Jednoduchá lineární regrese

Úvod

Jednoduchá lineární regrese je statistická metoda, která slouží k tomu, abychom získali předpis, pomocí kterého budeme schopni předpovědět hodnotu jedné proměnné ze znalosti hodnoty jiné proměnné, pokud mezi těmito dvěma proměnnými existuje příčinná souvislost.

Rovnice přímky

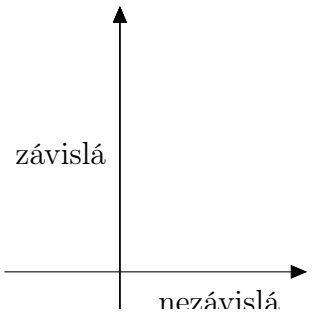
Základem jednoduché lineární regrese je rovnice přímky, která bývá nejčastěji reprezentována předpisem $y = mx + c$ nebo $y = a + bx$. Ve statistice používáme obvykle pro zápis této rovnice předpis obsahující parametry beta:

$$y = \beta_0 + \beta_1 x$$

Naším cílem je zkoumat vztah (lineární) proměnných x a y. Nazýváme je následovně:

- y: odezva nebo také závislá (vysvětlovaná) proměnná
- x: prediktor nebo též nezávislá (vysvětlující) proměnná

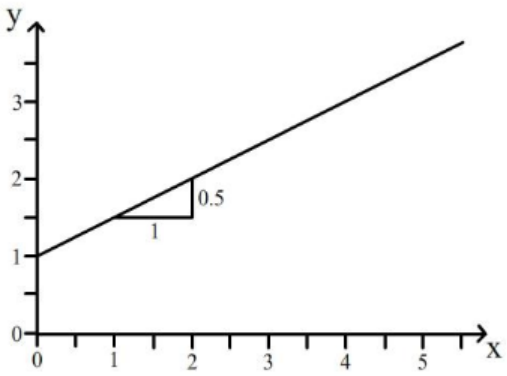
Je zvykem, že při grafickém znázornění dat vyneseme závislou proměnnou na osu y a nezávislou proměnnou na osu x:



parametry (koeficienty) β_0 a β_1 jsou neznámé konstanty, musíme je tedy odhadnout z dat. Jejich úlohy v rovnici přímky jsou následující:

- β_0 konstantní člen (udává posunutí přímky po ose y)
- β_1 směrnice přímky (určuje sklon přímky).

Demonstrujme si to na přímce $y = 1 + 0,5x$.



Předpoklady modelu

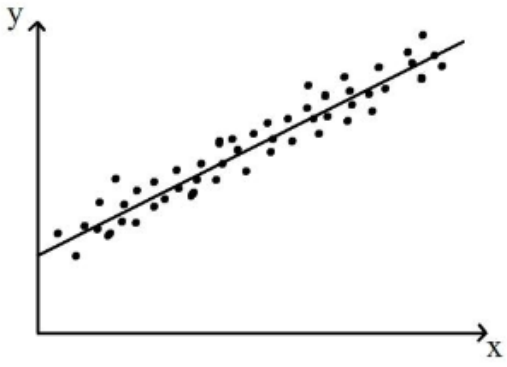
Při jednoduché lineární regresi jsme se zaměřili na to, abychom předpověděli odezvu pro i-té Y_i, s využitím konkrétní hodnoty prediktoru X_i. Tvar modelu je:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

tento model obsahuje deterministickou složku, která zahrnuje dva regresení koeficienty (β_0 a β_1) a náhodnou složku, kterou je reziduální (chybový) člen (ϵ_i).

Deterministická složka je ve tvaru přímky, pomocí ní získáme předpověď (střední/očekávané) odezvi při dané hodnotě prediktoru.

Reziduální část reprezentuje rozdíl mezi předpovězenou hodnotou a pozorovanou hodnotou v našem datovém souboru. Předpokládáme, že rezidua jsou nezávislá a pocházejí z normálního rozdělení s nulovou střední hodnotou a rozptylem σ^2 a tvoří přirozenou variabilitu souboru a mohou být chápána také jako odhad chybového členu. Náš datový soubor by měl tedy vypadat jako soubor bodů, které jsou náhodně rozmístěny kolem přímky s „konstantní proměnlivostí“.



