1. Modellalapú adatfeldolgozás

1.1. Vizsgálati eredmények statisztikai feldolgozásának módjairól

Megfigyelések és mérések eredményeit, az *adatokat* azért dolgozzák fel, hogy azok között *összefüggéseket* ismerjenek fel, feltéve, hogy ilyenek léteznek. Más szavakkal megkísérlik az adatok között esetleg fennálló valamilyan *adatstruktura* felismerését.

Ezek az összefüggések azon a szinten, amelyen vizsgálódunk, végső soron bizonyára *okságiak, kauzálisak*, alkalmasint azonban közvetettek, bonyolultak, kísérletileg végletesen nehezen vizsgálhatók, a vizsgálat idején gyakorlatilag felismerhetetlenek. Lényeges tehát, hogy a jelenségekrõl, amelyeket az adatok leírnak, ismeretesek-e bizonyos elõzetes, *a priori* ismeretek, vagy nem.

Ha nincsenek elõzetes ismeretek, azaz bizonyos oksági összefüggések csak sejthetõk, akkor a *feltáró, felderítõ statisztika* módszereit (leírás, csoportosítás, fõkomponens analízis, faktoranalízis, korreláció) alkalmazzák, amelyek eredményei *önmagukban is nagyon hasznosak,* mivel azonban a feltáró statisztika eredményei között az említett kauzális összefüggések nem nyilvánvalóak, értelmezésükhöz nagy figyelem és adott esetben további kutatás kell. Sokszor ezekre az eredményekre támaszkodik a további. már nem feltáró jellegû statisztikai feldolgozás.

Ha adottak elõzetes ismeretek, akkor kívánatos, hogy azok *matematikai* összefüggések alakjában legyenek jelen. Ezek az összefüggések szerencsés esetben az *oksági* viszonyokat tükrözõ *függvények*, gyakran differenciálegyenletek megoldásai. Egy-egy ilyen matematikai összefüggésrõl azonban nem tételezhetõ fel, hogy a vizsgált jelenséget hibátlanul, kimerítõen leírja, annál is inkább nem, mivel megfogalmazása során gyakorta hasznossági, egyszerûségi szempontokat is figyelembe vettek. Nem meglepő tehát, hogy a jelenséget a megkívánt vagy elérhetõ pontossággal több versengõ összefüggés is leírhatja.

Bizonyos esetekben (pl. “empirikus formulák” megalkotásánál, folyamat irányító algoritmusoknál) eleve lemondanak arról, hogy a jelenségeket, a vizsgált rendszert leíró összefüggés tagjainak fizikai tartalma legyen.

Mindezek miatt a jelenségek leírására felhasznált matematikai képleteket, ebben a szövegkörnyezetben éppúgy *modelleknek* tekintik, mint ahogyan a például a felderítő statisztika egyes osztályozó módszereinél feltételezett eloszlásfüggvényekkel is tették.

(Megjegyzés: matematikai modellen a matematikusok mást értenek).

Az oksági viszonyok kifejezésére törekvõ, fizikai, kémiai, biológiai tartalmú változókkal bíró modelleket az angol irodalom *hard model*-eknek nevezi, szembeállítva azokat a fizikai tartalommal nem bíró *soft model*-ekkel. Az utóbbin jobbára azokat a modelleket értik, amelyeknél a mért változókat számított *látens* változókkal helyettesítették, pl. fõkomponensekkel.

Ami az „adatstrukturát“ illeti: egymással össze nem függő objektumok mért tulajdonság értékei (feltéve, hogy azok relevánsak) egyenletesen, „statisztikusan homogén“ módon töltik ki azt a tulajdonságteret, amelyre a kisérletezés kiterjed. Ha a mérési pontok a térben csomósodnak, clusterekbe tömörülnek, bizonyos nemleges összefüggés már létezik, nevezetesen az, hogy objektumok különböznek egymástól, pontosabban tulajdonságaik várható értéke más és más. *Hipergömb* alakú pontfelhők arról árulkodnak, hogy a felhőbe tartozó objektumok tulajdonságai egymás között csak véletlenül különböznek, *hiperellipszoid* alakú (lencseszerü, szivaralakú) felhő eltérő szórású tulajdonságokra, ha pedig az ellipszoid a tulajdonságtér koordináta tengelyeihez képest el is fordul, az objektumok jellemzői valamiképpen korrelálnak. Természetesen ezek a „hiperalakok“ a síkban és a térben vetületeik alakjában szemlélhetők. Ha a vizsgált objektumok végül vonalakba*, síkokba*, *felületekbe* rendeződnek, felderíthető függvénykapcsolat sejthető.

1.2. A modellek jellemzői

A matematikai modell *függõ* és *független* változók között állapít meg összefüggést. A továbbiakban egy független változót általában *x*-szel, vektorukat **x**-szel, mátrixukat **X**-szel jelöjük, a függõ változókat *y*-nal, **y**-nal vagy **Y**-nal. A modell legáltalánosabban legyen

$$\mathbf{Y} = f(\mathbf{X}) \qquad (1.1)$$

A matematikai modell *x* független változóit adott esetekben szokás *prediktoroknak* (predictors) vagy *magyarázó* (explanatory) változóknak, a függõ változókat *válasznak (response)* is nevezni.

