

Korelační a regresní analýza

1. Pearsonův korelační koeficient
 2. jednoduchá regresní analýza
 3. vícenásobná regresní analýza
-

Pearsonův korelační koeficient

- u intervalových a poměrových dat můžeme jako míru asociace – vztahu mezi proměnnými použít **Pearsonův korelační koeficient**
 - **korelace**
 - ko = s, spolu, vzájemně
 - relace = vztah
 - korelace = vzájemný vztah proměnných
-

Pearsonův korelační koeficient

- absolutní hodnota koeficientu vyjadřuje **sílu (těsnotu) vztahu**
 - znaménko (+ nebo -) **směr vztahu**
 - rozsah -1 až +1**
 - označuje se **r**
-

Pearsonův korelační koeficient

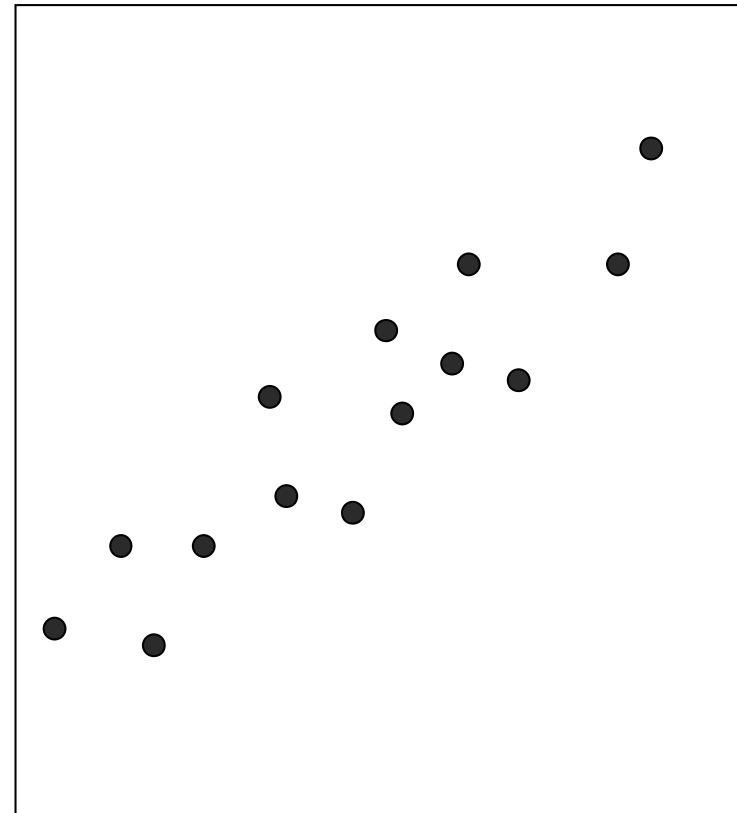
- sám o sobě je deskriptivní statistikou, ale podobně jako u ostatních měr asociace je možno spočítat **statistickou významnost**
 - závisí na velikosti výběru – čím vyšší, tím nižší koeficient vychází průkazný
-

Pearsonův korelační koeficient

- je mírou **pouze pro lineární vztahy**
 - před výpočtem je vhodné zobrazit vztah mezi proměnnými také graficky – tzv. **scatter** (dvourozměrný tečkový diagram)
-

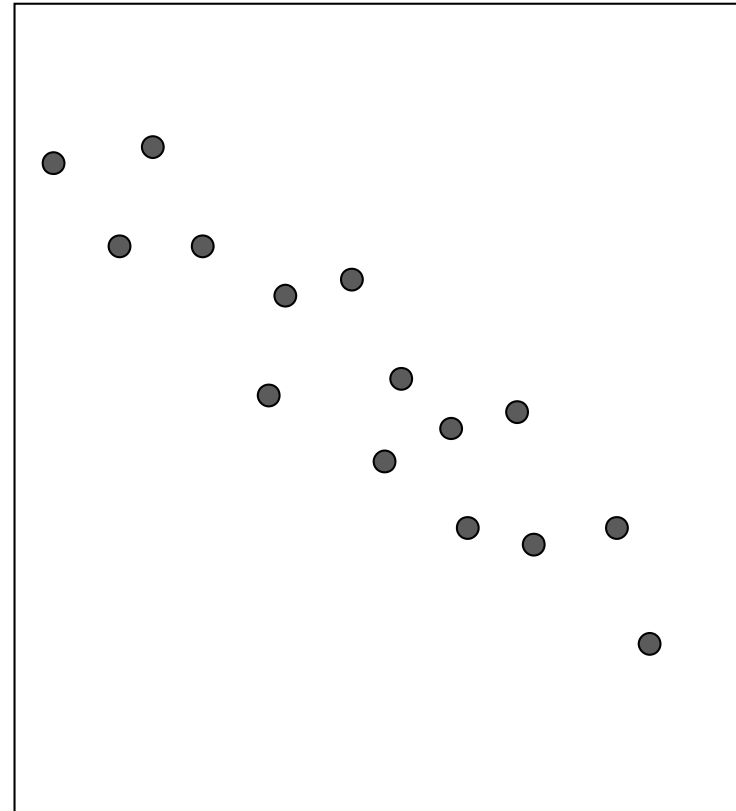
Scatter

- **pozitivní vztah** (přímá úměra) – čím vyšší hodnoty proměnné X, tím vyšší hodnoty proměnné Y
- $r > 0$



