Výpočet parametrov pre lineárnu regresiu

Spracované podľa:

- Prof. RNDr. Karel Rektorys, DrSc. a spolupracovníci: *Přehled užité matematiky*, 4. nezmenené vydanie, Praha 1981.
- http://en.wikipedia.org/wiki/Coefficient_of_determination

Lineárna závislosť typu $y = a + bx$

Predpokladajme, že v experimente meriame, ako veličina y závisí od veličiny x. Prvotným výsledkom merania nech je n usporiadaných dvojíc (x_i, y_i), $i = 1, 2, \ldots, n$. Ďalej predpokladajme, že medzi x a y by mala platiť všeobecná lineárna závislosť, teda

$$y = a + bx \tag{1}$$

Parametre tejto závislosti sú a (konštantný člen) a b (smernica). Sú to dve konštanty, ktoré často majú aj fyzikálny rozmer. Našou úlohou je pomocou nameraných údajov $\{x_i, y_i\}_{i=1}^n$ určiť tieto neznáme konštanty a tiež odhady s_a a s_b ich smerodajných odchýlok. Úloha nie je triviálna, pretože namerané body $\{x_i, y_i\}_{i=1}^n$ nikdy neležia presne na jednej priamke, a to jednak z dôvodov náhodného rozptylu nameraných údajov a často aj preto, že lineárny model (1) je len zjednodušením skutočnej situácie.

Najpoužívanejší spôsob určenia parametrov a a b a súvisiacich veličín je lineárna regresia metódou najmenších štvorcov. Z nej vyplývajú pre b a a nasledujúce vzťahy.

$$\boxed{b = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}} \qquad \boxed{a = \frac{1}{n}\sum_{i=1}^{n} y_i - b\,\frac{1}{n}\sum_{i=1}^{n} x_i} \tag{2}$$

Smerodajné odchýlky týchto parametrov sa počítajú zo vzťahov

$$\boxed{s_b = \frac{s}{\sqrt{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}}} \qquad \boxed{s_a = \sqrt{\frac{\frac{1}{n}\sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}}\; s} \tag{3}$$

kde s je vyjadrené výrazom

$$s = \sqrt{\frac{\chi^2}{n-2}} \tag{4}$$

χ^2 je súčet štvorcov rezíduí definovaný výrazom

$$\chi^2 = \sum_{i=1}^{n} \left[y_i - (a + bx_i)\right]^2 \tag{5}$$

Lineárna závislosť typu $y = bx$

Veľmi často je skúmaná situácia taká, že ak je veličina x nulová, tak s určitosťou vieme povedať, že aj veličina y musí byť tiež nulová. V takom prípade je konštantný člen a nulový a lineárnu regresiu môžme aplikovať na modelovú rovnicu jednoduchšieho tvaru

$$y = bx \tag{6}$$

V tomto jednoduchšom modeli treba určiť len smernicu b a ako dodatočné parametre aj jej smerodajnú odchýlku s_b a korelačný koeficient alebo koeficient determinácie. Výsledky sú

$$\boxed{b = \frac{\sum\limits_{i=1}^{n} x_i y_i}{\sum\limits_{i=1}^{n} x_i^2}} \tag{7}$$

Súčet štvorcov rezíduí je

$$\chi^2 = \sum_{i=1}^{n} (y_i - bx_i)^2 \tag{8}$$

Odhad odchýlky s potom je

$$s = \sqrt{\frac{\chi^2}{n-1}} \tag{9}$$

a odhad *smerodajnej odchýlky* smernice má tvar

$$\boxed{s_b = \frac{s}{\sqrt{\sum_{i=1}^{n} x_i^2}}} \tag{10}$$

Koeficient determinovanosti a korelačný koeficient

Výstupom z regresie zvyčajne býva aj parameter typicky označovaný R alebo jeho druhá mocnina. Nie vždy je však jasné, ako je R určené konkrétnym programom definované. Zvyčajne je to buď *koeficient determinovanosti* alebo *výberový korelačný koeficient.*

Koeficient determinovanosti

Existuje niekoľko definícií koeficientu determinovanosti. Najpoužívanejšia definícia (jeho druhej mocnicny) je

$$\boxed{\mathcal{R}^2 = 1 - \frac{\chi^2}{S_{\text{tot}}^2}} \tag{11}$$

kde χ^2 je súčet štvorcov rezíduí definovaný vo všeobecnosti ako

$$\chi^2 = \sum_{i=1}^{n} (y_i - f_i)^2 \tag{12}$$

Pritom $f_i = f(x_i)$, kde $f(x)$ je funkcia (teraz bližšie nešpecifikovaná), ktorú prekladáme nameranými bodmi. [V prípade lineárnej funkcie je $f(x) = a+bx$.] Len v ideálnom prípade, keď by namerané body presne zodpovedali modelu $y = f(x)$, by sme dostali zhodu hodnôt y_i a f_i. Hodnota S^2_{tot} je definovaná nasledovne:

$$S^2_{\text{tot}} = \sum_{i=1}^{n} (y_i - \overline{y})^2 \tag{13}$$

kde $\overline{y}$ je aritmetický priemer súboru hodnôt $\{y_i\}_{i=1}^n$.

