

Parciální korelace

Regresní analýza

Parciální korelace

- parciální (dílčí) korelace nám umožňují při výpočtu **uměle vyloučit vliv některých proměnných**
 - a odhalit tak případné zkreslení při zkoumání vztahů mezi proměnnými (viz přednášky z metodologie)
-

Parciální korelace

- příklad – zkoumáme vztah mezi proměnnými X a Y , a zajímá nás, zda tento vztah nějak není ovlivněn proměnnou Z
-

Parciální korelace - příklad

□ zjistíme následující **korelace**

$$X \text{ versus } Y: \mathbf{r}_{XY} = 0.50 \quad (\mathbf{r}^2_{XY} = 0.25)$$

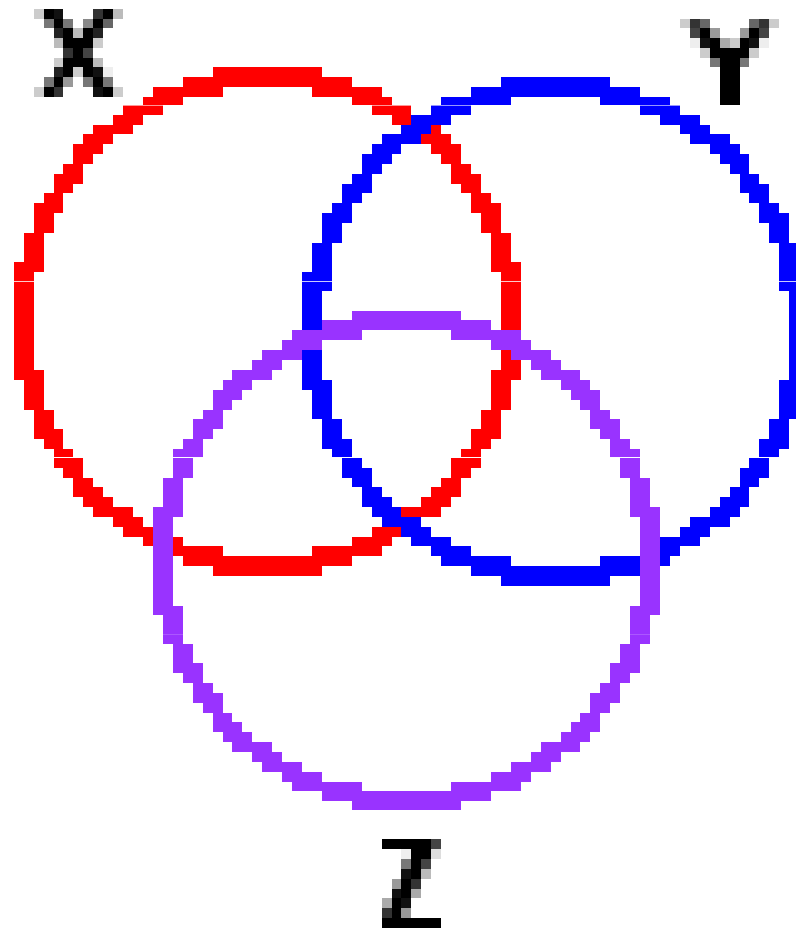
$$X \text{ versus } Z: \mathbf{r}_{XZ} = 0.50 \quad (\mathbf{r}^2_{XZ} = 0.25)$$

$$Y \text{ versus } Z: \mathbf{r}_{YZ} = 0.50 \quad (\mathbf{r}^2_{YZ} = 0.25)$$

Parciální korelace - příklad

- r^2 = koeficient determinace, tj. podíl společného rozptylu obou proměnných
 - pro každý pár proměnných je v tomto případě 25% (25% rozptylu proměnné X můžeme vysvětlit rozptylem v proměnné Y, atd.)
-

Parciální korelace - příklad



Parciální korelace - příklad

- z obrázku je zřejmé, že nastává určité prolínání rozptylů všech tří proměnných
 - to znamená, že určitá část korelace mezi každou dvojicí proměnných (např. X a Y) souvisí s korelacemi těchto dvou proměnných se třetí proměnnou (tj. např. X se Z a Y se Z)
 - tj. z 25% společného rozptylu proměnných X a Y se určitá část (odhadem z obrázku více než polovina) prolíná s rozptylem proměnné Z
-

Parciální korelace

- výpočet parciální korelace nám umožní „změřit“ tuto oblast překrývajících se rozptylů přesně
 - a tak určit, jaká by byla korelace mezi dvěma proměnnými v případě, že by (hypoteticky) ani jedna z nich nekorelovala s touto třetí proměnnou (nebo také můžeme říct – kdyby byly hodnoty třetí proměnné konstatní)
-

