

A REGRESSZIÓS EGYÜTTHATÓ MEGHATÁROZÁSA ADOTT ALAPPONTOK ESETÉN

VINCZE ISTVÁN

ÖSSZEFOGLALÁS

Az x és y változók között lineáris kapcsolat áll fenn: $y = Rx + Q$. Az R (és Q) állandók meghatározására x -nek rögzített $a_1, a_2, \dots, a_s$ értékei mellett összesen N számú mérést végzünk. Kérdés, hogyan osszuk el az a_i pontok között az N számú mérést, hogy a R tapasztalati regressziós együttható szórásnégyzete minimum legyen. Feltesszük, hogy az a_i -k pontos értékek, tehát nem valószínűségi változók, továbbá, hogy az a_i -ben mért y értékek azonos eloszlást követnek, szórásuk az a_i ponttól függ: $\sigma(a_i)$.

Bebizonyítjuk, hogy e feltételek mellett a regressziós egyenest az s számú a_i pont közül elég két vagy legfeljebb három pontban végzett mérés alapján meghatározni. A két- és hárompontos extrémálisok.

Ha σ_i = állandó ($i = 1, 2, \dots, s$), akkor mindig két pontban és pedig a két szélső a_i pontban — kell méréseket végeznünk, egyenlő arányban osztva el a mérések számát.

A dolgozat foglalkozik még a regressziós együttható módosított definíciójával, amelynél a szélsőértékfeladat egyszerűbben tárgyalható és egyszerűbben kezelhető — ha nem is feltétlenül jobb — eredményre vezet.

Bevezetés

Az alábbiakban tárgyalt matematikai statisztikai probléma a Biokémiai Ipari Kutató Laboratórium munkája során merült fel egy koncentráció-mérési eljárás pontosságának vizsgálata és az eljárás egyszerűsítésére irányuló kutatás során. A feladatot a matematikai statisztika módszereivel megvizsgáltuk, eredményeink alapján a Kutató Laboratórium a mérési eljárást jelentősen egyszerűsítette és ugyanakkor pontosabbá tette. A mérésnél az eddiginél sokkal kisebb számú különböző hígítású oldattal kell dolgozni és ennek megfelelően az értékelés munkája is lényegesen egyszerűsödött. Ez lehetővé tette, hogy azok a kisebb intézetek, amelyek munkájához ez szükséges, a mérést önállóan elvégezzék.

A probléma felvetése *Tettamanti Károly*, a Kutató Laboratórium igazgatója részéről történt, aki a feladat matematikai megfogalmazásának munkájában is segítségemre volt.

E helyen köszönöm meg *Rényi Alfréd*-nek a probléma megoldása során tett értékes megjegyzéseit.

1. A probléma matematikai statisztikai megfogalmazása

Legyen az y változó mennyiség az x változó mennyiségnek — annak valamely (α, β) intervallumában — lineáris függvénye :

$$y = Rx + Q.$$

Az ismeretlen R (és Q) állandók meghatározására végezzünk y értékeire összesen n számú mérést az x változó rögzített $a_1 \leq a_2 \leq \dots \leq a_s$ értékei mellett és pedig az $x = a_1$ helyen n_1 számú, az $x = a_2$ helyen n_2 számú mérést, s. i. t., ahol

$$(1) \quad n_1 + n_2 + \dots + n_s = n$$

Kérdés, hogyan osszuk el az egyes a_i helyek között az összesen n számú mérést ahhoz, hogy a tapasztalati regressziós együttható a »legjobb közelítését« nyújtsa az elméleti értéknek. Pontosabban : *hogyan válasszuk az $n_1, n_2, \dots, n_s$ egész számokat az (1) feltétel mellett úgy, hogy a tapasztalati regressziós együttható szórásnégyzete minimális legyen*

$$\sigma^2(\varrho) = \min$$

Feladatunk feltételeit a következőkben foglaljuk össze :

a) Az $a_1, a_2, \dots, a_s$ mennyiségek pontosan beállíthatók és mérhetők, tehát nem valószínűségi változók.

b) Az a_i pontban mért n_i számú y -ra nyert $\eta_{i1}, \eta_{i2}, \dots, \eta_{in_i}$ értékek független mérési adatok eredményei, eloszlásuk ugyanaz $y = Rx + Q$ várható értékkel és $\sigma_i(a_i) = \sigma_i$ szórással. Ez a követelmény fennáll $i = 1, 2, \dots, s$ -re. Az η_{ik} -k ($k = 1, 2, \dots, s$) közös eloszlása egyébként tetszőleges lehet.

2. A feltételek gyakorlati vonatkozásai

Azt, hogy az y és x változó mennyiségek között lineáris kapcsolat áll fenn, akár a tapasztalat támaszthatja alá, akár elméleti megfontolások nyújthatják. A regressziós egyenes lehető pontos meghatározása azt a célt szolgálja, hogy segítségével a változók további összetartozó értékeit határozhassuk meg numerikus vagy grafikus úton.

Az x változó mennyiség rögzített $a_1, a_2, \dots, a_s$ értékeinek pontos, szórásmentes beállíthatósága a gyakorlatban pl. olyan esetekben teljesül, ha az a_i -k maguk természetes számok : atomok rendszámai, bizonyos egyedek darabszámai stb. Ilyennek tekinthetők bizonyos feltételek mellett a fizikában vagy kémiában előforduló igen pontosan mérhető mennyiségek is, amelyeknél a mérési pontossághoz képest az a_i -k beállításánál elkövetett hibák elenyészőek, vagy pedig az x értékek beállításánál felmerülő szórások elhanyagolhatóan csekélyek a hozzájuk tartozó y értékek szórásaival szemben.

Az a feltevés, hogy valamely rögzített vagy rögzítettnek tekinthető $x = a_i$ érték mellett az y -ra kapott mérési eredmények ugyanazt az eloszlást követik, kézenfekvő, ennek ellenére ezt tapasztalatilag ellenőrizni kell. Ennek a feltevésnek már következménye, hogy az y értékek szórását a megfelelő x érték függvényének tekinthetjük : $\sigma(\eta_{ik}) = \sigma(a_i) = \sigma_i$ adott i mellett $k = 1, 2, \dots, n_i$ -re, $i = 1, 2, \dots, s$ -re. Erre nézve még a következőket jegyezzük

meg: az y és x kémiai vagy fizikai mennyiségek lineáris összefüggése rendszerint az x -nek valamely (α, β) intervallumban áll fenn, ezenkívül vagy már nem érvényes a mennyiségek lineáris összefüggése, vagy egyáltalán nincsenek értelmezve a mennyiségek. Előfordulhat ilyen esetben, hogy az y értékek szórása az intervallum határai felé növekszik annak belsejében felvett értékeihez képest.

Különlegesen egyszerű a helyzet, mint azt alább megmutatjuk, ha a szórások azonosak x minden értékére.

3. A szélsőértékfeladat felállítása

Az 1. pontban használt jelölések figyelembevételével a tapasztalati regressziós együttható — a legkisebb négyzetek módszerének az ordináták különbségére való alkalmazásával — a következő:

$$\varrho = \frac{\sum_{i=1}^s \sum_{k=1}^{n_i} (\eta_{ik} - \bar{\eta}_i) (a_i - \omega)}{\sum_{i=1}^s n_i (a_i - \omega)^2},$$

ahol

$$\bar{\eta}_i = \frac{\sum_{i=1}^s \sum_{k=1}^{n_i} \eta_{ik}}{n}, \quad \omega = \frac{\sum_{i=1}^s n_i a_i}{n}, \quad \sum_{i=1}^s n_i = n.$$

Az

$$\bar{\eta}_i = \frac{\sum_{k=1}^{n_i} \eta_{ik}}{n_i}$$

jelölés bevezetésével, valamint a $\sum_{i=1}^s n_i (a_i - \omega) = 0$

összefüggés figyelembevételével ϱ kifejezése a következő

$$\varrho = \frac{\sum_{i=1}^s n_i (a_i - \omega) \bar{\eta}_i}{\sum_{i=1}^s n_i (a_i - \omega)^2}$$

ϱ e kifejezésben csupán az $\bar{\eta}_i$ valószínűségi változók szerepelnek, melyek szórásnégyzete — η_i előbbi értelmezése és az 1. §. β .) feltételei értelmében

$$\sigma^2(\bar{\eta}_i) = \frac{\sigma^2(\eta_{ik})}{n_i} = \frac{\sigma_i^2}{n_i},$$

ahol a σ -k a megfelelő elméleti eloszlások szórásai.

