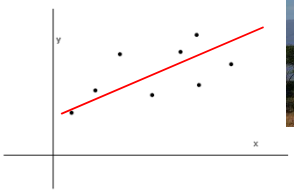
MASARYKOVA UNIVERZITA

Z1069 Statistické metody a zpracování dat

VII. Regresní počet

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

Regresní analýza



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

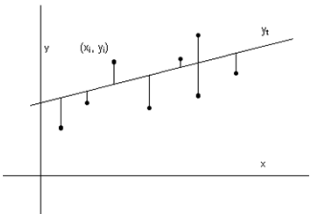
Regresní analýza řeší :

- odhady neznámých parametrů regresní funkce
- testování hypotéz o těchto parametrech
- ověřování předpokladů regresního modelu

Určení lineární regresní závislosti

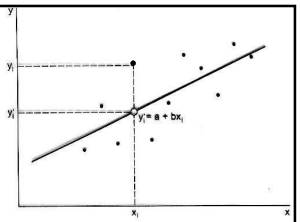
Nejjednodušším případem regresní závislosti je případ, kdy regresní funkce je přímkou. Rovnice regresní přímky má tvar:

$$y' = a + bx$$



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.

Metoda nejmenších čtverců



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:

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

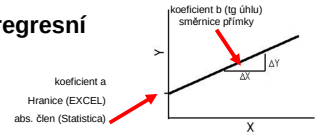
Součet čtverců svislých vzdáleností y_i od regresní přímky je potom:

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

Pro MNČ musí platit

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

Výpočet koeficientů regresní přímky $y' = a + bx$



Z výše uvedených vztahů lze následnými úpravami obdržet výrazy pro výpočet koeficientů regresní přímky *a,b*

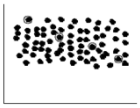
$$b = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2} \quad \text{případně} \quad b = \frac{s_{xy}}{s_x^2} \qquad a = \bar{y} - b\bar{x}$$

kovariance s_{xy} lomeno druhá mocnina směrodatné odchylky s_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.

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

Intervaly a pásy spolehlivosti lineární regresní závislosti



Koeficienty (parametry) regresní přímky jsou **bodovými odhady** !

- 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 vhodné 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 hladinu 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:

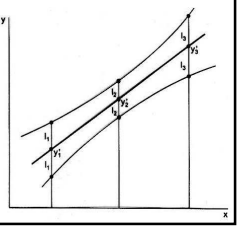
$$l = t_{1-p} \cdot \frac{h\sqrt{A}}{\sqrt{n-2}} \qquad h = \sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{(n-1)s_x^2}}$$

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: $y' + l$

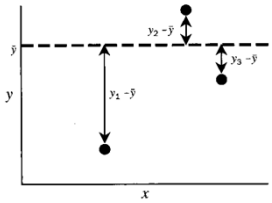
dolní mez: $y' - l$

Pásy spolehlivosti vzniknou spojením krajních bodů intervalů spolehlivosti.



Testování vhodnosti regresní závislosti

- Nejčastěji se k testování používá **analýzy rozptylu (ANOVA).**
- **Princip:** Zjistíme celkovou proměnlivost hodnot *y* a následně vypočteme, z jaké části je tato celková variabilita objasněna proměnlivostí v hodnotách *x*.

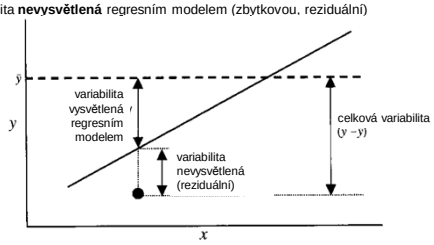


Celková variabilita: celková suma čtverců: od každé hodnoty y odečteme průměr, výsledek povýšíme na druhou a sečteme pro všechna y.

