

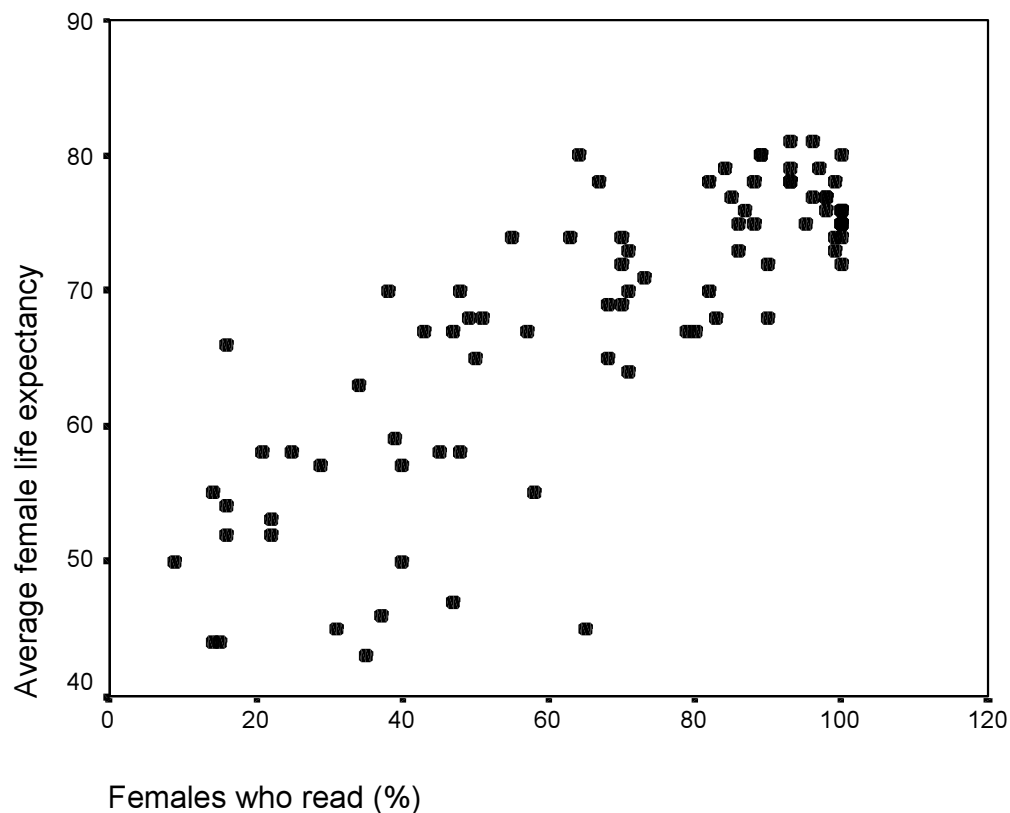
# LEKCE11 ZÁKLADY LINEÁRNÍ REGRESE - VZTAH SPOJITÝCH PROMĚNNÝCH

Velmi často nás zajímá jaký je VZTAH SPOJITÉ VELIČINY k ostatním veličinám, neboli to, co se ve statistice nazývá REGRESE.

Cílem REGRESE je vyjádřit VZTAH SPOJITÉ VELIČINY k ostatním veličinám prostřednictvím:

- REGRESNÍ ROVNICE (nějaké funkce), která by umožnila predikovat hodnotu určité proměnné na základě znalosti hodnoty jiné proměnné.
- REGRESNÍ ČÁRY, která je grafickým vyjádřením regresního vztahu (regresní rovnice) ve formě:
  - Regresní KŘIVKY (jako vyjádření nelineárního vztahu).
  - Regresní PŘÍMKY (jako vyjádření lineárního vztahu - **lineární regrese**).

**GRAPHS**  $\Rightarrow$  **SCATTERPLOT**  $\Rightarrow$  **SIMPLE**  
pro osu x „Average female life expectancy“  
pro osu x „Females who read“

