Statistické metody a zpracování dat

VI. Korelační a regresní počet

Petr Dobrovolný

K čemu to je dobré?

Analýza závislostí

• V řadě geografických disciplín studujeme jevy, u kterých vyšetřujeme ne jednu jejich vlastnost (znak), ale znaků několik.

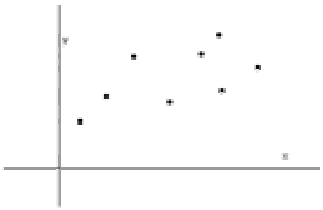
• Tyto znaky mohu být navzájem závislé.

• Cílem této části statistiky je vyšetřovat, do jaké míry spolu dva či více statistických znaků souvisí.

• Do jaké míry změna hodnoty jednoho znaku podmiňuje změnu hodnot znaku jiného.

Příklady použití

Př. *Vztah mezi teplotou vzduchu a nadmořskou výškou, mezi množstvím srážek a velikostí odtoku, mezi výnosy a hodnotami několika meteorologických prvků, mezi počtem dojíždějících a vzdáleností od centra dojížďky, …*



Analýza závislostí

• Předmětem statistické analýzy v tomto případě bude stanovení **síly závislosti** a **druhu závislosti**

• Analýzou síly závislosti statistických znaků se zabývá **korelační počet**

• Analýzou druhu závislosti statistických znaků se zabývá **regresní počet**

• Budeme tedy pracovat s dvourozměrnými soubory

• **Korelační i regresní počet** však lze využít i pro studium vícerozměrných souborů, pro studium znaků kvantitativních i kvalitativních.

Druhy závislostí

• **Vztahy jednostranné:** Změna statistického znaku jednoho souboru náhodné veličiny - tzv. **nezávisle** proměnné (x) podmiňuje změnu statistického znaku souboru druhé náhodné veličiny - tzv. **závisle** proměnné (y).

• V tomto případě jde o vztahy příčiny a následku

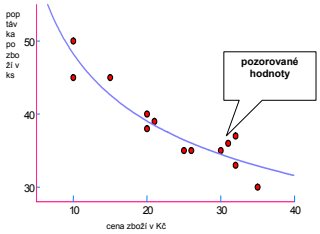
• **Vztahy vzájemné**: Nelze rozlišit mezi souborem závisle a nezávisle proměnné (např. vztah hodnot teploty vzduchu na dvou sousedních stanicích)

Vztahy závislosti podle stupně závislosti statistických znaků

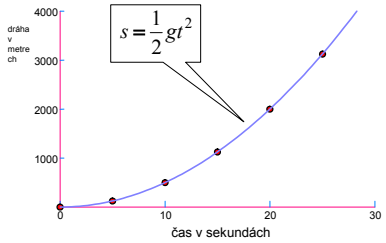
• Závislost funkční

• Závislost statistická

• Závislost korelační

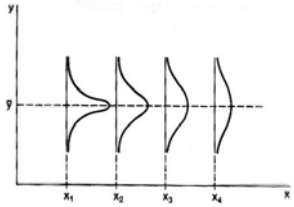


Závislost funkční



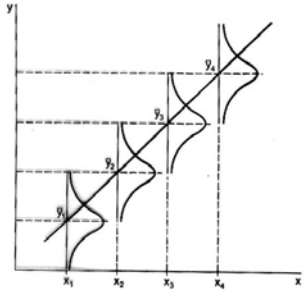
Každé hodnotě znaku nezávisle proměnné náhodné veličiny *x* odpovídá vždy pouze jediná určitá hodnota závisle proměnné veličiny *y*

Závislost statistická



- Každé hodnotě znaku nezávisle proměnné náhodné veličiny *x* odpovídá více hodnot závisle proměnné veličiny *y*,
- Hodnoty *y* mají své rozdělení
- Při změně znaku nezávisle proměnné *x* mění podmíněná rozdělení relativních četností závisle proměnné *y*

Závislost korelační



Se změnou hodnoty znaku nezávisle proměnné *x* se mění podmíněná rozdělení relativních četností hodnoty znaku závisle proměnné *y* tak, že změna *x* podmiňuje změnu průměru $\bar{y}$ souborů hodnot *y*, odpovídajících daným hodnotám *x*.