Az $\bar{\eta}_i$ -k függetlensége miatt ϱ szórásnégyzete

$$(I) \quad \sigma^2(\varrho) = \frac{\sum_{i=1}^s n_i (a_i - \omega)^2 \sigma_i^2}{\left[\sum_{i=1}^s n_i (a_i - \omega)^2 \right]^2}.$$

E kifejezést kell tehát minimummá tenni, figyelembe véve az

$$(I.1) \quad \frac{\sum_{i=1}^s a_i n_i}{n} = \omega$$

$$(I.2) \quad \sum_{i=1}^s n_i = n$$

összefüggéseket.

Szélsőértékfeladatunknak az analízis eszközei segítségével való megoldhatósága érdekében egyrészt az n_i egészszámokat folytonos változóknak tekintjük, másrészt pozitivitásuk biztosítására helyettük a következő új változókat vezetjük be:

$$u_i^2 = \frac{n_i}{n}, \quad i = 1, 2, \dots, s.$$

A szórásnégyzet mármost a következő alakba megy át:

$$\sigma^2(\varrho) = \frac{1}{n} \frac{\sum_{i=1}^s (a_i - \omega)^2 \sigma_i^2 u_i^2}{\left[\sum_{i=1}^s (a_i - \omega)^2 u_i^2 \right]^2} = \frac{1}{n} \cdot f(u_1, u_2, \dots, u_s, \omega).$$

Szélsőértékfeladatunkat végülis következőképpen fogalmazzuk:

Az

$$(I') \quad f(u_1, u_2, \dots, u_s, \omega) = \frac{\sum_{i=1}^s (a_i - \omega)^2 \sigma_i^2 u_i^2}{\left| \sum_{i=1}^s (a_i - \omega)^2 u_i^2 \right|^2}$$

$s + 1$ változós függvény az $(u_1, u_2, \dots, u_s, \omega)$ mely értékrendszerei mellett veszi fel legkisebb értékét, a

$$(I'.1) \quad g(u_1, u_2, \dots, u_s, \omega) \equiv \sum_{i=1}^s a_i u_i^2 - \omega = 0$$

$$(I'.2) \quad h(u_1, u_2, \dots, u_s) \equiv \sum_{i=1}^s u_i^2 - 1 = 0$$

mellékfeltételek figyelembevételével.

4. Az eredmények összefoglalása

Az előző §-ban megfogalmazott szélsőértékfeladat vizsgálata a következő eredményre vezet:

Az (I') -ben megadott függvénynek az $(I'.1.)$ és $(I'.2.)$ feltételek mellett — hacsak az (a_i, σ_i^2) $(i = 1, 2, \dots, n)$ értékpárok között bizonyos speciális összefüggések nem állanak fenn — csakis olyan $(u_1, u_2, \dots, u_s)$ értékrendszerek mellett lehet minimuma, amelyekre nézve az u_i -k közül legfeljebb három különbözik 0-tól.

Az (a_i, σ_i^2) értékpárok különleges megválasztása mellett más értékrendszerek is tekintetbe jöhetnek szélsőérték szempontjából, de ezek nem feltétlenül adnak minimumot.

Ha $\sigma_i = \sigma$ $(i = 1, 2, \dots, s)$ állandó, akkor csakis az

$$u_1^2 = \frac{1}{2}, \quad u_s^2 = \frac{1}{2}, \quad u_i = 0, \quad i = 2, 3, \dots, s-1$$

értékrendszer adja a minimumot.*

Különösen egyszerű eredmény adódik abban a speciális esetben, ha a regressziós egyenlet alakja

$$y = Rx.$$

* Az a_1 és a_s pontokban egyenlő számú mérést természetesen csak akkor végezhetünk, ha n páros. Amennyiben az n_i -kre így kapott kifejezések nem egész számok, úgy azokat természetesen ki kell kerekíteni egész számmá.

Ekkor ugyanis az egyenes egy pontja — $x = 0, y = 0$ — ismeretes és abban a további egy pontban kell mérni, amelyre a $\left(\frac{\sigma_i}{a_i}\right)^2$ kifejezés — $i = 1, 2, \dots, s$ — értéke minimum.

Eredményeinket gyakorlati szempontból a következőképpen fogalmazhatjuk:

Ahhoz, hogy az 1. §-ban megadott feltételek mellett keresett regressziós együtthatót olyan mérésekkel határozzuk meg, amelyek mellett az együttható szórásnégyzete minimális, az a_i helyek közül két, vagy legfeljebb három pontban kell mérést végezni.

Ha az y értékek szórása minden a_i pontban ugyanaz és az összes mérések száma páros, akkor az egymástól legtávolabbra eső két pontban kell a méréseket fele-fele arányban megosztani, tehát

$$n_1 = u_1^2 \cdot n = \frac{1}{2} n, \quad n_s = u_s^2 n = \frac{1}{2} n$$

mellett lesz a szórásnégyzet minimum:

$$\min \sigma^2(\varrho) = \left(\frac{2\sigma}{a_1 - a_s} \right)^2.$$

Ha az y értékek szórása a_i -vel változik, akkor minden pontpárra és minden ponthármásra — az alább megadandó eljárás szerint — meg kell határozni azt a szórást, amely mint minimális tekintetbe jöhet és ezek közül kell a legkisebbet kiválasztani.

A kétpontos mérések közül azon (a_j, a_k) értékpár mellett kell a méréseket végezni, amelyre

$$\left(\frac{\sigma_j + \sigma_k}{a_j - a_k} \right)^2 = \text{minimum} \quad (j \neq k; j, k = 1, 2, \dots, s)$$

ekkor a mérések számát az a_j és a_k pontok között a szórások arányában kell megosztani

$$n_j = \frac{\sigma_j}{\sigma_j + \sigma_k} \cdot n \quad n_k = \frac{\sigma_k}{\sigma_j + \sigma_k} \cdot n$$

Amennyiben az n_i -re így kapott kifejezések nem egész számok, úgy azokat természetesen ki kell kerekíteni egész számmá.

Minden (a_j, a_k, a_l) értékhármashoz tartozik olyan (u_j^*, u_k^*, u_l^*) értékrendszer, vagyis a mérések számának olyan eloszlása, amely a regressziós együttható szórásnégyzetének minimuma szempontjából tekintetbe jöhet; ezek meghatározási módját az alábbi 5. §-ban ismertetjük. Ez az eljárás konkrét esetekben alkalmazható s a szélsőérték jellegét minden esetben külön meg kell állapítani. Az ott konstruált numerikus példában a három ponthoz kapott szórásnégyzet nagyobb, mint a három pont közül kiragadható kétpontokhoz tartozó szórásnégyzetek közül a legkisebb. Ha ez általában igaz — amit ezideig csupán a $\sigma_i = \text{állandó}$ esetében sikerült bizonyítani — akkor kétpontos mérés nyújtja a szórásnégyzetre az abszolút minimumot.

