Regressziószámítás

A gyakorlatban gyakran vizsgálunk két mennyiség közötti összefüggést. Például tekintsük a főiskola hallgatói ξ testsúlyát és az előző félévi η kreditindexüket. n elemű véletlen mintát veszünk, kérdés, hogy van-e a két változó között lineáris összefüggés? A kapott adatok $(x_1, y_1), \ldots, (x_n, y_n)$, ezeket (x, y) pontdiagramon ábrázolhatjuk.

A már ismert **elméleti korrelációs együttható**

$$R = \frac{E(\xi - E\xi)(\eta - E\eta)}{D\xi D\eta} = \frac{E(\xi\eta) - E\xi E\eta)}{D\xi D\eta},$$

ennek becslése a Pearson-féle **tapasztalati korrelációs együttható**

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}}.$$

Könnyen látható, hogy

A korrelációs együtthatók abszolútértéke legföljebb 1.

A tapasztalati korrelációs együttható az elméleti korrelációs együttható konzisztens becslése.

Amennyiben az elméleti korrelációs együttható abszolút értéke 0, lineáris kapcsolat nincs (más függvénykapcsolat még lehet, függetlenség csak normális esetben következik); amennyiben 1, lineáris kapcsolat van a két változó között.

Milyen $r^2 = \frac{(\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y}))^2}{n^2 S_x^2 S_y^2}$ eloszlása kétváltozós normális populáció esetén? Mint ismeretes, $nr^2 S_y^2/\sigma_y^2$ és $n(1-r^2)S_y^2/\sigma_y^2$ független χ^2-eloszlásúak rendre 1 illetve $n-2$-szabadságfokkal. Ennélfogva,

$$r^2 = \frac{nr^2 S_y^2}{nS_y^2} = \frac{nr^2 S_y^2/\sigma_y^2}{nr^2 S_y^2/\sigma_2^2 + n(1-r^2)S_y^2/\sigma_y^2},$$

$\chi_1^2/(\chi_1^2 + \chi_{n-2}^2)$ alakú valószínűségi változó. Nyilván $r^2/(1-r^2)\ \chi_1^2/\chi_{n-2}^2$ alakú valószínűségi változó, és a

$$t = \sqrt{n-2}\frac{r}{\sqrt{1-r^2}}$$

statisztika (függetlenség esetén, de emlékezzünk vissza, normális esetben $R = 0$ maga után vonja a függetlenséget) $n-2$-szabadságfokú Student-eloszlású. Ennek alapján próbát konstruálhatunk a H_0: ξ és η korrelálatlanok nullhipotézisra a H_1:

korreláltak ellenében. Ha $t_c = F^{-1}_{n-2}((1+c)/2)$, ahol $F_{n-2}(x)$ az $n-2$-szabadságfokú Student eloszlás eloszlásfüggvénye, akkor elfogadva a nullhipotézist ha $|t| < t_c$, a próba terjedelme $1 - c$ lesz.

Ha a korrelációs együttható értéke nemnulla, eloszlása egyszerűen nem kezelhető. Fisher megmutatta, hogy a

$$z = \frac{1}{2} \ln \frac{1+r}{1-r}, \quad Z = \frac{1}{2} \ln \frac{1+R}{1-R}$$

transzformációval z eloszlása aszimptotikusan normális Z várható értékkel és $1/(n-3)$ szórásnégyzettel. Elég nagy mintanagyság esetén két minta korrelációs együtthatójának H_0 : egyezőségére H_1 : a korrelációs együtthatók különbözőek hipotézis ellenében kétmintás u-próbát konstruálhatunk. Az

$$u = \frac{z_1 - z_2}{\sqrt{1/(n_1 - 3) + 1/(n_2 - 3)}}$$

statisztika közelítőleg standard normális. Ha $u_c = \Phi^{-1}((1 + c)/2)$, akkor $|u| < u_c$ esetén elfogadva a nullhipotézist $1 - c$ terjedelmű próbát kapunk. Az egyoldali $H_1 : R_1 > R_2$ hipotézis ellenében $u_c = \Phi^{-1}(c)$ választással lesz a próba terjedelme $1 - c$.