Určení těsnosti korelační závislosti

- Úkolem korelačního počtu je vyjádřit tendenci změn hodnoty znaku závisle proměnné při změně hodnoty znaku nezávisle proměnné **matematickou funkcí**
- Tato funkce představuje tzv. **regresní čáru** a vyjadřuje, jaká hodnota znaku závisle proměnné odpovídá s největší pravděpodobností určité hodnotě znaku nezávisle proměnné.
- Odhad regresní závislosti je tím přesnější, čím větší je **těsnost korelační závislosti**.
- Určení těsnosti korelační závislosti je prvním krokem analýzy.

Charakteristiky korelační závislosti

Máme dva výběrové soubory náhodných veličin *X*, *Y*. Proměnlivost hodnot znaku obou výběrů můžeme vyjádřit odchylkami d_{xi} a d_{yi} prvků od jejich průměrů:

$$d_{xi} = x_i - \bar{x} \qquad d_{yi} = y_i - \bar{y}$$

Vzájemnou proměnlivost obou výběrových souborů charakterizuje součin odchylek :

$$(x_i - \bar{x}) \cdot (y_i - \bar{y})$$

Suma součinů odchylek vydělaná rozsahem výběrů n určuje tzv. **kovarianci** výběrových souborů s_{xy} – tedy první společnou charakteristiku proměnlivosti obou souborů:

$$s_{xy} = \frac{\sum (x_i - \bar{x}) \cdot (y_i - \bar{y})}{n}$$

Charakteristiky korelační závislosti

- Kovariance je obdobou rozptylu
- Omezenost - je mírou **absolutní** – nelze jí použít k porovnání těsnosti vztahu dvou či více dvojic výběrových souborů.

Relativní míra – kovariance dělená součinem směrodatných odchylek s_x a s_y obou výběrů - **korelační koeficient** r_{xy}:

$$r_{xy} = \frac{s_{xy}}{s_x \cdot s_y} = \frac{\frac{1}{n}\sum (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\frac{1}{n}\sum (x_i - \bar{x}) \cdot \frac{1}{n}\sum (y_i - \bar{y})}}$$

Podmínky použitelnosti r_{xy}

Výpočet r_{xy} se opírá o rozptyl a směrodatnou odchylku

Jeho použití tedy předpokládá splnění tří následujících podmínek:

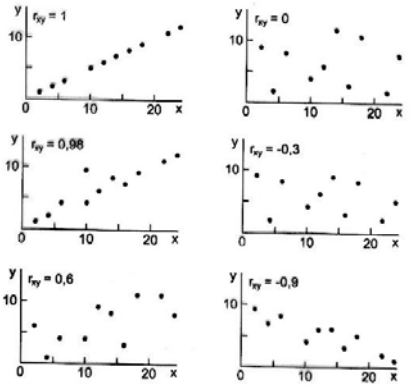
- normální rozdělení použitých výběrů
- dvojrozměrnost normálního rozdělení (každé hodnotě znaku veličiny x odpovídá soubor hodnot znaku y, který má normální rozdělení a naopak
- linearita vztahu hodnot x a y (regresní čára je přímka)

Hodnota r_{xy} nás informuje o druhu a těsnosti závislosti

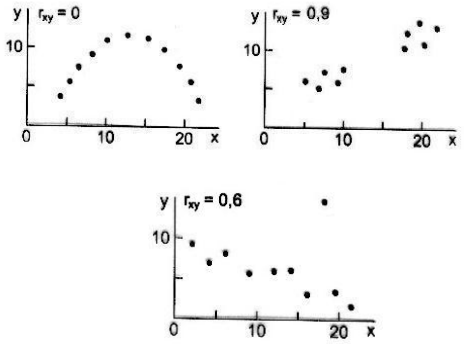
Dokonalá korelační závislost přímá $r_{xy} = 1$

Dokonalá korelační závislost nepřímá $r_{xy} = -1$

Graf korelačního pole pro různá r_{xy}



Graf korelačního pole pro různá r_{xy} ???



Koeficient pořadové korelace (Spearmanův) (r_s)

Používá se k určení závislosti **kvalitativních znaků.**

Každé hodnotě x_i a y_i přiřadíme pořadové číslo px_i a py_i podle velikosti hodnot x_i a y_i.

Určíme rozdíly D_i dvojic pořadových čísel odpovídajících si hodnot.

$$r_s = 1 - \frac{6\sum D_i^2}{n \cdot (n^2 - 1)}$$

Koeficient pořadové korelace - příklad

Příklad: Kvantifikujte vztah mezi dobou, po kterou jsou pole ponechána ladem a počtem rostlinných druhů (na m²).

