REGRESSZIÓANALÍZIS

Bolla Marianna

2020. május 12.

Az alapprobléma a következő: Az X, Y v.v. együttes eloszlásának ismeretében közelíteni szeretnénk Y-t X mérhető t fv.-évl legkisebb négyzetes értelemben:

$$\mathbb{E}(Y - t(X))^2 \to \min . \quad t - \text{ben.}$$

Tudjuk, hogy az optimumot az ún. *regressziós görbe* szolgáltatja, melynek egyenlete:

$$t_{opt}(x) = \mathbb{E}(Y \mid X = x),$$

azaz Y feltételes várható értéke a $X = x$ feltétel mellett. Amennyiben X, Y együttes eloszlása 2-dimenziós normális, a regressziós görbe egyenes lesz. Egyéb esetekben is szokták a a legkisebb négyzetes értelemben legjobb lineáris közelítést keresni, különösen ha az elméleti együttes eloszlás nem ismert, csak egy 2-dimenziós, folytonos eloszlásból vett minta áll rendelkezésünkre.

1. Egyváltozós lineáris regresszió

1.1. Elméleti megoldás

Tegyük fel, hogy az X, Y v.v.-k (általában ismeretlen) együtte eloszlása abszolút folytonos, továbbá a változók első, második és vegyes második momentumai léteznek, ezeket külön jelöljük is:

$$\mathbb{E}(X) = m_1,\ \mathbb{E}(Y) = m_2,\ \mathrm{Var}(X) = \sigma_1^2,\ \mathrm{Var}(Y) = \sigma_2^2,\ \mathrm{Cov}(X, Y) = c,\ \mathrm{Corr}(X, Y) = r,$$

feltehető, hogy $\sigma_1 > 0$.

Keressük az $l(x) = ax + b$ *regressziós egyenes*t, mellyel

$$h(a, b) = \mathbb{E}(Y - l(X))^2 = \mathbb{E}(Y - aX - b)^2 \to \min . \quad a, b - \text{ben.}$$

Ez egy kétváltozós szélsőérték feladat, a stacionárius megoldás az alábbi egyenletrendszerből kapható:

$$\frac{\partial h}{\partial a} = -2\mathbb{E}[(Y - aX - b)X] = 0$$
$$\frac{\partial h}{\partial b} = -2\mathbb{E}[Y - aX - b] = 0$$

(ui. a Cramér–Rao egyenlőtlenségnél tanult regularitási feltételek mellett a paraméter szerinti deriválás és az integrálást jelentő várható érték képzés felcserélhető), vagy ami ezzel ekvivalens:

$$a \cdot \mathbb{E}(X^2) + b \cdot \mathbb{E}(X) = \mathbb{E}(XY)$$
$$a \cdot \mathbb{E}(X) + b = \mathbb{E}(Y).$$

Az ismeretlenek a és b, az együtthatómátrix:

$$\mathbf{H} = \begin{pmatrix} \mathbb{E}(X^2) & \mathbb{E}(X) \\ \mathbb{E}(X) & 1 \end{pmatrix},$$

melynek determinánsa: $|\mathbf{H}| = \mathbb{E}(X^2) - \mathbb{E}^2(X) = \sigma_1^2 > 0$, így a Cramér-szabállyal:

$$a = \frac{\mathbb{E}(XY) - \mathbb{E}(X) \cdot \mathbb{E}(Y)}{\sigma_1^2} = \frac{c}{\sigma_1^2} = \frac{r\sigma_1\sigma_2}{\sigma_1^2} = r\frac{\sigma_2}{\sigma_1}.$$

Ezt a 2. egyenletbe helyettesítve:

$$b = \mathbb{E}(Y) - a\mathbb{E}(X) = m_2 - \frac{c}{\sigma_1^2}m_1.$$

A másodrendű deriváltakat tartalmazó Hesse-mátrix a stacionárius megoldás elyén szintén $\mathbf{H}$, ennek mindkét főminora pozitív, így a fenti a, b valóban lokális minimumot szolgáltat, ami a tartományok nyíltsága, és a differenciálhatósági feltételek teljesülése miatt globális minimumot is ad. A regressziós egyenes egyenlete tehát:

$$y = ax + b = \frac{c}{\sigma_1^2}(x - m_1) + m_2,$$

vagy még könnyebben megjegyezhető formában:

$$\frac{y - m_2}{\sigma_2} = r\frac{x - m_1}{\sigma_1}.$$

Az is látható, hogy a kovariancia (korreláció) előjele adja meg a regressziós egyenes iránytangensének előjelét.

Néhány szó a regresszió (=visszatérés) fogalom jelentéséről. Sir Francis Galton brit orvos a XIX. század második felében apa–fiú testmagasság kapcsolatát vizsgálta. Feltételezte, hogy $\sigma_1 = \sigma_2 = \sigma$. Akkor a fiú testmagassága (Y) az apa testmagasságával (X) a (6) összefüggés a következőképpen predikálható lineárisan:

$$Y = m_2 + r(X - m_1),$$

ahol r az X és Y közti korrelációt jelöli. Ha $|r| < 1$, akkor nyilván

$$|Y - m_2| < |X - m_1|.$$

Ebből látható, hogy az $r > 0$ esetben: amennyiben az apa az átlagnál magasabb, a gyerek is az lesz, de az utód magassága kevesebbel múlja felül az átlagot, mint a szülőé. Hasonlóan, ha az apa az átlagnál alacsonyabb, a gyerek is az lesz, de az utód magassága kevesebbel van alatta az átlagnak, mint a szülőé. (Az átlagtól való abszolút eltérésre negatív korreláció esetén is hasonló mondható.) Ezt a jelenséget nevezte el Galton az átlaghoz való „visszatérés”nek, latinul regressziónak.

A feladat átfogalmazható a következő *lineáris modell*el: az

$$Y = l(X) + \varepsilon = (aX + b) + \varepsilon \tag{1}$$

előállítást keresük úgy, hogy a hibatagot képező ε v.v.-ra $\mathbb{E}(\varepsilon^2)$ minimális legyen. Könnyen látható, hogy a minimumot adó a, b-vel vett $\varepsilon = Y - aX - b$ hibatagra $\mathbb{E}(\varepsilon) = 0$, továbbá a kovariancia bilinearitása miatt

$$\mathrm{Cov}(l(X), \varepsilon) = \mathrm{Cov}(aX + b, Y - (aX + b)) = \mathrm{Cov}(aX, Y - aX) =$$
$$= ac - a^2\sigma_1^2 = \frac{c}{\sigma_1^2}c - \frac{c^2}{(\sigma_1^2)^2}\sigma_1^2 = 0.$$

Tehát a (1)-beli összeg tagjai korrelálatlanok (2-dimenziós normális esetben függetlenek is egymástól), ezért szórásnégyzetük összeadódik:

$$\mathrm{D}^2(Y) = \mathrm{D}^2(l(X)) + \mathrm{D}^2(\varepsilon). \tag{2}$$

Ebből a minimum értékére

$$\mathrm{D}^2(\varepsilon) = \sigma_2^2 - \mathrm{D}^2(aX + b) = \sigma_2^2 - \frac{r^2\sigma_1^2\sigma_2^2}{\sigma_1^2} = \sigma_2^2(1 - r^2)$$

adódik. Innen is látható, hogy $|r| \leq 1$ és egyenlőség pontosan akkor teljesül, ha $\mathrm{D}^2(\varepsilon) = 0$, ami $\mathbb{E}(\varepsilon) = 0$ miatt csak úgy lehetséges, hogy $\varepsilon = 0$ (1 vsz.-el), azaz $Y = l(X)$ (1 vsz.-el). A másik extrém esetben, ha $r = 0$, akkor $c = 0$, $a = 0$, a regressziós egyenes meredeksége 0, így egyenlete $y = b = m_2$ lesz. A fentiekből az is következik, hogy

$$1 - r^2 = \frac{\mathrm{D}^2(\varepsilon)}{\mathrm{D}^2(Y)} \Longrightarrow r^2 = \frac{\mathrm{D}^2(l(X))}{\mathrm{D}^2(Y)}.$$

Tehát a korrelációs együttható négyzete megadja, hogy Y szórásnégyzetének hányad részét magyarázza a lineáris regresszió; nyilván annál „jobb" a lineáris regresszió, minél „nagyobb" az r^2 érték.