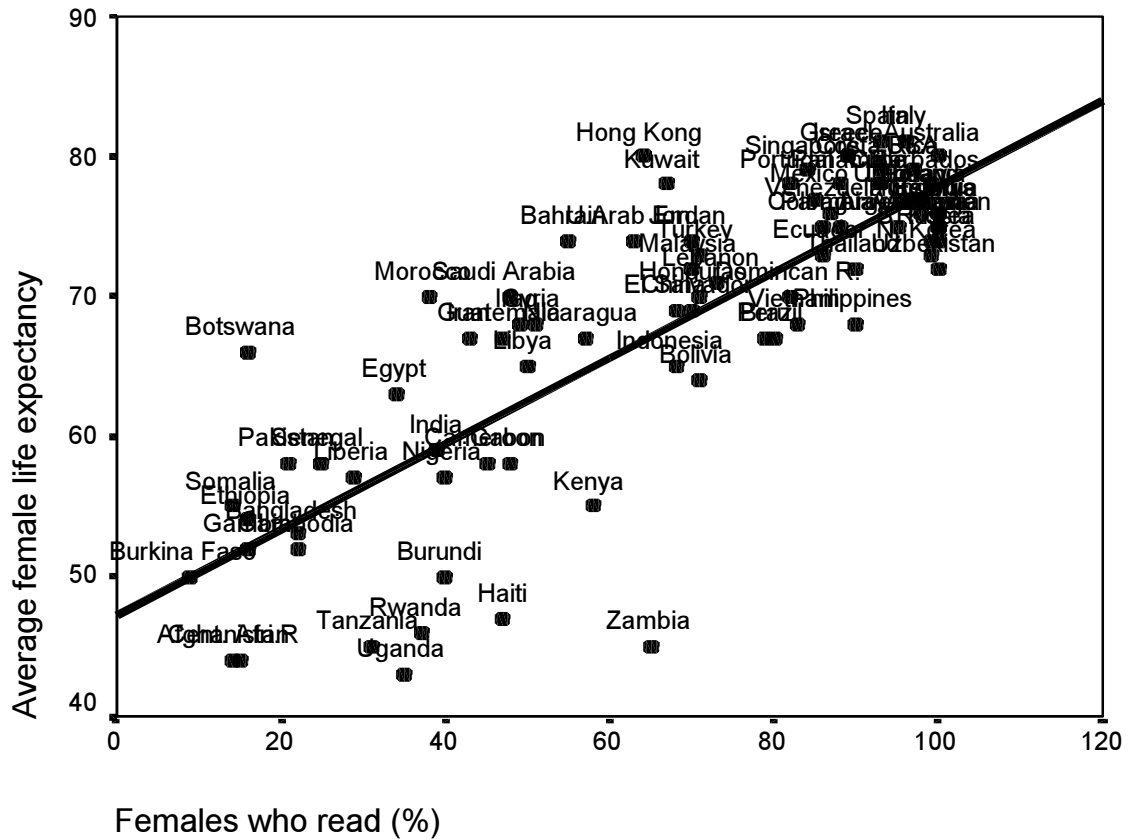


GRAPHS  $\Rightarrow$  SCATTERPLOT  $\Rightarrow$  SIMPLE

pro osu x „Average female life expectancy“  
 pro osu x „Females who read“

EDITOVAT GRAF:

- SCATTERPLOT OPTIONS (změnit CASE LABELS OFF na CASE LABELS ON)
- FIT LINE TOTAL (zadat)



## REGRESNÍ MODEL V JEDNODUCHÉ LINEÁRNÍ REGRESI

Základní informace, o kterou usilujeme je rovnice regresní (predikční) přímky. V případě JEDNODUCHÉ LINEÁRNÍ REGRESE je její obecný tvar:

$$y = b_0 + b_1x$$

Ve složitějším případě bychom mohli uvažovat i o tzv. náhodné chybě (random error)  $e$ , protože ne všechny body leží přímo na přímce:

$$y = b_0 + b_1x + e$$

$y$  = ZÁVISLE PROMĚNNÁ - závisle proměnná neboli výsledek (outcome). Je to ta proměnná, jejíž hodnotu chceme predikovat.

$x_1$  = NEZÁVISLE PROMĚNNÁ - neboli prediktor. Je to ta proměnná, jejíž hodnota slouží k predikci hodnoty  $y$ .

$b_0$  = Konstanta neboli INTERCEPT, bod ve kterém přímka protne osu  $y$  (hodnota  $y$  pro  $x_i = 0$ ).

$b_1$  = SMĚRNICE (sklon) přímky neboli SLOPE, která určuje o kolik jednotek se změní hodnota  $y$ , když se hodnota  $x$  změní o 1 jednotku

$e$  = náhodná chyba (variance nevysvětlitelné regresní rovnicí – zahrnutými nezávislými proměnnými).