Megjegyzés: *x* azonosító nem szükségszerűen csak magukat a független változókat jelenti, jelölheti azok valamely függvényét (logaritmusát, hatványát, reciprokát is). A két lehetőséget azonban az egyszerűség kedvéért most a képletekben nem különböztetjük meg.

A matematikai modell alakja függvényseregeket határoz meg, ahol a függvények paramétereikben különböznek. Egy-egy függvényt tehát paraméterei is jellemeznek:

$$\mathbf{Y} = f(\mathbf{X},\mathbf{A}) \qquad (1.2)$$

ahol *a*, **a**, ill **A** jelöli a függvény skaláris-, vektor- vagy mátrix paramétereit.

A megismerés elsõ lépése az adott, *ismert* vagy beállított *x változók* mellett *mért* vagy megfigyelt *y változók* alapján a jelenséget legjobban leíró partikuláris modell meghatározása. Ez a tevékenység a modellfüggvény *paramétereinek* megállapítását jelenti. Mivel az *y* változók kísérleti eredmények, a paramétereket csak pontatlanul lehet meghatározni. Ezért nevezik a megismerésnek ezt a lépését paraméter*becslésnek*. Bár a *regresszió* eljárása történetileg a korreláció számítás kapcsán, tehát lineáris esetek vizsgálatánál alakult ki, a regresszió elnevezést többé kevésbé a paraméterbecslés szinonimájaként használják. Hasonlóan szinonímája a paraméterbecslésnek a *modellillesztés* vagy *görbeillesztés* (model fitting, curve fitting).

A becsült (más szóhasználattal *számított*) paraméterek jele legyen $\hat{a}$, $\hat{\mathbf{a}}$, vagy $\hat{\mathbf{A}}$, aszerint hogy skaláris- vektor- vagy mátrixváltozókról van szó. Sok esetben csak a becsült paraméterek meghatározása a cél, természetesen *varianciájukkal* és *kovarianciájukkal* együtt. Utóbbi mennyiségeket var($\hat{a}$) ill. cov($\hat{a}$) jellel azonosítják. Vektorba rendezett számított paraméterek varianciáit és kovarianciáit $\hat{\mathbf{C}}$ mátrix tartalmazza.

A becsült paramétereket (2) modellbe helyettesítve megkaphatók a *becsült, illesztett* vagy *számított y* értékek, amelyeket $\hat{y}$ -nal jelölünk. Ezek varianciáját is kiszámítják.

Adott független változók *i*-edik értéke mellett mért és az illesztett paraméterekkel ugyanott számított *y* értékek

$$d_i = y_i - \hat{y}_i \tag{1.3}$$

különbségét *reziduális különbségnek* vagy röviden *reziduálisnak* szokás nevezni.

A számított $\hat{y}$ értékek és az *d* reziduálisok kiszámítása már azért is fontos, mert ezekkel ítélhetõ meg mennyire *alkalmas* (*adekvát*) egy modell.

A reziduális különbségek ugyanis két részbõl állhatnak: *rendszeres eltérésbõl* (*bias*) és mérésekre visszavezethetõ *véletlen hibából.* Más szavakkal az *i*-edik mért érték a modell alapján várt értéken kivül b_i rendszeres eltérést és a valószínûségi változó jellegû rendszeres hiba e_i realizációját is tartalmazhatja:

$$y_i = f(\mathbf{X}_i, \mathbf{A}) + b_i + e_i \tag{1.4}$$

azaz az *i*-edik reziduális:

$$d_i = y_i - \hat{y}_i = (f(\mathbf{X}_i, \mathbf{A}) + b_i + e_i) - \hat{y}_i \tag{1.5}$$

Ha a reziduálisok becsült *várható értéke* (átlaga) nem tér el szignifikánsan 0-tól, akkor ez azt jelenti, hogy a rendszeres hiba és a véletlen hibák várható értéke 0, tehát

$$f(\mathbf{X}, \mathbf{A}) = \hat{y} \tag{1.6}$$

tehát *torzítatlan, alkalmas, adekvát*, rendszeres hiba (bias) nincs. Ha nem ez a helyzet, vagy kísérleti hibát kell keresni, vagy a vizsgált jelenséget nem a választott modell írja le.

Az illesztéssel kapott $\hat{a}$ paraméterek és az azokkal számolt $\hat{y}$ értékek egy adott kísérlet eredményei. Nem helyes emiatt az eredményekbõl levont következtetések kiterjesztése a jövõ esetekre. Az eredmények jellemzik az *illesztés jóságát,* de nem jellemzik a modell *jóslási* (predikciós) képességét. A paraméterek akkor lesznek predikcióra alkalmasak, ha alkalmas módon kiszámítják azok *jóslási varianciáját*, a jövõben várható *y* értékek pedig akkor ismerhetõk meg, ha az $\hat{y}$ értékek mellé meg van adva azok *jóslási* szórása. Ezeknek a mennyiségeknek számítási módszereit *keresztellenõrzésnek* (*cross*

validation) is nevezik. A keresztellenõrzés –amelyrõl késõbb lesz szó– lineáris modelleknél viszonylag egyszerû, ellenkezõ esetben általában kísérleti feladat.