1.2. A regressziós együtthatók becslése mintából

Legyen most $(X_1, Y_1), \ldots, (X_n, Y_n)$ i.i.d. minta az (X, Y) háttérváltozóra. A fenti lineáris modell a, b együtthatóit becsüljük a legkisebb négyzetek módszerével:

$$h(a, b) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (Y_i - aX_i - b)^2 \to \min\,. \quad a, b - \text{ben}.$$

Miután az a, b szerinti parciális deriváltakat 0-val tesszük egyenlővé, a következő lineáris egyenletrendszert kapjuk:

$$a \cdot \sum_{i=1}^{n} X_i^2 + b \cdot \sum_{i=1}^{n} X_i = \sum_{i=1}^{n} X_i Y_i$$
$$a \cdot \sum_{i=1}^{n} X_i + b \cdot n = \sum_{i=1}^{n} Y_i.$$

A Cramér-szabály itt is alkalmazható, hiszen feltehető, hogy az együtthatómátrix determinánsa $n^2 S_X^2 > 0$. Teljesen hasonló számolással, mint az 1. részben kijön, hogy

$$\hat{a} = \frac{C}{S_X^2} = R\frac{S_Y}{S_X}, \quad \hat{b} = \bar{Y} - \hat{a}\bar{X} = \bar{Y} - R\frac{S_Y}{S_X}\bar{X}, \tag{3}$$

ahol S_X ill. S_Y jelöli X ill. Y (korrigálatlan) empírikus szórását, C ill. R pedig az X és Y közti empírikus kovarianciát ill. korrelációt. Mivel az egyenletrendszer megoldásakor ugyanazokat a lépéseket követjük el, mint az 1. részben, nem meglepő, hogy a és b becslésénél az elméleti első és második momentumok helyébe a mintából számolt empírikus momentumok lépnek, azaz momentum becslést kapunk.

A regressziós egyenes meredekségének előjelét most az empírikus korrelációs együttható, R határozza meg. Előzetesen vizsgálni szokták R segítségével az $r = 0$ null-hipotézist (ami 2-dimenziós normális esetben függetlenségvizsgálatot jelent), ezt itt most nem részletezzük. Ugyancsak végrehajtható, (2)-nak megfelelő szórásfelbontás is:

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \frac{[\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})]^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2} + \sum_{i=1}^{n}(Y_i - \hat{a}X_i - \hat{b})^2,$$

vagy a programcsomagokban szokásos jelöléssel:

$$SST = SSR + SSE,$$

ahol SST (Sum of Squares Total)$=nS_Y^2$ jelöli a függő változó (Y) „teljes ingadozását", azaz négyzetes eltéréseinek összegét saját átlagától, SSE (Sum of Squares due to Error) pedig a függő változó (Y) „ingadozását" jelöli a regressziós egyenes körül, azaz Y_i-k négyzetes eltéréseinek összegét a regressziós egyenesen levő $\hat{a}X_i + \hat{b}$ második koordinátákkal rendelkező pontoktól. Statisztika könyvek használják a következő jelölést is:

$$S_{XY} = \sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}), \quad S_{XX} = \sum_{i=1}^{n}(X_i - \bar{X})^2.$$

Nyilván $S_{XY} = nC$ és $S_{XX} = nS_X^2$. Előbbit *product moment*-nek is nevezik.

A többváltozós statisztika kurzuson tanulandó szórásfelbontási technikával kijön, hogy $SSR = nC^2/S_X^2$ a regresszió okozta szóródás (Sum of Squares due to linear Regression). Az 1. részben tárgyaltakhoz hasonlóan

$$R^2 = 1 - \frac{SSE}{SST} = \frac{SSR}{SST}.$$

Ez a mennyiség megmutatja, hogy a lineáris regresszió mennyit magyaráz a teljes varianciából, ezért R^2-et (az empírikus korrelációs együttható négyzetét) *meghatározottsági együttható*nak is szokták nevezni. A Fisher–Cochran tétel segítségével majd belátjuk (ld. többváltozós statisztika), hogy amennyiben mintánk 2-dimenziós normális eloszlásból származik, a fenti SSR, SSE mennyiségek χ^2-eloszlásúak, így szabadsági fokaikkal leosztott hányadosaikkal, mint F-eloszlású statisztikákkal próbákat hajthatunk végre a regressziós együtthatók és maga a regresszió szignifikanciájának vizsgálatára. Mindezt általánosabban, több független változó esetén tárgyaljuk majd többváltozós regresszió címszó

alatt. Egyváltozós regresszió esetében t-statisztikával is tesztelhetők hipotézisek, mint látni fogjuk a 3. rész végén.

Megjegyezzük, hogy lineáris regresszióra vezethetők vissza a következő approximációs feladatok:

(a) $Y \sim ae^{bX} \iff \ln Y \sim \ln a + bX$

(b) $Y \sim aX^b \iff \ln Y \sim \ln a + b \ln X$

(c) $Y \sim 1/(aX+b) \iff 1/Y \sim aX + b$

Mintából becslésnél (a) esetben az $(X_i, \ln Y_i)$, (b) esetben az $(\ln X_i, \ln Y_i)$, (c) esetben az $(X_i, 1/Y_i)$ $(i = 1, \ldots, n)$ 2-dimenziós mintákon hajtjuk végre a 2. részben leírt lineáris regressziót, és a végén néha még a becsült paramétert is transzformálni kell.

(d) *Polinomiális regresszió*: r-edfokú polinomiális regressziónál keressük az $Y \sim a_r X^r + \cdots + a_1 X + a_0$ közelítést legkisebb négyzetes értelemben:

$$\mathbb{E}(Y - a_r X^r - \cdots - a_1 X - a_0)^2 \to \min\,. \quad a_i - \text{kben.}$$

Az $a_r, \ldots, a_1, a_0$ együtthatók meghatározásához deriváljuk célfv.-ünket mindegyik együttható szerint parciálisan. A deriváltakat 0-val egyenlővé téve $r+1$ db. lineáris egyenletből álló egyenletrendszert kapunk, mely megoldható Cramér-szabállyal. A megoldásokba $2r$ rendig jönnek be momentumok (ezek létezését fel kell tenni). Amennyiben 2-dimenziós minta alapján szeretnénk becsülni az együtthatókat, a becslésekbe a megfelelő empírikus momentumok jönnek be ($2r$ rendig). Megjegyezzük, hogy itt az $r \geq 1$ egész szám értékét előre meg kell adni, bár egyes programcsomagokban elég a szóbajöhető maximális r-t megadni, és automatikusan megtörténik az ennél alacsonyabb fokú polinomokhoz való illesztés is az illeszkedés szignifikanciájának vizsgálatával együtt, ha a felhasználó kéri. (Az $r = 1$ eset a lineáris regresszió.)