Testování významnosti regresní závislosti

Celkovou variabilitu lze rozdělit na dvě části:

- variabilitu **vysvětlenou** regresní čarou
- variabilita **nevysvětlená** regresním modelem (zbytkovou, reziduální)



- testování je založeno na porovnání množství vysvětlené a nevysvětlené variability
- je tedy **obdobou F-testu**: testujeme, zda variabilita vysvětlená modelem se významně liší od variability reziduální, tedy: **H0: neliší se**
- Vypočte se testovací kritérium (F) a pokud jemu příslušející p-hodnota je menší než alfa (0,05) potom **zamítáme nulovou hypotézu** a konstatujeme, že **regresní model je vhodný**

Příklad regresní analýzy v EXCELu

Nástroje – Analýza dat - Regrese

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

Zjistěte, jak souvisí hladina hluku se vzdáleností od komunikace.

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 — **model je vhodný**

	Rozdíl	*SS*	*MS*	*F*	*Významnost F*
Regrese	1	240,1428571	240,1429	77,11009	0,000317688
Rezidua	5	15,57142857	3,114286		
Celkem	6	255,7142857			

95% int. odhad hladiny hluku ve vzdálenosti 0 metrů

95% int. odhad poklesu hl. hluku na každý metr

	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

$y' = 94{,}2857 - 0{,}1464x$

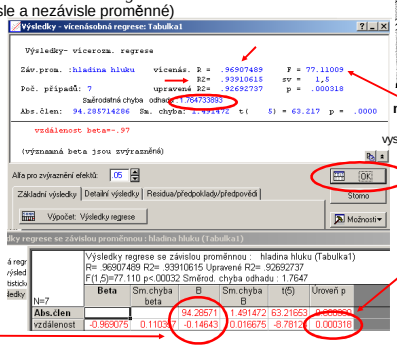
pokles hl. hluku na každý metr

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

Řešení v programu Statistica

1) Statistika – Vícerozměrná regrese

(zvolení závisle a nezávisle proměnné)



model je vhodný (změna vzdálenosti vysvětluje téměř 94 % variability hladiny hluku)

Parametry regresní přímky

Pokles hladiny hluku s rostoucí vzdáleností **je statisticky významný**

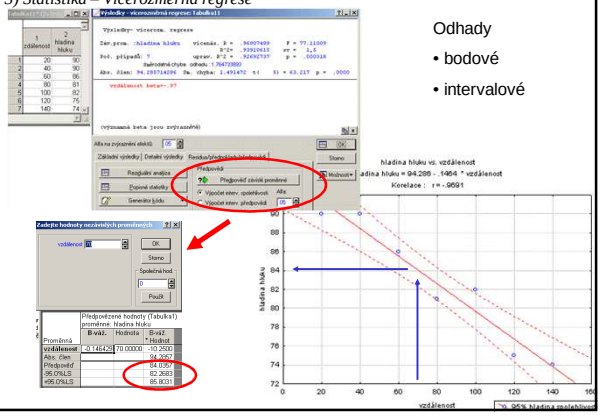
U regresního modelu se často testuje, zda se **směrnice přímky významně liší od nuly**. Používá se **t-testu**, kterým lze prokázat statisticky **významný růst či pokles hodnot** závisle proměnné při změně hodnot nezávisle proměnné.

Řešení v programu Statistica

2) OK – Bodové grafy – Korelace 2 proměnných

Výpočet neznámé hodnoty - předpovědi

3) Statistika – Vícerozměrná regrese



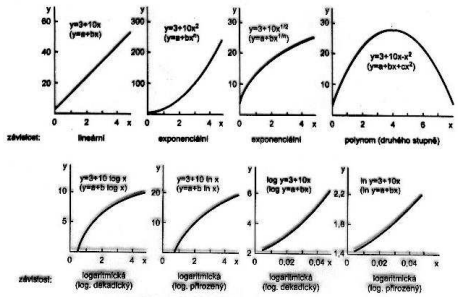
Odhady

- bodové
- intervalové

Další typy regresních funkcí

Regresní vztah dvou proměnných často nelze vhodně vyjádřit přímkou – jiné typy funkcí.