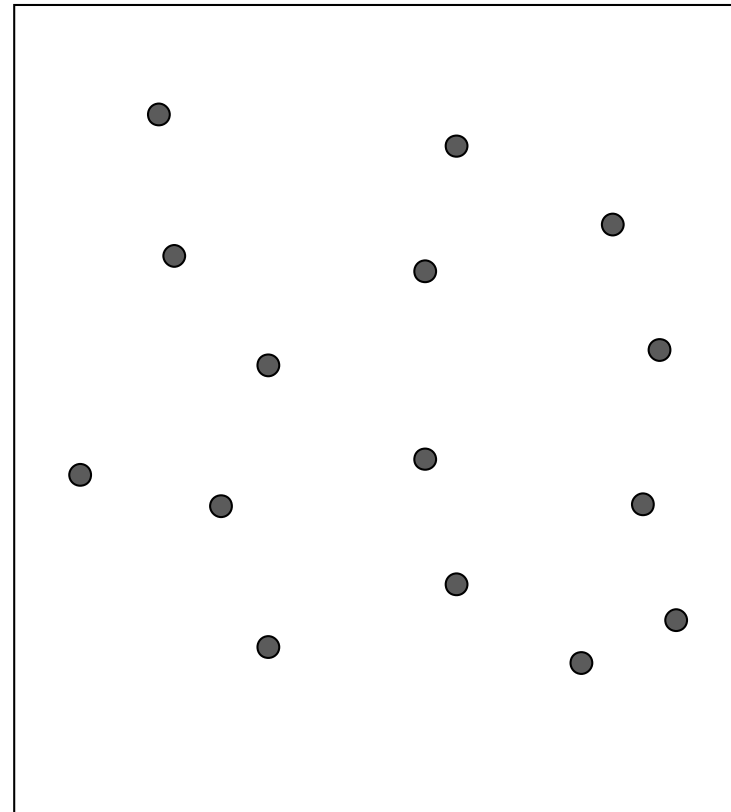
Scatter

- **negativní vztah** (nepřímá úměra) – čím vyšší hodnoty proměnné X, tím nižší hodnoty proměnné Y
- $r < 0$



Scatter

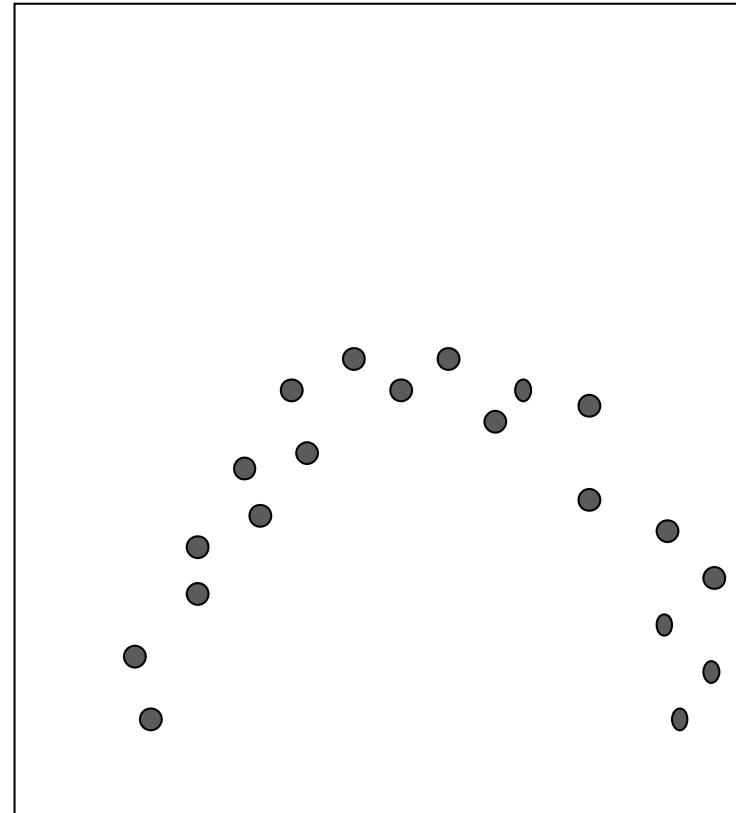
- **žádný vztah** -
hodnoty
proměnné X
nesouvisí s
hodnotami
proměnné Y
- $r = 0$



