Egyenesillesztés

Bevezető:

Ha van N darab adatpontunk, melynek x és y koordinátái adottak, akkor a következőképpen lehet rájuk egyenest illeszteni. Legyenek x_i-k és y_i-k a pontok adott koordinátái (az i index 1-től N-ig fut ugye), az y_i-k hibái legyenek σ_i-k. (Az x_i-ről tegyük most fel, hogy nincsenek hibáik, vagyis őket pontosan ismerjük). Kérdés, hogy mi az az egyenes, ami a legjobban illeszkedik ezekre az adatpontokra? Az egyenes egyenlete: $y = ax + b$, tehát meg kéne mondani, hogy hogyan válasszuk az a meredekséget és a b tengelymetszetet úgy, hogy az adott N darab x_i, y_i pontot a lehető legjobban megközelítse ez az egyenes.

Ezt a legkisebb négyzetek módszerével szokás csinálni. Megnézzük, hogy az a meredekség, illetve a b tengelymetszet egy adott választásánál egy adott x_i pontra mennyi lenne az y, ha tökéletesen eltaláltuk volna az egyenessel az adatpontunkat: nyilván $ax_i + b$ lenne. A mért érték, y_i általában persze nem ennyi. Az eltérés: $y_i - (ax_i + b)$. Ez lehet pozitív, vagy negatív, mindenesetre a négyzete biztosan pozitív.

Azt mondjuk, hogy az az egyenes illeszkedik legjobban az adatokra, amelyre ezen eltérések négyzeteinek az összege a legkisebb. Még egy dolog van hátra: minél pontosabban ismerünk egy adatpontot, annál inkább valószínűtlen, hogy a feltételezett egyenes azt messze elkerüli, ezért nem az $[y_i - (ax_i + b)]^2$-eket, hanem ezeknek az y_i mért értékeknek megadott hibáikkal való súlyozott változatát adjuk össze. Tehát az, amit minimalizálni szeretnénk, a következő:

$$\chi^2(a,b) = \sum_{i=1}^{N} \frac{[y_i - (ax_i + b)]^2}{\sigma_i^2}.$$

Minél nagyobb ez, annál rosszabb az egyenes, és ez minél kisebb, annál inkább „eltaláltuk” az egyenessel az adatokat. A hibákkal való súlyozást így is lehet érteni: az a pont, amit csak nagy hibával (nagy σ_i-vel) ismerünk, kevésbé „számít”, mint az, amit pontosabban ismerünk.

Számolás:

Írjuk fel még egyszer $\chi^2(a,b)$-t, elvégezve a négyzetreemelést, valamint kicsit rendezve:

$$\chi^2(a,b) = \sum_{i=1}^{N} \frac{[y_i - (ax_i + b)]^2}{\sigma_i^2} = \sum_{i=1}^{N} \left\{ \frac{y_i^2}{\sigma_i^2} - \frac{2y_i\,(ax_i + b)}{\sigma_i^2} + \frac{(ax_i + b)^2}{\sigma_i^2} \right\} =$$

$$= \sum_{i=1}^{N} \left\{ \frac{y_i^2}{\sigma_i^2} - \frac{2ay_ix_i}{\sigma_i^2} - \frac{2by_i}{\sigma_i^2} + \frac{a^2x_i^2}{\sigma_i^2} + \frac{2abx_i}{\sigma_i^2} + \frac{b^2}{\sigma_i^2} \right\}.$$

Itt az összegzést külön-külön is írhatjuk:

$$\chi^2(a,b) = \sum_{i=1}^{N} \frac{y_i^2}{\sigma_i^2} - \sum_{i=1}^{N} \frac{2ay_ix_i}{\sigma_i^2} - \sum_{i=1}^{N} \frac{2by_i}{\sigma_i^2} + \sum_{i=1}^{N} \frac{a^2x_i^2}{\sigma_i^2} + \sum_{i=1}^{N} \frac{2abx_i}{\sigma_i^2} + \sum_{i=1}^{N} \frac{b^2}{\sigma_i^2}.$$

Sőt, az összegzésből ki lehet emelni az a-kat és a b-ket:

$$\chi^2(a,b) = \sum_{i=1}^{N} \frac{y_i^2}{\sigma_i^2} - 2a \sum_{i=1}^{N} \frac{y_ix_i}{\sigma_i^2} - 2b \sum_{i=1}^{N} \frac{y_i}{\sigma_i^2} + a^2 \sum_{i=1}^{N} \frac{x_i^2}{\sigma_i^2} + 2ab \sum_{i=1}^{N} \frac{x_i}{\sigma_i^2} + b^2 \sum_{i=1}^{N} \frac{1}{\sigma_i^2}.$$