A hárompontos lokális szélsőérték meghatározása eléggé fáradságos számolást igényel, amit — ha a legjobb szórásnégyzetet keressük — bármely három pontra el kell végezni. Ezért az alábbiakban egy módszert ismertetünk,

melynek segítségével három vagy többpontos mérések esetén is meghatározhatjuk a mérések számának a pontok közti olyan eloszlását, amely a regressziós együttható szórásnégyzetének szempontjából viszonylag jó eredményre vezet, számítástechnikailag pedig igen jól kezelhető. Sőt, mint azt a 6. §-ban numerikus példán látni fogjuk, e módosított regressziós együttható minimális szórásnégyzete kisebb lehet hárompontos mérés esetén annál, mintha a regressziós együtthatót a szokott alakban — minimális szórásnégyzettel — határoztuk volna meg. Ez a módszer a regressziós egyenes következő módosított meghatározása alapján sikerül:

Tekintsük az a_i helyen mért $\eta_{i1}, \eta_{i2}, \dots, \eta_{in_i}$ értékek számtani közepét

$$\bar{\eta}_i = \frac{\sum_{k=1}^{n_i} \eta_{ik}}{n_i}.$$

Keressük most az n számú (a_i, η_{ik}) ($k = 1, 2, \dots, n_i$; $i = 1, 2, \dots, s$) pont helyett az $s < n$ számú $(a_i, \bar{\eta}_i)$ ponthoz tartozó regressziós egyenest. Ekkor a regressziós együttható

$$\varrho^* = \frac{\sum_{i=1}^s (a_i - \bar{a})(\eta_i - \bar{\eta})}{\sum_{i=1}^s (a_i - \bar{a})^2},$$

ahol most $\bar{a}$ az a_i -k számtani közepe:

$$\bar{a} = \frac{\sum_{i=1}^s a_i}{s}.$$

Ebben az esetben ahhoz, hogy

$$\sigma^2(\varrho^*) = \text{minimum}$$

legyen, n_j -t az a_j -ben végzett mérések számát — a következőképpen kell kiszámítani:

$$n_j = \frac{|a_j - \bar{a}| \sigma_j}{\sum_{i=1}^s |a_i - \bar{a}| \sigma_i} \cdot n,$$

emellett a szórásnégyzet minimuma:

$$\min \sigma^2(\varrho^*) = \frac{1}{n} \frac{\left[\sum_{i=1}^s |a_i - \bar{a}| \sigma_i \right]^2}{\left[\sum_{i=1}^s (a_i - \bar{a})^2 \right]}.$$

Ez a módszer tehát tetszőleges s mellett minimumot ad a regressziós együtthatókra. A pontok száma redukálódik, ha valamelyik a_i összeesik az $\bar{a}$ súlyponttal, ahol — mint azt formulánk mutatja — mérést nem kell végezni.

Ha $s = 2$, akkor a mérések számának eloszlását és a szórásnégyzet minimumát illetően ugyanarra az eredményre jutunk, ami megegyezik a regressziós együttható fenti — szokott — értelmezése alapján nyert eredménnyel.

A gyakorlati alkalmazások szempontjából megjegyezzük, hogy rendszerint elég az (a_j, a_k) értékpárokhoz tartozó extrémális szórások közül a legkisebbet venni és ennek megfelelően a mérést elvégezni; ezzel általában biztosíthatjuk, hogy nem túlságosan nagy számú összmérés igénybevételével a szórás a gyakorlat számára megfelelő kis értékkel bír.

5. A szórásnégyzet szélsőértékeinek meghatározása

Vizsgáljuk meg mindenekelőtt az $f(u_1, u_2, \dots, u_s, \omega)$ függvényt a mellékfeltételek figyelembevételével. Az I'2. feltétel szerint e függvénynek az s dimenziós egységgömb zárt felszínén kell minimumát keresni, éspedig az (I' 2.) mellékfeltétel figyelembevételével, mely szerint ω az x tengelyen az a_i pontoknak u_i^2 súlyok szerint vett súlypontja. E függvény pozitív és amint azt alább α) alatt megmutatjuk, a következő s számú pontban válik végtelenné :

$$u_i = 1, \quad u_1 = u_2 = \dots = u_{i-1} = u_{i+1} = \dots = u_s = 0,$$

ahol $i = 1, 2, \dots, s$. Minden egyéb helyen a függvény folytonos és a benne szereplő változók szerint akárhányszor differenciálható.

Megmutatjuk, hogy a függvény határozottan pozitív alsó korláttal bír. Legyen ugyanis a σ_i -k közül a legkisebb pl. a σ_1 . Ekkor, minthogy ω az a_i értékek pozitív súlyok szerint vett súlypontja, tehát

$$a_1 < \omega < a_s \quad \text{és így} \quad |a_i - \omega| \leq |a_1 - a_s|$$

$i = 1, 2, \dots, s$ -re. Felírhatjuk tehát — a σ_i -ket legkisebb értékükkel helyettesítve, majd $(a_i - \omega)^2$ értékeit a nevezőben nagyobbítva — a következő egyenlőséget :

$$f(u_1, u_2, \dots, u_s, \omega) \geq \frac{\sigma_1^2}{(a_1 - a_s)^2} > 0,$$

amivel állításunkat bebizonyítottuk.

Függvényünk tehát felveszi a mondott tartományban határozottan pozitív abszolút minimumát (esetleg több helyen is).

Az alábbi tárgyalásunkban arra szorítkozunk, hogy a Lagrange-féle multiplikátor eljárás segítségével meghatározzuk a változók ama értékrendszereit, amelyek a szélsőérték szempontjából figyelembe jöhetnek. Ezek, mint mondtuk, olyan tulajdonságúak, amelyek a változónak legfeljebb három 0-tól különböző értékét tartalmazzák. Ezek az értékrendszerek véges számban vannak. Meghatározásuk alapján lehet tehát a legkisebb szórásnégyzetet nyerni és így azt az u_i értékrendszert kiválasztani, ami a követendő mérési eljárást meghatározza.

Feladatunk tehát az (I')-ben adott $s + 1$ -változós függvény szélsőértékeinek meghatározása az (I'1.) és (I' 2.) mellékfeltételek figyelembevételével.

Alkalmazzuk a Lagrange-féle multiplikátor módszert. Keresnünk kell az $u_1, u_2, \dots, u_s, \omega$ változóknak, továbbá a λ és μ számoknak azon értékeit, amelyek mellett a következő egyenletek teljesülnek :

$$(II) \quad \frac{\partial f}{\partial u_j} + \lambda \frac{\partial g}{\partial u_j} + \mu \frac{\partial h}{\partial u_j} = 0, \quad j = 1, 2, \dots, s$$

$$(II') \quad \frac{\partial f}{\partial \omega} + \lambda \frac{\partial g}{\partial \omega} + \mu \frac{\partial h}{\partial \omega} = 0.$$

Ennek az $s + 1$ számú egyenletnek az I'. 1. és I'. 2. egyenletekkel való kiegészítése nyújt az $n + 3$ számú ismeretlen meghatározására kellő számú egyenletet.