Scatter

□ **nelineární
vztah**

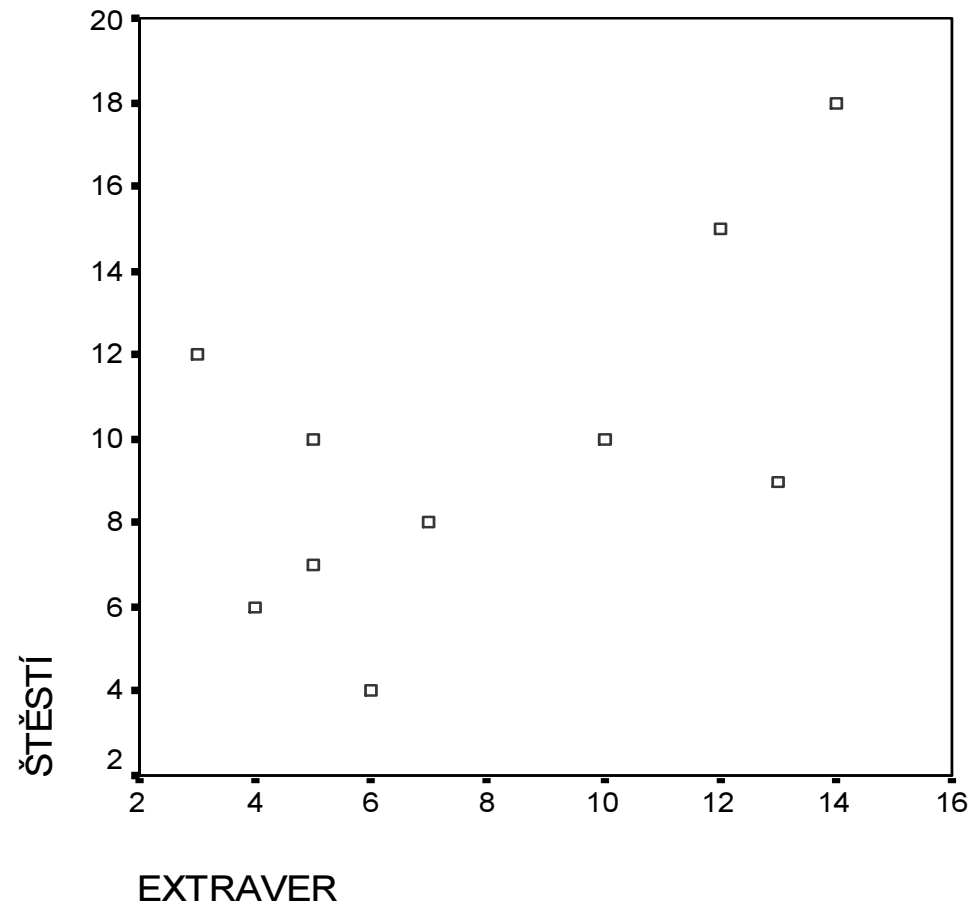
□ $r = 0$



Příklad

- jak spolu souvisí pocit štěstí a míra extraverze?
 - 10 osob, 2 proměnné – skór z dotazníku štěstí a skór ze škály extraverze
-

Příklad



Příklad

šťěstí	15	8	7	18	4	12	10	10	6	9
extraverze	12	7	5	14	6	3	5	10	4	13

Příklad

□ výpočet r

$$r = \frac{SP_{XY}}{\sqrt{SS_X \cdot SS_Y}}$$

Příklad

$$\square \mathbf{SP}_{XY} = \Sigma(X - \bar{X})(Y - \bar{Y}) = \Sigma XY - (\Sigma X)(\Sigma Y)/N$$

$$\mathbf{SS}_X = \Sigma(X - \bar{X})^2 = \Sigma X^2 - (\Sigma X)^2/N$$

$$\mathbf{SS}_Y = \Sigma(Y - \bar{Y})^2 = \Sigma Y^2 - (\Sigma Y)^2/N$$

Příklad

$$\square SP_{xy} = 91,9$$

$$SS_x = 158,9$$

$$SS_y = 144,9$$

$$r = 91,9 / (\sqrt{158,9 * 144,9})$$

$$r = \mathbf{0,606}$$

Výstup ve Statistice

Proměnná	Korelace (data příklad přednáška 2) Označ. korelace jsou významné na hlad. $p < ,0500$ N=10 (Celé případy vynechány u ChD)	
	šťěstí	extraverze
šťěstí	1,00	0,61
extraverze	0,61	1,00

Interpretace r

- není shoda v tom, jaká hodnota r je považována za těsný vztah
 - interpretace navržená Guilfordem:
 - <0.20 zanedbatelný vztah
 - $0.20-0.40$ nepříliš těsný vztah
 - $0.40-0.70$ středně těsný vztah
 - $0.70-0.90$ velmi těsný vztah
 - >0.90 extrémně těsný vztah
-

Interpretace r

- pro lepší interpretaci je vhodné převést koeficient korelace na **koeficient determinace (r^2)**
 - ukazuje, kolik rozptylu v jedné proměnné může být vysvětleno rozptylem ve druhé proměnné
-

Interpretace r

- v našem příkladu
 - $r = 0,606$
 - $r^2 = 0,367$
 - 36,7% rozdílů v míře štěstí můžeme vysvětlit rozdíly v míře extraverze
-

