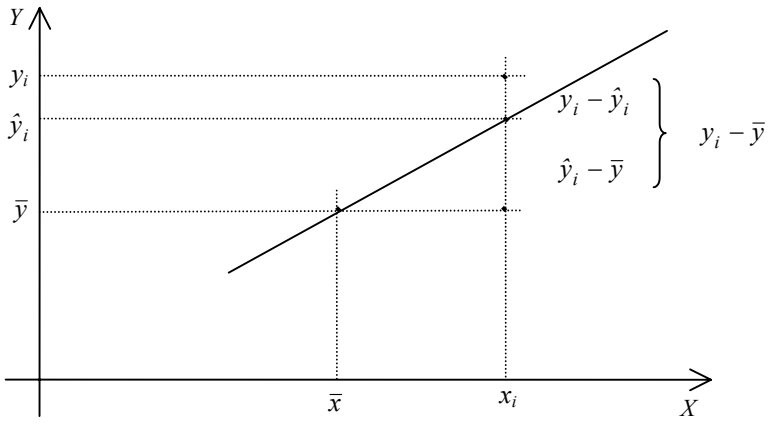
Představu o síle závislosti můžeme získat

a) z bodového diagramu (podle rozložení bodů okolo regresní křivky),
b) pomocí měr těsnosti závislosti $\rightarrow$ poměr determinace p^2 (viz podmíněné charakteristiky), index determinace i^2 a koeficient determinace r^2 (viz dále).

Východiskem pro konstrukci indexu determinace i^2 a koeficientu determinace r^2 je podíl variability vyrovnaných hodnot $\hat{y}_i$ okolo průměru na celkové variabilitě proměnné Y.

1) <u>index determinace i^2</u>

Vychází ze známé regresní funkce, tedy udává, jaké % celkové variability lze vysvětlit zvoleným regresním modelem.



Protože $y_i - \bar{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})$, potom platí (bez důkazu)

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \quad \rightarrow \quad S_C(y) = S_R(y) + S_T(y), \text{ kde}$$

$S_C(y) = n \cdot s^2(y)$ … celkový součet čtverců, $s^2(y) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$,

$S_R(y) = (n-c) \cdot s_R^2(y)$ ... reziduální součet čtverců, $s_R^2(y) = \frac{1}{n-c}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$,

$S_T(y) = n \cdot s^2(\hat{y})$ ... teoretický součet čtverců, $s^2(\hat{y}) = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$.

Index determinace je potom definovaný jako poměr rozptylu vyrovnaných hodnot a celkového rozptylu: $i_{yx}^2 = \frac{s^2(\hat{y})}{s^2(y)} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$, tj. $i_{yx}^2 = \frac{S_T(y)}{S_C(y)}$.

Protože $s^2(y) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{1}{n}\sum_{i=1}^{n}y_i^2 - \bar{y}^2$, lze obecně vyjádřit celkový součet čtverců ve tvaru $S_C(y) = \sum_{i=1}^{n}y_i^2 - \frac{1}{n}(\sum_{i=1}^{n}y_i)^2$. Teoretický součet čtverců $S_T(y)$ určíme pro každou regresní křivku.

Např. pro regresní hyperbolu $\hat{y} = b_0 + b_1 \frac{1}{x}$ dostaneme

$$S_T(y) = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = \sum_{i=1}^{n}(b_0 + b_1\frac{1}{x_i} - \frac{1}{n}\sum_{i=1}^{n}y_i)^2 = b_0\sum_{i=1}^{n}y_i + b_1\sum_{i=1}^{n}\frac{y_i}{x_i} - \frac{1}{n}(\sum_{i=1}^{n}y_i)^2 .$$

Pro ostatní regresní modely se postupuje analogicky $\rightarrow$ viz přehled vzorců.

Poznámky k indexu determinace:

~ Teoretický součet čtverců $S_T(y)$ je ta část celkového součtu čtverců $S_C(y)$, která je vysvětlená použitou regresní funkcí; naopak $S_R(y)$ vysvětlená regresní funkcí není $\rightarrow$ i^2 tedy udává, jaké % celkové variability lze vysvětlit použitým regresním modulem $\rightarrow$ $i^2 \in \langle 0, 1 \rangle$

~ Čím více se i^2 blíží k 1, tím považujeme danou závislost za silnější, a tedy dobře vystiženou použitou regresní funkcí; naopak čím více se bude blížit k 0, tím považujeme danou závislost za slabší a regresní funkci za méně výstižnou.

~ Velikost i^2 je zcela ovlivněna tím, zda se podařilo nalézt vhodný typ regresní funkce pro

popis sledované závislosti → nízká hodnota i^2 ještě nemusí znamenat nízký stupeň závislosti mezi proměnnými, ale může to signalizovat chybnou volbu regresní funkce.

~ Kritéria vhodnosti použité regresní funkce pro popis závislosti:

1. kritérium: čím je i^2 blíže k 1, tím vhodnější je použitý model,
2. kritérium: obecně platí $i^2 \leq p^2$, potom čím je diference $p^2 - i^2$ menší, tím je použitý model vhodnější.

