Lineární regresní analýza

Regresní analýza

- výsledkem regresní analýzy je **model vztahu mezi** dvěma nebo více **proměnnými**
- umožňuje přesněji než korelace popsat tvar vztahu mezi proměnnými
- snažíme se z hodnot 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 také jako **regresand**, prediktor jako **regresor**
- predikovaná proměnná se označuje **Y**, prediktory $\mathbf{X_1}$, $\mathbf{X_2}$ …$\mathbf{X_n}$
- pouze 1 prediktor – **jednoduchá regrese**
- více prediktorů – **mnohonásobná regrese**

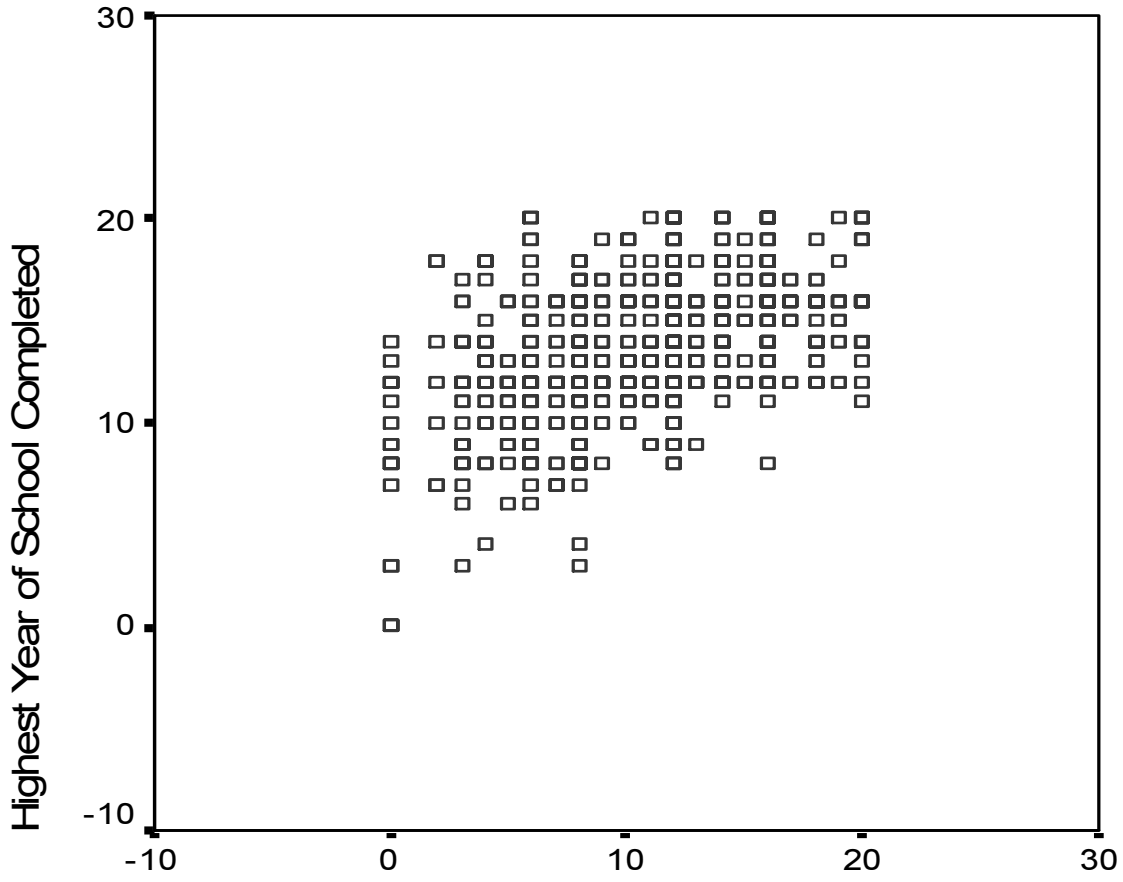
Regresní analýza

- 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?