Interpretace r

□ **korelace neznamená příčinný vztah mezi proměnnými!!**

- ten můžeme ověřovat pouze experimentem, kdy jsou všechny ostatní proměnné udržovány konstantní, proměnná X předchází Y v čase atd.
-

Faktory ovlivňující r

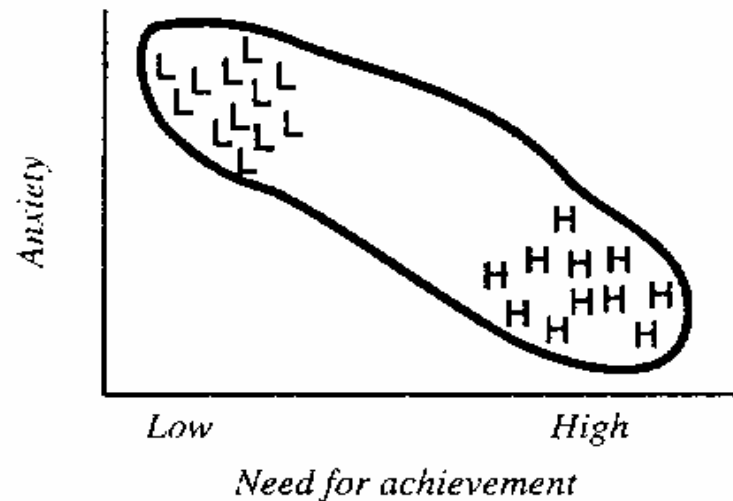
- omezený rozsah hodnot proměnné
 - použití extrémních skupin
 - nehomogenní soubor
 - extrémní hodnoty (outliers)
 - nelineární vztahy
 - reliabilita použitých nástrojů
-

Omezený rozsah hodnot

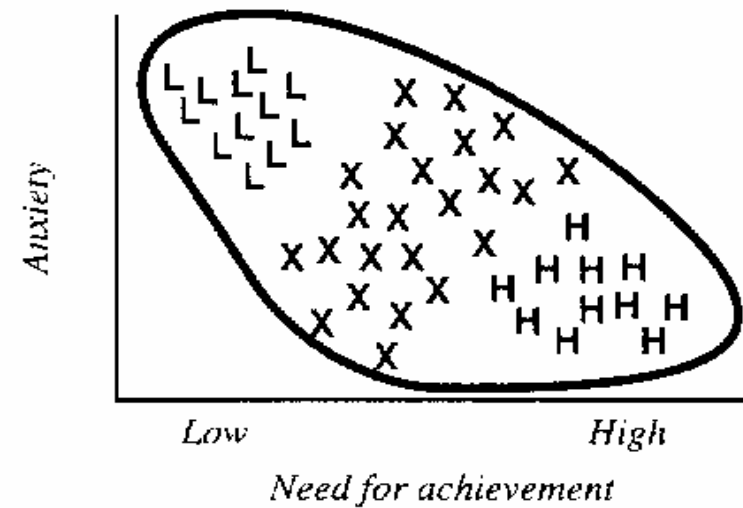
- omezený rozsah hodnot jedné nebo obou proměnných snižuje hodnotu r
 - stejně tak nízká variabilita (extrémní případ: pokud by všechny hodnoty 1 proměnné byly stejné, zákonitě $r=0$)
-

Použití extrémních skupin

- použití extrémních skupin (např. jen osob s vysokým IQ) vede k vyššímu r



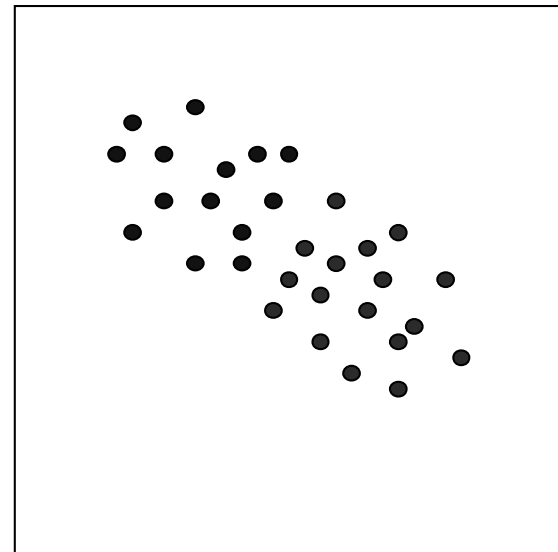
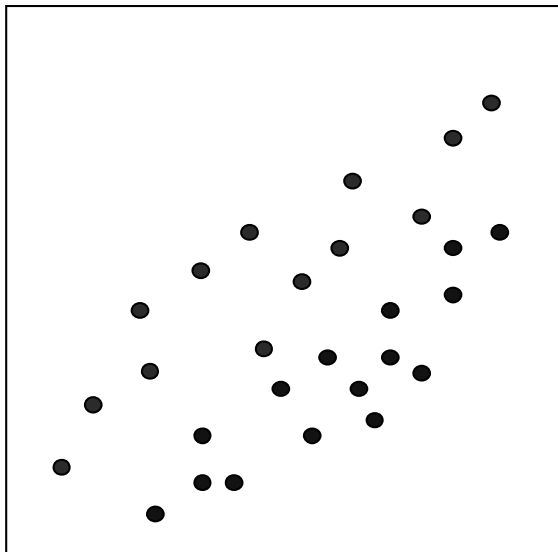
(a)