Parciální korelace - příklad

□ výpočet parciální korelace mezi X a Y s kontrolou proměnné Z

$$\square r_{XY \cdot Z} = \frac{(r_{XY} - (r_{XZ}r_{YZ}))}{\sqrt{(1 - r_{XZ}^2)(1 - r_{YZ}^2)}}$$

$$\square r_{XY \cdot Z} = \frac{(0.50 - (0.50)(0.50))}{\sqrt{(1 - 0.25)(1 - 0.25)}}$$

$$\square r_{XY \cdot Z} = 0.25/0.75 = \mathbf{0.33}$$

$$\square (r_{XY \cdot Z}^2 = 0.11)$$

Parciální korelace

- příklad „ze života“
 - chceme zjistit, jaký je vztah mezi počtem bodů ze závěrečného testu ze statistiky (Y) a celkovým počtem hodin stráveným během semestru studiem (X)
 - zjistíme, že $r_{XY} = \mathbf{0.20}$
-

Parciální korelace

- zajímá nás, jak je tento vztah ovlivněn třetí proměnnou – strachem studenta ze zkoušky ze statistiky
 - zjistíme, že
 - $r_{XZ} = \mathbf{0.80}$ (tj. čím větší strach, tím více se student připravoval)
 - a $r_{YZ} = \mathbf{-0.40}$ (tj. čím větší strach, tím horší výsledek testu)
-

Parciální korelace

$$r_{xy \cdot z} = \frac{0,20 - (0,80 \times (-0,40))}{\sqrt{(1 - 0,80^2)(1 - (-0,40)^2)}}$$

$$r_{xy \cdot z} = \frac{0,20 - (-0,32)}{\sqrt{0,36 * 0,84}}$$

$$r_{xy \cdot z} = \frac{0,52}{0,55} = 0,95$$

Parciální korelace

- parciální korelace mezi dobou studia a počtem bodů v testu s kontrolou míry strachu ze zkoušky je
 $r_{xy \cdot z} = 0.95$
 - tj. pokud „odstraníme“ vliv strachu, je vztah mezi dobou strávenou přípravou na zkoušku a jejím výsledkem mnohem těsnější (0.20 vs 0.95)
-