A *regresszió* kifejezést először Sir Francis Galton használta 1886-ban, aki apák és elsőszülött fiaik testmagasságát vizsgálva azt találta, hogy ha apák egy csoportjában az átlag testmagasságtól való eltérés d *cm* akkor elsőszülött fiaik testmagasságának tekintetében az átlagtól való eltérés mindössze $(1/2)d$ *cm*, tendencia fedezhető fel az átlaghoz való visszatérésre, azaz regresszióra.

Képzeljük el a kétváltozós normális eloszlás

$$f(x, y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1 - R^2}} e^{-\frac{1}{2(1-R^2)}q(x,y)}$$

sűrűségfüggvénye által meghatározott felületet (az úgynevezett korrelációs felületet), itt az R paraméter éppen az elméleti korrelációs együttható. Ha elmetszük a felületet az x tengelyre merőleges síksereggel, mindegyik síkmetszet haranggörbe, várható értékeik egyenest határoznak meg, amelynek egyenlete $y = R\sigma_y x/\sigma_x$, ez az η regressziós egyenese ξ-re. A ξ regressziós egyenese η-ra analóg módon kapható, egyenlete $x = R\sigma_x y/\sigma_y$. Ezek $R = \pm 1$-re egybeesnek, $R = 0$-ra a tengelyek.

A **lineáris statisztikai modell** fogalmával folytatjuk. Legyen $D \subseteq \mathbb{R}$ halmaz, és legyen minden $x \in D$ esetén η_x $f_{\eta_x}(t)$ sűrűségfüggvényű valószínűségi változó

σ szórással. A D halmazon értelmezett $e(x) = ax + b$ függvényről azt mondjuk, hogy lineáris statisztikai modell, ha η_x várható értéke $e(x)$ minden $x \in D$-re.

A D halmazból megfigyelünk egy $x_1, \ldots, x_n$ mintát, és az $f_{\eta_{x_i}}$ sűrűségfüggvényű populációból egy egyelemű $\eta_{x_i} = y_i$ mintát. Tételezzük fel, hogy az η_{x_i}-k normális valószínűségi változók és teljesen függetlenek. Az ismeretlen a és b együtthatókra és a σ szórásra a megfigyelt (x_i, y_i) mintából következtetünk a maximum likelihood elvnek megfelelően. A likelihood függvény

$$L(a, b, \sigma^2) = \prod_{i=1}^{n} 1/\sqrt{2\pi}\sigma e^{(-1/2)(y_i - ax_i - b)^2)/\sigma^2},$$

$$\ln L(a, b, \sigma^2) = -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - ax_i - b)^2.$$

Ekkor

$$\frac{\partial \ln L}{\partial a} = \frac{1}{\sigma^2}\sum_{i=1}^{n} x_i(y_i - ax_i - b),$$

$$\frac{\partial \ln L}{\partial b} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - ax_i - b),$$

$$\frac{\partial \ln L}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(y_i - ax_i - b)^2.$$

Ezeket 0-val egyenlővé téve kapjuk a paraméterek maximum likelihood egyenleteit:

$$\sum_{i=1}^{n} x_i a + \sum_{i=1}^{n} 1b = \sum_{i=1}^{n} y_i,$$

$$\sum_{i=1}^{n} x_i^2 a + \sum_{i=1}^{n} x_i b = \sum_{i=1}^{n} x_i y_i,$$

az a, b együtthatókra vonatkozó **normálegyenletek**, illetve

$$\sum_{i=1}^{n}(y_i - ax_i - b)^2 = n\sigma^2.$$

A normálegyenletekből kapjuk a b és σ maximum likelihood pontbecslését:

Az $e(x) = ax + b$ regressziós egyenes paramétereinek maximum likelihood pontbecslése

$$\hat{a} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = r\frac{S_y}{S_x},$$

$$\hat{b} = \overline{y} - \hat{a}\overline{x},$$

$$\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{a}x_i - \hat{b})^2.$$

(Világos, hogy legalább 2 különböző x_i-nek lennie kell).