(b)

Nehomogenní soubor

- může zkreslit r jak směrem nahoru, tak dolů



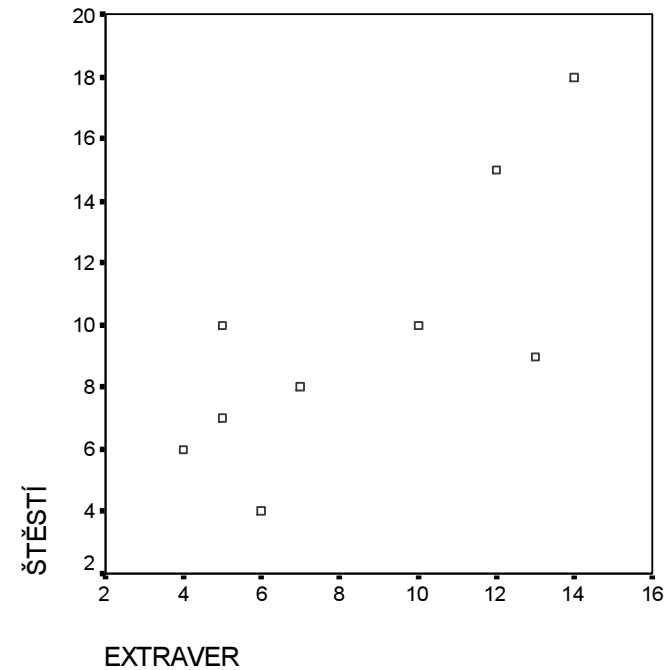
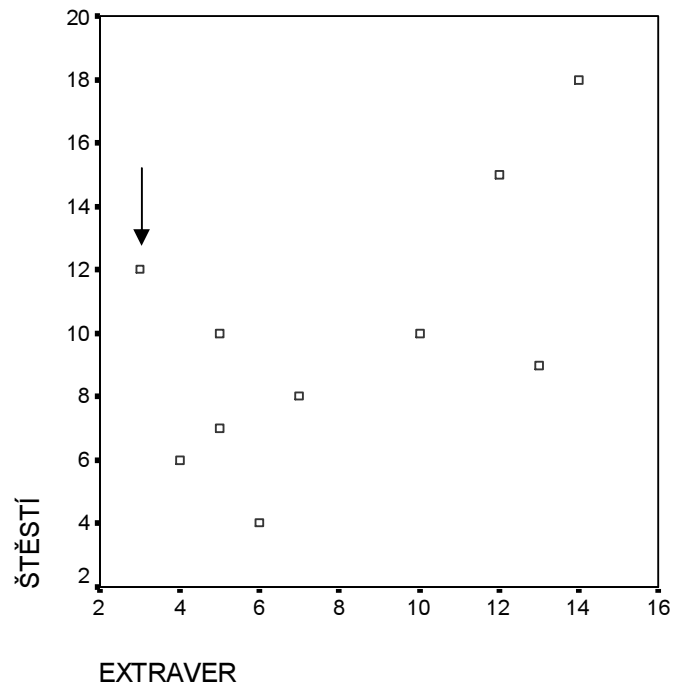
Extrémní hodnoty

- extrémní hodnoty v jedné nebo obou proměnných mohou r výrazně zkreslit (nejen hodnotu, ale i směr), zvláště když je počet osob v souboru nízký
-

Extrémní hodnoty

□ $r = 0,606$

□ $r = 0,766$



Neparametrický koeficient

- pro ordinální data je možno spočítat **Spearmanův koeficient pořadové korelace** (ρ)
 - počítá se tak, že
 - hodnoty obou proměnných se seřadí od nejnižší po nejvyšší a přidělí se jim pořadí
 - z pořadí se pak počítá Pearsonův koeficient korelace
-

Parciální korelace

- parciální korelace je taková korelace mezi dvěma proměnnými, kdy kontrolujeme vliv třetí proměnné na obě z nich
 - např. chceme zjistit, jaký je vztah mezi prospěchem na SŠ a prospěchem na VŠ; obě proměnné jsou nejspíš ovlivněny IQ
-

Regresní analýza

- výsledkem regresní analýzy je **matematický model vztahu mezi dvěma nebo více proměnnými**
 - snažíme se z jedné proměnné nebo lineární kombinace více proměnných predikovat hodnoty další proměnné
-

Regresní analýza

- dva typy proměnných: **predikovaná** (závislá) **proměnná** a **prediktory** (nezávisle proměnné)
 - predikovaná proměnná se označuje **Y**, prediktory **$X_1, X_2 \dots X_n$**
 - pouze 1 prediktor – **jednoduchá regrese**
 - více prediktorů – **vícenásobná regrese**
-