Koeficient determinovanosti teda vo všeobecnosti závisí od toho, ako ďaleko (vo zvislom smere) sa namerané body nachádzajú od modelovej závislosti $f(x)$. Čím sú namerané hodnoty y_i bližšie k modelovým hodnotám f_i, tým viac sa hodnota $\mathcal{R}$ priblíži zdola k hodnote 1. Keď zmeníme typ prekladanej funkcie (napr. miesto priamky použijeme parabolu), tak sa hodnota $\mathcal{R}$ zmení. Koeficient determinovanosti je teda *modelovo závislý.*

Výberový korelačný koeficient

Výberový korelačný koeficient pre súbor usporiadaných dvojíc $\{x_i, y_i\}_{i=1}^n$ vypočítame zo vzťahu

$$\boxed{r_{xy} = \frac{\sum\limits_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum\limits_{i=1}^{n} (x_i - \overline{x})^2 \sum\limits_{i=1}^{n} (y_i - \overline{y})^2}}} \tag{14}$$

kde $\overline{x}$ a $\overline{y}$ sú aritmetické priemery súborov hodnôt $\{x_i\}_{i=1}^n$ a $\{y_i\}_{i=1}^n$. Z definície (14) je vidieť, že korelačný koeficient r_{xy} je *modelovo nezávislý*, čiže závisí výlučne od daných hodnôt $\{x_i, y_i\}_{i=1}^n$. Nezávisí od typu funkcie, ktorú prekladáme pomedzi dané body. Jeho hodnota nám vo všeobecnosti nenapovie, do akej miery sú body $\{x_i, y_i\}_{i=1}^n$ blízke navrhnutej modelovej funkcii (tu priamky). Hodnoty r_{xy} spadajú do intervalu $\langle -1; 1 \rangle$. Praktické je namiesto hodnoty r_{xy} udávať jeho druhú mocninu r^2_{xy}.

Vzťah medzi $\mathcal{R}$ a r_{xy}

Ako sme napísali vyššie, koeficient determinovanosti a korelačný koeficient sú vo všeobecnosti dva odlišné pojmy. V prípade, že naším modelom je lineárna funkcia typu $y = a + bx$, sa však hodnota koeficientu determinovanosti $\mathcal{R}$ zhoduje s hodnotou výberového korelačného koeficientu r_{xy}.

Praktické poznámky

Program Excel (funkcia LINEST) pre prípad modelovej funkcie $y = bx$ vypočíta hodnotu koeficientu determinovanosti (aktuálne jeho štvorca) podľa vzťahu

$$\mathcal{R}_{\text{spec}} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\left(\sum_{i=1}^{N} x_i^2\right)\left(\sum_{i=1}^{N} y_i^2\right)}} \tag{15}$$

čiže odlišne než je uvedené vyššie. Ohodnotenie týmto koeficientom však *neodporúčame uvádzať*, lebo jeho hodnoty sú príliš blízke 1 aj pre prípady, keď namerané body očividne neležia na prekladanej priamke.

V prípade všeobecnej lineárnej funkcie $y = a + bx$ nám LINEST v Exceli vypočíta koeficient determinovanosti podľa vzťahu (11).

Program OpenOffice.org vypočíta (pomocou svojej implementácie funkcie LINEST) hodnotu r_{xy}^2, čiže štvorec korelačného koeficientu, a to tak pre jednoduchšiu závislosť $y = bx$ ako aj pre všeobecnú priamku $y = a + bx$. Ako sme uviedli vyššie, hodnota $\mathcal{R}$ sa pre prípad všeobecnej priamky zhoduje s hodnotou r_{xy}. Preto pre prekladanie všeobecnej priamky dostaneme tak z Excelu ako aj z OpenOffice.org tú istú hodnotu koeficientu. Pre prekladanie funkcie $y = bx$ sa však hodnota koeficientu z OpenOffice.org líši od hodnoty z Excelu, čiže funkcia LINEST *v OpenOffice.org je definovaná čiastočne odlišne ako rovnomenná funkcia v Exceli.*

Tieto aj iné programy zvyčajne vypočítavajú druhé mocnicy koeficientov, pretože ich hodnoty sa od 1 líšia zvyčajne už na nižšom desatinnom mieste, čím sa uľahčí prečítanie takto vypísaných čísiel.