Hogyan válasszuk meg tehát az egyenes a, b paramétereit, hogy ez a χ^2 mennyiség a lehető legkisebb legyen? Az a-t és b-t változtatva az összegek (szummák) itt konstansoknak számítanak: ha valaki

megadta a mért pontok x_i, y_i koordinátáit és σ_i hibáit, minden további nélkül kiszámíthatjuk az itt szereplő összegeket. Be is vezetünk rájuk jelöléseket:

$$S_{yy} \equiv \sum_{i=1}^{N} \frac{y_i^2}{\sigma_i^2}, \quad S_{xy} \equiv \sum_{i=1}^{N} \frac{y_i x_i}{\sigma_i^2}, \quad S_y \equiv \sum_{i=1}^{N} \frac{y_i}{\sigma_i^2}, \quad S_{xx} \equiv \sum_{i=1}^{N} \frac{x_i^2}{\sigma_i^2}, \quad S_x \equiv \sum_{i=1}^{N} \frac{x_i}{\sigma_i^2}, \quad S_1 \equiv \sum_{i=1}^{N} \frac{1}{\sigma_i^2}. \tag{1}$$

Ezzel a $\chi^2(a,b)$ minimalizálandó mennyiség:

$$\chi^2(a,b) = S_{yy} - 2aS_{xy} - 2bS_y + a^2 S_{xx} + 2abS_x + b^2 S_1.$$

Erre vagyunk kíváncsiak, hogy milyen a és b választással lesz a lehető legkisebb.

Emlékezzünk a másodfokú egyenletes tanulmányokra: egy $Y = AX^2 + 2BX + C$ egyenletű parabola minimumhelye $X_{min} = -B/A$, minimumértéke pedig $Y_{min} = C - B^2/A$. A $\chi^2(a,b)$ egyenlete a függvényében egy parabola:

$$\chi^2(a,b) = S_{xx}a^2 + 2\left(bS_x - S_{xy}\right)a + \left(S_{yy} - 2bS_y + b^2 S_1\right),$$

tehát ha b-t már megtaláltuk volna, akkor erre az adott b-re az az $a_{min}(b)$ hely, ahol ez az előző kifejezés a függvényében felveszi a minimumát, illetve a $\chi^2_{min}(b)$ minimális érték a következő lenne az előzőek alapján:

$$a_{min}(b) = -\frac{bS_x - S_{xy}}{S_{xx}}, \quad \chi^2_{min}(b) = -\frac{\left(bS_x - S_{xy}\right)^2}{S_{xx}} + \left(S_{yy} - 2bS_y + b^2 S_1\right).$$

Meg kell még néznünk, hogy ez a $\chi^2_{min}(b)$, amiben most már egy adott b értékre az a paramétert a lehető legjobban választottuk meg, milyen b-re lesz végül tényleg a lehető legkisebb. A $\chi^2_{min}(b)$ iménti kifejezését átalakítva:

$$\chi^2_{min}(b) = \left(S_1 - \frac{S_x^2}{S_{xx}}\right) b^2 + 2\left(\frac{S_x S_{xy}}{S_{xx}} - S_y\right) b + \left(S_{yy} - \frac{S_{xy}^2}{S_{xx}}\right).$$

Ez most meg a b paraméter függvényében láthatóan egy parabola: tehát a legjobb b választás ennek a minimuma:

$$b_{min} = \frac{S_y - \frac{S_x S_{xy}}{S_{xx}}}{S_1 - \frac{S_x^2}{S_{xx}}} = \frac{S_y S_{xx} - S_x S_{xy}}{S_1 S_{xx} - S_x^2}.$$

Ahhoz, hogy az a paraméter legjobb értékét is megkapjuk, ezt most vissza kell helyettesíteni az előző kifejezésbe, ahol $a_{min}(b)$-t kaptuk meg adott b-re; most már tudjuk, hogy mennyi legyen a b. Ezt elvégezve némi egyszerűsítés után az adódik, hogy:

$$a_{min} = \frac{S_1 S_{xy} - S_x S_y}{S_1 S_{xx} - S_x^2}.$$

Megvan tehát az a_{min} és a b_{min} értéke, amik a $\chi^2(a,b)$-t minimalizálják; az ezekkel a paraméterekkel meghatározott $y = a_{min}x + b_{min}$ egyenletű egyenes lesz az, amire tehát azt mondjuk, hogy a legjobban illeszkedik a megadott pontokra.

Az eredmények összefoglalása:

Összefoglalva: ha tehát adottak a mérési x_i, y_i és σ_i értékek, akkor azokból ki lehet számolni az S_1, S_x, S_y, S_{xx}, S_{xy} mennyiségeket a fentebbi módon (ld. az (1) egyenletet), és ezekből az előző két képlet alapján a_{min}-t és b_{min}-t; az ilyen paraméterű egyenes illeszkedik a legjobban az adatokra. Az a és a b paraméterek hibáit is meg lehet kapni, ennek levezetését majd később írom le. Összegyűjtve a képleteket:

$$a = \frac{S_1 S_{xy} - S_x S_y}{S_1 S_{xx} - S_x^2}, \quad b = \frac{S_y S_{xx} - S_x S_{xy}}{S_1 S_{xx} - S_x^2}, \quad \Delta a = \frac{1}{\sqrt{S_{xx}}}, \quad \Delta b = \frac{1}{\sqrt{S_1}}. \tag{2}$$