### Může jít nejen o:

- JEDNODUCHOU LINEÁRNÍ REGRESI, kdy jde o vliv jediné nezávisle proměnné na sledovanou závislou proměnnou.

#### Příklad:

Souvislost mezi velikostí inflace (vyjádřené mírou inflace) a velikostí nezaměstnanosti (vyjádřené mírou nezaměstnanosti).

$$\text{míra nezaměstnanosti} = a + b \cdot \text{míra inflace}$$

ale též o:

- MNOHONÁSOBNOU LINEÁRNÍ REGRESI, kdy jde o současný kombinovaný vliv více nezávisle proměnných na sledovanou závislou proměnnou

#### Příklad:

Subjektivní hranice chudoby jako vyjádření

$$\text{SPL} = a + b_1 \cdot \text{příjem rodiny} + b_2 \cdot \text{počet dospělých v rodině} + b_3 \cdot \text{počet dětí v rodině}$$

## CÍL REGRESNÍ ANALÝZY

Najít koeficienty, které pomohou

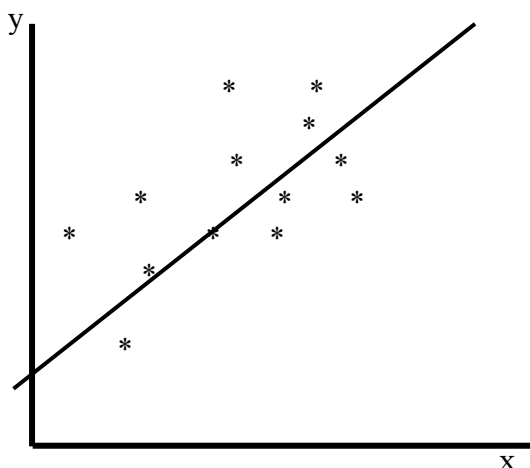
- odhadnout hodnotu predikované proměnné
- za pomoci hodnoty predikátoru pro nové případy.

Těmito koeficienty jsou již zmíněné:

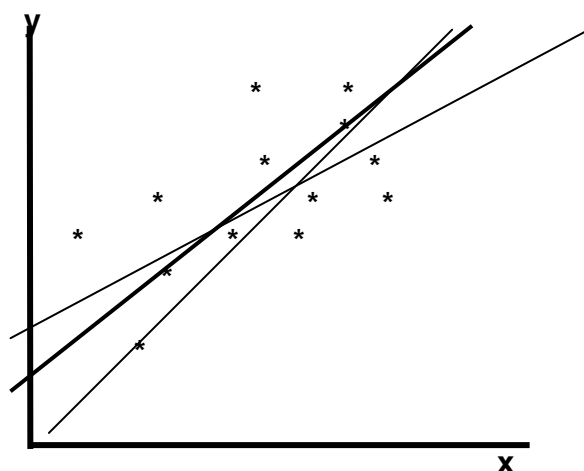
- Konstanta (intercept)  $b_0$  což je bod, ve kterém přímka protíná osu  $y$  ( $x=0$ ).
- Sklon (směrnice) přímky (slope)  $b_1$  (respektive  $b$ ) což je poměr mezi vertikální změnou a horizontální změnou podél přímky. Jinak řečeno je to změna  $y$ , která je způsobena změnou  $x$  o jednotku.

## PŘÍMKA JE MODELEM ROZLOŽENÍ DAT

V sociální realitě se nesetkáváme s případy ideální lineární regrese. Data jsou více či méně rozptýlena a linearita vztahu je vyjádřena tím, že přímka je jen vhodným modelem pro proložení daty (vyjadřuje tendenci v datech).