~ i_{yx}^2 představuje výběrový index determinace, který lze použít jako odhad teoretického indexu determinace I_{yx}^2 v základním souboru, tj. $\hat{I}_{yx}^2 = i_{yx}^2$ → tento odhad je asymptoticky nestranný, avšak

- pro malé výběry nadhodnocuje skutečnou těsnost závislosti,
- záleží i na počtu parametrů regresní funkce.

Proto provádíme korekci:

$i_{kor}^2 = 1 - (1 - i^2)\,\frac{n-1}{n-c}$ → $\hat{I}_{yx}^2 = i_{kor}^2$ splňuje podmínku nestrannosti.

2) <u>koeficient determinace r^2</u>

Zvláštním případem indexu determinace pro závislost popsanou regresní přímkou je koeficient determinace $r_{yx}^2 = \frac{s^2(\hat{y})}{s^2(y)} = \frac{S_T(y)}{S_C(y)}$.

Tato míra těsnosti závislosti má obdobné vlastnosti jako i^2 → pokud se r^2 blíží k 1, tím o silnější lineární závislosti je možné hovořit; pokud se ale r^2 blíží k 0, nemusí to nutně znamenat slabou závislost, protože korelované proměnné mohou být nelineárně silně závislé. Další vlastnosti r^2: $r_{yx}^2 = r_{xy}^2$ … $\in \langle 0,1 \rangle$, $r = \sqrt{r^2} \in \langle -1,1 \rangle$ … korelační koeficient, kde

$$r = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}\sqrt{n\sum_{i=1}^{n} y_i^2 - (\sum_{i=1}^{n} y_i)^2}},$$

pro $r > 0$ … přímá závislost resp. pro $r < 0$ … nepřímá závislost

- r_{yx}^2 představuje výběrový koeficient determinace → obdobně jako u indexu determinace lze provést korekci:

  $r_{kor}^2 = 1 - (1 - r^2)\,\frac{n-1}{n-2}$ → $\hat{\rho}^2 = r_{kor}^2$ je nestranným odhadem ρ^2 (ρ je korelační koeficient v základním souboru)

- Test o významnosti korelačního koeficientu:

  $H : \rho = 0 \rightarrow A : \rho \neq 0$

  $t = \frac{r}{\sqrt{1-r^2}}\sqrt{n-2} \quad \sim \quad t(n-2)$

  $W_\alpha : |t| > t_{1-\alpha/2}(n-2)$

Pozn.: Jestli-že náhodné veličiny X a Y mají dvourozměrné normální rozdělení, potom korelační koeficient ρ vyjadřuje míru lineární závislosti veličin X a Y. Lze ukázat, že nezávislé veličiny jsou nekorelované, ale obráceně toto tvrzení neplatí! [11]

Odhady v lineární regresi

Intervalový odhad parametrů

Odhady $b_0, b_1, \ldots, b_p$ získané metodou nejmenších čtverců jsou nestrannými odhady regresních parametrů $\beta_0, \beta_1, \ldots, \beta_p$. Platí tedy $E(b_j) = \beta_j$. Představu o velikosti chyby, kterou lze u bodových odhadů parametrů očekávat, získáme pomocí *směrodatné chyby odhadů* b_j.

~ Při platnosti podmínek klasického regresního modelu lze vyjádřit odhad směrodatné chyby odhadů b_0, b_1 (pouze pro přímku, pro ostatní modely viz přehled vzorců) ve tvaru

$$s(b_0) = \sqrt{s_R^2 \cdot \frac{\sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}} \quad \text{a} \quad s(b_1) = \sqrt{s_R^2 \cdot \frac{n}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}}.$$

~ Určení intervalů spolehlivosti pro regresní parametry $\beta_0, \beta_1, \ldots, \beta_p$ je založeno na tom, že při platnosti podmínek klasického regresního modelu mají statistiky $t_j = \frac{b_j - \beta_j}{s(b_j)}$ Studentovo rozdělení $t(n-c)$, kde $c = p + 1$ (počet odhadovaných parametrů v modelu):

$$b_j - t_{1-\alpha/2}(\nu) \cdot s(b_j) < \beta_j < b_j + t_{1-\alpha/2}(\nu) \cdot s(b_j), \quad \text{kde } \nu = n - c.$$

Testy hypotéz v lineární regresi

Za stejných předpokladů jako v předchozí části je možné ověřovat statistickou významnost jednotlivých regresních parametrů i celého modelu jako celku.