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



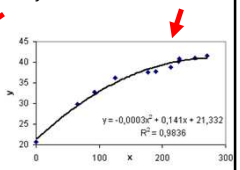
Regresní závislost není přímka

Příklad (viz. Brázdil a kol., 1995, str. 139, cvič. 8.5)

Stanice	x	y
Ivančice	0	20,7
Sen. Mlýn	65	29,9
Jakub	92	32,7
Senorady	125	36,2
Zastávka	177	37,6
Kratochvilka	190	37,7
Příbram	213	38,8
Bučín	226	40,2
Hlína	237	40,9
Ketkovice	252	41
Výs. Popovice	271	41,5

nevhodný model

vhodný model



VÝSLEDEK

Regresní statistika

Násobné R	0,956724
Hodnota spc	0,915321
Nastavená h	0,905912
Chyba stř. h	1,921192
Pozorování	11

koeficient determinace

směrodatná chyba odhadu

ANOVA

	Rozdíl	SS	MS	F	Významnost F
Regrese	1	359,0703	359,0703	97,28322	4,01237E-06
Rezidua	9	33,21881	3,690979		
Celkem	10	392,2891			

	Koeficienty	chyba stř. hod.	t stat	Hodnota P	Dolní 95%	Horní 95%	Dolní 95,0%	Horní 95,0%
Hranice	24,43416	1,31782	18,54135	1,77E-08	21,45303879	27,41527	21,45304	27,41527
x	0,069872	0,007084	9,863225	4,01E-06	0,053846464	0,085897	0,053846	0,085897

EXCEL:

pravým tlačítkem myši na graf –

přidat spojnici trendu

Hledání vhodného regresního modelu

Lze postupovat dvěma způsoby:

1. Volba vhodného modelu na základě praktické zkušenosti či teoretických předpokladů
2. Posouzením bodového grafu a interpretací nástrojů regresní analýzy

Způsoby hodnocení vhodnosti regresního modelu

- analýza reziduálních hodnot
- výpočet směrodatné chyby odhadu (s_e)
- výpočet koeficientu determinace (r^2_{xy}).

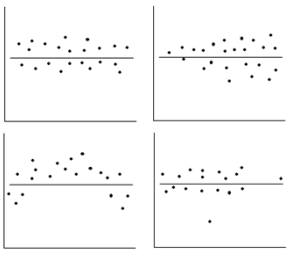
Hledání vhodného regresního modelu

Analýza reziduálních hodnot

Rezidua jsou vzdálenosti skutečných hodnot y_i od modelem odhadnutých hodnot y_j`

Zvolený regresní model považujeme za vhodný, pokud reziduální hodnoty splňují všechny následující podmínky:

- rezidua jsou ***náhodná*** a ***nezávislá***
- mají ***normální rozdělení*** s nulovým průměrem a konstantním rozptylem
- rozptyl reziduí je ***konstantní***.



Hledání vhodného regresního modelu

Směrodatná chyba odhadu – je vyjádřením směrodatné odchylky resp. rozptylu reziduálních hodnot a vhodnou mírou pro posouzení vhodnosti použité regresní závislosti

$$s_e = \sqrt{\frac{\sum_{i=1}^{n}(y_i - y_i')^2}{n-2}}$$

Čím je hodnota reziduálního rozptylu nižší, tím je model vhodnější.