További számításaink áttekinthetőbbé tételére vezessük be a következő jelöléseket :

$$(III.1) \quad M = \sum_{i=1}^s \sigma_i^2 (a_i - \omega)^2 u_i^2,$$

$$(III.2) \quad N = \sum_{i=1}^s (a_i - \omega)^2 u_i^2,$$

$$(III.3) \quad R = \sum_{i=1}^s \sigma_i^2 (a_i - \omega) u_i^2.$$

Tehát

$$\sigma^2(\varrho) = \frac{1}{n} \frac{M}{N^2}.$$

E jelölések figyelembevételével (II) és (II') egyenleteink — a parciális deriváltak kiszámítása után — a következő alakot öltik :

$$(II.1) \quad \frac{2\sigma_j^2(a_j - \omega)^2 u_j N^2 - 4(a_j - \omega)^2 u_j M N}{N^4} + 2a_j u_j \lambda + 2u_j \mu = 0$$

$j = 1, 2, \dots, s,$

$$(II'1) \quad \frac{-2 \left[\sum_{i=1}^s \sigma_i^2 (a_i - \omega) u_i^2 \right] N^2 - 4 \left[\sum_{i=1}^s (a_i - \omega) u_i^2 \right] M N}{N^4} - \lambda = 0.$$

A (II' 1.)-ből λ azonnal kifejezhető; a számlálóban a második zárójelben lévő kifejezés az (I'1) és (I'2) feltételek miatt 0 s így

$$\lambda = - \frac{2R}{N^2}$$

μ kiszámításához szorozzuk a II. 1. egyenletek mindegyikét a megfelelő u_j -vel s összegezzük az s számú egyenletet. Ekkor a III. jelölések, valamint az (I'1) és (I'2) feltételek figyelembevételével a következőt nyerjük :

$$\frac{2MN^2 - 4MN^2}{N^4} + 2\lambda\omega + 2\mu = 0.$$

S ebből λ fenti kifejezésének behelyettesítésével

$$\mu = \frac{M + 2R\omega}{N^2}$$

λ és μ e kifejezéseit (II. 1.)-be behelyettesítve, az $M' = \frac{M}{N}$ jelölés bevezetésével egyszerű átalakítás után a következőt nyerjük :

$$(II. 2) \quad u_j [(a_j - \omega)^2 (\sigma_j^2 - 2M') - 2(a_j - \omega) R + M] = 0, \quad j = 1, 2, \dots, s.$$

A keresett u_j és ω változóknak teljesíteniük kell még az (I'1) és (I'2) feltételeket. (II. 2.) egyenleteink mindegyikében tehát az u_j vagy a szögletes zárójelben álló kifejezés legalább egyikének el kell tűnnie.

Szélsőérték szempontjából tekintetbe jövő értékrendszereket tehát a következő módon kell meghatározunk : az s számú u_j közül l számú meghatározottat 0-tól különbözőnek, míg az $s - l$ számú többi 0-nak vesszük. Ekkor az l számú u_j -re a (II. 2.) egyenletek l egyenletet szolgáltatnak, továbbá az (I'1) és (I'2) feltételeket kell figyelembe venni. Azonban az l számú (II. 2.)-es egyenlet közül $l - 1$ -ből, valamint az (I. 2.) és (I'1) feltételekből már következik az l -ik. Így az $l + 1$ számú ismeretlen meghatározására $l + 1$ számú független egyenlet szolgál. Mégis — mint azt alább kimutatjuk — általában, vagyis az a_i és σ_i^2 ($i = 1, 2, \dots, s$) tetszőleges értékeire, ezek az egyenletek nem elégíthetők ki a felvetett kérdésnek megfelelő reális értékekkel, ha $l > 3$.

Most rátérünk azoknak az értékrendszereknek meghatározására, amelyek a (II. 2.) egyenleteket, valamint (I'1) és (I'2) feltételeket kielégítik.

a) Legyen az u_j -k közül egy — pl. az u_1 — 0-tól különböző, míg a többi 0. Megmutatjuk, hogy a feltételek mellett az $f(u_1, u_2, \dots, u_s, \omega)$ függvénynek ∞ helye van.

Végezzük ugyanis a következő határátmenetet :

$$u_1^2 \rightarrow 1$$

$$u_j^2 \rightarrow 0, \quad j = 2, 3, \dots, s$$

ekkor

$$\omega \rightarrow a_1$$

s könnyen meggyőződhetünk arról, hogy az $f(u_1, u_2, \dots, u_s, \omega)$ kifejezésének számlálója és nevezője egyaránt 0-hoz tart, azonban a nevező negyedrendben, míg a számláló csupán másodrendben, q. e. d.

A továbbiakban a számunkra tekintetbe jövő értékrendszerek meghatározásához tehát legalább két u_j -t kell 0-tól különbözőnek tekinteni. Ebből viszont következik, hogy a (III)-ban értelmezett M és N mennyiségek határozottan pozitívak :

$$M > 0, \quad N > 0.$$

Ugyanis (III)-ból látható, hogy ω legfeljebb egy a_i -val eshet össze, de ugyanakkor van még egy u_j ($i \neq j$), amely 0-tól különbözik.

β) Legyen most $l \geq 5$ és különbözzék pl. az $u_1, u_2, \dots, u_l$ a 0-tól, míg legyen $u_j = 0$, ha $j = l+1, l+2, \dots, s$.

Tekintsük e § első részében jelzett $l+1$ egyenletből álló egyenletrendszert $u_1, u_2, \dots, u_l, \omega$ -ra megoldva s számítsuk ki segítségükkel az M, R, M' mennyiségeket. Ekkor teljesülnie kell az l számú

$$(a_j - \omega)^2 (\sigma_j^2 - 2M') - 2(a_j - \omega) R + M = 0, \quad j = 1, 2, \dots, l$$

egyenletnek. Ez azonban azt jelenti, hogy a (ξ, η) koordinátarendszerben értelmezett

$$(\xi - \omega)^2 (\eta - 2M') - 2(\xi - \omega) R + M = 0$$

harmadrendű algebrai görbének az $l \geq 5$ számú $(\xi = a_j, \eta = \sigma_j^2)$ ($j = 1, 2, \dots, l$) ponton kell áthaladnia. Márpedig e négyparaméteres (M, M', R, ω) görbét 4 pontja teljesen meghatározza s előre tetszőlegesen megadott 5 pont általában nem fekszik rajta. Ezzel azonban megmutattuk, hogy — hacsak az (a_j, σ_j^2) ($j = 1, 2, \dots, s$) mennyiségek bizonyos feltételeknek eleget nem tesznek — a tárgyalt eset nem szolgáltat feladatunk számára lokális szélsőértéket.

γ) Az előző meggondolás azonban lényegében azt az állítást is tartalmazza, hogy általában már az $l = 4$ számú nullától különböző u_j felvétele nem vezet szélsőértékfeladatunk szempontjából reális eredményhez. Ugyanis az (a_j, σ_j^2) ($j = 1, 2, 3, 4$) mennyiségek már egyértelműen meghatározzák az ω, M, M', R mennyiségeket.

Ekkor a következő egyenletrendszernek kell teljesülnie:

$$(a_j - \omega)^2 (\sigma_j^2 - 2M') - 2(a_j - \omega) R + M = 0, \quad j = 1, 2, 3, 4.$$

Vezessük be még az L ismeretlent és írjuk egyenletrendszerünket a következő alakban:

$$\sigma_j^2 (a_j - \omega)^2 L - 2(a_j - \omega)^2 M' - 2(a_j - \omega) R + M = 0, \quad j = 1, 2, 3, 4.$$

E homogén egyenletrendszernek az L, M, M', R ismeretlenekre nézve kell legyen nem triviális megoldása (sőt $L = 1, M > 0, M' > 0$ feltételeknek eleget tevő megoldásnak is kell léteznie), amiből a determináns eltűnése következik. Így némű átalakítás után ω -ra nézve a következő egyenletet nyerjük:

$$\begin{vmatrix} \sigma_1^2 (a_1 - \omega)^2 & (a_1 - \omega)^2 & (a_1 - \omega) & 1 \\ \sigma_2^2 (a_2 - \omega)^2 & (a_2 - \omega)^2 & (a_2 - \omega) & 1 \\ \sigma_3^2 (a_3 - \omega)^2 & (a_3 - \omega)^2 & (a_3 - \omega) & 1 \\ \sigma_4^2 (a_4 - \omega)^2 & (a_4 - \omega)^2 & (a_4 - \omega) & 1 \end{vmatrix} = 0.$$

Ezt a determinánst egyszerű átalakítással olyan alakra hozhatjuk, amely ω -ra nézve másodfokú egyenletet szolgáltat.

