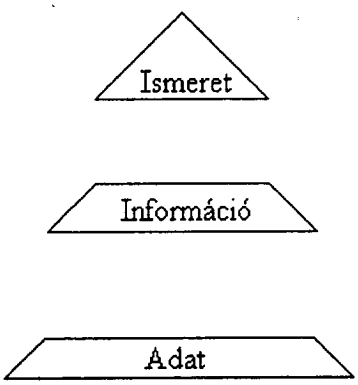
KOVÁCS PÉTER – PETRES TIBOR

A Petres-féle *Red*-mutató

1. Bevezetés

Az évezred elején, a globalizálódó világunkban nagy mértékben növekszik mindannyiunk információigénye. Az adatok mennyiségének robbanásszerű növekedése nem jár együtt a megfelelő mértékű információ-növekedéssel. A két fogalom közötti jelentős különbséget az alábbi ábra szemlélteti.



Rendszerezett információk összessége, problémák megoldását teszi lehetővé.

Döntéshozatalt szolgáló hasznos tartalmat hordozó adatok összessége. Minőségét az határozza meg, hogy milyen mértékben alkalmazható.

Tárolt formájában független, tényszerű szám vagy szöveg. Minőségét pontossága, elérhetősége határozza meg.

Igazából a döntéshozóknak nem az adatok hiányával, hanem azok bőségével kell szembenézniük, ugyanis (még a legóvatosabb becslések szerint is) az elektronikusan tárolt adatok volumene évente legalább megkétszereződik. A rendelkezésre álló adatok nagy mennyisége növeli ezek elemzésének összetettségét és az adatelemzőkkel szemben támasztott elvárásokat. Mivel az adatok információvá alakítása kisebb sebességgel történik, mint azok rendelkezésre bocsátása, ezért a felhasználóknak egyre inkább adatelemzési szakértővé kell válniuk, ismerniük kell azokat a módszereket, amelyekkel az adatok értékelhetőek és hasznosíthatóak.

A többváltozós statisztikai elemzéseknél két nézőpont ismeretes. Az egyik szerint az összes rendelkezésre álló változót szerepeltetjük, míg a másik szerint csak kevesebb változót használunk, amik azonban sűrítve tartalmazzák az (eredeti) adatállományban rejlő információt. Vagyis, képletesen szólva, az első szerint egy „narancs" egészét tekintjük, míg az utóbbi szerint ennek csak kivonatát, a „narancslét".

Mindezekből következően az alkalmazott modellek két csoportját lehet megkülönböztetni.

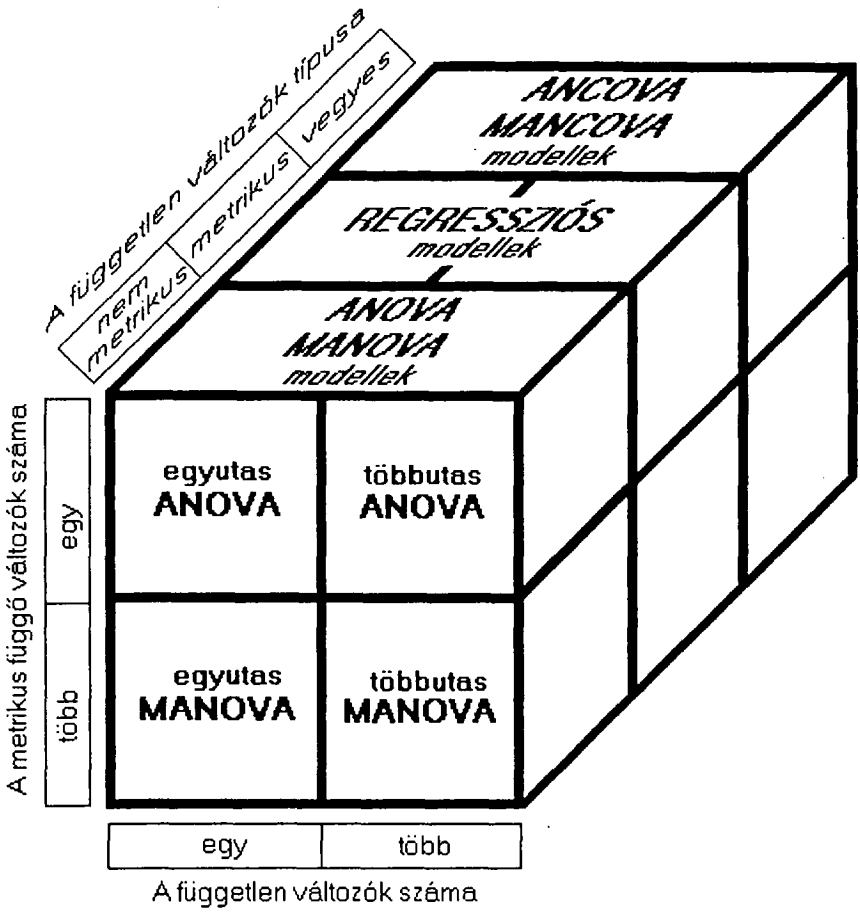
- Az ún. adatredukciós modellek esetén a változók számának csökkentésére törekszünk úgy, hogy ez a lehető legkevesebb információveszteséggel járjon. (Ebben az esetben nincs értelme a változók függő-független megkülönböztetésének.)
- Az ún. magyarázó modellek alkalmazásakor összefüggések feltárására törekszünk, vagy az összes rendelkezésre álló változó alapján, vagy az ezekből származtatott (kevesebb számú) változó segítségével. Ebből következően megkülönböztetünk független (magyarázó-) és függő (eredmény-) változókat.