Látni fogjuk, hogy a polinomiális regresszió többváltozós regresszióra is visszavezethető.

1.3. Tervezett (determinisztikus) megfigyelés

Az előző részben tárgyalt problémákat úgy kell elképzelni, hogy lineáris összefüggést keresünk pl. a testmagasság és a testsúly, vagy a vérnyomás és a koleszterinszint között, az egyiket kinevezzük függő, a másikat pedig független változónak; a *célváltozó* és *prediktor változó* elnevezés szerencsésebb. Mintánkat páciensek adják, akiken egyidejűleg mérünk meg két véletlen dolgot. Fizikai, kémiai kísérleteknél gyakran előfordul, hogy egy Y v.v. értékeit adott x beállításoknál mérik meg. Pl. különböző (előre beállított) hőmérsékleten nézik huzalok szakítószilárdságát, vagy előre beállított gyógyszer-dózisok mellett mérik kísérleti állatok vérében valamely kémiai anyag koncentrációját. A beállítás pontos (determinisztikus), a reakció azonban véletlen (az első esetben mérési hibával, a második esetben egyedi érzékenységgel terhelt). Amennyiben az x_i (hiba nélküli) beállítás mellett az Y_i mérési eredményt kapjuk $(i = 1, \ldots, n)$, lineáris modellünk a következő alakban írható:

$$Y_i = ax_i + b + \varepsilon_i \qquad (i = 1, \ldots, n), \tag{4}$$

ahol az ε_i hibatagok korrelálatlanok, továbbá feltesszük, hogy $\mathbb{E}(\varepsilon_i) = 0$, $\mathrm{D}^2(\varepsilon_i) = \sigma^2 < \infty$. Következésképpen $\mathbb{E}(Y_i) = ax_i + b$, $\mathrm{D}^2(Y_i) = \sigma^2$, és Y_i-k is korrelálatlanok ($i = 1, \dots, n$). Nagyon gyakran $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ fae., ún. *homoscedasztikus* hibák. Ilyenkor Y_i-k is függetlenek, persze várható értékük különböző.

Az a, b együtthatókat itt is a legkisebb négyzetek módszerével becsüljük:

$$h(a,b) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (Y_i - ax_i - b)^2 \to \min . \quad a, b - \text{ben}.$$

Parciális deriválással a és b becslésére alakilag az (3) képlet megfelelője jön ki:

$$\begin{aligned} \hat{a} &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \sum_{i=1}^{n} \frac{x_i - \bar{x}}{\sum_{j=1}^{n}(x_j - \bar{x})^2} Y_i = \sum_{i=1}^{n} k_i Y_i \\ \hat{b} &= \bar{Y} - \hat{a}\bar{x} = \sum_{i=1}^{n} \left(\frac{1}{n} - \bar{x} k_i \right) Y_i = \sum_{i=1}^{n} l_i Y_i, \end{aligned} \tag{5}$$

tehát *lineáris becslés*eket kaptunk ($\hat{a}$ és $\hat{b}$ az Y_i v.v.-k lineáris kombinációi a k_i ill. az l_i együtthatókkal).

Tétel (Gauss–Markov). A (4) lineáris modellben az a, b együtthatók (5) legkisebb négyzetes becslései lineárisak, torzítatlanok és az összes lineáris torzítatlan becslés közt a leghatásosabbak (minimális szórásúak). Angolul **BLUE** becslések (Best Linear Unbiased Estimate).
Bizonyítás.

- A *linearitás*t már láttuk.
- Fel fogjuk használni, hogy

$$\sum_{i=1}^{n} k_i = \frac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \sum_{i=1}^{n} (x_i - \bar{x}) = 0,$$

$$\sum_{i=1}^{n} k_i^2 = \frac{1}{[\sum_{i=1}^{n}(x_i - \bar{x})^2]^2} \sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\sum_{i=1}^{n} k_i x_i = \frac{\sum_{i=1}^{n}(x_i - \bar{x})x_i - \bar{x}\sum_{i=1}^{n}(x_i - \bar{x})}{\sum_{j=1}^{n}(x_j - \bar{x})^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\sum_{j=1}^{n}(x_j - \bar{x})^2} = 1.$$

Megjegyezzük, hogy a fenti egyenlőségeknek a formális bizonyítás mellett a szemléletes tartalma a következő: Az (5) összefüggés alapján $\sum_{i=1}^{n} k_i Y_i$ nem más, mint az (x_i, Y_i) $(i = 1, \dots, n)$ pontokhoz illesztett egyenes iránytangense. Így $\sum_{i=1}^{n} k_i$ tekinthető az $(x_i, 1)$ $(i = 1, \dots, n)$ pontokhoz illesztett egyenes iránytangensének, ami – a fv. konstans lévén – nyilván 0.

Hasonlóan, $\sum_{i=1}^{n} k_i x_i$ nem más, mint az (x_i, x_i) $(i = 1, \dots, n)$ pontokhoz illesztett egyenes iránytangense, ami – az egyenes az identitás fv. gráfja lévén – nyilván 1.

Végül $\sum_{i=1}^{n} k_i^2$ az (x_i, k_i) $(i = 1, \dots, n)$ pontokhoz illesztett egyenes iránytangense, ami – mivel

$k_i = x_i / \sum_{j=1}^{n}(x_j - \bar{x})^2$+egy konstans – egyenlő $1/\sum_{j=1}^{n}(x_j - \bar{x})^2$-tel.

- A *torzítatlanság* bizonyítása: a fentiek miatt

$$\mathbb{E}(\hat{a}) = \mathbb{E}(\sum_{i=1}^{n} k_i Y_i) = \sum_{i=1}^{n} k_i \mathbb{E}(Y_i) = \sum_{i=1}^{n} k_i (a x_i + b) =$$
$$= a \sum_{i=1}^{n} k_i x_i + b \sum_{i=1}^{n} k_i = a \cdot 1 + b \cdot 0 = a,$$

és

$$\mathbb{E}(\hat{b}) = \mathbb{E}(\sum_{i=1}^{n} l_i Y_i) = \sum_{i=1}^{n} l_i \mathbb{E}(Y_i) = \sum_{i=1}^{n} \left(\frac{1}{n} - \bar{x} k_i \right) (a x_i + b) =$$
$$= \frac{1}{n} a n \bar{x} - \bar{x} a \sum_{i=1}^{n} k_i x_i + \frac{1}{n} n b - \bar{x} b \sum_{i=1}^{n} k_i = b.$$

- A *hatásosság* bizonyítása lineáris, torzítatlan becslések körében:

$$\mathrm{D}^2(\hat{a}) = \sum_{i=1}^{n} k_i^2 \mathrm{D}^2(Y_i) = \sigma^2 \sum_{i=1}^{n} k_i^2 = \frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}, \tag{6}$$

és

$$\mathrm{D}^2(\hat{b}) = \sum_{i=1}^{n} l_i^2 \mathrm{D}^2(Y_i) = \sigma^2 \sum_{i=1}^{n} l_i^2 = \sigma^2 \sum_{i=1}^{n} \left(\frac{1}{n} - \bar{x} k_i \right)^2 =$$
$$= \sigma^2 \left(\frac{1}{n} + \bar{x}^2 \sum_{i=1}^{n} k_i^2 - 2\bar{x} \frac{1}{n} \sum_{i=1}^{n} k_i \right) = \tag{7}$$
$$= \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right).$$