Válasszuk most a következő értékeket:

$$a_1 = -2, \quad a_2 = -1, \quad a_3 = 1, \quad a_4 = 2$$

$$\sigma_1^2 = 1, \quad \sigma_2^2 = 11, \quad \sigma_3^2 = 19, \quad \sigma_4^2 = 5;$$

ω -ra a 0 és 8 értékeket nyerjük, minthogy azonban ω az x tengelyen az a_i pontoknak pozitív súlyokkal vett súlypontja, az $\omega = 0$ jöhet csak tekintetbe. Ehhez az értékhez azonban az

$$M = -16, \quad M' = -\frac{1}{2}, \quad R = 2$$

értékeket nyerjük, amelyek az

$$M > 0, \quad M' > 0,$$

feltételeknek nem tesznek eleget s így nem szolgálhatnak alapul az u_j -k kiszámítására, q. e. d.

δ) Térjünk most át az $l = 2$ esetre. Legyenek

$$u_1 \neq 0, \quad u_2 \neq 0, \quad u_3 = u_4 = \dots = u_s = 0.$$

Ekkor az

$$\begin{aligned} a_1 u_1^2 + a_2 u_2^2 &= \omega \\ u_1^2 + u_2^2 &= 1 \end{aligned}$$

és egyszerű számítással nyerjük az

$$\begin{aligned} N &= (a_1 - a_2)^2 u_1^2 u_2^2, \\ M &= (\sigma_1^2 u_2^2 + \sigma_2^2 u_1^2) N, \\ R &= \frac{\sigma_1^2 - \sigma_2^2}{a_1 - a_2} N, \\ M' &= \sigma_1^2 u_2^2 + \sigma_2^2 u_1^2 \end{aligned}$$

összefüggéseket. E kifejezéseket a $j = 1$ és 2-re felírt (II. 2.) egyenletekbe helyettesítve, megfelelő egyszerűsítés után u_1^2 -re a következő egyenletet nyerjük:

$$(\sigma_1^2 - \sigma_2^2) u_1^4 - 2\sigma_1^2 u_1^2 + \sigma_1^2 = 0.$$

Innen

$$u_1^2 = \frac{\sigma_1^2 \pm \sigma_1 \sigma_2}{\sigma_1^2 - \sigma_2^2}.$$

A négyzetgyök pozitív előjele $u_1^2 > 1$ eredményhez vezetne s így a negatív előjelet véve bevezetésünkben állított

$$u_1^2 = \frac{\sigma_1}{\sigma_1 + \sigma_2}$$

eredményünkhöz jutottunk. Ebből

$$u_2^2 = 1 - u_1^2 = \frac{\sigma_2}{\sigma_1 + \sigma_2}$$

s a szórásnégyzet — némi számolás eredményeként —

$$\sigma^2(\varrho) = \frac{1}{n} \frac{M}{N^2} = \frac{1}{n} \left(\frac{\sigma_1 + \sigma_2}{a_1 - a_2} \right)^2.$$

ε) Bebizonyítjuk $\sigma_i = \sigma$ ($i = 1, 2, \dots, s$) esetre vonatkozó állításunkat.
Ekkor

$$M = \sigma^2 \sum_{i=1}^s (a_i - \omega)^2 u_i^2 = \sigma^2 \cdot N$$

$$R = \sigma^2 \sum_{i=1}^s (a_i - \omega) u_i^2 = 0.$$

(II.) egyenleteink így alakulnak :

$$\sigma^2 N u_j [-(a_j - \omega)^2 + N] = 0.$$

Tegyük fel most, hogy az u_j -k közül három különbözik 0-tól :

$$u_1 \neq 0, u_2 \neq 0, u_3 \neq 0, u_4 = u_5 = \dots = u_s = 0.$$

$\sigma > 0$, $N > 0$ miatt kell tehát; hogy

$$-(a_j - \omega)^2 + N = 0, \quad j = 1, 2, 3$$

teljesüljön. Ebből

$$(a_1 - \omega)^2 = (a_2 - \omega)^2 = (a_3 - \omega)^2$$

következnék, minthogy azonban az a_i -k valósak és egymástól különbözőek, ezek az egyenlőségek nem teljesülhetnek.

i) Az $y = Rx$ esetben eredményünket legegyszerűbben közvetlenül vezethetjük le. Ekkor ugyanis a regressziós együttható a már használt jelöléseinkkel :

$$R = \frac{\sum_{i=1}^s n_i a_i \overline{\eta_i}}{\sum_{i=1}^s n_i a_i}.$$

Ennek szórásnégyzete pedig az u_i változók bevezetésével :

$$\sigma^2(R) = \frac{1}{n} \frac{\sum_{i=1}^s \sigma_i^2 a_i^2 u_i^2}{\left(\sum_{i=1}^s a_i^2 u_i^2 \right)^2}.$$

E kifejezést kell minimummá tenni a $\sum_{i=1}^s u_i^2 = 1$ mellékfeltétel figyelembevételével.

A *Lagrange* multiplikátor módszer a következő feltételi egyenletekhez vezet:

$$u_i \frac{a_i^2(\sigma_i^2 - 2M') + M}{N^2} = 0, \quad i = 1, 2, \dots, s.$$

Itt

$$M = \sum_{i=1}^s \sigma_i^2 a_i^2 u_i^2$$

$$N = \sum_{i=1}^s a_i^2 u_i^2$$

$$M' = \frac{M}{N}.$$

Azonban a β) és γ) pontokban alkalmazott megfontolásaink itt arra az eredményre vezetnek, hogy az u_i -k közül legfeljebb egy választható 0-tól különbözőnek, hacsak az (a_i, σ_i^2) ($i = 1, 2, \dots, s$) értékpárok között bizonyos speciális összefüggések nem állanak fenn. (Ilyen összefüggés pl. az $a_i = -a_j$ és $\sigma_i = \sigma_j$, amely esetben u_i és u_j is választható 0-tól különbözőnek.) Legyen tehát

$$u_i \neq 0, \quad u_j = 0, \quad \text{ha } j \neq i.$$

Ekkor

$$a_i^2(\sigma_i^2 - M') + M = 0$$

kifejezésből $u_i^2 = 1$ adódik s ebből

$$\sigma^2(\varrho) = \frac{1}{n} \left(\frac{\sigma_i}{a_i} \right)^2.$$

Tehát az összes mérést abban az a_i pontban kell végezni, amelyre $\sigma^2(\varrho)$ itt kapott kifejezése minimum.