Zjištěná data		Pořadová čísla		Diference	
Počet roků	Počet druhů	Počet roků	Počet druhů	D	D²
1	2	1	1	0	0
2	3	2	2	0	0
3	5	3	4	-1	1
4	4	4	3	-1	1
8	7	5	6,5	-1,5	2,25
10	6	6	5	1	1
> 10	7	7	6,5	0,5	0,25
					$\sum D_i^2 = 5,5$

$$r_s = 1 - \frac{6\sum D_i^2}{n \cdot (n^2 - 1)} = 1 - \frac{6 \times 5,5}{7 \times (49 - 1)} = 0,902$$

V tabulkách vyhledáme pro n=7 a p=0,05 kritickou hodnotu:

$r_{krit}=0,786$

Závěr: Existuje statisticky významný vztah mezi dobou, po kterou jsou pole ponechána ladem a počtem rostlinných druhů, které se na nich vyskytují.

Koeficient determinace

- Koeficient korelace se často ve výpočtech doplňuje hodnotou koeficientu determinace (r^2_{xy}).
- Jeho hodnota kolísá v intervalu 0 až 1
- Vynásoben 100 udává v procentech tu část rozptylu závisle proměnné *y*, která je vysvětlena (podmíněna) změnami hodnot nezávisle proměnné *x*.

Hodnocení významnosti koeficientu korelace

Významnost r_{xy} závisí na povaze řešeného problému

Jeho hodnota je mírou relativní a posouzení těsnosti je do značné míry subjektivní.

Významnost r_{xy} lze též zjistit objektivně – testováním:

Ze dvou základních jednorozměrných souborů lze provést sérii dvojic výběrů, které mají koeficienty korelace r_{xy}.

Soubor těchto výběrových koeficientů korelace má při velkých výběrech a při hodnotě korelačního koeficientu základního souboru (ρ) blízké nule tzv. normální rozdělení.

Jeho průměr je $\bar{r}_{xy} = \rho$ a směrodatná odchylka s_r se vypočte podle vztahu:

$$s_r = \frac{1-\rho^2}{\sqrt{n-1}}$$

Hodnocení významnosti koeficientu korelace

Při testování r_{xy} vycházíme z nulové hypotézy, která je ρ = 0 (tedy mezi dvěma základními soubory nepředpokládáme žádný korelační vztah). Testovací kritérium se vypočte podle vztahu:

$$t = \frac{r_{xy}}{\sqrt{1-r_{xy}^{\ 2}}} \cdot \sqrt{n-2}$$

Přísluší mu *t*-rozdělení s ν = n - 2 stupni volnosti.

S určitou pravděpodobností - tedy na určité hladině významnosti předpokládáme, že hodnota *t* nepřekročí kritickou hodnotu t_p (při správnosti nulové hypotézy).

V opačném případě zamítáme nulovou hypotézu – mezi výběry náhodných veličin vztah existuje.

Nelineární závislost dvou výběrových souborů

V případě, kdy regresní čára není přímka, ale je vyjádřena složitější matematickou funkcí, se jako míry korelační závislosti používá tzv. korelační poměr (η_{yx}).

Prvky výběru závisle proměnné y_i rozdělíme podle hodnot nezávisle proměnné x_j do skupin označených y_j a pro každou skupinu vypočteme průměr $\bar{y}_j$. Korelační poměr se vypočte podle vztahu:

$$\eta_{yx} = \sqrt{\frac{\sum(\bar{y}_j - \bar{y}) \cdot n_j}{\sum(y_i - \bar{y})^2}} = \sqrt{\frac{\sum(\bar{y}_j n_j - n\bar{y}^2}{\sum y_i^{\ 2} - n\bar{y}^2}}$$

V uvedeném vzorci je n_j četnost v y_j. Při výpočtu **záleží** na tom, kterou proměnou zvolíme za závislou a kterou za nezávislou.

Porovnání hodnot korelačního koeficientu a korelačního poměru lze použít jako kritéria linearity vztahu.

Pokud se hodnoty přibližně rovnají, jedná se o závislost lineární, pokud je r_{xy} výrazně větší, jde o závislost nelineární.