Legyen most $\tilde{a} = \sum_{i=1}^{n} c_i Y_i$ tetszőleges lineáris, torzítatlan becslés a-ra. De

$$\mathbb{E}(\tilde{a}) = \sum_{i=1}^{n} c_i (a x_i + b) = a \sum_{i=1}^{n} c_i x_i + b \sum_{i=1}^{n} c_i = a$$

csak úgy lehetséges, hogy

$$\sum_{i=1}^{n} c_i x_i = 1 \quad \text{és} \quad \sum_{i=1}^{n} c_i = 0.$$

Legyen $d_i := c_i - k_i$. Ezzel és az előbbiek figyelembevételével

$$\sum_{i=1}^{n} k_i d_i = \sum_{i=1}^{n} k_i (c_i - k_i) = \sum_{i=1}^{n} c_i \frac{x_i - \bar{x}}{\sum_{j=1}^{n} (x_j - \bar{x})^2} - \sum_{i=1}^{n} k_i^2 =$$
$$= \frac{1}{\sum_{j=1}^{n} (x_j - \bar{x})^2} \left[\sum_{i=1}^{n} c_i x_i - \bar{x} \sum_{i=1}^{n} c_i \right] - \frac{1}{\sum_{j=1}^{n} (x_j - \bar{x})^2} = 0,$$

és így

$$\mathrm{D}^2(\tilde{a}) = \sigma^2 \sum_{i=1}^{n} c_i^2 = \sigma^2 \sum_{i=1}^{n} (k_i + d_i)^2 = \sigma^2 \sum_{i=1}^{n} k_i^2 + 2\sigma^2 \sum_{i=1}^{n} k_i d_i + \sigma^2 \sum_{i=1}^{n} d_i^2 =$$
$$= \mathrm{D}^2(\hat{a}) + 0 + \sigma^2 \sum_{i=1}^{n} d_i^2 \geq \mathrm{D}^2(\hat{a}),$$

amit bizonyítani akartunk.

Másrészt, ha $\tilde{b} = \sum_{i=1}^n w_i Y_i$ tetszőleges lineáris, torzítatlan becslés b-re, akkor

$$\mathbb{E}(\tilde{b}) = \sum_{i=1}^n w_i(ax_i + b) = a\sum_{i=1}^n w_i x_i + b\sum_{i=1}^n w_i = b.$$

Ez csak úgy lehetséges, hogy

$$\sum_{i=1}^n w_i x_i = 0 \quad \text{és} \quad \sum_{i=1}^n w_i = 1.$$

Legyen $d_i := w_i - l_i$. Ezzel és az előbbiek figyelembevételével

$$\begin{aligned}\sum_{i=1}^n l_i d_i &= \sum_{i=1}^n l_i(w_i - l_i) = \sum_{i=1}^n w_i l_i - \sum_{i=1}^n l_i^2 = \sum_{i=1}^n w_i(\frac{1}{n} - \bar{x}k_i) - \sum_{i=1}^n(\frac{1}{n} - \bar{x}k_i)^2 = \\ &= \frac{1}{n}\sum_{i=1}^n w_i - \bar{x}\frac{1}{\sum_{j=1}^n (x_j - \bar{x})^2}\sum_{i=1}^n (x_i - \bar{x})w_i - \frac{1}{n} - \bar{x}^2\sum_{i=1}^n k_i^2 + 2\frac{1}{n}\bar{x}\sum_{i=1}^n k_i = \\ &= \frac{1}{n} + \frac{\bar{x}^2}{\sum_{j=1}^n (x_j - \bar{x})^2} - \frac{1}{n} - \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} = 0,\end{aligned}$$

és így

$$\begin{aligned}\mathrm{D}^2(\tilde{b}) &= \sigma^2\sum_{i=1}^n w_i^2 = \sigma^2\sum_{i=1}^n (l_i + d_i)^2 = \sigma^2\sum_{i=1}^n l_i^2 + 2\sigma^2\sum_{i=1}^n l_i d_i + \sigma^2\sum_{i=1}^n d_i^2 = \\ &= \mathrm{D}^2(\hat{b}) + 0 + \sigma^2\sum_{i=1}^n d_i^2 \geq \mathrm{D}^2(\hat{b}),\end{aligned}$$

amivel a tételt bebizonyítottuk.

Tétel. Ha a (4) modellben még azt is feltesszük, hogy $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ fae. v.v.-k ($i = 1, \ldots, n$), akkor az a, b paraméterekre legkisebb négyzetes becslést szolgáltató, (5)-beli $\hat{a}, \hat{b}$ egyben maximum likelihood becslések is; továbbá a σ^2 paraméter maximum likelihood becslése:

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

Bizonyítás. A tétel feltételei miatt $Y_i \sim \mathcal{N}(ax_i + b, \sigma^2)$ független minta ($i = 1, \ldots, n$), így a likelihood-fv.:

$$L_{a,b,\sigma^2}(Y_1, \ldots, Y_n) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^n (Y_i - ax_i - b)^2\right],$$

a loglikelihood-fv. pedig:

$$l_{a,b,\sigma^2}(Y_1, \ldots, Y_n) = -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^n (Y_i - ax_i - b)^2.$$

Ezt deriválva az a, b, σ^2 paraméterek szerint, a következő egyenletrendszert kapjuk:

$$\frac{\partial l}{\partial a} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i - ax_i - b)x_i = 0,$$

$$\frac{\partial l}{\partial b} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i - ax_i - b) = 0,$$

$$\frac{\partial l}{\partial \sigma^2} = -\frac{n}{2}\frac{1}{\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(Y_i - ax_i - b)^2 = 0.$$

Innen az $\hat{a}, \hat{b}$ maximum likelihood becslések ugyanazok, mint az (5)-beli legkisebb négyzetes becslések voltak. (Ez nem véletlen, hiszen a, b a likelihood fv.-ben csak az exponensben van benne, így a likelihood fv. maximalizálása ekvivalens az exponensben álló négyzetösszeg minimalizálásával, ami éppen a legkisebb négyzetes becslésnél minimalizálandó célfv.)

A harmadik egyenletből az is kijön, hogy

$$\hat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{a}x_i - \hat{b})^2 = \frac{1}{n}SSE,$$

ha a 2. részben használt jelölést aktualizáljuk erre az esetre.

A tétel tanulsága az, hogy normális eloszlású ε_i „hibák" esetén (ami a gyakorlatban a centrális határeloszlás tétel miatt sokszor feltehető, pl. ha a hibák sok apró tényező eredői) a fenti legkisebb négyzetes becslések magukon viselik a maximum likelihood becslések „jó" tulajdonságait (ld. Cramér–Dugue tétel). Ugyancsak ilyenkor (6) és (7) alapján:

$$\hat{a} \sim \mathcal{N}(a, \frac{\sigma^2}{s_{xx}}), \qquad \hat{b} \sim \mathcal{N}(b, \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{s_{xx}}\right)).$$

Megjegyzés: σ^2 torzítatlan becslése $s^2 := \frac{1}{n-2}SSE$ lenne. Belátható, hogy $\frac{(n-2)s^2}{\sigma^2} \sim \chi^2(n-2)$ és független $\hat{a}, \hat{b}$-tól. Így a, b-re konfidenciaintervallumokat szerkeszthetünk és hipotéziseket vizsgálhatunk. A fenti statisztikák függetlensége **Basu tétel**éből (1955) következik: ha T egy elégséges és teljes statisztika (a θ paraméterre) és S egy másik olyan statisztika, mely nem függ θ-tól, akkor T és S függetlenek egymástól.