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**
-

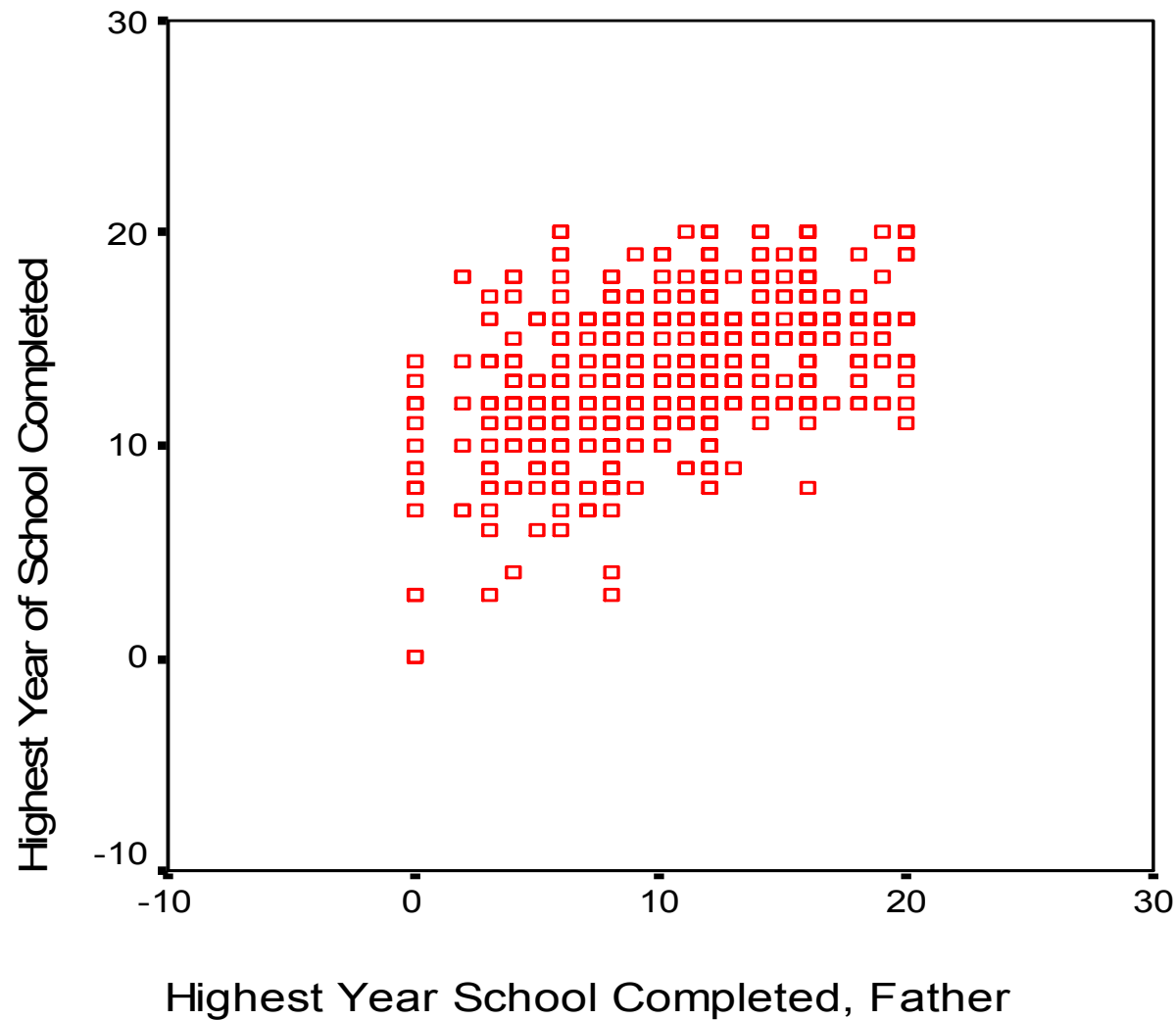
Regresní analýza

- 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?
-

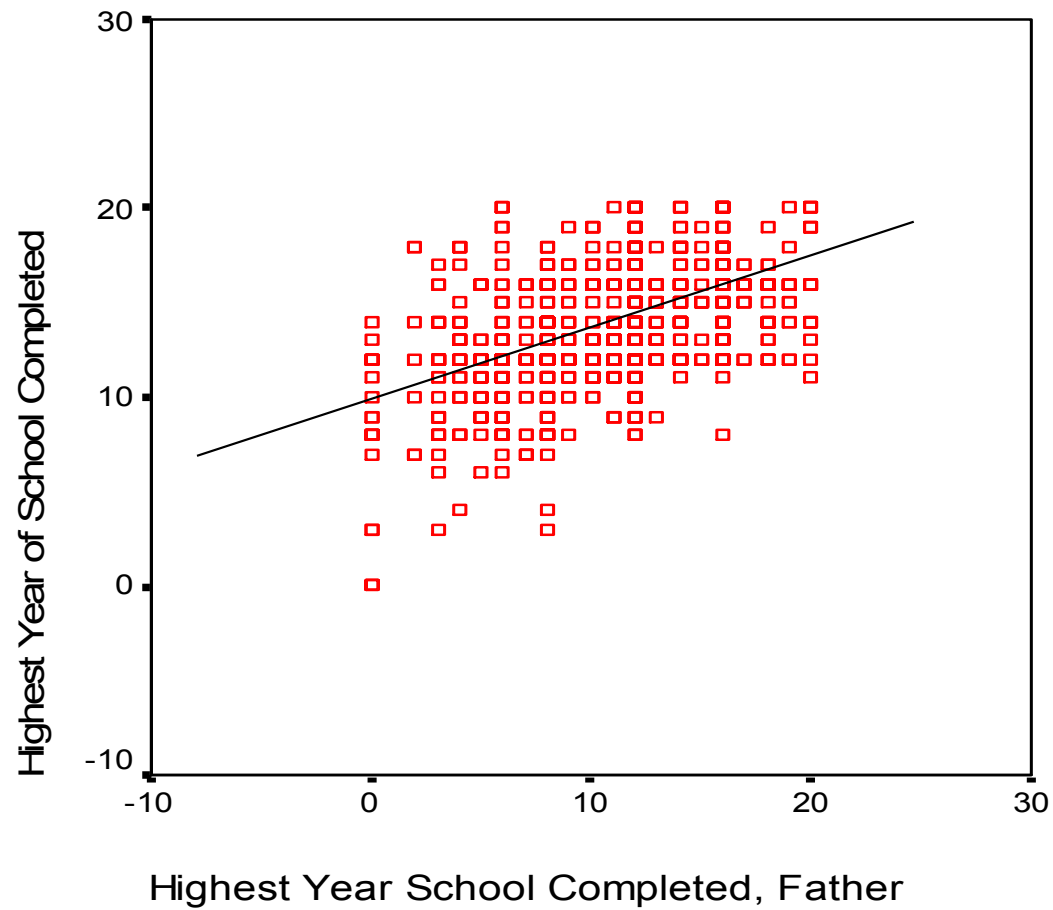
Jednoduchá regresní analýza



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 v SPSS

Coefficients^a

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
1 (Constant)	9.926	.219		45.260	.000
Highest Year School Completed, Father	.322	.019	.463	17.050	.000

a. Dependent Variable: Highest Year of School Completed

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)
- regresní rovnice se dvěma prediktory

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

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 v SPSS

Coefficients^a

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	
	B	Std. Error	Beta			
1	(Constant)	10.675	.271		39.408	.000
	Highest Year School Completed, Father	.296	.019	.427	15.225	.000
	Number of Brothers and Sisters	-.128	.028	-.129	-4.595	.000

a. Dependent Variable: Highest Year of School Completed

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** (angl. R square)
-

Výstup v SPSS

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.479 ^a	.229	.228	2.512

a. Predictors: (Constant), Number of Brothers and Sisters, Highest Year School Completed, Father

Vícenásobná regresní analýza

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

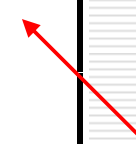
Testování hypotéz v regresní analýze

- 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 v SPSS

ANOVA^b

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1992.838	2	996.419	157.949	.000 ^a
	Residual	6693.297	1061	6.308		
	Total	8686.134	1063			



a. Predictors: (Constant), Number of Brothers and Sisters, Highest Year School Completed, Father

b. Dependent Variable: Highest Year of School Completed

Výstup v SPSS

Coefficients^a

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	
	B	Std. Error	Beta			
1	(Constant)	10.675	.271		39.408	.000