κ) Oldjuk most meg a hárompontos feladatot. A (II. 2.) egyenleteket $j = 1, 2, 3$ -ra felírva, ezekből M, M', R kifejezhetők ω segítségével:

$$M = - \frac{(a_1 - \omega)(a_2 - \omega)(a_3 - \omega)}{A} [\sigma_1^2(a_2 - a_3)(a_1 - \omega) + \sigma_2^2(a_3 - a_1)(a_2 - \omega) + \sigma_3^2(a_1 - a_2)(a_3 - \omega)],$$

$$M' = \frac{1}{2A} [\sigma_1^2(a_2 - a_3)(a_1 - \omega)^2 + \sigma_2^2(a_3 - a_1)(a_2 - \omega)^2 + \sigma_3^2(a_1 - a_2)(a_3 - \omega)^2]$$

$$R = - \frac{1}{2A} [\sigma_1^2(a_2 - a_3)(a_1 - \omega)^2(a_2 + a_3 - 2\omega) + \sigma_2^2(a_3 - a_1)(a_2 - \omega)^2(a_1 + a_3 - 2\omega) + \sigma_3^2(a_1 - a_2)(a_3 - \omega)^2(a_1 + a_2 - 2\omega)],$$

ahol

$$A = a_1^2(a_2 - a_3) + a_2^2(a_3 - a_1) + a_3^2(a_1 - a_2).$$

Innen $N = \frac{M}{M'}$ és felírhatjuk u_1, u_2, u_3 és ω -ra a következő egyenleteket:

$$\sigma_1^2(a_1 - \omega)^2 u_1^2 + \sigma_2^2(a_2 - \omega)^2 u_2^2 + \sigma_3^2(a_3 - \omega)^2 u_3^2 = M$$

$$(a_1 - \omega)^2 u_1^2 + (a_2 - \omega)^2 u_2^2 + (a_3 - \omega)^2 u_3^2 = N$$

$$\sigma_1^2(a_1 - \omega) u_1^2 + \sigma_2^2(a_2 - \omega) u_2^2 + \sigma_3^2(a_3 - \omega) u_3^2 = R$$

$$(a_1 - \omega) u_1^2 + (a_2 - \omega) u_2^2 + (a_3 - \omega) u_3^2 = 0$$

$$u_1^2 + u_2^2 + u_3^2 = 1.$$

Az itt felírt öt egyenlet közül bármely négyből kiszámíthatók az ismeretlenek s azok az ötödiket kielégítik.

ω meghatározására legegyszerűbben úgy járhatunk el, hogy a második egyenletet elhagyjuk s a jobboldalon álló $M, R, 0, 1$ -hez u_i^2 tényezőt, mint további ismeretlent írunk. Így homogén egyenletrendszerhez jutunk, amelynek nem triviális megoldását a determináns eltűnése biztosítja:

$$\begin{vmatrix} \sigma_1^2(a_1 - \omega)^2 & \sigma_2^2(a_2 - \omega)^2 & \sigma_3^2(a_3 - \omega)^2 & -M \\ \sigma_1^2(a_1 - \omega) & \sigma_2^2(a_2 - \omega) & \sigma_3^2(a_3 - \omega) & -R \\ (a_1 - \omega) & (a_2 - \omega) & (a_3 - \omega) & 0 \\ 1 & 1 & 1 & -1 \end{vmatrix} = 0.$$

Így általában hatodfokú egyenlethez jutottunk, amely némi átalakítás után egy legfeljebb harmadfokú egyenletbe vihető át. Gyökei közül a (a_1, a_3) intervallumba esők jönnek figyelembe szélsőértékfeladatunk szempontjából.

Példaképpen legyenek

$$a_1 = -1, \quad a_2 = 1, \quad a_3 = 2,$$

$$\sigma_1^2 = 1, \quad \sigma_2^2 = 1, \quad \sigma_3^2 = 4,$$

ekkor

$$M = (1 - \omega)(1 + \omega)(2 - \omega)^2,$$

$$M' = \frac{1}{2}(5 - 4\omega + \omega^2),$$

$$R = \omega(2 - \omega)^2.$$

ω -ra nézve a következő egyenletet kapjuk:

$$2\omega^3 - 11\omega^2 + 14\omega - 3 = 0.$$

A $(-1, 2)$ intervallumba eső gyökök: $\omega = 0,26794\dots$ és $\omega' = 1,5$. Kiszámítva az ezekhez az értékekhez tartozó u_i -ket, azt találjuk, hogy feladatunknak csupán az $\omega = 0,26794\dots$ felel meg, ω' -nél ugyanis u_i^2 -ra nézve negatív mennyiségeket is kapunk.

ω ezen értéke mellett

$$u_1^2 = 0,4433\dots$$

$$u_2^2 = 0,4019\dots$$

$$u_3^2 = 0,1546\dots$$

A regressziós együttható szórásnégyzete:

$$\sigma^2(\varrho) = \frac{1}{n} \cdot 1,43649 \dots$$

Hasonlítsuk össze a szórásnégyzetre kapott eredményünket a három pontból választható két-két ponthoz tartozó extrémális szórásnégyzetekkel:

$$a_1 = -1, a_2 = 1, \sigma^2(\varrho) = \frac{1}{n} \cdot 1$$

$$a_1 = -1, a_3 = 2, \sigma^2(\varrho) = \frac{1}{n} \cdot 1$$

$$a_2 = 1, a_3 = 2, \sigma^2(\varrho) = \frac{1}{n} \cdot 9.$$

Tehát a három ponthoz kapott szórásnégyzetnél jobb eredményt kapunk, ha vagy az $a_1 = -1$ és $a_2 = 1$ pontokban mérünk, a mérések számát fele fele arányban osztva meg, vagy pedig az $a_1 = -1$ és $a_3 = 2$ pontokban mérünk, az elsőben a mérések

$$\frac{\sqrt{1}}{\sqrt{1} + \sqrt{4}} = 33,3 \dots \% \text{-át},$$

a másodikban

$$\frac{\sqrt{4}}{\sqrt{1} + \sqrt{4}} = 66,6 \dots \% \text{-át}$$

végezve.

6. A regressziós együttható módosított alakjának vizsgálata

A regressziós egyenest tehát az egyes (a_i, η_{ik}) $k = 1, 2, \dots, n_i$, $i = 1, 2, \dots, s$ összesen n számú pont helyett csupán az s számú $(a_i, \bar{\eta}_i)$ ponton vezetjük keresztül, ahol az s számú

$$\bar{\eta}_i = \frac{\sum_{k=1}^{n_i} \eta_{ik}}{n_i}, \quad i = 1, 2, \dots, s$$

egymástól független értékek, rendre

$$y_i = \varrho^* a_i + c, \quad i = 1, 2, \dots, s$$

várható értékkel és

$$\sigma^2(\bar{\eta}_i) = \frac{\sigma^2(\eta_{i,k})}{n_i} = \frac{\sigma_i^2}{n_i}$$

szórásnégyzettel. A tapasztalati regressziós együttható

$$\varrho^* = \frac{\sum_{i=1}^s (a_i - \bar{a}) (\bar{\eta}_i - \bar{\eta})}{\sum_{i=1}^s (a_i - \bar{a})^2},$$

ahol most

$$\bar{a} = \frac{\sum_{i=1}^s a_i}{s}$$

tehát $\bar{a}$ — az a_i pontok súlypontjának koordinátája — most állandó. Tekintve, hogy

$$\sum_{i=1}^s (a_i - \bar{a}) = 0,$$

tehát

$$\varrho^* = \frac{\sum_{i=1}^s (a_i - \bar{a}) \bar{\eta}_i}{\sum_{i=1}^s (a_i - \bar{a})^2}.$$

ϱ^* várható értéke nyilván a regressziós együttható elméleti értéke.

A szórásnégyzet, tekintettel az $\bar{\eta}_i$ ($i = 1, 2, \dots, s$) valószínűségi változók függetlenségére, továbbá az egyes η értékek szórásnégyzetére vonatkozó feltevésekre

$$\sigma^2(\varrho^*) = \frac{\sum_{i=1}^s (a_i - \bar{a})^2 \frac{\sigma_i^2}{n_i}}{\left[\sum_{i=1}^s (a_i - \bar{a})^2 \right]^2}.$$

Feladatunk ennek a kifejezésnek minimummá tétele, ahol a változókra a

$$\sum_{i=1}^s n_i = n$$

feltétel kell teljesüljön.

Vezessünk be ismét új változókat, legyen

$$\lambda_i = \frac{n_i}{n}, \quad i = 1, 2, \dots, s$$

amelyeket ismét folytonos változóknak tekintünk. Ekkor szélsőértékfeladatunk így fogalmazható:

Meghatározandók az

$$f(\lambda_1, \lambda_2, \dots, \lambda_s) = \frac{1}{n} \frac{\sum_{i=1}^s (a_i - \bar{a})^2 \frac{\sigma_i^2}{\lambda_i}}{\left[\sum_{i=1}^s (a_i - \bar{a})^2 \right]^2}$$

függvény szélső értékei a

$$g(\lambda_1, \lambda_2, \dots, \lambda_s) = \sum_{i=1}^s \lambda_i - 1 = 0$$

mellékfeltétel figyelembevételével.