Belátható, hogy $\widehat{\sigma^2}$ torzított becslés, a korrigált becslés $\frac{n}{n-2}\widehat{\sigma^2}$.

Legyen $S_y^2 = (1/n)\sum_{i=1}^{n}(y_i - \overline{y})^2$ a megfigyelt értékek szórásnégyzete, a **teljes szórásnégyzet**, $\hat{S}^2 = (1/n))\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2$ a számított értékek szórásnégyzete, a **külső szórásnégyzet**, $S_{rez}^2 = (1/n))\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ a **reziduum szórásnégyzet**. Könnyen látható, hogy

$$S_{rez}^2 = S_y^2(1 - r^2).$$

Milyen a regressziós egyenes meredekségének eloszlása? $\hat{a}$ független normális eloszlású valószínűségi változók lineáris kombinációja, s mint ismeretes, ez normális eloszlású. Várható értéke (mivel $\sum_{i=1}^{n}(x_i - \overline{x}) = 0$)

$$E\hat{a} = \sum_{i=1}^{n}((x_i - \overline{x})/s)E\eta_{x_i} = \sum_{i=1}^{n}((x_i - \overline{x})/s)(ax_i + b) =$$

$$(a/s)\sum_{i=1}^{n}(x_i - \overline{x})^2 + (b/s)\sum_{i=1}^{n}(x_i - \overline{x}) = a,$$

vagyis:

A regressziós egyenes meredekségének maximum likelihood pontbecslése torzítatlan.

A szórásnégyzet

$$D^2\hat{a} = \sum_{i=1}^{n}((x_i - \overline{x})^2/s^2)D^2\eta_{x_i} = \sigma^2/s = \sigma^2/(nS_x^2).$$

Az $u = \sqrt{n}S_x(\hat{a} - a)/\sigma$ valószínűségi változó standard normális eloszlású. A normálegyenletekből egyszerű meggondolások alapján kapjuk, hogy

Az $y = ax + b$ lineáris modellben a teljes szórásnégyzet a belső és a reziduum szórásnégyzetek összege.

Azaz $S_y^2 = \hat{S}^2 + S_{rez}^2$. Minthogy $S_{rez}^2 = S_y^2(1 - r^2)$, így a $D = \hat{S}^2/S_y^2$ **determinációs együttható**ra kaptuk:

A determinációs együttható értéke egybeesik a tapasztalati korrelációs együttható négyzetével.

A determinációs együttható megmutatja, hogy a teljes szóródás mekkora hányada magyarázható a korrelációval.

Láttuk, hogy a szórásnégyzet $D^2\hat{a} = \sigma^2/(nS_x^2)$, és a $\sqrt{n}S_x(\hat{a}-a)/\sigma$ valószínűségi változó standard normális eloszlású. Továbbá,

$$n(1 - r^2)S_y^2/\sigma^2 = n(S_{rez}^2/S_y^2)S_y^2/\sigma^2 = nS_{rez}^2/\sigma^2$$

χ^2-eloszlású $n - 2$ szabadságfokkal. Meg lehet mutatni, hogy függetlenek, ennélfogva a

$$t = \sqrt{n-2}(\hat{a} - a)S_x/S_{rez}$$

statisztika Student-eloszlású $n - 2$-szabadságfokkal. Ha $t_c = F_{n-2}^{-1}((1 + c)/2)$, ahol $F_{n-2}(x)$ az $n - 2$-szabadságfokú Student eloszlás eloszlásfüggvénye, akkor $P(-t_c < t < t_c) = c$, és kapjuk az

$$(\hat{a} - t_c\frac{S_{rez}}{\sqrt{n-2}S_x}, \hat{a} + t_c\frac{S_{rez}}{\sqrt{n-2}S_x})$$

$100c\%$-os szintű konfidencia intervallumot az a meredekségre. Megjegyezzük, hogy a fenti konfidencia intervallum alkalmazásával t-próbákat is szerkeszthetünk, például a $H_0 : a = a_0$ nullhipotézisre a $H_1 : a \neq a_0$ ellenében.