Koeficient mnohonásobné korelace (r_{xyz})

Používá se pro hodnocení korelační závislosti tří nebo více výběrů náhodných veličin.

Při jeho určení se vychází z jednotlivých korelačních koeficientů pro dva výběry (r_{xy}, r_{xz}, r_{yz}) a jejich hodnoty se dosazují do vzorce pro r_{xyz}:

$$r_{xyz} = \sqrt{\frac{r_{xz}^{\ 2} + r_{yz}^{\ 2} - 2r_{xy} \cdot r_{xz} \cdot r_{yz}}{1-r_{xy}^{\ 2}}}$$

Dílčí (parciální) korelace:

Řeší otázku vlivu jedné nebo více nezávisle proměnných na závisle proměnnou při vyloučení vlivu zbývajících nezávisle proměnných, u nichž předpokládáme konstantní hodnotu.

Jedná se o zvláštní případ mnohonásobné korelace.

Hodnota koeficientu dílčí korelace $r_{xy.z}$ se vypočte podle vztahu:

$$r_{xy \cdot z} = \frac{r_{xy} - r_{xz} \cdot r_{yz}}{\sqrt{(1-r_{xz}^{\ 2}) \cdot (1-r_{yz}^{\ 2})}}$$

Tečkou v indexu se označuje nezávisle proměnná, jejíž hodnotu považujeme za konstantní.

Poznámky k aplikaci korelačního počtu:

Použití korelačního počtu je nevhodné např. v těchto případech:

- Korelace je způsobena formálními vztahy mezi veličinami (hodnoty x a y se doplňují do 100%)
- Korelace je způsobena nehomogenitou studovaného materiálu (obsahuje tzv. subpopulace – viz. obr. bodového grafu)
- Korelace je výsledkem působení třetí veličiny (korelace mezi počtem čapích hnízd a počtem novorozenců)

Měření závislosti kvalitativních znaků

• Kvalitativní znaky mají slovní charakter a získáváme je v sociologických průzkumech, při terénním šetření apod.

• Slovní charakter mají odpovědi na otázky týkající se např. pohlaví, vzdělání nebo povolání respondenta atd.

• K popsání vztahu závislosti spojitých kvantitativních veličin slouží **korelační koeficient**.

• K charakterizování závislostí kvalitativních znaků slouží tzv. **kontingenční tabulky**

Klasifikace kvalitativních znaků:

• Podle počtu možných obměn dělíme znaky na **alternativní** (také dvojné) nabývající pouze dvou obměn a znaky **množné,** nabývající více než dvou obměn,

• Podle možnosti určit objektivní pořadí obměn na znaky, které **mají pořadový charakter** (např. vzdělání, stupeň souhlasu či nesouhlasu apod.) a znaky, které tento charakter nemají (např. povolání, typ absolvovaného vzdělání, značka výrobku) a u nichž tedy objektivní uspořádání není možné,

• Podle toho zda lze jednoznačně vymezit kde „začíná" a „končí" každá obměna znaku nebo nelze (např. u barevných odstínů) dělíme znaky na **nespojité a spojité**.

Statistická analýza kvalitativních znaků:

• Statistické zpracování jednoho slovního znaku spočívá jednak v jeho třídění

• Nejčastěji se jedná o prosté třídění podle jednotlivých obměn slovního znaku a o stanovení absolutních nebo relativních četností.

• V omezené míře lze určovat charakteristiky úrovně (modus, u pořadových znaků medián, nikdy aritmetický průměr).

• Existují i speciální charakteristiky proměnlivosti.

• O měření závislosti má smysl uvažovat, je-li k dispozici dvojice slovních znaků.

Měření závislosti kvalitativních znaků

Spočívá v sestavení tzv. kontingenční tabulky

Z kontingenční tabulky lze určit intenzitu závislosti ve dvojici slovních znaků.

Nelze z ní však určit průběh závislosti. O směru závislosti má smysl se vyslovit pouze v případě pořadových slovních znaků.

Máme-li dva alternativní znaky dostaneme tzv. čtyřpolní tabulku.

Měření závislosti kvalitativních znaků

Obecně může mít každý kvalitativní znak A *r* tříd a znak B *s* tříd. Výsledky šetření potom sestavujeme do kontingenční tabulky *r x s*.

Pozorované četnosti v jednotlivých buňkách označujeme dvěma indexy – obecně n_{ij}.

Také marginální četnosti mají dva indexy.