Az egyes magyarázó modellek alapvetően abban különböznek egymástól, hogy hány változóból állnak, illetve milyen mérési szintű adatokat tartalmaznak.

Kizárólag egy független- és egy függő változót tartalmazó modellek a legegyszerűbbek, leggyakrabban azonban több független és csak egy függő változónk van.

A függő változó szempontjából két nagy csoport létezik: az egyiknél a függő változó metrikus, míg a másiknál nemmetrikus. A független változók is lehetnek metrikus és nemmetrikus mérési szintűek, illetve egyszerre mindkét típusú változó szerepeltetése is előfordulhat.

A fentiek szerint a metrikus függő változó(ka)t tartalmazó modellek grafikus ábrája a következő.



Az ábrán szereplő esetek közül a redundancia mérésének szempontjából kizárólag a metrikus adatok relevánsak. A metrikus adatok információtartalma az empirikus elemzéseknél lényeges kérdés, mert a nagyon nagy mennyiségű adat gyakran kevés információt hordoz, azaz nagymértékű a redundancia. Ez utóbbi alatt a vizsgálat szempontjából újabb információt, érdemleges közlést már nem tartalmazó, „felesleges" adatokat értjük. Ennek a problematikának a bemutatása céljából a továbbiakban a regresszió-számítást alkalmazzuk.

2. Előzmények

Többváltozós empirikus elemzéseknél a statisztikai módszerek közül leggyakrabban a regressziós modell kerül alkalmazásra, melynek legismertebb

típusa a standard lineáris regressziós modell. Ez mátrixalgebrai jelöléssel az $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ formában is felírható.

A modellben szereplő ismeretlen paraméterek – n megfigyelésből álló minta alapján meghatározott – becslőfüggvénye a közönséges legkisebb négyzetek módszere (OLS) szerint $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$.

A becsült paraméterek varianciáit a $Var(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ képlet alapján tudjuk kiszámítani.

Mivel a fenti összefüggésnél a $\boldsymbol{\varepsilon}$ hibatagok σ^2 szórásnégyzete számunkra ismeretlen, ezért ennek értékét az OLS szerint az $s_e^2 = \frac{\mathbf{e}'\mathbf{e}}{n-m-1}$ képlettel adott reziduális szórásnégyzettel tudjuk torzítatlanul becsülni.

A regresszió-számítás gyakorlati alkalmazásakor ügyelnünk kell arra, hogy a standard lineáris regressziós modellt ne használjuk, ha valamelyik feltétele szignifikánsan nem teljesül! Ezért, a továbbiakban figyelmünket a modell specifikációjában szereplő egyik feltétel hiányára, a magyarázóváltozók együttmozgásának jelenlétére fordítjuk. Ezt azért tesszük, mert ha a tényezőváltozók átlagos együttmozgása szignifikáns, akkor a $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ becslőfüggvénnyel kapott becslések – ceteris paribus – instabillá válnak.