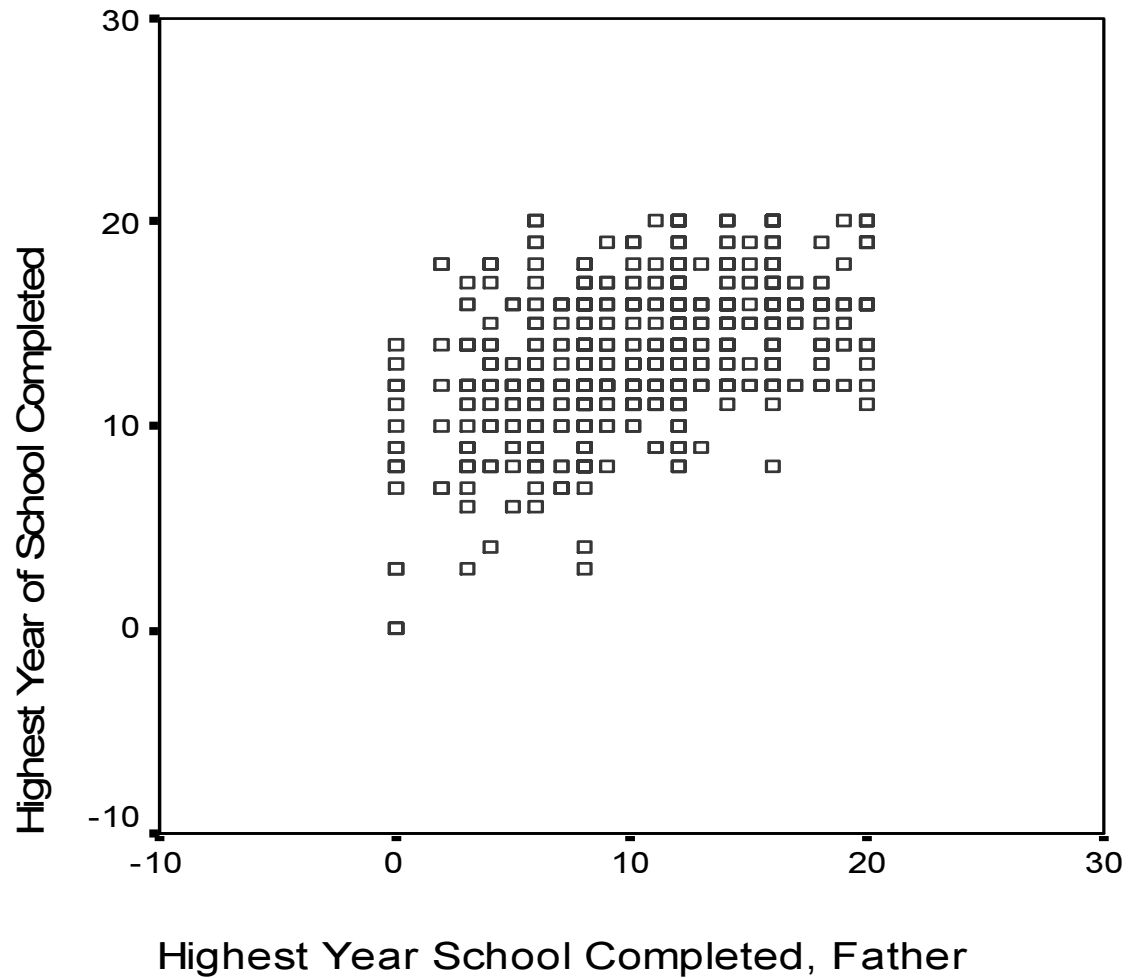
Regresní analýza

- regresní analýza umožňuje
 - porozumět vztahům mezi proměnnými,
 - predikovat hodnoty proměnné Y z hodnot proměnné X (s určitou přesností) – např. z hodnot známek na střední škole nebo z počtu bodů u přijímacího testu předpovědět úspěšnost na VŠ
-

Jednoduchá regresní analýza

- **příklad** – Jak souvisí vzdělání respondenta se vzděláním otce?
 - tj. jak dobře můžeme předpovědět počet let formálního vzdělání respondenta z údaje o počtu let vzdělání jeho otce?
-