	Highest Year School Completed, Father	.296	.019	.427	15.225	.000
	Number of Brothers and Sisters	-.128	.028	-.129	-4.595	.000

a. Dependent Variable: Highest Year of School Completed



Reziduály

- výsledkem regresní analýzy jsou **predikované skóry** (na základě regresní rovnice)
 - z nich je možno odvodit **reziduální skóry** – rozdíl mezi skutečnou a predikovanou hodnotou proměnné
-

Předpoklady regresní analýzy

- skóry v proměnných jsou nezávislými měřeními
 - 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 mulikolinearity)
 - pokud ano, je vhodné buď některou z nich vyřadit, nebo z nich vytvořit např. faktorovou analýzou jeden skór
-

Předpoklady regresní analýzy

- rozdělení hodnot reziduálů je normální
 - zkontrolovat analýzou reziduálů – histogramem, pravděpodobnostním grafem
 - dostatečně velký počet osob ve výběru vzhledem k počtu prediktorů v modelu (nejméně 10-20x více osob než prediktorů)
-

Příklad prezentace výsledků

TABLE 3. Multiple Regression Analysis: The Effects of IQ and PD scores on Motor Assessment Battery for Children (MABC) Total Score in 11-12-year-old Children ($n = 150$)

Dependent variable: MABC total score

	Standardized coefficient (β)	<i>p</i>-value
IQ	-.346	.000
Paranoid	-.046	.704
Borderline	.060	.674
Dependent	-.277	.043
Histrionic	.110	.475
Obsessive-compulsive	-.034	.767
Schizotypal	-.185	.198
Schizoid	.113	.226
Narcissistic	-.041	.793
Conduct	.145	.128
Avoidant	.516	.001
Passive-Aggressive	-.046	.756
Depressive	.108	.412

Note: $F = 5.04$; $p < .001$; $R^2 = .33$, adjusted $R^2 = .26$

Zápis výsledků - příklad

- Regresní analýzou bylo zjištěno, že počet let formálního vzdělání respondenta je ovlivněn především vzděláním otce ($\beta=0,43$), zčásti také počtem sourozenců respondenta ($\beta=-0,13$). Vzdělání otce a vzdělání respondenta je v pozitivním vztahu; naopak čím vyšší počet sourozenců, tím nižší vzdělání respondenta. Regresní model vysvětloval celkem 23% rozptylu v počtu let vzdělání respondenta ($F=157,9$, $p<0,001$).
-

Literatura

- Hendl, kapitoly 9 a 10
-