A multikollinearitás szélsőséges esete – azaz a lineáris függőség – a gyakorlatban nem okoz gondot, de a tényezőváltózók között különböző mértékű sztochasztikus kapcsolat – különösen fogyasztáselemzésnél – szinte mindig jelentkezik, ezért fontos számszerűsíteni az említett kapcsolat(ok) erősségét, azaz a modellben szereplő magyarázóváltozók együttmozgásának mértékét. A szakirodalomban erre vonatkozóan több mutató is ismeretes. Egy mutatóval szemben különböző követelmények támaszthatóak: például normáltság, szintetikusság stb. Az eddig ismert, multikollinearitást számszerűsítő mutatók közül egyidejűleg egyik sem tesz eleget az említett két követelménynek, míg az alábbiakban ismertetett, módszertanilag teljesen új mutató már egyben szintetikus, normált és százalékosan is értelmezhető.

3. A PETRES-féle Red-mutató

Ennek definiálásakor a tényezőváltozók $\mathbf{R}$ korrelációs mátrixának λ_j (j=1,2,...,m) sajátértékeit alkalmazzuk. Ha a magyarázóváltozók forrásául szolgáló adatállomány a $\hat{\boldsymbol{\beta}}$ becslőfüggvény szempontjából redundáns, azaz nagymértékű az adatok együttmozgása, akkor nem mindegyik adat hordoz hasznos tartalmat. Minél kisebb a hasznos tartalmat hordozó adatok aránya, annál nagyobb a redundancia mértéke. Minél nagyobb mértékben szóródnak a

sajátértékek, annál nagyobb mértékű az adatállományban szereplő magyarázóváltozók együttmozgása. Két szélsőséges eset létezik: minden sajátérték egyenlő egymással (azaz értékük egy), illetve egy sajátérték kivételével mindegyik sajátérték nullával egyenlő. A diszperzió mértékét számszerűsíthetjük a sajátértékek relatív szórásával vagy (ebben az esetben az ezzel egyenlő) szórásával.

$$v_\lambda = \frac{\sigma_\lambda}{\bar{\lambda}} = \frac{\sqrt{\frac{\sum_{j=1}^{m}(\lambda_j - \bar{\lambda})^2}{m}}}{\frac{\sum_{j=1}^{m}\lambda_j}{m}} = \frac{\sqrt{\frac{\sum_{j=1}^{m}(\lambda_j - \bar{\lambda})^2}{m}}}{\frac{m}{m}} = \sqrt{\frac{\sum_{j=1}^{m}(\lambda_j - 1)^2}{m}} = \sigma_\lambda$$

Különböző adatállományok redundanciájának összevethetősége végett a fenti mutatót normálni kell. Mivel a sajátértékek nemnegatívak, ezért a relatív szórásra vonatkozó $0 \le v_\lambda \le \sqrt{m-1}$ összefüggés[1] miatt, a normálás $\sqrt{m-1}$ értékével történik.

Az így kapott mutatót a továbbiakban a redundancia mértékének számszerűsítésére fogjuk használni, és segítségével a *Red*-mutatót az alábbiak szerint definiáljuk.

$$Red = \frac{v_\lambda}{\sqrt{m-1}}$$

A redundancia hiánya esetén a fenti mutató értéke nulla, illetve nulla százalék, míg maximális redundancia esetén egy, illetve száz százalék. A *Red*-mutató a vizsgált, adott méretű adatállomány redundanciáját méri. Két vagy több különböző méretű adatállomány redundanciájának összevetésekor a *Red*-mutatók alapján csak annyi állítható, hogy az egyes adatállományok mennyire redundánsak, de arra vonatkozó közvetlen kijelentés nem tehető, hogy ezek közül melyiknek van több hasznosítható adata.

A *Red*-mutató számszerűsíthető a sajátértékek ismerete nélkül is, ha az eredeti adatokat tartalmazó adatállományban a tényezőváltozókat az

---

[1] A relatív szórás két szélső korlátjára (ha $x_i \ge 0$) felírhatjuk a $0 \le v \le \sqrt{N-1}$ összefüggést. Az alsó korlát $v = 0$ minden esetben fennáll, ha $x_i = x$ (i=1,2,...,N). A felső korlát $v = \sqrt{N-1}$ csak akkor áll fenn, ha $x_i = 0$ (i=1,2,...,N-1) és $x_N = N \cdot \bar{x}$.

