Puuttuvan tiedon ongelma pitkittäisaineistoissa

Tapio Nummi

tan@uta.fi

Matematiikan, tilastotieteen ja filosofian laitos

Tampereen yliopisto

mtl.uta.fi/tilasto/sekamallit/puupitkit.pdf

Johdanto

- Pitkittäisaineistoissa on varsin tavallista, että kaikille vastemuuttujille ei saada mitattuja arvoja.
- Puuttuva tieto voi syntyä monella tavalla. Esimerkiksi,

  mittaus voi puuttua jonakin ajankohtana ja jonakin myöhempänä ajankohtana mittaus saadaan tai

  mittauksia saadaan johonkin ajanhetkeen asti, jonka jalkeen mittauksia ei enää saada (ns. dropout).

Puuttuva tieto tekee analyysin vaikeaksi monella tavalla:

- Mittauksia ei saada kaikille yksilöille samoissa aikapisteissä **(imbalance)** $\rightarrow$ monia tilastollisia menetelmiä ei voida suoraan käyttää.
- **Informaatiota menetetään** $\rightarrow$ estimoinnin tarkkuus heikkenee.
- **Voi aiheuttaa tuloksiin harhaa** $\rightarrow$ voi johtaa vääriin johtopäätöksiin.

Puuttuvan tiedon generoiva mekanismi on siten aina huolellisesti tutkittava.

Puuttuvan tiedon tyypit

Otetaan käyttöön seuraavat merkinnät:

$Y = (Y_1, \ldots, Y_n)'$ mahdollisten arvojen vektori ja

$R = (R_1, \ldots, R_n)'$ indikaattori-vektori.

Nyt siis

$R_j = 1,$ jos Y_j on havaittu ja

$R_j = 0,$ jos Y_j puuttuu

Lisäksi merkitään

Y^O havaittu osa (observed)

Y^M puuttuva osa (missing)

Puuttuva tieto voidaan jaotella kolmeen päätyyppiin R:n ja Y:n keskinäisen suhteen perusteella

- Täysin satunnainen (**MCAR**, missing completele at random)
- Satunnainen (**MAR**, missing at random)
- Ei-satunnainen (**NMAR**, not missing at random)

Käytettäessä tilastollisia menetelmiä on huomioitava minkä tyyppisestä puuttuvasta tiedosta kulloinkin on kysymys.

Täysin satunnainen (MCAR)

Puuttuva tieto on täysin satunnaista, jos R on riippumaton sekä Y^O:sta että Y^M:stä.

Esimerkiksi jos $Y = (Y_1, Y_2)'$, missä Y_1 on täysin havaittu ja Y_2 voi sisältää puuttuvia. Nyt jos Y_2 on

MCAR, niin

$$P(R_2 = 1 \mid Y_1, Y_2, X) = P(R_2 = 1 \mid X),$$

eli tn, että Y_2 puuttuu ei riipu muuttujista Y_1 eikä Y_2.

Huomaa riippuvuus kovariaateista X.

Oletus on, että mukana ovat kaikki muuttujien R ja Y ennustamisen kannalta relevantit kovariaatit. Jos jokin tärkeä kovariaatti puuttuu, ei MCAR pidä paikkansa.

Jos puuttuva on MCAR, niin saatu aineisto voidaan olettaa otokseksi "täydellisestä" aineistosta.

Analyysin tekeminen vain niille havainnoille, joilta on saatu kaikki mittaukset antaa periaatteessa oikean tuloksen, mutta pienemmällä otoskoolla.

Satunnainen (MAR)

Puuttuminen on satunnaista, jos voidaan olettaa, että puuttuminen riippuu havaituista arvoista (mutta ei riipu arvoista joita oltaisiin voitu havaita).

Saadaan siis

$$P(R \mid Y^O, Y^M, X) = P(R \mid Y^O, X)$$

Huom. annetuilla Y^O:n arvoilla puuttuminen on satunnaista eikä riipu arvoista Y^M.

Esimerkkinä voidaan mainita tilanne, jossa tutkimusprotokolla edellyttää, että koe keskeytetään, jos vasteen arvot ylittävät jonkin ennalta asetetun rajan.

Nyt puuttuminen on kontrolloitua ja riippuu ainoastaan havaituista arvoista Y^O.

Havaintoja ei nyt voida pitää satunnaisotoksena kohdepopulaatiosta.

Eräs tärkeä seuraus on, että analyysin tekeminen vain "täydellisille" havainnnoille saattaa johtaa harhaisiin tuloksiin.