Ten, přes který je sčítáno je označen hvězdičkou – tedy n_{2*} značí součet četností v druhé řádce, n_{*1} značí součet četností v prvním sloupci.

Tabulka bývá doplněna hodnotami procentuálních (relativních) četností. Častým požadavkem je konstantní délka intervalů tvořících třídy.

Stejně jako v případě kvantitativních znaků ověřujeme i zde existenci vztahu testy významnosti a hodnotíme ho vhodnou mírou závislosti.

Kontingenční tabulka typu r x s

Třídený znak		Znak B						Součet
		b_1	b_2	...	b_j	...	b_s	
Znak A	a_1	n_{11}	n_{12}				n_{1s}	n_{1*}
	a_2	n_{21}						n_{2*}
	⋮							⋮
	a_i				n_{ij}			n_{i*}
	⋮							⋮
	a_r	n_{r1}					n_{rs}	n_{r*}
Součet		n_{*1}	n_{*2}	...	n_{*j}	...	n_{*s}	$n_{**} = n$

Posuzování závislosti v kontingenčních tabulkách

Podmíněné četnosti uvnitř kontingenční tabulky mají podobný význam jako body korelačního diagramu — jejich rozmístění umožňuje usuzovat na charakter závislosti tříděných znaků.

Pro posouzení nezávislosti obou znaků můžeme vedle pozorovaných četností stanovit pro jednotlivá pole také očekávané (teoretické) četnosti :

$$n'_{ij} = \frac{n_{i*} n_{*j}}{n}$$

tedy jako součin okrajových četností příslušného řádku a sloupce dělený rozsahem souboru.

Pro každé pole kontingenční tabulky existuje dvojice četností - četnost pozorovaná a četnost vypočtená.

Hypotéza nezávislosti

Ukazatel, který pro tabulku jako celek měří rozdílnost pozorovaných a vypočtených četností v jednotlivých polích tabulky se nazývá čtvercová kontingence χ^2

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(n_{ij} - n'_{ij})^2}{n'_{ij}}$$

Je to bezrozměrná hodnota a platí: $\chi^2 \geq 0$

Hodnoty nula nabývá pouze v případě, že znaky v kontingenční tabulce jsou nezávislé.

Vypočtená hodnota χ^2 se porovnává na zvolené hladině významnosti p s kritickou hodnotou χ^2rozdělení pro (r-1)(s-1) stupňů volnosti.

Hypotézu zamítáme, jestliže vypočtená hodnota je větší než tabulková, případě, když jí příslušející p-hodnota je menší než zvolená hladina významnosti.

Koeficienty kontingence

Maximální možná hodnota čtvercové kontingence závisí na rozměrech kontingenční tabulky a rozsahu souboru - z toho důvodu není nejvhodnějším ukazatelem intenzity závislosti.

Na bázi čtvercové kontingence jsou konstruovány vhodnější ukazatele - koeficienty kontingence.

Jsou konstruovány tak, aby jejich hodnota závisela pouze na intenzitě závislosti.

Koeficienty kontingence měří intenzitu závislosti pro dvojici slovních znaků.

Pearsonův koeficient kontingence:

$$P = \sqrt{\frac{\chi^2}{\chi^2 + n}}$$

nabývá hodnot $0 \leq P < 1$

Příklad analýzy závislosti v tabulce r x s

Pro výběr 234 studentů zjišťujeme, zda existuje vztah mezi sportem, který provozují a sportovními pořady, které sledují v televizi.

Sestavíme tabulku typu 4 x 4:

Oblíbenost při sledování televize	Oblíbenost při sportování				Řádkové součty
	hry	atletika	gymnastika	plavání	
hry	133	6	2	4	145
atletika	15	10	4	3	32
gymnastika	4	1	25	0	30
plavání	9	0	1	17	27
Sloupcové součty	161	17	32	24	234

Hypotéza nezávislosti H_0: Neexistuje vztah mezi provozovaným sportem a sportem sledovaným v TV.

Vypočtená hodnota testovacího kritéria $\chi^2 = 273{,}3$

Kritická hodnota z tabulek pro p=0,05 a (4-1)x(4-1)=9 stupňů volnosti: $\chi^2 = 16{,}9$

Závěr: H_0 zamítáme, existuje významný vztah.