$$\hat{x}_{ij} = \frac{x_{ij} - \bar{x}_j}{\sqrt{n\sigma_j^2}} \qquad i = 1,2,...,n \qquad j = 1,2,...,m$$

szerint standardizáljuk, ahol σ_j a j-edik magyarázóváltozó tapasztalati szórását jelöli. Ugyanis, ekkor az így standardizált változókra vonatkozóan fennáll az $\hat{\mathbf{X}}'\hat{\mathbf{X}} = \mathbf{R}$ összefüggés. Mivel szimmetrikus mátrixok esetén a mátrix sajátértékeinek négyzetösszege megegyezik a mátrix elemeinek négyzetösszegével, ezért a *Red*-mutató értéke nem más, mint az $\mathbf{R}$ korrelációs mátrix főátlón kívüli elemeinek négyzetes átlaga:

$$Red = \frac{{}^{V}\lambda}{\sqrt{m-1}} = \frac{\sqrt{\frac{\sum_{j=1}^{m}\lambda_j^2}{m} - 1}}{\sqrt{m-1}} = \sqrt{\frac{\sum_{i=1}^{m}\sum_{j=1}^{m} r_{ij}^2 - m}{m\cdot(m-1)}} = \sqrt{\frac{\sum_{i=1}^{m}\sum_{\substack{j=1 \\ j\neq i}}^{m} r_{ij}^2}{m\cdot(m-1)}},$$

azaz

$$Red = \sqrt{\frac{tr(\mathbf{R}^2 - \mathbf{I})}{m\cdot(m-1)}} = \sqrt{\frac{tr\left((\hat{\mathbf{X}}'\hat{\mathbf{X}})(\hat{\mathbf{X}}'\hat{\mathbf{X}}) - \mathbf{I}\right)}{m\cdot(m-1)}}.$$

A mutató további előnye az, hogy segítségével mérni lehet az elemzés alapjául szolgáló adatállományon belül a – regresszió-számítás becslőfüggvényének alkalmazása szempontjából – hasznos tartalmat hordozó adatok arányát is. Ugyanis, egy adott méretű adatállományban a hasznos tartalmat hordozó adatok aránya az azonos méretű, minimális redundanciájú adatállományhoz viszonyítva $100\cdot(1 - Red)$ százalék, míg az adatok átlagos együttmozgásának a maximálishoz viszonyított mértéke $100\cdot Red$ százalék.

4. Összefoglaló

Összefoglalva a következőket állapíthatjuk meg. Nagymennyiségű adatot tartalmazó adatállományok empirikus elemzésekor különösen fontos a redundancia mértékének számszerűsítése, illetve annak ismerete, hogy a sok adat milyen mértékben tartalmaz érdemleges közlést. A problémára a regresszió-számítás segítségével mutattunk rá, ahol ez a magyarázóváltozók együttmozgásaként jelenik meg. Ennek mérésére a szakirodalomban többféle mutató ismert, de ezek többsége vagy nem szintetikus, vagy értelmezése

szubjektív és meglehetősen ellentmondásos. A redundancia általunk bemutatott, új megközelítésű – normált és százalékosan is kifejezhető – mérőszáma biztosítja a $\hat{\beta}$ becslőfüggvény szempontjából újabb információt, érdemleges közlést már nem tartalmazó adatok részarányának olyan számszerűsítését, amely objektíven értelmezhető. Ráadásul, azonos méretű adatállományok redundanciájának mértéke közvetlenül is összehasonlítható.

Irodalom

KOVÁCS P. – PETRES T. – TÓTH L. [2004]. Adatállományok redundanciájának mérése. Statisztikai Szemle 82. évf. 6–7. szám, Központi Statisztikai Hivatal, Budapest, 595–604. p.

KOVÁCS P. – PETRES T. – TÓTH L. [2005]. A new measure of multicollinearity in linear regression models. International Statistical Review Volume 73 Number 3, International Statistical Institute, Voorburg The Netherlands, 405–412. p.

PETRES T. – TÓTH L. [2004]. Piaci információk és a multikollinearitás. SZTE Gazdaságtudományi Kar Közleményei, Szeged. 382–392. p.

KOVÁCS P. – PETRES T. – TÓTH L. [2006]. Válogatott fejezetek Statisztikából. Többváltozós statisztikai módszerek. JATEPress, Szeged.

PÉTER KOVÁCS – TIBOR PETRES

PETRES' RED INDEX

(Summary)

Databases with a lot of data very often mean little information. It is because of the collinearity of variables which consist of the data of the database. This collinearity is in fact a kind of redundancy of the database.

In the study a new indicator is given. With this indicator, which contains the eigenvalues of the variables' correlation matrix, it is possible to quantify the percentage of collinearity: from 0% (all the eigenvalues are equal to 1) to 100% (all the eigenvalues, except the first, are equal to 0).