Hipotézisvizsgálat: a

$$H_0 : a = a_0 \quad \text{vers.} \quad H_1 : a \neq a_0$$

alternatíva vizsgálatára konstruálunk egy statisztikát, melynek eloszlása H_0 fennállásakor Student t. Ui. (6) miatt H_0 fennállásakor

$$\frac{\hat{a} - a_0}{\mathrm{D}(\hat{a})} = \frac{\hat{a} - a_0}{\sqrt{\frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}} = \frac{\hat{a} - a_0}{\sigma}\sqrt{s_{xx}} \sim \mathcal{N}(0,1)$$

és tőle függetlenül $\frac{(n-2)s^2}{\sigma^2} \sim \chi^2(n-2)$. Ezért

$$t = \frac{\frac{\hat{a}-a_0}{\sigma}\sqrt{s_{xx}}}{\sqrt{\frac{(n-2)s^2}{\sigma^2}/(n-2)}} = (\hat{a} - a_0)\sqrt{\frac{s_{xx}}{s^2}} \sim t(n-2),$$

így ha $|t| \geq t_{\alpha/2}$, akkor elutasítjuk H_0-t. $a_0 = 0$ esetén H_0 azt jelenti, hogy a regressziós egyenes meredeksége 0, elutasítása pedig azt, hogy a regresszió szignifikáns (van értelme Y előrejelzésének x-el). A

$$H_0 : a \leq a_0 \quad \text{vers.} \quad H_1 : a > a_0$$

egyoldali alternatíva vizsgálatára is a fenti t statisztikát használjuk, de akkor utasítunk el, ha $t \geq t_\alpha$, azaz az I. fajú hiba val.sége α lesz.

Hasonlóan, a

$$H_0 : b = b_0 \quad \text{vers.} \quad H_1 : b \neq b_0$$

alternatíva vizsgálatára is konstruálunk egy statisztikát, melynek eloszlása H_0 fennállásakor Student t. Ui. (7) miatt H_0 fennállásakor

$$\frac{\hat{b} - b_0}{\mathrm{D}(\hat{b})} = \frac{\hat{b} - b_0}{\sigma\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}} \sim \mathcal{N}(0,1)$$

és tőle függetlenül $\frac{(n-2)s^2}{\sigma^2} \sim \chi^2(n-2)$. Ezért

$$t = \frac{\frac{\hat{b} - b_0}{\sigma\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{s_{xx}}}}}{\sqrt{\frac{(n-2)s^2}{\sigma^2}/(n-2)}} = \frac{\hat{b} - b_0}{s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{s_{xx}}}} \sim t(n-2),$$

így ha $|t| \geq t_{\alpha/2}$, akkor elutasítjuk H_0-t. Hasonló mondható egyoldali alternatívára is.

$1-\alpha$ szintű konfidenciaintervallum is szerkeszthető az $\mathbb{E}(Y|X = x^*) = \hat{a}x^* + \hat{b}$ várható válasz köré egyetlen x^* értékre:

$$\hat{a}x^* + \hat{b} \pm t_{\alpha/2} \cdot s\sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{s_{xx}}},$$

ami a legszűkebb $\bar{x}$ közeli x^*-okra.

2. Többváltozós lineáris regresszió

2.1. Elméleti megoldás

Az $Y, X_1, \ldots, X_p$ val. változók együttes eloszlásáról (tegyük fel, hogy ez abszolút folytonos, az együttes sűrűségfüggvényt jelölje $f(y, x_1, \ldots, x_p)$). Akkor

$$\mathbb{E}(Y - g(X_1, \ldots, X_p))^2$$

minimumát a p-változós g függvények körében Y-nak az $X_1, \ldots, X_p$ változók adott értéke mellett vett feltételes várható értéke szolgáltatja:

$$g_{opt}(x_1, \ldots, x_p) = \mathbb{E}(Y|X_1 = x_1, \ldots, X_p = x_p) = \frac{\int_{-\infty}^{\infty} y f(y, x_1, \ldots, x_p) dy}{\int_{-\infty}^{\infty} f(y, x_1, \ldots, x_p) dy},$$

ezt nevezzük regressziós függvénynek.

Adott f sűrűségfüggvény mellett sem mindig triviális a fenti integrál kiszámolása, általában azonban f nem adott, csak egy statisztikai mintánk van a

függő és független változókra az $(Y^{(m)}, X_1^{(m)}, \ldots, X_p^{(m)})$, $(m = 1, \ldots, n)$ független, $(p+1)$-dimenziós megfigyelések formájában. A legegyszerűbb ilyenkor a fenti minimumot a lineáris függvények körében keresni, ezt nevezzük *lineáris regresszió*nak. Erre az esetre vezethető vissza olyan függvényekkel való közelítése Y-nak, amely az X_i változók lineáris függvényének monoton (például exponenciális, logaritmikus) transzformációja. Ilyenkor az inverz transzformációt alkalmazva Y-ra, az így kapott új függő változón hajtunk végre lineáris regressziót az eredeti független változók alapján.

A másik érv a lineáris regresszió mellett az, hogy amennyiben $Y, X_1, \ldots, X_p$ együttes eloszlása $(p+1)$-dimenziós normális, akkor a regressziós fv. valóban lineáris. A többdimenziós normalitás pedig a centrális határeloszlás tételre hivatkozva elég általánosan feltehető, vagy legalábbis közelíthető vele eloszlásunk. Még akkor is, ha eloszlásunk nem közelíthető normálissal, de abszolút folytonos, előfordulhat, hogy pusztán a változók között második momentumokra akarunk hagyatkozni, azaz az együttes kovarianciával akarunk csak dolgozni (amely mintánkból becsülhető). Ilyenkor is alkalmazható az alább ismertetendő módszer, hiszen ennek is – a főkomponens- és faktoranalízishez hasonlóan – az a sajátossága, hogy csak a második momentumig bezárólag használ momentumokat.

Térjünk rá tehát a lineáris regresszióra. A legjobb

$$Y \sim l(\mathbf{X}) = a_1 X_1 + \cdots + a_p X_p + b$$

lineáris közelítést keressük legkisebb négyzetes értelemben, azaz minimalizálni akarjuk az

$$\mathbb{E}(Y - (a_1 X_1 + \cdots + a_p X_p + b))^2$$

kifejezést az $a_1, \ldots, a_p$ és b együtthatókban.

Deriválás útján,

$$b = \mathbb{E}Y - \sum_{j=1}^{p} a_j \mathbb{E}X_j,$$

az $\mathbf{a} = (a_1, \ldots, a_p)^T$ vektor pedig megoldása a

$$\mathbf{C}\mathbf{a} = \mathbf{d}$$

lineáris egyenletrendszernek, ahol $\mathbf{C}$ jelöli az $\mathbf{X} = (X_1, \ldots, X_p)^T$ véletlen vektor $p \times p$-s kovarianciamátrixát, a $\mathbf{d} \in \mathbb{R}^p$ vektor pedig az Y változónak $\mathbf{X}$ komponenseivel vett (kereszt)kovarianciáit tartalmazza. A fenti lineáris egyenletrendszernek létezik egyértelmű megoldása, ha $|\mathbf{C}| \neq 0$, ami teljesül, ha az $X_1, \ldots, X_p$ változók között nincsen lineáris kapcsolat (például nem-elfajult p-dimenziós normális eloszlásúak). A megoldás

$$\mathbf{a} = \mathbf{C}^{-1}\mathbf{d}$$

lesz, összhangban az egyváltozós regressziónál tanultakkal.