Az $\hat{y}_i$ számított értékekre is szerkeszthető egyszerű de hosszadalmas számításokkal $100c\%$-os szinten konfidencia intervallum

$$\hat{y}_i \pm t_c S_{rez}^* \sqrt{1 + 1/n + (x_i - \overline{x})^2/(n-1)S_x^{*2}}$$

konfidencia korlátokkal, ahol $t_c = F_{n-2}^{-1}((1+c)/2)$ és $F_{n-2}(x)$ az $n-2$-szabadságfokú Student eloszlás eloszlásfüggvénye.

Abban az esetben, amikor nem tesszük föl az eloszlás ismeretét, a *legkisebb négyzetek elve* alkalmazható, amit 1805-ben Legendre javasolt asztronómiai megfigyelésekhez. Legyen $x_1, \ldots, x_{t-1}$ független változók, x_t a függő változó. Továbbá legyen $(x_{1j}, x_{2j}, \ldots, x_{t-1\,j}, x_{tj})$ $(j = 1, \ldots, n)$ a megfigyelt minta, a statisztikai modell az $x_t = b + a_1x_1 + \ldots + a_{t-1}x_{t-1}$ **regressziós hipersík**. Az a_i **együtthatók legkisebb négyzetek szerinti pontbecslései** azok az értékek, amelyekre

$$Q(a_1, \ldots, a_{t-1}, b) = \sum_{j=1}^{n}(x_{tj} - b - a_1x_{1j} - \ldots - a_{t-1}x_{t-1\,j})^2$$

minimális. A

$$\frac{\partial Q}{\partial a_1} = -2\sum_{j=1}^{n} x_{1\,j}(x_{t\,j} - b - a_1 x_{1\,j} - \ldots - a_{t-1}x_{t-1\,j})\ldots$$

$$\frac{\partial Q}{\partial a_{t-1}} = -2\sum_{j=1}^{n} x_{t-1\,j}(x_{t\,j} - b - a_1 x_{1\,j} - \ldots - a_{t-1}x_{t-1\,j})$$

$$\frac{\partial Q}{\partial b} = -2\sum_{j=1}^{n}(x_{t\,j} - b - a_1 x_{1\,j} - \ldots - a_{t-1}x_{t-1\,j})$$

parciális deriváltaknak el kell tünniük, ahonnan kapjuk a **normálegyenlete**ket:

$$\sum_j x_{1\,j}a_1 + \sum_j x_{2\,j}a_2 + \sum_j x_{3\,j}a_3 + \cdots + \sum_j x_{t-1\,j}a_{t-1} + \sum_j 1b = \sum_j x_{t\,j}$$

$$\sum_j x_{1\,j}^2 a_1 + \sum_j x_{1\,j}x_{2\,j}a_2 + \sum_j x_{1\,j}x_{3\,j}a_3 + \cdots + \sum_j x_{1\,j}x_{t-1\,j}a_{t-1} + \sum_j x_{1\,j}b = \sum_j x_{1\,j}x_{t\,j}$$

$$\sum_j x_{2\,j}x_{1\,j}a_1 + \sum_j x_{2\,j}^2 a_2 + \sum_j x_{2\,j}x_{3\,j}a_3 + \cdots + \sum_j x_{2\,j}x_{t-1\,j}a_{t-1} + \sum_j x_{2\,j}b = \sum_j x_{2\,j}x_{t\,j}$$

$$\ldots\ldots\ldots$$

$$\sum_j x_{t-1\,j}x_{1\,j}a_1 + \sum_j x_{t-1\,j}x_{2\,j}a_2 + \sum_j x_{t-1\,j}x_{3\,j}a_3 + \cdots$$

$$+ \sum_j x_{t-1\,j}^2 a_{t-1} + \sum_j x_{t-1\,j}b = \sum_j x_{t-1\,j}x_{t\,j},$$

amelyek lineáris algebrai vagy numerikus módszerekkel oldhatók meg.