Sílu tohoto vztahu lze posoudit Pearsonovým koeficientem kontingence

$$P = \sqrt{\frac{\chi^2}{\chi^2 + n}} = \sqrt{\frac{273{,}3}{273{,}3 + 274}} = 0{,}71$$

Testování nezávislosti v tabulce 2 x 2

	Zájem o statistiku		řádkové součty
	ano	ne	
chlapci	a	b	a + b
dívky	c	d	c + d
sloupcové součty	a + c	b + d	n

Pro výpočet testovacího kritéria χ^2 v tabulce 2 x 2 můžeme využít zjednodušený vzorec:

$$\chi^2 = \frac{n(ad - bc)^2}{(a+b)(c+d)(a+c)(b+d)}$$

Protože v 2x2 tabulce můžeme uvažovat i směr poruchy nulové hypotézy – proto musíme rozhodnout, zda použijeme test jednostranný či dvoustranný.

Kritické hodnoty jsou uvedeny v tabulce χ^2- rozdělení o jednom stupni volnosti.

Příklad analýzy závislosti v tabulce 2 x 2

	Zájem o statistiku		řádkové součty
	ano	ne	
chlapci	30	36	66
dívky	11	63	74
sloupcové součty	41	99	140

Hypotéza nezávislosti H_0: Relativní četnost studentů se zájmem o statistiku je nezávislá na pohlaví.

Vypočtená hodnota testovacího kritéria: $\chi^2 = \frac{140(30 \times 63 - 11 \times 36)^2}{41 \times 99 \times 66 \times 74} = 15{,}8$

Kritická hodnota χ^2-rozdělení z tabulek pro p=0,05: 3,84

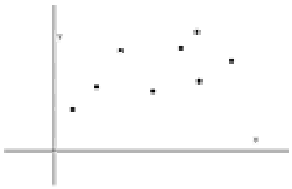
Závěr: H_0 zamítáme, existuje významný rozdíl.

Zájem u chlapců: 30/66 = 0,45

Zájem u dívek: 11/74 = 0,14

Chlapci mají zhruba 3x větší zájem o statistiku než dívky.

Regresní analýza



• Úkolem regresní analýzy je sestavit vztah (model) závislosti mezi závisle a nezávisle proměnnou.

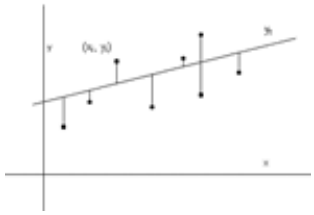
• Stejně jako v případě korelačního počtu je prvním indikátorem možného vztahu obou studovaných veličin graf pole hodnot.

• Z grafu je patrný typ závislosti (tato může být lineární či nelineární, …)

Určení lineární regresní závislosti

Nejjednodušším případem regresní závislosti je případ, kdy regresní funkce je přímkou. Rovnice regresní přímky má tvar:

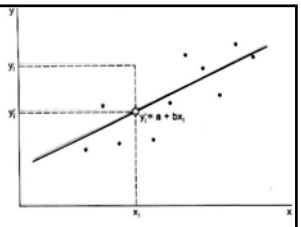
$y' = a + bx$



Symbol *y'* se používá pro označení nejpravděpodobnější teoretické hodnoty y odpovídající danému x, která leží na regresní přímce a která se odlišuje od konkrétních hodnot y_i, které se nacházejí mimo ni.

MNČ

Průběh regresní přímky je určen tzv. metodou nejmenších čtverců, kdy musí být splněna podmínka takového průběhu přímky, při kterém je součet čtverců vzdálenosti všech bodů pole od přímky minimální, tedy platí:



$$\sum (y_i - y_i')^2 = \min$$

Výpočet vertikální vzdálenosti bodů korelačního pole od regresní přímky se provádí podle uvedeného obrázku. Z něho je zřejmé, že pro vzdálenost konkrétní hodnoty závisle proměnné y_i od bodu regresní přímky y_i' musí platit:

$$y_i - y_i' = y_i - (a + bx_i) = y_i - a - bx_i$$

Součet čtverců svislých vzdáleností y_i od regresní přímky je potom:

$$\sum (y_i - y_i')^2 = \sum (y_i - a - bx_i)^2 = A$$

MNČ

Pro MNČ musí platit

$$A = \sum (y_i - a - bx_i)^2 = \min$$

Následnými úpravami lze obdržet vztahy pro výpočet koeficientů regresní přímky *a,b*

$$b = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2} \qquad a = \bar{y} - b\bar{x}$$