Mivel a kovarianciák eltolásinvariánsak, a továbbiakban az összes valószínűségi változó várható értékét 0-nak tekintjük. Így $b = 0$ és $l(\mathbf{X}) = \sum_{i=1}^{p} a_i X_i$ az optimális közelítést adó lineáris kombináció az

$$Y = l(\mathbf{X}) + \varepsilon$$

modellben, ahol kijön $l(\mathbf{X})$ és ε korrelálatlansága (az optimális paraméterek mellett). Az fenti felbontásban a szórásnégyzetek összeadódnak:

$$\mathrm{Var}(Y) = \mathrm{Var}l(\mathbf{X}) + \mathrm{Var}\varepsilon.$$

Ez úgy is írható, hogy

$$\mathrm{Var}Y = r^2_{Y(X_1,\ldots,X_p)}\mathrm{Var}Y + (1 - r^2_{Y(X_1,\ldots,X_p)})\mathrm{Var}Y,$$

ahol $r_{Y(X_1,\ldots,X_p)}$ jelöli az Az Y független- és az $X_1,\ldots,X_p$ függő változók közötti *többszörös korrelációs együtthatót*, azaz Y és az optimális lineáris közelítést adó $l(\mathbf{X})$ korrelációját. A $p = 1$ esetben ez a szokásos korrelációs együttható.

Innen látható, hogy $r_{Y(X_1,\ldots,X_p)} = \pm 1$ ekvivalens azzal, hogy $\mathrm{Var}\varepsilon = 0$, ami 1 valószínűségű lineáris függvénykapcsolatot takar Y és $X_1,\ldots,X_p$ között; $r_{Y(X_1,\ldots,X_p)} = 0$ pedig azt jelentené, hogy $\mathrm{Var}l(\mathbf{X}) = 0$, ami nem-konstans X_i-k esetén csak úgy lehetséges, hogy az összes $a_i = 0$, azaz nincs is igazából regresszió. ($\mathbf{a} = \mathbf{0}$ ekvivalens $\mathbf{d} = \mathbf{0}$-val, ami azt jelenti, hogy Y korrelálatlan az összes X_i-vel.)

2.2. A regressziós együtthatók becslése mintából

Most a célváltozóra és a prediktorokra az az $(Y_i, X_{i1},\ldots,X_{ip})$, $(i = 1,\ldots,n)$ független, $(p+1)$-dimenziós megfigyelések állnak rendelkezésünkre, és a legkisebb négyzetek módszerével keressük

$$\sum_{i=1}^{n}(Y_i - (a_1X_{i1} + \cdots + a_nX_{ip} + b))^2$$

minimumát. Ezt

$$\hat{\mathbf{a}} = \hat{\mathbf{C}}^{-1}\hat{\mathbf{d}} \tag{8}$$

adja az, ahol az empirikus kovarianciákat és keresztkovarianciákat a mintából becsüljük. Továbbá

$$\hat{b} = \bar{Y} - \sum_{j=1}^{p}\hat{a}_j\bar{X}_j.$$

Itt az empirikus többszörös korrelációs együttható négyzetét, R^2-et nevezik meghatározottsági együtthatónak.

Többváltozós regresszióra vezethető vissza az

$$Y \sim a_1X + a_2X^2 + \cdots + a_pX^p + b$$

polinomiális regresszió. A megoldást az $X_i = X^i$ $(i = 1,\ldots,p)$ prediktorokra vonatkozó többváltozós lineáris regresszióval kaphatjuk.

2.3. Tervezett (determinisztikus) megfigyelés

Legyenek most $x_1,\ldots,x_p$ mérési pontok, melyek hiba nélkül beállíthatók (tehát nem valószínűségi változók), méréseink pedig ezek valamely ismeretlen $a_1,\ldots,a_p$ paraméterekkel való lineáris kombinációira vonatkoznak, és mérési hibával terheltek. Jelölje ε a mérési hibát, Y a mért értéket, ezek valószínűségi változók.

Feltehető, hogy $\mathbb{E}(\varepsilon) = 0$. Tegyük fel, hogy a konstans tagtól már megszabadultunk (az átlagok levonásával). Modellünk tehát a következő:

$$Y = a_1 x_1 + \cdots + a_p x_p + \varepsilon$$

ami hasonlít a többváltozós regressziójéhoz, csak ott X_i-k valószínűségi változók voltak, azért is jelöltük őket nagybetűvel. Itt $\mathbb{E}(Y) = \sum_{j=1}^{p} a_j x_j$.

Célunk az ismeretlen $\mathbf{a} = (a_1, \ldots, a_p)^T$ paramétervektor (oszlopvektor) legkisebb négyzetes becslése n mérés alapján ($n \geq p$, általában n sokkal nagyobb, mint p). Az i-edik mérés az $(x_{i1}, \ldots, x_{ip})$ p-dimenziós pontban történik, a mért értéket jelölje Y_i, a mérési hibát pedig ε_i, $(i = 1, \ldots, n)$. Vezessük be még a következő jelöléseket is:

$$\mathbf{Y} := (Y_1, \ldots, Y_n)^T, \quad \boldsymbol{\varepsilon} := (\varepsilon_1, \ldots, \varepsilon_n)^T$$

n-dimenziós oszlopvektorok, az x_{ij} $(i = 1, \ldots, n;\ j = 1, \ldots, p)$ mérési pontokat pedig az $n \times p$-s $\mathbf{X}$ mátrixban gyűjtjük össze. $\mathbf{X}$ oszlopvektorait jelölje $\mathbf{x}_1, \ldots, \mathbf{x}_p$! Ezekkel a jelölésekkel a fenti rendszeregyenlet

$$\mathbf{Y} = \mathbf{X}\mathbf{a} + \boldsymbol{\varepsilon}$$

alakban írható, ahol tehát $\mathbb{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$, továbbá tegyük fel, hogy a mérési hibák korrelálatlanok (normális eloszlás esetén függetlenek) és azonos szórásúak, azaz $\boldsymbol{\varepsilon}$ kovarianciamátrixa $\sigma^2 \mathbf{I}_n$ alakú. Ekkor persze a mérések is korrelálatlanok, és ugyanaz a kovarianciamátrixszuk, mint $\boldsymbol{\varepsilon}$-é.

Az $\mathbf{a}$ ismeretlen paraméter legkisebb négyzetes becslése alatt azt az $\hat{\mathbf{a}}$ vektort értjük, amelyre a mérési hibák négyzetösszege, azaz

$$\sum_{i=1}^{n} \varepsilon_i^2 = \|\boldsymbol{\varepsilon}\|^2 = \|\mathbf{Y} - \mathbf{X}\mathbf{a}\|^2$$

minimális.