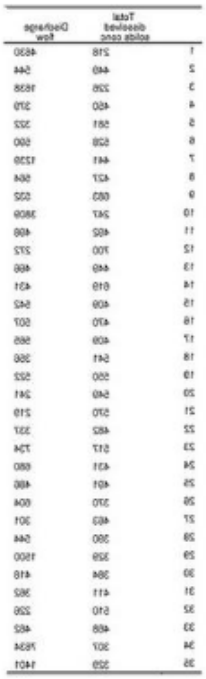
Deterministická složka je lineární funkce proměnné x s neznámými regresními koeficienty, které musíme odhadnout. Tyto koeficienty odhadujeme tak, aby model „co nejlépe“ popisoval náš datový soubor. Tohoto dosáhneme tak, že minimalizujeme součet čtverců reziduí (metoda nejmenších čtverců). Takto také získáme odhad rozptylu chybového členu, který je nezbytný pro testování významnosti regresních koeficientů a pro určení konfidenčních/ predikčních intervalů. Výše popsané úvahy jsou demonstrovány v dokumentu „Lineární regrese řešené příklady.pdf“.

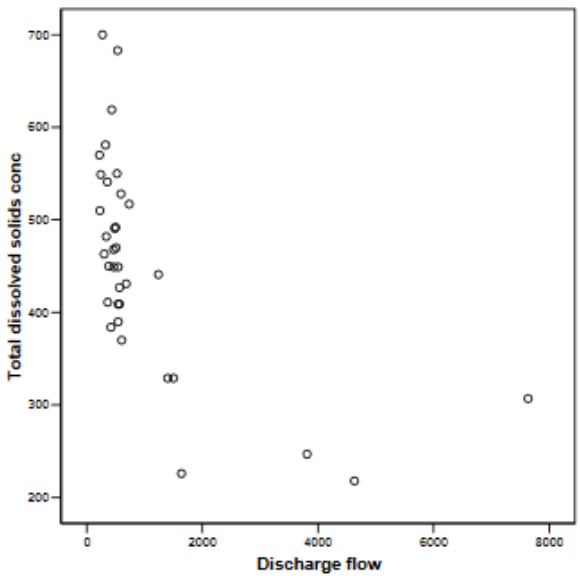
Příklad

Na tomto příkladu si ukážeme, jak mohou být porušeny předpoklady a nastíníme, co v takovém případě dělat. Předpokládejme, že naším cíle je odhadnout celkovou koncentraci (mg/L) rozouštěné pevné látky (TDS) k řece na základě průtoku (m^3/s). Máme k dispozici vzorek 35 pozorování nasbíraných za poslední rok.

Nejdříve pozorně prozkoumejme data:

- Je v datech rostoucí/ klesající trend nebo bychom jimi mohli proložit horizontální přímku?
- Je trend lineární nebo „zakřivený“?
- Je konstantní rozptyl kolem regresní přímky nebo se rozptyl systematicky mění s měnícími se hodnotami prediktoru?





Na základě bodového grafu předpokládáme, že v datech je klesající trend, který není lineární ale „zakřivený“. Rozptyl kolem hypotetické křivky se jeví přibližně konstantní.

Poznámka

Jednoduchá lineární regrese je vhodná pro modelování lineárního trendu pro data, která jsou stejně rozdělena kolem přímky. Pokud tomu tak v našem datovém souboru není, můžeme využít jisté „modelovací techniky“ nebo transformovat náš datový soubor abychom dosáhli výše zmíněných vlastností. Zaměřme se nyní na transformování datového souboru:

- jestliže je trend „zakřivený“ transformujme vysvětlující proměnnou x.
- jestliže rozptyl v našem datovém souboru není konstantní (může jít i o „zakřivený“ trend) můžeme buď transformovat vysvětlovanou proměnnou y nebo transformovat obě proměnné x i y.