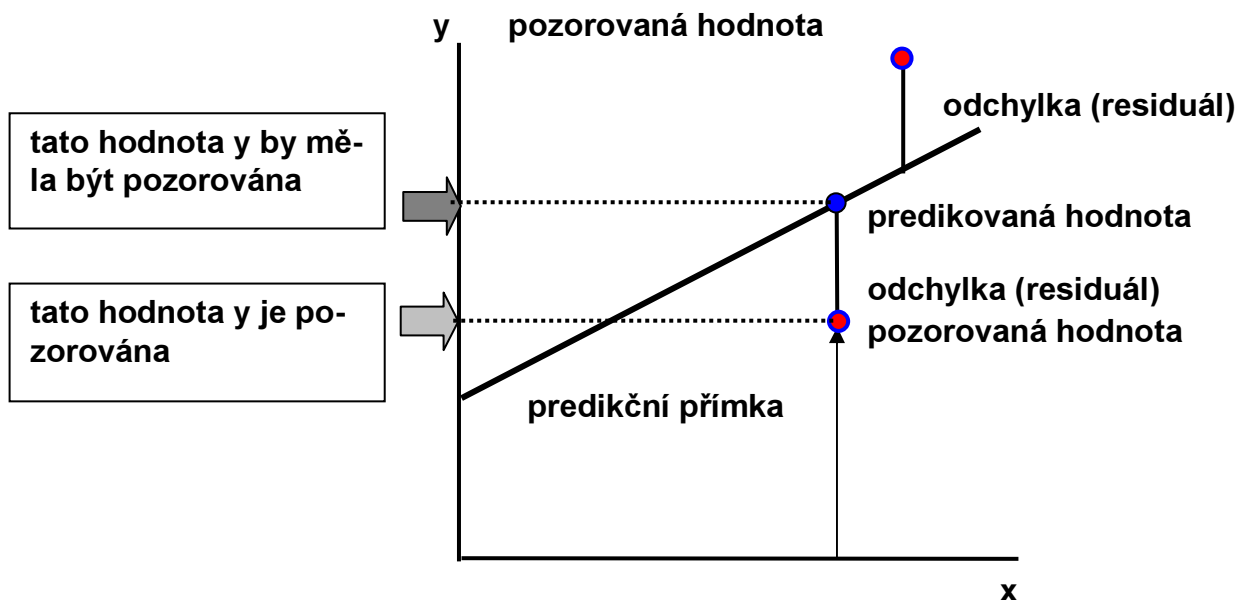
Daty lze proložit řadu přímek. Nejjednodušší způsob jak stanovit regresní přímku je metoda nejmenších čtverců odchylek (residuálů). Jen u jedné z přímek je totiž suma čtverců odchylek minimální.



## IDENTIFIKACE REGRESNÍ PŘÍMKY

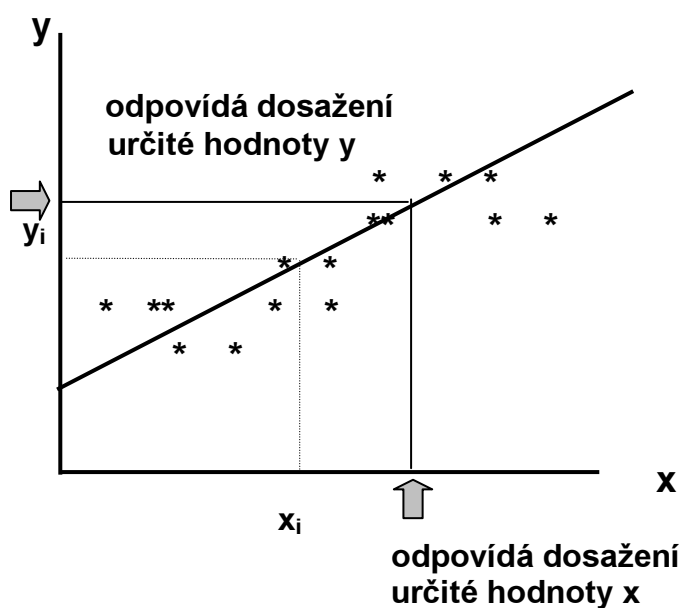
Nejjednodušší způsob identifikace regresní respektive predikční přímky představuje METODA NEJMENŠÍCH ČTVERCŮ