Gauss a következő geometriai szemlélet alapján oldotta meg a problámát: $\|\mathbf{Y} - \mathbf{X}\mathbf{a}\|^2$ nyilván akkor minimális $\mathbf{a}$-ban, ha $\mathbf{X}\mathbf{a}$ az $\mathbf{Y}$ vektornak az F altérre való merőleges vetülete, ahol az $F = \mathrm{Span}(\mathbf{x}_1, \ldots, \mathbf{x}_p) \subset \mathbb{R}^n$ alteret $\mathbf{X}$ oszlopvektorai (az $\mathbf{x}_1, \ldots, \mathbf{x}_p$ vektorok) feszítik ki, $\dim(F) = r \leq p$ (tipikusan p-vel egyenlő, ha az $\mathbf{x}_i$ vektorok lineárisan függetlenek). Jelölje $\boldsymbol{P}$ ennek az r-rangú ortogonális projekciónak az $n \times n$-es mátrixát! Ezzel az optimális $\mathbf{a}$-ra $\mathbf{X}\mathbf{a} = \boldsymbol{P}\mathbf{Y}$. Mivel $\mathbf{X}\mathbf{a} \in F$ és $\mathbf{Y} - \mathbf{X}\mathbf{a} \perp F$, ezért $\mathbf{Y} - \mathbf{X}\mathbf{a}$ merőleges F tetszőleges vektorára, ami $\mathbf{X}\mathbf{b} = \sum_{j=1}^{p} b_j \mathbf{x}_j \in F$ alakú lesz valamely $\mathbf{b} \in \mathbb{R}^p$ vektorral. Így

$$(\mathbf{X}\mathbf{b})^T \cdot (\mathbf{Y} - \mathbf{X}\mathbf{a}) = 0, \quad \forall \mathbf{b} \in \mathbb{R}^p.$$

Ebből

$$\mathbf{b}^T \mathbf{X}^T (\mathbf{Y} - \mathbf{X}\mathbf{a}) = 0, \quad \forall \mathbf{b} \in \mathbb{R}^p.$$

Ez csak úgy lehetséges, ha

$$\mathbf{X}^T (\mathbf{Y} - \mathbf{X}\mathbf{a}) = \mathbf{0},$$

azaz

$$\mathbf{X}^T \mathbf{X}\mathbf{a} = \mathbf{X}^T \mathbf{Y}$$

adódik, amit *Gauss normálegyenlet*nek nevezünk.

A normálegyenlet mindig konzisztens, hiszen az $\mathbf{X}^T\mathbf{Y}$ vektor benne van az $\mathbf{X}^T$ mátrix oszlopvektorai által kifeszített altérben, és ugyanezt az alteret feszítik ki az $\mathbf{X}^T\mathbf{X}$ mátrix oszlopai is. A megoldás pontosan akkor egyértelmű, ha az $\mathbf{X}^T\mathbf{X}$ mátrix rangja $r = p(\leq n)$, ilyenkor a megoldás

$$\hat{\mathbf{a}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

alakban írható. Ha pedig az $\mathbf{X}^T\mathbf{X}$ mátrix szinguláris, azaz rangja $r < p$, akkor egy lehetséges megoldás az $\mathbf{X}^T\mathbf{X}$ mátrix pszeudoinverzével az

$$\hat{\mathbf{a}} = (\mathbf{X}^T\mathbf{X})^{+}\mathbf{X}^T\mathbf{Y}$$

formában adható meg. (Ha a pszeudoinverz helyett másik általánosított inverzzel képezzük a becslést, másik $\hat{\mathbf{a}}$ vektort kapunk, az $\mathbf{X}\hat{\mathbf{a}}$ vetület azonban egyértelmű.) A gyakorlatban tipikusabbak azok az esetek, mikor $\mathbf{X}$ oszlopvektorai lineárisan függetlenek. A továbbiakban ezt feltesszük. Beláthatók a következők.

Állítás Ha $r = p$ és $\boldsymbol{\varepsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$, akkor $\hat{\mathbf{a}} \sim \mathcal{N}_p(\mathbf{a}, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1})$. Ez egyben ML-becslés is $\mathbf{a}$-ra.

Így az $r = p$ esetben az $\hat{\mathbf{a}}$ lineáris becslés tehát torzítatlan és meg lehet mutatni, hogy minimális kovarianciamátrixú az $\mathbf{a}$-ra vonatkozó lineáris, torzítatlan becslések között.

Gauss–Markov tétel általánosan. A fenti lineáris modellben $\hat{\mathbf{a}}$ lineáris torzítatlan becslés $\mathbf{a}$-ra. Legyen $\tilde{\mathbf{a}}$ az $\mathbf{a}$ paramétervektor tetszőleges lineáris torzítatlan becslése. Ekkor $\tilde{\mathbf{a}}$ kovarianciamátrixa - $\hat{\mathbf{a}}$ kovarianciamátrixa pozitív szemidefinit. Így $\hat{\mathbf{a}}$ megint BLUE.

3. Varianciaanalízis (ANOVA)

Az ANOVA, bár látszólag más problémát old meg, a fenti lineáris modellből táplálkozik. A varianciaanalízis speciális lineáris modelleket vizsgál, kísérlettervezésben és minőségellenőrzésben felmerülő hipotézisek tesztelésére. A tekintett modellek specifikuma az, hogy a beállítható mérési pontok mátrixa helyett 0-1 elemekből álló ún. *struktúramátrix*szal dolgozunk, amelyet úgy állítunk össze, hogy bizonyos megfigyelések csak bizonyos paraméterektől függjenek,

Gyakorlati alkalmazásokban olyan mintákat vizsgálunk, melyeket különböző körülmények közt figyeltünk meg, és célunk éppen annak a megállapítása, vajon ezek a körülmények jelentősen befolyásolják-e a mért értékeket. Tehát mintánkat eleve csoportokba osztottan kapjuk, feltesszük azonban, hogy a különböző csoportokban felvett minták egymástól függetlenek, normális eloszlásúak és azonos szórásúak.

Például egy-egy minta betegek vérnyomását jelenti különböző dózisban adagolt gyógyszer hatására (ilyenkor maga a dózis is kvantitatív változó, a vérnyomás pedig normális eloszlású), vagy különböző vérnyomáscsökkentő szerek hatására (a gyógyszerféleség változója most kvalitatív). Más példa: több gépen, vagy többféle technológiával gyártott alkatrészek valamilyen mérhető jellemzőjét (pl. szakítószilárdság) vizsgáljuk, és az érdekel bennünket, vajon a gyártó gép vagy a gyártási technológia befolyásolja-e az alkatrész mért tulajdonságát. Ha egyszerre mindkét hatás, esetleg azok kölcsönhatása is érdekel bennünket,

akkor kétszempontos varianciaanalízisről beszélünk, ha külön vizsgáljuk az egyes tényezők hatását, akkor egyszempontos a varianciaanalízis.

Természetesen bevezethetnénk további szempontokat is. Az első példánál a kor, nem, egyéb kezelések is lehetnek szempontok. Általában háromnál több tényezőt (szempontot) nem szoktak vizsgálni, mert az túlságosan elbonyolítaná a számításokat, három szempont esetén pedig felmerül a kísérletek ésszerű és gazdaságos tervezésének problámája is, amit szintén érinteni fogunk. Itt csak az egyszempontos (egyfaktoros) varianciaanalízist ismertetjük.

3.1. Egyszempontos varianciaanalízis

Valamilyen szempont alapján (például különböző kezelések) k csoportban külön végzünk megfigyeléseket. Az egyes csoportokban a mintaelemek száma általában nem egyenlő: jelölje n_i az $i.$ csoportbeli mintaelemek számát, $n = \sum_{i=1}^{k} n_i$ pedig az összminta elemszámát. Az $i.$ csoportban az $X_i \sim \mathcal{N}(b_i, \sigma^2)$ valószínűségi változóra vett mintaelemeket

$$X_{ij} \sim \mathcal{N}(b_i, \sigma^2), \quad (j = 1, \dots, n_i)$$

jelöli. Ezek egymás közt és különböző i-kre is függetlenek, azonos szórásúak. A várható értékekre a $b_i = \mu + a_i$ felbontást alkalmazzuk, ahol μ a várható értékek súlyozott átlaga, a_i pedig az $i.$ csoport hatása:

$$\mu = \frac{1}{n} \sum_{i=1}^{k} n_i b_i, \qquad a_i = b_i - \mu \quad (i = 1, \dots, k).$$

Könnyen látható, hogy

$$\sum_{i=1}^{k} n_i a_i = 0. \tag{9}$$

Ezzel a

$$H_0 : b_1 = \dots = b_k$$

null-hipotézis ekvivalens a

$$H_0 : a_1 = \dots = a_k = 0$$

null-hipotézissel, melyek azt vizsgálják, hogy a k csoportban a várható értékek megegyeznek-e. Valójában a kétmintás t-próba kiterjesztéséről van szó kettőnél több csoportra. A szórások egyenlőségét a Bartlett-próbával kell tesztelni előtte.