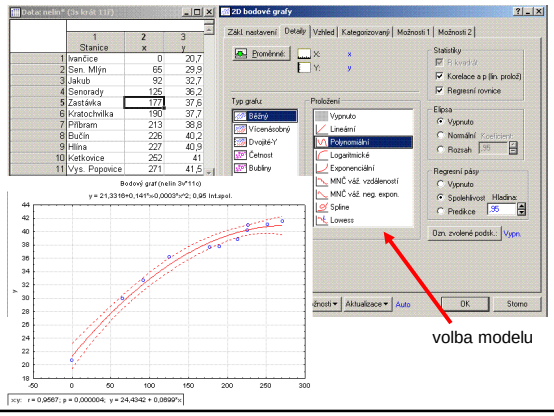
Koeficient determinace (r^2_{xy}) – viz. Korelační počet

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

Čím je hodnota koeficientu determinace větší, tím je model vhodnější.

Hledání vhodného regresního modelu

Grafy – Bodové grafy



volba modelu

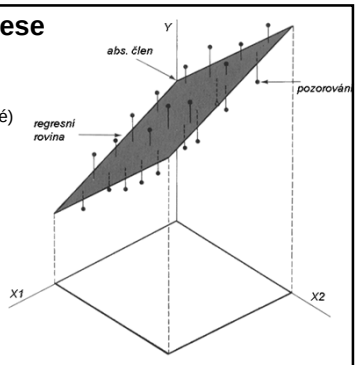
Vícerozměrná regrese

Popisuje závislost více proměnných z nichž více je příčinami (vysvětlující proměnné) a jen jedna je důsledek (vysvětlovaná proměnná).

Jsou–li dvě vysvětlující proměnné regresní model je rovina

Odhad parametrů se provádí MNČ

Výstupy a interpretace jsou **„obdobné“** jako u modelů jednorozměrné regrese



$y' = a + b_1x_1 + b_2x_2 + \dots$

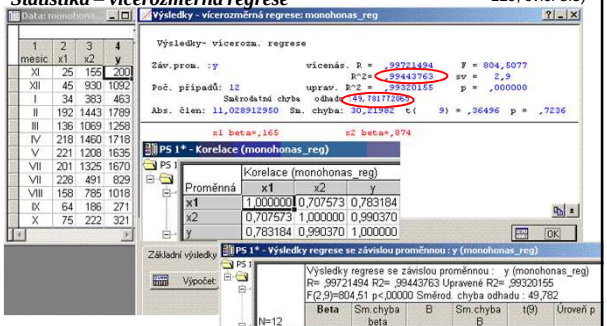
Např:

*Úhrn_srážek = 345,6 + 0,45*zem_délka + 1,23*nadm_výška*

Vícerozměrná regrese

Statistika – vícerozměrná regrese

(data viz. Brázdil a kol., 1995, str. 129, cvič. 8.3)



Punkva – pod Sk. Mlýnem (y) model odtoku:

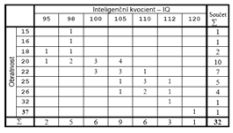
$y = 11{,}0289 + 1{,}2487x_1 + 1{,}0497x_2$

regresní koeficienty

standardizované regresní koeficienty

Měření závislosti kvalitativních znaků

- Kvalitativní znaky mají slovní charakter a získáváme je v sociologických průzkumech, při terénním šetření apod.
- K charakterizování závislostí kvalitativních znaků slouží tzv. **kontingenční tabulky**



- Z kontingenční tabulky lze určit intenzitu závislosti ve dvojici slovních znaků.
- Máme-li dva alternativní znaky dostaneme tzv. **čtyřpolní tabulku**.

	praváci	leváci	celkem
muži	43	9	52
ženy	44	4	48
celkem	87	13	100

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

Tříděný znak		b_1	b_2	...	b_j	...	b_s	Součet
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 α s kritickou hodnotou χ^2 rozdělení pro *(r-1)(s-1)* stupňů volnosti.

Hypotézu (H0) o nezávislosti dvou studovaných znaků zamítáme, jestliže vypočtená hodnota χ^2 je větší než tabulková; případně, když jí příslušející *p-hodnota* je menší než zvolená hladina významnosti.

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.

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 α=*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.

Čyřpolní tabulka - řešení v programu Statistica

Statistiky – Neparametrická statistika – Tabulka 2 x 2