~ Individuální testy o významnosti parametrů $\beta_0, \beta_1, \ldots, \beta_p$:

  $H: \beta_j = 0 \rightarrow A: \beta_j \neq 0$

$$t = \frac{b_j - \beta_j}{s(b_j)} = \frac{b_j - 0}{s(b_j)} \sim t(n-c)$$

$W_\alpha : |t| > t_{1-\alpha/2}(n-c)$

~ Test o významnosti modelu = celkový F-test:

$H : \beta_0 = k, \beta_1 = \beta_2 = \ldots = \beta_p = 0 \rightarrow A : \beta_j \neq 0$ pro $j = 1, 2, \ldots, p$ a $k \neq 0$

$$F = \frac{\frac{S_T(y)}{c-1}}{\frac{S_R(y)}{n-c}} \sim F(c-1,\ n-c)$$, kde $c = p+1$ je počet parametrů

$W_\alpha : F > F_{1-\alpha}(c-1, n-c)$

Výsledky testů o jednotlivých parametrech modelu a celkového testu vedou k praktickému rozhodování o tom, zda je použitý model pro popis závislosti přijatelný, případně který z regresorů je vhodné z modelu vypustit.

Přednáška 18 - Nelineární regresní analýza

Regresní analýza užívá řadu dalších funkcí, které nejsou lineární vzhledem k parametrům → nelineární regresní modely → lze je rozdělit do 2 tříd:

1. nelineární modely, které lze linearizovat, např.
   - ~ regresní exponenciální funkce $Y = \beta_0 \beta_1^X$; $Y = \beta_0 e^{\beta_1 X}$
   - ~ regresní mocninná funkce $Y = \beta_0 X^{\beta_1}$
   - ~ Törnquistova křivka I $Y = \frac{\beta_0 X}{\beta_1 + X}$
2. nelineární modely, které nelze linearizovat, např.
   - ~ regresní exponenciální funkce $Y = \beta_0 \beta_1^X + \beta_2$; $Y = \beta_0 e^{\beta_1 X} + \beta_2$
   - ~ regresní mocninná funkce $Y = \beta_0 X^{\beta_1} + \beta_2$
   - ~ Törnquistovy křivky II a III $Y = \frac{\beta_0 (X - \beta_1)}{\beta_2 + X}$; $Y = \frac{\beta_0 X (X - \beta_1)}{\beta_2 + X}$

   pozn.: Törnquistovy křivky vyjadřují závislost poptávky po určitém zboží v závislosti na příjmu.

Odhad parametrů těchto a dalších nelineárních regresních funkcí se neprovádí metodou nejmenších čtverců. Postupuje se tak, že se nejprve najde vhodný tzv. *počáteční odhad*, který se dále numerickými (iteračními) metodami postupně zlepšuje:

metody počátečních odhadů → linearizující transformace,

→ metoda vybraných bodů.

Linearizující transformace

Spočívá v tom, že se vhodnou transformací převede nelineární funkce Y na lineární funkci Y^* → bodovým odhadem transformované funkce je výběrová regresní funkce $\hat{y}^*$, jejíž parametry se určí metodou nejmenších čtverců a pomocí nich se potom odhadnou parametry původní funkce $\hat{y}$:

př.1: $Y = \beta_0\, \beta_1^X \quad \rightarrow \quad \hat{y} = b_0\, b_1^x$

transformace $\ln \hat{y} = \ln b_0 + x \cdot \ln b_1$

lineární model $y^* = b_0^* + b_1^* \cdot x^*$

užité substituce $y^* = \ln \hat{y}$; $x^* = x$, potom

$b_0^* = \ln b_0 \quad \Rightarrow \quad b_0 = \exp(b_0^*)$ a $b_1^* = \ln b_1 \quad \Rightarrow \quad b_1 = \exp(b_1^*)$

př.2: $Y = \dfrac{\beta_0 X}{\beta_1 + X} \quad \rightarrow \quad \hat{y} = \dfrac{b_0 x}{b_1 + x}$

transformace $\dfrac{1}{\hat{y}} = \dfrac{b_1 + x}{b_0 x} = \dfrac{b_1}{b_0} \cdot \dfrac{1}{x} + \dfrac{1}{b_0}$

lineární model $y^* = b_0^* + b_1^* \cdot x^*$

užité substituce $y^* = \dfrac{1}{\hat{y}}$; $x^* = \dfrac{1}{x}$, potom

$$b_0^* = \frac{1}{b_0} \quad \Rightarrow \quad b_0 = \frac{1}{b_0^*} \quad \text{a} \quad b_1^* = \frac{b_1}{b_0} \quad \Rightarrow \quad b_1 = b_0 \cdot b_1^* \ .$$

Metoda vybraných bodů

Tato metoda spočívá v tom, že z řady empirických hodnot vybereme malý počet bodů, ve kterých položíme teoretickou hodnotu příslušné regresní funkce rovnu empirické hodnotě. Vybíráme tolik bodů, kolik parametrů má daná regresní funkce. Získáme tak soustavu nelineárních rovnic, jejíž řešením jsou odhadované parametry.

Při volbě vhodných bodů vycházíme z bodového diagramu. Pokud vybíráme dva body, vybereme jeden z oblasti nízkých a druhý z oblasti vysokých hodnot proměnné X. Pokud potřebujeme vybrat tři body, je vhodné vzít po jednom z oblasti nízkých, středních a vysokých hodnot proměnné X a analogicky postupujeme při výběru většího počtu bodů.