Ezekkel a jelölésekkel az egyszempontos modell

$$X_{ij} = \mu + a_i + \varepsilon_{ij} \qquad (j = 1, \dots, n_i;\ i = 1, \dots, k)$$

alakban írható, ahol az $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$ független valószínűségi változók véletlen hibák.

Lineáris modellről van szó, hiszen ha megfigyeléseinket és hibáinkat az

$$\mathbf{Y} := (X_{11}, \dots, X_{1n_1}, X_{21}, \dots, X_{2n_2}, \dots, X_{k1}, \dots, X_{kn_k})^T$$
$$\boldsymbol{\varepsilon} := (\varepsilon_{11}, \dots, \varepsilon_{1n_1}, \varepsilon_{21}, \dots, \varepsilon_{2n_2}, \dots, \varepsilon_{k1}, \dots, \varepsilon_{kn_k})^T$$

$\sum_{i=1}^{k} n_i = n$-dimenziós vektorokban, a_i paramétereinket pedig az $\mathbf{a} = (a_1, \ldots, a_k)^T$ vektorban helyezzük el, akkor a fenti modell az

$$\mathbf{Y} = \boldsymbol{B} \cdot \mathbf{a} + \mathbf{1} \cdot \mu + \boldsymbol{\varepsilon}$$

alakban írható, ahol $\mathbf{1} \in \mathbb{R}^n$ az azonosan 1 koordinátájú vektor, $\boldsymbol{B}$ pedig a következő struktúramátrix, melyet a $k = 3$, $n_1 = 3$, $n_2 = 4$, $n_3 = 5$ esetben szemléltetünk:

$$\boldsymbol{B} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}.$$

Látható, hogy $\operatorname{rank}\boldsymbol{B} = k$, az oszlopok által kifeszített k-dimenziós alteret jelölje F; nyilván $\mathbf{1} \in F$. A paramétereket közvetlenül a legkisebb négyzetek módszerével becsüljük, azaz keressük a

$$\sum_{i=1}^{k} \sum_{j=1}^{n_i} \varepsilon_{ij}^2 = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \mu - a_i)^2$$

kifejezés minimumát a $\mu, a_1, \ldots, a_k$ paraméterekben az (9)) kényszerfeltétel mellett. Vezessük be a csoportátlagokra ill. a teljes mintaátlagra az

$$\bar{X}_{i\cdot} = \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij} \quad (i = 1, \ldots, k) \quad \text{ill.} \quad \bar{X}_{\cdot\cdot} = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} X_{ij}$$

jelöléseket! Könnyen látható, hogy a paraméterek legkisebb négyzetes becslései

$$\hat{\mu} = \bar{X}_{\cdot\cdot} \quad \text{és} \quad \hat{a}_i = \bar{X}_{i\cdot} - \bar{X}_{\cdot\cdot} \quad (i = 1, \ldots, k)$$

lesznek. (A Lagrange-multiplikátor módszerrel ellenőrízhető a fenti megoldás helyessége.)

A minimum értéke

$$Q_e = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \hat{\mu} - \hat{a}_i)^2 = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_{i\cdot})^2$$

lesz. Itt Q_e az $\boldsymbol{\varepsilon}$ változó ún. reziduális varianciája. A lineáris becslés varianciája:

$$Q_a = \|\boldsymbol{B}\hat{\mathbf{a}}\|^2 = \sum_{i=1}^{n_i} n_i \hat{a}_i^2 = \sum_{i=1}^{k} n_i (\bar{X}_{i\cdot} - \bar{X}_{\cdot\cdot})^2.$$

A gyakorlati alkalmazók terminológiájával élve: a fenti kvadratikus alakok segítségével a mintaelemek teljes mintaátlagtól vett eltéréseinek négyzetösszega

(Q) felbomlik csoportok közötti (between, Q_a) ill. csoportokon belüli (within, Q_e) részre a következőképpen:

$$Q = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_{..})^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}[(X_{ij} - \bar{X}_{i.}) + (\bar{X}_{i.} - \bar{X}_{..})]^2 =$$
$$= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_{i.})^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{X}_{i.} - \bar{X}_{..})^2 =$$
$$= \sum_{i=1}^{k} n_i(\bar{X}_{i.} - \bar{X}_{..})^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_{i.})^2 = Q_a + Q_e.$$

A fenti felbontást az alábbi ún. ANOVA táblázatban foglaljuk össze:

A szóródás oka	Négyzetösszeg	Szabadsági fok	Empirikus szórásnégyzet
Csoportok között	$Q_a = \sum_{i=1}^{k} n_i(\bar{X}_{i.} - \bar{X}_{..})^2$	$k-1$	$s_a^2 = \frac{Q_a}{k-1}$
Csoportokon belül	$Q_e = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_{i.})^2$	$n-k$	$s_e^2 = \frac{Q_e}{n-k}$
Teljes	$Q = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_{..})^2$	$n-1$	-

A fenti modellben először a $\mu = 0$ hipotézist teszteljük. Ha ezt elutasítjuk (az összes várható érték nem 0, azaz van ún. főhatás), akkor a

$$H_0 : a_1 = \cdots = a_k = 0, \quad \text{tömören} \quad \mathbf{a} = \mathbf{0}$$

hipotézist vizsgáljuk. Bevezetve az

$$s_a^2 = \frac{Q_a}{k-1} \qquad \text{ill.} \qquad s_e^2 = \frac{Q_e}{n-k}$$

kifejezéseket, ezek azonos (σ^2) szórásúak, függetlenek, hányadosuk pedig H_0 fenállása esetén F-eloszlást követ $k-1$ ill. $n-k$ szabadsági fokkal:

$$F = \frac{s_a^2}{s_e^2} = \frac{Q_a}{Q_e} \cdot \frac{n-k}{k-1} \sim \mathcal{F}(k-1, n-k).$$

Az indoklás a Fisher–Cochran tétellel (MSc tananyag) következik, ui. a fenti χ^2-eloszlások szabadsági fokai összeadódnak:

$$n-1 = (k-1) + (n-k).$$

Így H_0 tesztelésére F-próba alkalmazható, mellyel tulajdonképpen arról döntünk, hogy a csoportok közötti eltéréseket mérő s_a^2 szignifikánsan nagyobb-e, mint a csoportokon belüli ingadozásokat mutató s_e^2 (utóbbi ingadozásokat csak a véletlen eltérések hozzák létre). Ha a fenti F-statisztika nagyobb vagy egyenlő, mint az $\mathcal{F}(k-1, n-k)$-eloszlás $1-\alpha$ szinthez tartozó kritikus értéke, akkor $1-\alpha$ szinten elutasítjuk H_0-t, azaz az a_i várható értékek között van olyan, ami nem egyenlő a többivel; különben pedig $1-\alpha$ szinten elfogadjuk H_0-t. Az ANOVA programokban általában feltüntetik azt a legkisebb α értéket, amely mellett a csoportok közti eltérés még szignifikáns.