Visszatérve az egyszeres regresszióhoz, nemcsak lineáris összefüggést kereshetünk, hanem, elméleti megfontolások vagy a pontdiagram alapján, kereshetjük az $y = b + a_1 h_1(x) + \ldots + a_{t-1}h_{t-1}(x)$ **regressziós görbét**, ahol a $h_i(x)$ függvények tetszőlegesen adottak. Legyen (x_j, y_j) $(j = 1, \ldots, n)$ a megfigyelt minta. A többszörös regressziónál alkalmazott megfontolásokat megismételve $x_{ij} = h_i(x_j)$ $(i = 1, \ldots, n\ j = 1, \ldots, t-1)$, $y = x_t$ és $x_{t\,j} = y_j$ helyettesítésekkel élve kapjuk azonnal a **normálegyenlete**ket:

$$\sum_j h_1(x_j)a_1 + \sum_j h_2(x_j)a_2 + \sum_j h_3(x_j)a_3 + \cdots + \sum_j h_{t-1}(x_j)a_{t-1} + \sum_j 1b = \sum_j y_j$$

$$\sum_j h_1(x_j)^2 a_1 + \sum_j h_1(x_j)h_2(x_j)a_2 + \sum_j h_1(x_j)h_3(x_j)a_3 + \cdots$$

$$+\sum_j h_1(x_j)h_{t-1}(x_j)a_{t-1}+\sum_j h_1(x_j)b=\sum_j h_1(x_j)y_j$$

$$\sum_j h_2(x_j)h_1(x_j)a_1+\sum_j h_2(x_j)^2a_2+\sum_j h_2(x_j)h_3(x_j)a_3+\cdots$$

$$+\sum_j h_2(x_j)h_{t-1}(x_j)a_{t-1}+\sum_j h_2(x_j)b=\sum_j h_2(x_j)y_j$$

$$\cdots\cdots$$

$$\sum_j h_{t-1}(x_j)h_1(x_j)a_1+\sum_j h_{t-1}(x_j)h_2(x_j)a_2+\sum_j h_{t-1}(x_j)h_3(x_j)a_3+\cdots$$

$$+\sum_j h_{t-1}(x_j)^2a_{t-1}+\sum_j h_{t-1}(x_j)b=\sum_j h_{t-1}(x_j)y_j.$$

Gyakorta alkalmazott speciális eset a polinomillesztés, amikor is $h_i(x)=x^i$, ekkor a **normálegyenletek** alakja

$$\sum_j x_ja_1+\sum_j x_j^2a_2+\sum_j x_j^3a_3+\cdots+\sum_j x_j^{t-1}a_{t-1}+\sum_j 1b=\sum_j y_j$$

$$\sum_j x_j^2a_1+\sum_j x_j^3a_2+\sum_j x_j^4a_3+\cdots+\sum_j x_j^ta_{t-1}+\sum_j x_jb=\sum_j x_jy_j$$

$$\sum_j x_j^3a_1+\sum_j x_j^4a_2+\sum_j x_j^5a_3+\cdots+\sum_j x_j^{t+1}a_{t-1}+\sum_j x_j^2b=\sum_j x_j^2y_j$$

$$\cdots\cdots$$

$$\sum_j x_j^ta_1+\sum_j x_j^{t+1}a_2+\sum_j x_j^{t+2}a_3+\cdots+\sum_j x_j^{2t-2}a_{t-1}+\sum_j x_j^{t-1}b=\sum_j x_j^{t-1}y_j.$$

Ha a regressziós felületet $x_t=h(b+a_1x_1+\ldots+a_{t-1}x_{t-1})$ alakban keressük, ahol $h(x)$ adott invertálható függvény, akkor $h^{-1}(x_t)=b+a_1x_1+\ldots+a_{t-1}x_{t-1}$, és alkalmazható a regressziós hipersíkra vonatkozó normálegyenletrendszer a függő változóra vonatkozó mintaértékek $h^{-1}(x_t)$ transzformációjával.