Puuttuvan tiedon käsittelystä pitkittäisaineistoissa

Tapio Nummi

tan@uta.fi

Matematiikan, tilastotieteen ja filosofian laitos

Tampereen yliopisto

Johdanto

- Pitkittäisaineistoissa on varsin yleistä, että kaikille vastemuuttujille ei saada mitattuja arvoja.
- Puuttuva tieto voi syntyä monella tapaa, 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:

$Y_i = (Y_{i1}, \ldots, Y_{in})'$ täydellinen data

$R_i = (R_{i1}, \ldots, R_{in})'$ indikaattori-vektori.

Nyt siis

$R_{ij} = 1,$ jos Y_{ij} on havaittu ja

$R_{ij} = 0,$ jos Y_{ij} puuttuu

Lisäksi merkitään

Y_i^O havaittu osa (observed)

Y_i^M puuttuva osa (missing)

Puuttuva tieto voidaan jaotella kolmeen päätyyppiin R_{ij}:n ja Y_i keskinäisen suhteen perusteella

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

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

Täysin satunnainen (MCAR)

Puuttuva tieto on täysin satunnaista, jos R_i on riippumaton sekä Y_i^O:sta että Y_i^M:sta.

Esimerkiksi jos $Y_i = (Y_{i1}, Y_{i2})'$ (2-ulotteinen tilanne) Y_{i1} on taysin havaittu ja Y_{i2} voi sisältää puuttuvia.

Nyt jos Y_{i2} on MCAR, niin

$$P(R_{i2} = 1 \mid Y_{i1}, Y_{i2}, X_i) = P(R_{i2} = 1 \mid X_i),$$

eli tn, että Y_{i2} puuttuu ei riipu muuttujista Y_{i1} tai Y_{i2} (arvoista jotka "pitäisi" havaita).

Huom. em määritelmässä riippuvuus kovariaateista X_i kuitenkin hyväksytään.

Itse asiassa oletus on, että mukana ovat kaikki muuttujien R_i ja Y_i ennustamisen kannalta relevantit kovariaatit. Jos jokin tärkeä kovariaatti puuttuu, niin MCAR ei pidä paikkansa.

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

Voidaan ajatella, että 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 (periaatteessa) oltaisiin voitu havaita.

Saadaan siis

$$P(R_i \mid Y_i^O, Y_i^M, X_i) = P(R_i \mid Y_i^O, X_i)$$

2- ulotteisessa tapauksessa saamme

$$P(R_{i2} = 1 \mid Y_{i1}, Y_{i2}, X_i) = P(R_{i2} = 1 \mid Y_{i1}, X_i)$$

Annetuilla Y_{i1}:n arvoilla puuttuminen on siten satunnaista eikä riipu arvoista Y_{i2}

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

Nyt siis puuttuminen on kontrolloitua ja riippuu ainoastaan Y_i:n havaituista arvoista.

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.

Yksi mielenkiitoinen ominaisuus on kuitenkin se, että havaitun datan suhteen lasketut ehdolliset jakaumat ovat samat kuin kohdepopulaatiossa.

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

Jos esimerkiksi oletetaan normaalijakauma, niin

$$E(Y_i^M \mid Y_i^O) = \mu_i^M + \Sigma_i^{MO}\Sigma_i^{O-1}(Y_i^O - \mu_i^O).$$

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

$$P(R_i \mid Y_i^O, X_i)$$

ei erikseen tarvitse mallintaa. Analyysit voidaan perustaa yhteisjaumasta $f(Y_i \mid X_i)$ 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 arvoista, jotka olisi pitänyt havaita. Nyt siis

$$P(R_i \mid Y_i^O, Y_i^M, X_i)$$

riippuu ainakin jostakin puuttuvasta arvosta Y_i^M.

2-ulotteisessa tapauksessa saadaan

$$P(R_{i2} = 1 \mid Y_{i1}, Y_{i2}, X_i),$$

mikä siis riippuu Y_{i2}:n potentiaalisesta arvosta.

Esimerkkinä NMAR aineistosta mainittakoon lasten lihavuustutkimus, jossa lihavien lasten vanhemmat voisivat olla myönteisempiä tai kielteisempiä kuin muiden lasten vanhemmat antamaan suostumuksensa tutkimukseen osallistumiselle. Näin siis lapsen paino ja pituus voisivat olla yhteydessä puuttuvan tiedon (lihavuusindeksi) syntymiseen aineistossa.

Huom. Puuttuvan tiedon jakauma riippuu nyt arvoista Y_i^O sekä todennäköisyydestä $P(R_i \mid Y_i, X_i)$. Myös puuttuvan tiedon malli $P(R_i)$ tulee sisällyttää analyysiin.

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).

Vaikutus analyyseihin

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

Tädellisiin 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.