Koeficient b (angl. slope) se označuje jako koeficient regrese a je směrnicí regresní přímky (tangentou úhlu, který přímka a svírá s osou x). Je-li b>0, mluvíme o regresi pozitivní, je-li b<0 o regresi negativní**.**

Výpočet koeficientů regresní přímky

Vzorec pro výpočet koeficientu b lze zjednodušit pomocí vztahů pro kovarianci s_{xy}a směrodatnou odchylku s_x, tedy:

$$b = \frac{s_{xy}}{s_x^2}$$

Hodnota koeficientu *a* (angl. intercept) představuje y-ovou souřadnici průsečíku regresní přímky s osou y (tedy při x=0).

Dosazením výrazu pro koeficient a $a = \bar{y} - b\bar{x}$ do rovnice přímky y' = a + bx dostaneme:

$$y' = bx + \bar{y} - b\bar{x}$$

$$y' - \bar{y} = b(x - \bar{x})$$

Tohoto vztahu lze využít pro konstrukci regresní přímky – pro dvě zvolená x_1, x_2 vypočteme y_1 a y_2 a souřadnice obou bodů vyneseme do korelačního diagramu. Regresní přímka vznikne proložením oběma body.

Intervaly a pásy spolehlivosti lineární regresní závislosti

• Konstrukci regresní přímky provádíme na základě výběrových souborů.

• Proto se její rovnice může u různých výběrů ze stejných základních souborů lišit.

• Z tohoto důvodu je potřebné doplnit průběh regresní přímky také tzv. **intervaly spolehlivosti.**

• Výpočtem intervalů spolehlivosti určujeme pro vybraná x interval, v němž se mohou s určitou pravděpodobností vyskytovat hodnoty y s tím, že jejich nejreprezentativnější hodnota je y'.

Intervaly a pásy spolehlivosti

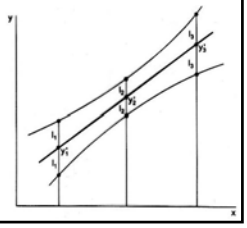
Nejprve je zapotřebí zvolit interval spolehlivosti – tedy pravděpodobnost, s níž očekáváme výskyt hodnot *y* v určených mezích *1-p (p=0,05 či 0,01*). Poloviční šířka intervalu spolehlivosti *l* je dána výrazem:

$$l = t_{1-p} \cdot \frac{h\sqrt{A}}{\sqrt{n-2}} \qquad h = \sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{(n-1)s_x^2}}$$

Hodnota t_p je kritická hodnota rozdělení pro *n-2* stupňů volnosti a hladinu významnosti *p*. Meze intervalů spolehlivosti určíme pomocí hodnot *y'* z rovnice $y' - \bar{y} = b(x - \bar{x})$

horní mez: *y' + l*

dolní mez: *y' - l*

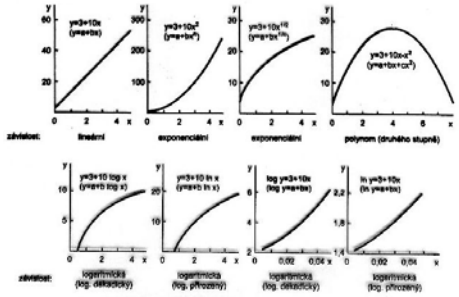


Pásy spolehlivosti vzniknou spojením krajních bodů intervalů spolehlivosti.

Nelineární regrese

Popisuje regresní vztah dvou proměnných, který nelze vyjádřit přímkou.

Může mít tvar např. logaritmických či exponenciálních funkcí a nebo je vztah vyjádřen rovnicí polynomu m-tého stupně.



Nelineární regrese

Volbu vhodné funkce, která by nejlépe vystihovala povahu studované závislosti provádíme na základě výpočtu směrodatné chyby aritmetického průměru $c_{\bar{y}}$ (viz. – Odhady parametrů a intervaly spolehlivosti).

Určení hodnoty směrodatné chyby aritmetického průměru spočívá v určení sumy čtverců odchylek *A* konkrétních hodnot y_i závisle proměnné od teoretických hodnot y'_i tedy:

$$\sum (y_i - y_i')^2 = \sum (y_i - a - bx_i)^2 = A$$

$$c_{\bar{y}} = \sqrt{\frac{A}{n}} = \sqrt{\frac{\sum (y_i - y_i')^2}{n}}$$

Povaze studované závislosti vyhovuje nejlépe ta z uvažovaných funkcí, která má hodnotu směrodatné chyby minimální.