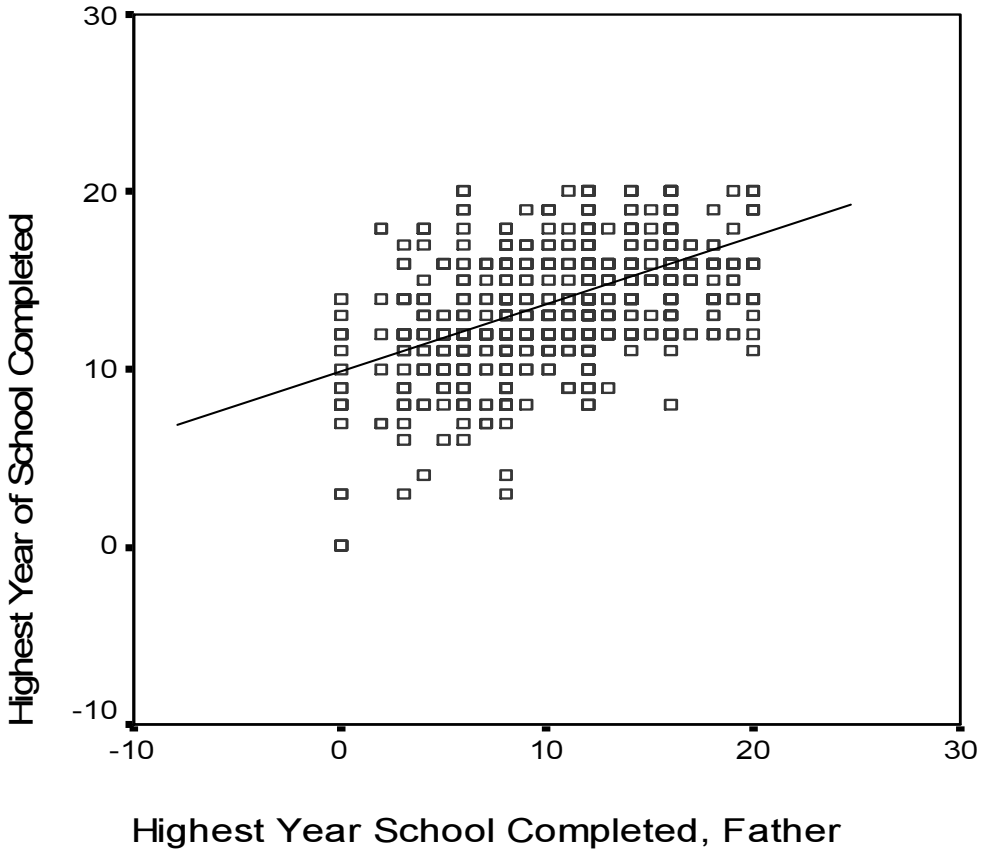
Jednoduchá regresní analýza



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

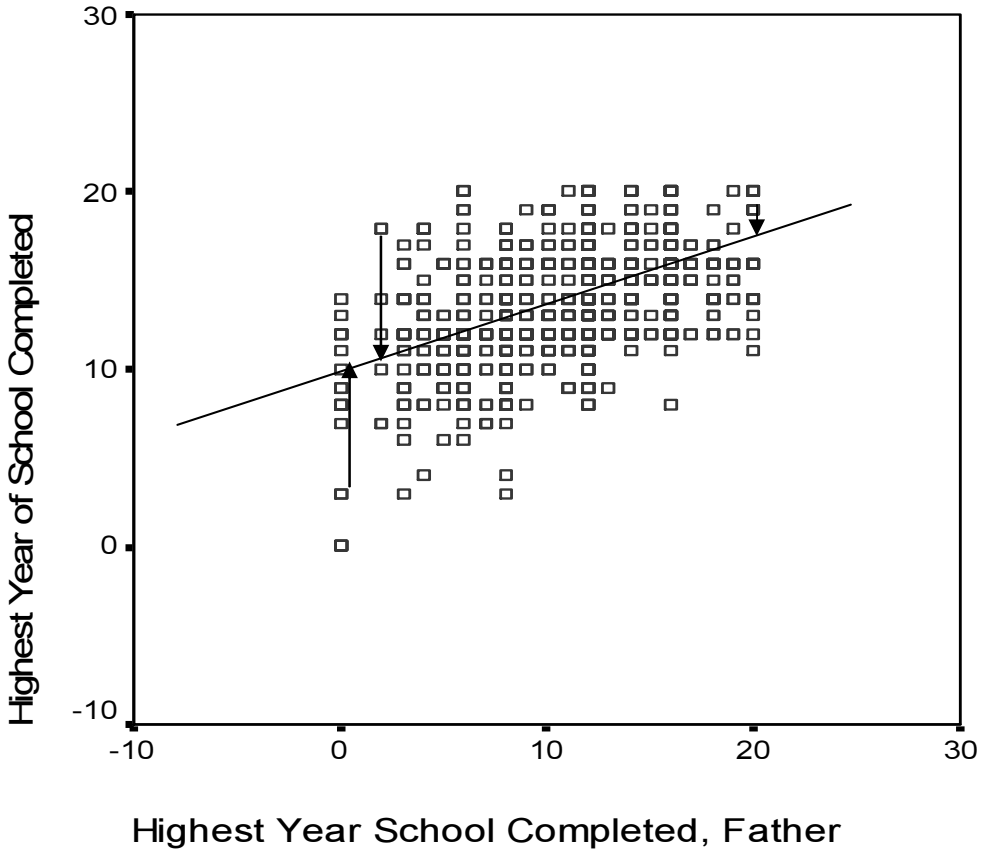
Jednoduchá regresní analýza



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



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

- rozdíl mezi naměřenou a predikovanou hodnotou = **reziduální hodnota predikce**, chyba predikce (e)

Jednoduchá regresní analýza

- $b = r_{xy} * (s_y/s_x)$
- $a = \bar{y} - b*\bar{x}$

Jednoduchá regresní analýza

- v příkladu vychází rovnice regresní přímky ***Y' = 9,93 + 0,32*X***
- pro děti otců s 0 lety vzdělání předpovídáme necelých 10 let vzdělání
- s každým dalším rokem otcova vzdělání předpovídáme 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

Jednoduchá regresní analýza

- pokud proměnné standardizujeme pomocí směrodatných odchylek a průměrů na z-skóry, pak
  - regresní přímka prochází počátkem os
  - regresní koeficient se rovná korelačnímu koeficientu

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

Mnohonásobná 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 „očištěný“ od vlivů ostatních proměnných - počítáme tzv. **parciální** koeficienty)

Mnohonásobná regresní analýza

- **příklad** – kromě vzdělání otce *(X_1)* může mít na dosažené vzdělání vliv také počet dalších dětí v rodině *(X_2)*
- rovnice regresní přímky je

$$\boldsymbol{Y' = a + b_1X_1 + b_2X_2}$$

Mnohonásobná 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 dalších dětí v rodině, který je zřejmě mírně ovlivněn také vzděláním otce
- vliv počtu dětí v rodině je záporný – tj. čím více dětí, tím nižší vzdělání

Mnohonásobná regresní analýza

- mnohonásobná 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?

Mnohonásobná regresní analýza

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

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

Mnohonásobná 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

Mnohonásobná regresní analýza

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

Mnohonásobná regresní analýza

- koeficient **mnohonásobné determinace** –% vysvětleného rozptylu (závislé proměnné) lineární kombinací prediktorů; ozn. $\mathbf{R^2}$

Výstup v SPSS

Model Summary[b]

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

b. Dependent Variable: Highest Year of School Completed

Mnohonásobná regresní analýza

- u jednoduché regresní analýzy je **koeficient mnohonásobné 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

- 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)

Kontrolní otázky

- účel regresní analýzy
- obecná rovnice regresní funkce
- jak se interpretují regresní koeficienty
- co je to koeficient mnohonásobné korelace?
- předpoklady regresní analýzy

Literatura

- Hendl, kapitoly 7.3 a 10