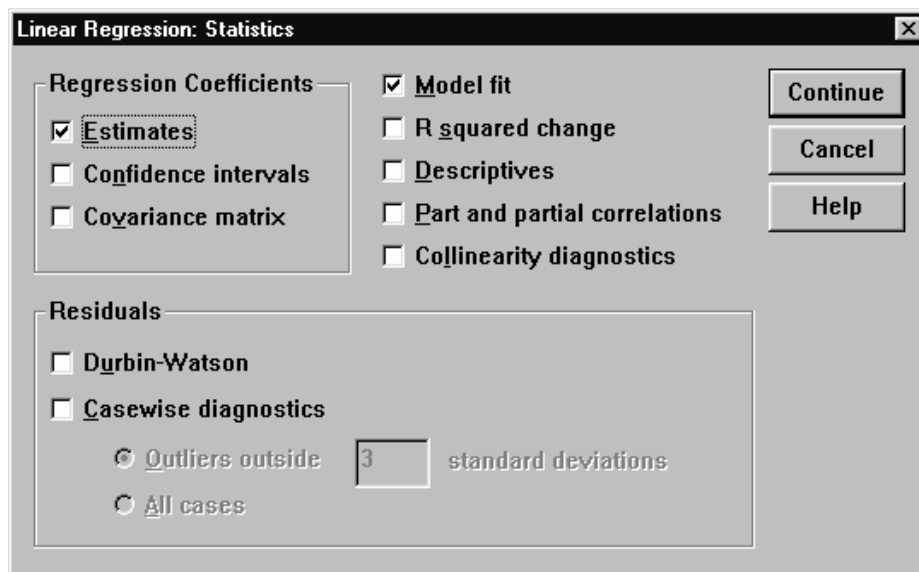
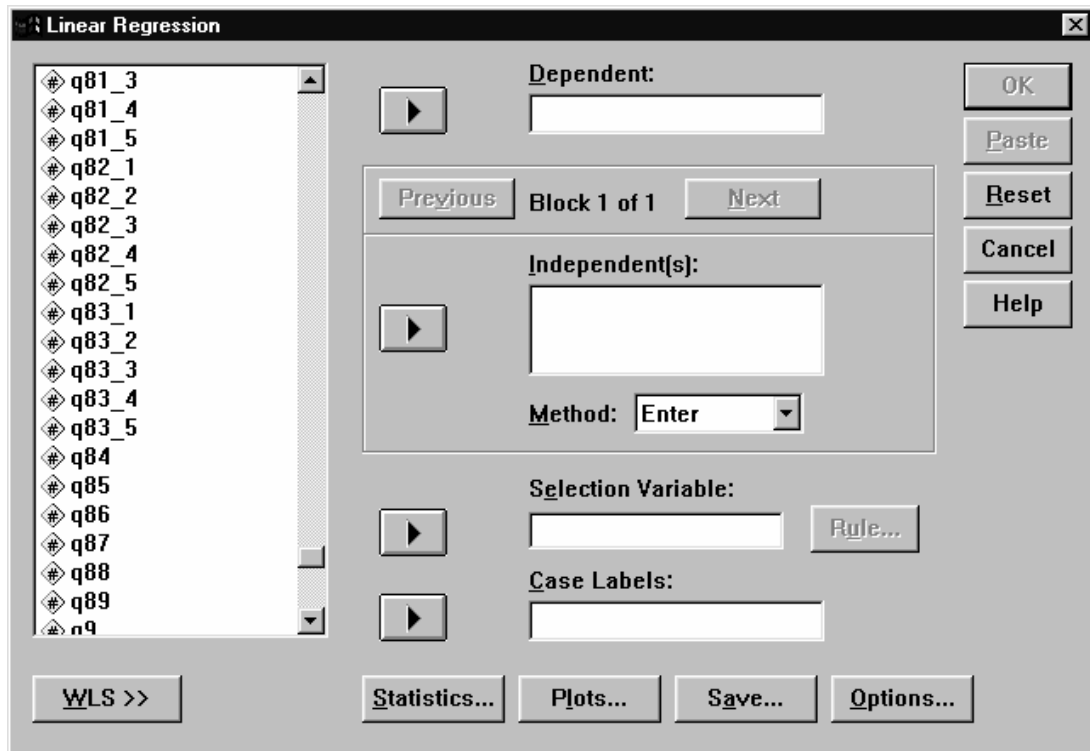
Predikované a pozorované hodnoty se liší (predikční přímka je pozorovanými hodnotami proložena) o tzv. RESIDUÁLY.



Součet čtverců všech residuálu musí být nejmenší možný.

PŘÍMKA NENÍ JEN MODELEM ROZLOŽENÍ DAT má též PREDIKČNÍ HODNOTU (predikční přímka). Z každé hodnoty  $x$  odvodíme příslušnou hodnotu  $y$ .





## PŘÍKLAD

## VÝPOČET KONSTANT

Coefficients<sup>a</sup>

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	89,985	1,765		50,995	,000
	BIRTHRAT Births per 1000 population, 1992	-,697	,050	-,968	-13,988	,000

a. Dependent Variable: LIFEEXPF Female life expectancy 1992

intercept

směrnice (slope)

URČENÍ ROVNICE:

$$y = 89,985 - 0,697 \cdot x$$