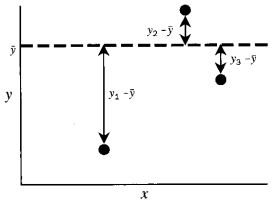
Konkrétní balíky statistických programů obsahují obvykle řadu nástrojů pro zvolení vhodné regresní závislosti.

Testování významnosti regresní čáry

• K testování významnosti zjištěné regresní závislosti lze využít **t-testu**, kterým lze zjistit, zda se gradient (směrnice) významně liší od nuly

• Nejčastěji se však používá techniky označované jako **analýza rozptylu (ANOVA).**

• **Princip:** Zjistíme celkovou proměnlivost hodnot *y* a následně vypočteme, z jaké části je tato celková variabilita objasněna proměnlivostí v hodnotách *x*.



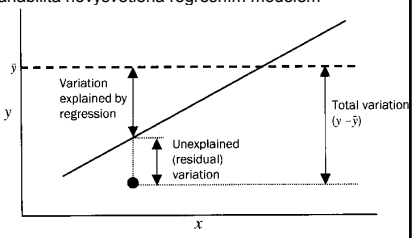
SS_{total} - celková variabilita: celková suma čtverců: od každé hodnoty y odečteme průměr, výsledek povýšíme na druhou a sečteme pro všechna y.

Testování významnosti regresní čáry

Celkovou variabilitu SS_{total} lze rozdělit na dvě části:

$SS_{regrese}$ - variabilitu vysvětlenou regresní čarou

$SS_{reziduální}$ – zbytková variabilita nevysvětlená regresním modelem



$$SS_{total} = \sum y^2 - \frac{(\sum y)^2}{n}$$

$$SS_{regrese} = \frac{\left(\sum xy - \frac{\sum x \sum y}{n}\right)^2}{\sum x^2 - \frac{(\sum x)^2}{n}}$$

$$SS_{reziduální} = SS_{total} - SS_{regrese}$$

Testování významnosti regresní čáry

Tabulka ANOVA

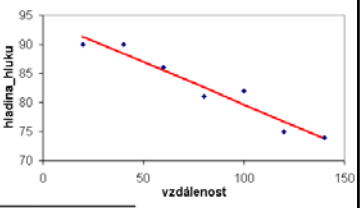
Variabilita	stupně volnosti (df)	Suma čtverců (SS)	Průměr sumy čtverců (MS)	F hodnota	p hodnota
Regresní	*1*	$SS_{regrese}$	$\frac{SS_{regres}}{1}$	$\frac{MS_{regres}}{MS_{resid}}$	F hodnota pro df = 1 a n-2
Reziduální	*n-2*	$SS_{reziduál}$	$\frac{SS_{residual}}{n-2}$		
Celková	*n-1*	SS_{total}			

Koeficient determinace regresní závislosti:

$$r^2 = \frac{SS_{regres}}{SS_{total}}$$

Příklad regresní analýzy v EXCELu

vzdálenost	hladina_hluku
20	90
40	90
60	86
80	81
100	82
120	75
140	74



VÝSLEDEK

Regresní statistika	
Násobné R	0,969074891
Hodnota spolehlivosti R	0,939106145
Nastavená hodnota spolehlivosti R	0,926927374
Chyba stř. hodnoty	1,764733893
Pozorování	7

ANOVA

	Rozdíl	*SS*	*MS*	*F*	*Významnost F*
Regrese	1	240,1428571	240,1429	77,11009	0,000317686
Rezidua	5	15,57142857	3,114286		
Celkem	6	255,7142857			

	Koeficienty	*Chyba stř. hodnoty*	*t stat*	*Hodnota P*	*Dolní 95%*	*Horní 95%*	*Dolní 95,0%*	*Horní 95,0%*
Hranice	94,2857	1,4915	63,2165	0,0000	90,4518	98,1197	90,4518	98,1197
vzdálenost	-0,1464	0,0167	-8,7812	0,0003	-0,1893	-0,1036	-0,1893	-0,1036

Existuje signifikantní pokles hladiny hluku se vzdáleností od komunikace.
Lineární regresní model vysvětluje 93,9 % variability hodnot hladiny hluku