"Täydellisestä" aineistosta lasketut estimaatit (keskiarvo, varianssi jne.) ovat nyt vastaavien perusjoukon parametrien harhaisia estimaatteja.

Havaitun datan suhteen lasketut ehdolliset jakaumat ovat kuitenkin samat kuin kohdepopulaatiossa.

Jos malli on oikein spesifiotu, niin havaittuja arvoja käyttäen puuttuvat arvot voidaan kuitenkin periaatteesa ennustaa.

Jos esimerkiksi oletetaan normaalijakauma, niin

$$E(Y^M \mid Y^O) = \mu^M + \Sigma^{MO}\Sigma^{O-1}(Y^O - \mu^O).$$

Huomaa riippuvuus odotusarvon μ ja kovarianssimatriisin Σ malleista.

Jos puuttuvat ovat MAR, niin arvot voidaan ennustaa havaittujen arvojen ja Y:n yhteisjakauman avulla. Puuttuvien generoivaa mekanismia

$$P(R \mid Y^O, X)$$

ei erikseen tarvitse mallintaa (ignorable mechanism). Analyysit voidaan perustaa yhteisjaumasta $f(Y \mid X)$ johdettuun uskottavuusfunktioon.

Huom. edellä sanottu pätee myös kun aineisto on MCAR, koska MCAR on MAR:n erikoistapaus.

Perusoletus pitkittäisaineistossa on yleensä MAR.

Ei-satunnainen (NMAR)

Puuttuminen on ei-satunnaista (NMAR), jos puuttumisen todennäköisyys riippuu myös arvoista, jotka "oltaisiin voitu" havaita. Nyt siis

$$P(R \mid Y^O, Y^M, X)$$

riippuu ainakin jostakin puuttuvasta arvosta Y^M.

Esimerkkinä lasten lihavuustutkimus, jossa lihavien lasten vanhemmat voisivat olla myönteisempiä tai kielteisempiä kuin muut antamaan suostumuksensa tutkimukseen osallistumiselle. Näin siis lapsen paino ja pituus voisivat olla yhteydessä puuttuvan tiedon syntymiseen (painoindeksissä).

Joskus ei-satunnaisesta (NMAR) puuttumisesta käytetään termiä informatiivinen (**informative**), koska puuttuvan tiedon jakaumaa voidaan yrittää päätellä indikaattorivektorin R jakaumasta.

Huom. Nyt myös puuttuvan tiedon malli $P(R)$ tulee nyt sisällyttää analyysiin (**nonignorable missingness**) ja sillä on ratkaiseva vaikutus lopputuloksiin.

Vaikutus analyyseihin

Jos aineisto on MCAR, niin havainnot voidaan olettaa satunnaisotokseksi perusjoukosta.

Tällöin periaatteessa melkein mitä tahansa tilanteeseen sopivaa tilastollista menetelmää voidaan käyttää (myös niitä, jotka edellyttävät täydellisen datan, ns. complete-case-analysis).

Jos aineisto on MAR, havaintoja Y ei enää voida pitää satunnaisotoksena alkuperäisestä populaatiosta.

Täydellisiin havaintohin perustuva analyysi antaa nyt harhaisia tuloksia.

Uskottavuusfunktioon pohjautuvia menetelmiä, joissa havaintojen yhteisjakauma on oikein spesifioitu, voidaan sitävastoin käyttää. Pitkittäisaineistossa riippuvuusrakenteen spesifiointiin (kovarianssirakenteeseen) tulisi kiinnittää erityistä huomiota.

Jos aineisto on NMAR, niin tilastollisia menetelmiä ei yleensä voida suoraan soveltaa.

Sekä täydellisten havaintojen analysointi, että uskottavuusfunktio-pohjaiset menetelmät antavat yleensä harhaisia tuloksia.

Analyysissä tulisi tällöin mallintaa sekä havainnot, että puuttuvia arvoja generoiva mekanismi.

Jos aineisto on NMAR, on sitä pelkän havaitun aineiston perusteella (ilman lisäinformaatiota) kuitenkin vaikea verifioida.

Käytännön mahdollisuudeksi jää tällöin tarkastella tulosten herkkyyttä erilaisille oletuksille puuttuvan tiedon mekanismeista.

Kirjallisuutta:

Fitzmaurice, Laird ja Ware (2004). Applied Longitudinal Analysis, Wiley.

Little ja Rubin (1987). Statistical Analysis with Missing Data, Wiley.