A *Lagrange* multiplikátor módszert alkalmazva nyerjük a

$$\frac{\partial f}{\partial \lambda_j} + \mu \frac{\partial g}{\partial \lambda_j} = -C(a_j - \bar{a})^2 \frac{\sigma_j^2}{\lambda_j^2} + \mu = 0$$

egyenleteket, ahol

$$C = \frac{1}{n} \frac{1}{\left[\sum_{i=1}^s (a_i - \bar{a})^2 \right]^2}.$$

Ezért az

$$\frac{(a_j - \bar{a})^2 \sigma_j^2}{\lambda_j^2} = C' = \text{állandó}, \quad j = 1, 2, \dots, s$$

összefüggéseknek kell teljesülniök. Tekintettel arra, hogy $C' > 0$,

$$\lambda_j = |a_j - \bar{a}| \cdot \sigma_j \cdot \frac{1}{C'}, \quad j = 1, 2, \dots, s.$$

λ_j ezen értékeinek j -re való összegezésével megkapjuk C' értékét s λ_j -re nyerjük, hogy

$$\lambda_j = \frac{|a_j - \bar{a}| \sigma_j}{\sum_{i=1}^s |a_i - \bar{a}| \sigma_i}, \quad j = 1, 2, \dots, s,$$

λ_j ezen értékei mellett a szórásnégyzet

$$\sigma^2(\varrho^*) = \frac{1}{n} \left(\frac{\sum_{i=1}^s |a_i - \bar{a}| \sigma_i}{\sum_{i=1}^s (a_i - \bar{a})^2} \right)^2.$$

Azt, hogy s adott értéke mellett a λ -k ily módon való megválasztása abszolút minimumot ad, a következőképpen láthatjuk be:

Függvényünk alulról határozottan pozitív korláttal bír. Vegyük ugyanis a legkisebbet a σ_i -k közül (legyen ez pl. σ_1) és a legnagyobbat a λ_i -k közül (legyen ez pl. λ_2 , a $\sum_{i=1}^s \lambda_i = 1$ miatt $\lambda_2 \geq \frac{1}{s}$). Ekkor

$$f(\lambda_1, \lambda_2, \dots, \lambda_s) \geq \frac{1}{n} \frac{\sigma_1^2}{\lambda_2} \frac{1}{\sum_{i=1}^s (a_i - \bar{a})^2}.$$

Másrészt a függvénydetermináns — a szabad feladatra is — határozottan pozitív; ugyanis.

$$\frac{\partial f}{\partial \lambda_j^2} = \frac{1}{n} \frac{(a_j - \bar{a})^2 \frac{\sigma_j^2}{\lambda_j}}{\left[\sum_{i=1}^s (a_i - \bar{a})^2 \right]^2} > 0, \quad \frac{\partial^2 f}{\partial \lambda \cdot \partial \lambda_k} = 0$$

s így a függvénydetermináns s számú pozitív mennyiség szorzata.

Vegyük példaképpen az 5. § 2)-ben tárgyalt példát. Eszerint

$$a_1 = -1, \quad a_2 = 1, \quad a_3 = 2, \quad \bar{a} = \frac{2}{3},$$

$$\sigma_1 = 1, \quad \sigma_2 = 1, \quad \sigma_3 = 2.$$

A regressziós együttható szórásnégyzetének minimuma:

$$\min \sigma^2(*) = \frac{1}{n} \cdot 1 = \frac{1}{n}.$$

Itt tehát

$$\min \sigma^2(\varrho^*) < \min \sigma^2(\varrho)$$

esete forog fenn, vagyis az $a_1 = 1, a_2 = 1, a_3 = 2$ pontokban a mérések számát rendre a fenti formulából nyert

$$\lambda_1 = 35,7 \dots \%, \quad \lambda_2 = 7,1 \dots \%, \quad \lambda_3 = 57,2 \dots \%$$

arányban megosztva ugyanolyan kis szórásnégyzettel nyerjük a regressziós együtthatót, mint az (a_1, a_3) vagy az (a_1, a_2) pontpárok esetében, ahol a regressziós együttható mindkét értelmezése szerint ugyanazt a

$$\min \sigma^2(\varrho) = \min \sigma^2(\varrho^*) = \frac{1}{n}$$

eredményt kapjuk.

7. A mérések számának meghatározása

Eddig elért eredményeink gyakorlati felhasználása szempontjából kiegészítésként az alábbiakban röviden két kérdéssel foglalkozunk:

a) hogyan kell a mérések összes számát — n -et — megválasztani ahhoz, hogy a konkrét feladatnak jól megfelelő értéket kapjunk a tapasztalati regressziós együtthatóra,

b) milyen eljárást kövessünk abban a gyakran felmerülő esetben, amikor az a) feladat alapján meghatározott n érték mellett az nu_i^2 mennyiségek nem egész számok.

a) Legyen a regressziós együttható szórása minimum az a_1, a_2, a_3 , ill. a_1, a_2 pontokban u_1^2, u_2^2, u_3^2 , ill. u_1^2, u_2^2 értékek mellett végzett mérések esetén. Legyen a minimum

$$\sigma(\varrho) = \frac{1}{\sqrt{n}} \frac{\left(\sum_{i=1}^r \sigma_i^2 (a_i - \omega)^2 u_i^2 \right)^{\frac{1}{2}}}{\sum_{i=1}^r (a_i - \omega)^2 u_i^2} = \frac{1}{\sqrt{n}} \cdot A,$$

ahol most $r = 3$, ill. 2 . E kifejezésben az A tényezőt az a_j és σ_j mennyiségek ($j = 1, 2 \dots s$) határozzák meg, így azt a követelményt, hogy $\sigma(\varrho)$ megfelelően kicsi legyen, n megválasztásával érhetjük el.

Ha a $\sigma(\varrho) < \varepsilon$ követelményt tesszük, akkor nyilván az $n > \frac{A^2}{\varepsilon^2}$ egyenlőtlenségnek kell teljesülnie a mérések számára nézve.

Megjegyezzük, hogy ha a mérések száma nem túl kicsi (gyakorlatilag, ha meghaladja pl. az 50-et), akkor a regressziós együttható jó közelítéssel normális eloszlást követ s ekkor pl. 99,7% valószínűséggel állíthatjuk, hogy a regressziós együttható valódi értéke — R — a következő egyenlőtlenségnek tesz eleget:

$$\varrho - 3\sigma(\varrho) < R < \varrho + 3\sigma(\varrho).$$

Ha ekkor ϱ a 0-tól határozottan különbözik, akkor n -nek a

$$\frac{3\sigma(\varrho)}{|\varrho|} = \frac{3A}{\sqrt{n}\varrho} < \varepsilon,$$

vagyis

$$n \geq \frac{9A^2}{\varepsilon^2 \varrho^2}$$

megválasztása mellett 99,7%-os valószínűséggel állíthatjuk, hogy a valódi és tapasztalati együttható hányadosa legfeljebb ε -nal különbözik az 1-től.

Vegyük az 5. § x)-ban tárgyalt példát.

$$a_1 = -1, \quad a_2 = 1, \quad a_3 = 2, \\ \sigma_1^2 = 1, \quad \sigma_2^2 = 1, \quad \sigma_3^2 = 4.$$

Válasszuk a méréseket az $a_1 = -1$ és $a_3 = 2$ pontokban, amikoris

$$u_1^2 = \frac{1}{3}, \quad u_3^2 = \frac{2}{3} \quad \text{és} \quad \sigma^2(\varrho) = \frac{1}{n}.$$

Ha az összes mérések számát 81-nek választjuk, akkor

$$\sigma(\varrho) = \frac{1}{9}$$

és 99,7% valószínűséggel állíthatjuk, hogy a regressziós együttható valódi értéke egy $6 \cdot \sigma(\varrho) = \frac{2}{3}$ terjedelmű intervallumba esik.