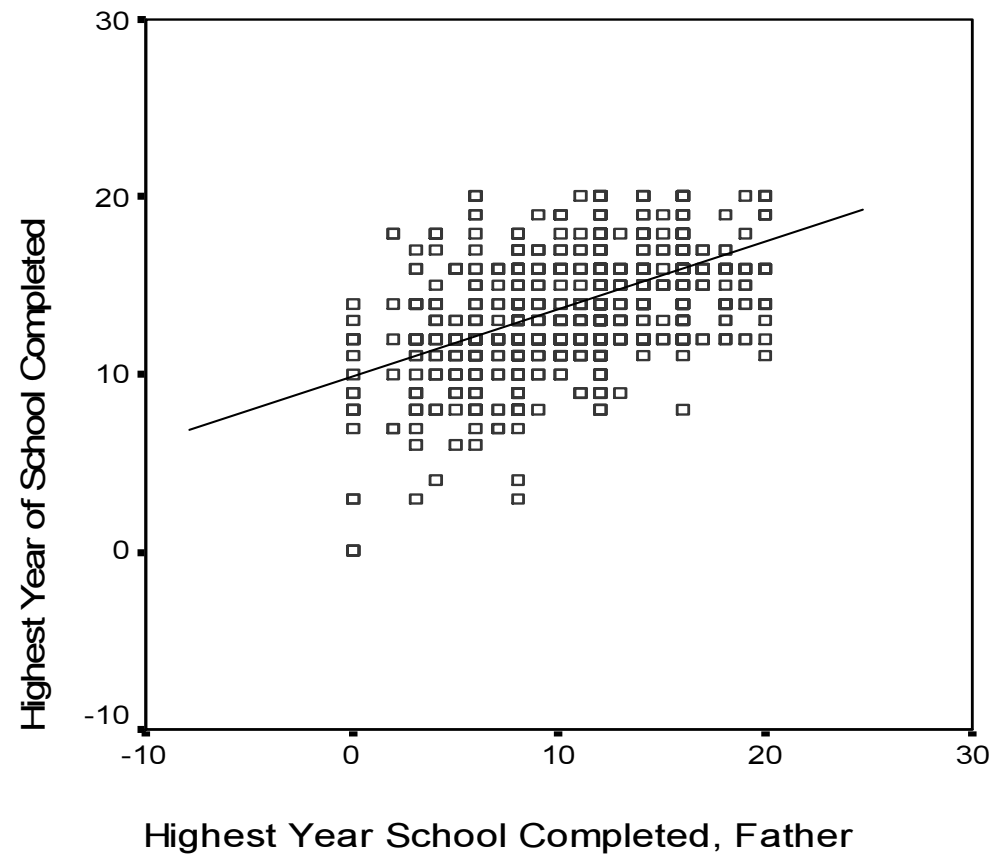
Jednoduchá regresní analýza



Jednoduchá regresní analýza

- snažíme se najít rovnici tzv. regresní přímky
 - **regresní přímka** je taková přímka, od které je vzdálenost bodů (představujících naměřená data) co nejmenší
 - taková přímka, která nejlépe vystihuje data
-

Jednoduchá regresní analýza



Jednoduchá regresní analýza

- jednou z metod, jak regresní přímku nalézt, je **metoda nejmenších čtverců**
 - je zvolena taková přímka, kdy platí, že součet čtverců vzdáleností jednotlivých bodů od přímky je minimální
-

Jednoduchá regresní analýza

- obecná rovnice regresní přímky

$$Y' = a + bX$$

- **a** je **konstanta** (predikovaná hodnota Y, když hodnota X je 0)
 - **b** je **směrnice** regresní přímky (úhel přímky vzhledem k ose; kolikrát se Y zvětší s každou jednotkou X);
-

Jednoduchá regresní analýza

- v příkladu vychází rovnice regresní přímky
 $Y' = 9,93 + 0,32 * X$
 - pro děti otců s 0 lety vzdělání
předpovíáme necelých 10 let vzdělání
 - s každým dalším rokem otcova vzdělání
předpovíáme o 0,32 roku vzdělání
respondenta více
 - např. pro děti otců s 12 lety vzdělání je
predikovaná hodnota jejich vlastního vzdělání
13,8 let
-

Výstup ve Statistice

N=1065	Výsledky regrese se závislou proměnnou : EDUC (us) R= ,46341505 R ² = ,21475351 uprav. R ² = ,21401480 F(1,1063)=290,72 p<0,0000 Směrod. chyba odhadu : 2,5348					
	Beta	Sm.chyba beta	B	Sm.chyba B	t(1063)	Úroveň p
	Abs.člen		9,925692	0,219305	45,25983	0,00
PAEDUC	0,463415	0,027179	0,321573	0,018860	17,05037	0,00



Vícenásobná regresní analýza

- predikujeme závislou proměnnou z více prediktorů
 - vliv každého z prediktorů na závislou proměnnou je **kontrolován** pro vliv všech ostatních prediktorů (jde tedy o vliv „očistěný od vlivů ostatních proměnných a tudíž počítáme **parciální** koeficienty)
-

Vícenásobná regresní analýza

□ **příklad** – kromě vzdělání otce (X_1) může mít na dosažené vzdělání vliv také počet dětí v rodině (X_2)

□ rovnice regresní přímky je

$$Y' = a + b_1X_1 + b_2X_2$$

Vícenásobná regresní analýza

- **$Y' = 10,68 + 0,30 * X_1 - 0,13 * X_2$**
 - vliv vzdělání otce ($b=0,30$) je o něco menší než u jednoduché regresní analýzy ($b=0,32$) – je kontrolován pro počet dětí v rodině, který je zřejmě ovlivněn také vzděláním otce
 - vliv počtu dětí v rodině je záporný – tj. čím více dětí, tím nižší vzdělání
-