Ugyanakkor végezzünk összesen 81 mérést az a_1, a_2, a_3 pontokban, a mérések számát egyenletesen osztva el. Ekkor

$$u_1^2 = u_2^2 = u_3^2 = \frac{1}{3}, \quad \omega = \frac{2}{3}$$

és

$$\sigma^2(\varrho') = \frac{1}{n} \cdot 1,38 \dots,$$

ami azt jelenti, hogy a mérések számát 38%-al nagyobbak, tehát 112-nek kell választani ahhoz, hogy a regressziós együtthatóra ugyanolyan biztonságot érjünk el, mint a kétpontos extrémális mérés mellett.

b) A mérések számának az előzőekben mondottak szerinti meghatározása mellett sok esetben n -nek további kevés változtatásával még azt is elérhetjük, hogy nu_i^2 ($i = 1, 2, 3$; ill. $i = 1, 2$) is egész szám legyen.

Előfordulhat azonban, hogy a mérések elvégzése költséges, vagy már eleve kevés számú megfigyelés áll rendelkezésére. Erre az esetre megvizsgáljuk, hogy kétpontos mérés esetén milyen mértékben tér el a szórás a minimálistól, ha az nu_i^2 értékek helyett a hozzájuk közeleső egész számokat választjuk.

Láttuk kétpontos mérés esetén

$$u_1^2 = \frac{\sigma_1}{\sigma_1 + \sigma_2}, \quad u_2^2 = \frac{\sigma_2}{\sigma_1 + \sigma_2},$$

a minimális szórásnégyzet

$$\sigma^2(\varrho) = \frac{1}{n} \left(\frac{\sigma_1 + \sigma_2}{a_1 - a_2} \right)^2.$$

Vezessük be az $u_1^2 = \alpha_1, u_2^2 = \alpha_2$ jelöléseket és legyen $\alpha_1 \leq \alpha_2$. n -et minden-

esetre úgy helyes megválasztani, hogy $n\alpha_1 \geq 1$ legyen, amikor is $[n\alpha_1] \geq 1$ (a szögletes zárójel $n \cdot \alpha_1$ egész részét jelenti), tehát

$$n \geq \frac{1}{\alpha_1}$$

kell legyen, hogy az α_1 pontba is jusson legalább egy mérés. Számítsuk most ki a szórásnégyzetet, ha

$$u_1^2 = [n\alpha_1], \quad u_2^2 = n - [n\alpha_1].$$

Itt az $n\alpha_1$ -et kerekítettük lefelé, de hasonló eredményre jutunk, ha $u_1^2 = [n\alpha_1] + 1$ -et választunk.

Ekkor — elemi számolással — a szórásnégyzetre a következő kifejezést nyerjük

$$\sigma^2(\varrho') = \sigma^2(\varrho) n \left(\frac{\alpha_1^2}{[n\alpha_1]} + \frac{\alpha_2^2}{[n\alpha_2]} \right),$$

ahol $\sigma(\varrho)$ a minimális szórási ; n megválasztásánál tehát arra kell ügyelnünk, hogy $\sigma^2(\varrho)$ tényezője megfelelően kevés haladjon meg az 1-et.

Lássuk példaképpen a $\sigma_1 = \sigma_2$ esetet, amikor is $\sigma_1 = \sigma_2 = \frac{1}{2}$. A mérések száma legyen páratlan $n = 2k + 1$. Ekkor $[n\alpha_1] = k$ és

$$\sigma^2(\varrho') = \sigma^2(\varrho) \frac{(2k+1)^2}{4k(k+1)}.$$

Könnyű belátni, hogy a $\frac{(2k+1)^2}{k(k+1)}$ kifejezés $k (\geq 1)$ növekedésével csökken. s legnagyobb értékét $k = 1$ -re veszi fel, tehát

$$\frac{(2k+1)^2}{4k(k+1)} \leq \frac{9}{8}, \quad \text{ha } k \geq 1,$$

amiből

$$\sigma(\varrho) < \sigma(\varrho') \leq \sigma(\varrho) \sqrt{\frac{9}{8}} = \sigma(\varrho) \cdot 1,06 \dots$$

Tehát abban a szélső esetben, amikor 3 mérés áll rendelkezésre s az egyik pontban egy, a másikban két mérést végzünk, a szórási mindössze 6%-kal növekszik, de a mérések számának növelésével ez a különbség tetszőlegesen csökkenthető.

ОБ ОПРЕДЕЛЕНИИ КОЭФФИЦИЕНТА РЕГРЕССИИ

И. ВИНЦЕ .

Р е з ю м е

Между переменными x и y существует линейная связь : $y = Rx + Q$. Для определения констант R и Q исполняем N измерений при фиксированных значениях от $a_1, a_2, \dots, a_s$ — Спрашивается, как разделять N измерений на точки a_i , чтобы дисперсия эмпирического коэффициента регрессии была минимальна. Полагается, что a_i —

точные значения, т. е. не случайные величины, далее : величины y , измеренные в a_i имеют тождественные распределения, и их дисперсия $\sigma(a_i)$ зависит от точки a_i .

Доказывается, что при этих условиях достаточно определить прямую регрессии на основе измерений, исполненных в двух или не более трех точках из s точек a_i . Для определения экстремальных двух и трех точек т. е. для выбора a_i и разделения числа измерений между точками вырабатывается метод.

Если σ_i — константа, ($i = 1, 2, \dots$), то мы всегда должны исполнять измерения в двух точках — именно в двух крайних точках a_i , — разделяя число измерений в равной пропорции.

Работа занимается тоже модифицированной определением коэффициента регрессии, при которой задача экстремума трактуется проще и приводит к результату, обслуживаемому проще, хотя он не безусловно лучший.

SUR LA DÉTERMINATION DU COEFFICIENT DE RÉGRESSION

E. VINCZE

Résumé

Entre les variables x et y existe une relation linéaire : $y = Rx + Q$. Pour déterminer les constantes R (et Q) on effectue des mesurages dont le nombre total est N , avec les valeurs $a_1, a_2, \dots, a_s$ fixées de x . La question se pose, comment il faut distribuer parmi les points a_i les mesurages de nombre total N , pour que la carré de dispersion du coefficient de régression empirique R soit un minimum. On suppose que les a_i sont des valeurs exactes c'est-à-dire qu'ils ne sont pas des variables stochastiques, et, en outre, que les valeurs y mesurées avec les valeurs a_i ont la même distribution leur dispersion dépendant de l'endroit a_i : $\sigma(a_i)$.

Nous prouverons qu'avec ces conditions il suffit de déterminer la droite de régression à l'aide des mesurages effectués dans deux, ou tout au plus dans trois endroits parmi les endroits a_i de nombre s . Pour déterminer les extrémales à deux et à trois points c'est-à-dire pour choisir les a_i et distribuer le nombre de mesurages parmi ces points, nous élaborons une méthode.

Si σ_i est constant, ($i = 1, 2, \dots, s$), c'est toujours dans deux points — notamment dans les deux points a_i extrêmes — qu'il faut effectuer des mesurages en distribuant le nombre des mesurages en proportion égale.

L'article s'occupe en outre d'une définition modifiée du coefficient de régression, dans laquelle le problème de valeur extrême peut être discuté plus simplement et conduit à un résultat qui — même s'il n'est pas certainement meilleur — peut être manié plus simplement.