Vícenásobná regresní analýza

- vícenásobná regresní analýza nám umožní srovnat vliv všech prediktorů na závislou proměnnou
 - můžeme dojít k závěru, že větší vliv na vzdělání respondenta má vzdělání otce než počet dětí v rodině?
-

Vícenásobná regresní analýza

- pokud chceme srovnávat vliv prediktorů měřených v různých jednotkách, je nutné použít tzv. **standardizované regresní koeficienty**
 - ukazují, kolikrát vzroste hodnota závislé proměnné, pokud se změní hodnota prediktoru o 1 směrodatnou odchylku a hodnoty ostatních prediktorů přitom zůstanou konstantní
-

Výstup ve Statistice

	Výsledky regrese se závislou proměnnou : EDUC (us) R= ,47898587 R ² = ,22942746 uprav. R ² = ,22797492 F(2,1061)=157,95 p<0,0000 Směrod. chyba odhadu : 2,5117					
N=1064	Beta	Sm.chyba beta	B	Sm.chyba B	t(1061)	Úroveň p
Abs.člen			10,67468	0,270874	39,40827	0,000000
SIBS	-0,128882	0,028046	-0,12766	0,027780	-4,59535	0,000005
PAEDUC	0,427009	0,028046	0,29631	0,019462	15,22516	0,000000



Vícenásobná regresní analýza

- beta pro vzdělání otce je 0,43
 - pro počet dětí v rodině -0,13
 - větší vliv má tedy vzdělání otce než počet dětí v rodině
-

Vícenásobná regresní analýza

- kromě regresních koeficientů je počítán také tzv. **koeficient vícenásobné korelace** – korelace všech prediktorů se závislou proměnnou; ozn. **R**
 - jde vlastně o korelaci mezi pozorovanými hodnotami závislé proměnné a hodnotami predikovanými na základě regresního modelu
-

Vícenásobná regresní analýza

- koeficient **vícenásobné determinace** – tzv. % vysvětleného rozptylu (závislé proměnné) lineární kombinací prediktorů; ozn. **R^2**
-

Výstup ve Statistice

	Statistické shrnutí; ZP: EDUC (us)	
Statist.	Hodnota	
Vícenás. R	0,4790	
Vícenás. R ²	0,2294	
Přizpůs. R ²	0,2280	
F(2,1061)	157,9491	
p	0,0000	
Sm. chyba odhadu	2,5117	

Vícenásobná regresní analýza


- u jednoduché regresní analýzy je **koeficient vícenásobné korelace** roven korelaci mezi oběma proměnnými
-

Testování hypotéz v regresní analýze

- jsou testovány 2 typy hypotéz
 - 1) zda se R průkazně liší od 0
 - testuje se analýzou rozptylu (porovnává rozptyl vysvětlený regresním modelem a reziduální rozptyl)
 - 2) zda se regresní koeficienty průkazně liší od 0
 - testuje se t-testem
-

Výstup ve Statistice

	Statistické shrnutí; ZP: EDUC (us)	
Statist.	Hodnota	
Vícenás. R	0,4790	
Vícenás. R ²	0,2294	
Přizpús. R ²	0,2280	
F(2,1061)	157,9491	
p	0,0000	
Sm. chyba odhadu	2,5117	



Výstup ve Statistice

Výsledky regrese se závislou proměnnou : EDUC (us)						
R= ,47898587 R ² = ,22942746 uprav. R ² = ,22797492						
F(2,1061)=157,95 p<0,0000 Směrod. chyba odhadu : 2,5117						
N=1064	Beta	Sm.chyba beta	B	Sm.chyba B	t(1061)	Úroveň p
Abs.člen			10,67468	0,270874	39,40827	0,000000
SIBS	-0,128882	0,028046	-0,12766	0,027780	-4,59535	0,000005
PAEDUC	0,427009	0,028046	0,29631	0,019462	15,22516	0,000000



Předpoklady regresní analýzy

- skóry v proměnných jsou nezávislé (nejde např. o opakovaná měření)
 - dostatečná variabilita všech proměnných
 - rozdělení hodnot proměnných je normální
 - u malých výběrů zkontrolovat extrémní hodnoty
-

Předpoklady regresní analýzy

- vztahy mezi Y a každou X jsou lineární
 - zkontrolovat scatterem
 - vzájemné korelace mezi prediktory nejsou příliš vysoké (tzv. problém multikolinearity)
 - pokud ano, je vhodné buď některou z nich vyřadit, nebo z nich vytvořit např. faktorovou analýzou jeden skór
-

Předpoklady regresní analýzy

- dostatečně velký počet osob ve výběru vzhledem k počtu prediktorů v modelu
-

Kontrolní otázky

- co vyjadřuje absolutní hodnota Pearsonova koeficientu korelace? a co jeho znaménko (+ nebo -)?
 - co je to koeficient determinace?
 - čím může být zkreslen korelační koeficient?
 - účel regresní analýzy
 - obecná rovnice regresní přímky
 - jak se interpretují regresní koeficienty
 - co je to koeficient vícenásobné korelace?
 - předpoklady regresní analýzy
-