Helsingin yliopisto
Imputointimenetelmät

Loppukoe Seppo Laaksonen
11.6.2009

Merkitse jokaiseen vastauspaperiin nimesi ja opiskelijanumerosi. Vastaa lyhyesti kahdeksaan (8) kysymykseen. Tulokset kerron sähköpostitse. Tarpeen mukaan selostan tehtäviä erillispalavereissa.

1. Alempi väritetty teksti on luennoista. Sieltä puuttuu sanoja joiden tilalle on asetettu **ISO KIRJAIN**. Imputoi ne. Anna vastaus tyyliin: **A**='imputointi.'

Imputointi tarkoittaa puuttuvan tai puuttuvaksi määritellyn **A** tiedon paikkaamista sellaisella korvikearvolla joka **B** estimaatin laatua verrattuna siihen mikä saataisiin **C** imputointia eli jättämällä tuo tieto käsittelystä pois. Käytännössä **D** imputointi tähtää yleensä useamman paremman estimaatin tuottamiseen samanaikaisesti. Käyttäjän on hyvä kuitenkin heti alkuun miettiä asiaa edes yhden estimaatin kannalta. Imputoinnin siis toivotaan vähentävän estimaatin **E** ja siten että **F** on hyvä ja tuotettu mahdollisimman oikealla menetelmällä (siis tarkkuus voidaan laskea imputoidusta datasta erityisen helposti harhauttavasti). **F** tässä kuten muulloinkin mitataan **G** ja/tai **H**, mutta on huomattava että sen laskeminen ei ole yhtä helppoa kuin tilanteessa jossa puuttuvuutta ei ole.

2. Toteuta imputointi vastaavalla tavalla kuin tehtävässä 1 tekstille:

Usein imputoinnissa on tavoitteena menestyä hyvin myös muuttujien välisten yhteyksien kuvaamisessa. Tämä onnistuu jos kuhunkin muuttujaan sovellettu imputointi onnistuu hyvin. Monet menetelmät (ja huono aineisto) eivät kuitenkaan takaa tätä. Seuraavanlaisia ratkaisuja on sovellettu:

*(i) Ei **I** ollenkaan vaan puuttuvat tiedot jätetään analyysistä pois. Tässä on ongelmana havaintomäärän **J** ja tästä johtuva tarkkuuden heikkeneminen sekä muussa kuin MCAR-tilanteessa myös tulosten vääristyminen.*
(ii) Käytetään analyysimenetelmää, jossa **K** on mukana (ehkä selittäjänä, sen teho ei aina vahva).
(iii) Puuttuvuudesta johtuva harha oikaistaan **L** (mainittu edelläkin).
(iv) Sovelletaan vastaajaluovuttaja -menetelmää siten, että samalta luovuttajalta otetaan tiedot **M** tai joukolle kiinnostuksen kohteena olevia muuttujia.
(v) Sovelletaan ns. **N** imputointia (mainittu jo edellä) jossa ensin imputoidaan **P** muuttuja, sitten **Q** läheinen käyttäen **R** imputoitua arvoa apumuuttujana, ja niin edespäin.

3. Alempana on SAS-ohjelma. Selosta mitä siinä tehdään, tietystikin kurssilla esitettyjen asioiden näkökulmasta. On olemassa muitakin imputointivaihtoehtoja saman imputointimallin tilanteeseen. Mainitse yksi?

```
proc genmod data=alku descending; class c1;
model y= x1 x2 c1 /link=probit type1;
output out=jatko p=pred; /*weight w;*/
data jatko2; set jatko;
if pred ne.;
u=ranuni(776);
if y ne . then yimp=y;
else if pred>u then yimp=1; else yimp=0;
run;
```

4. Sinulla on mikrodata jossa on puuttuvaa tietoa. Mitä teet ennen kuin mahdollisesti aloitat imputoinnit? Mistä syystä lähdet imputoimaan?

5. Seuraavassa on kaksi tulostetta joissa metodeilla i, ir, i2, ir2, iv1 ja iv2 on imputoitu muuttujaa liikevaihto_1. Tulokset koskevat vain imputoituja tilastoyksikköjä. Oikeat arvot on tässä tiedetty. Ensimmäisessä tulosteessa kohdassa obs=3 on oikea keskiarvo ja alla vastaavat eri imputointitulokset, sitten kohdassa 10 oikea suhteellinen keskihajonta ja kohdassa 17 oikea maksimiarvo. Jälkimmäistä tulostetta varten on muodostettu uusi muuttuja kullekin metodille siten että muuttujan arvo on absoluuttinen yksilötason poikkeama oikeasta arvosta. Sitten on laskettu tästä mainitut tunnusluvut.

Tehtävänäsi on eri näkökohtiin huomiota kiinnittäen valita paras metodi. Tee se sulkemalla ensin pois huonoin tai huonoimmat ja sitten valitse lopulta joku parhaaksi vaikkei ratkaisusi olisikaan kiistaton. Eli siis kerro valintasi syyt ajatellen asiakastasi jolle imputoinnin olet tehnyt.

Imputointivertailu estimaattitasolla 2

Obs	_NAME_	COL1
2	_FREQ_	278.00
3	Mean_liikevaihto_1	10700.50
4	mean_i_liikevaihto_1	26679.98
5	mean_ir_liikevaihto_1	25326.45
6	mean_i2_liikevaihto_1	10121.79
7	mean_ir2_liikevaihto_1	11271.34
8	mean_iv1_liikevaihto_1	10905.55
9	mean_iv2_liikevaihto_1	10830.59
10	cv_liikevaihto_1	220.08
11	cv_i_liikevaihto_1	0.00
12	cv_ir_liikevaihto_1	172.50
13	cv_i2_liikevaihto_1	215.50
14	cv_ir2_liikevaihto_1	233.73
15	cv_iv1_liikevaihto_1	225.16
16	cv_iv2_liikevaihto_1	233.48
17	max_liikevaihto_1	212325.00
18	max_i_liikevaihto_1	26679.98
19	max_ir_liikevaihto_1	87726.20
20	max_i2_liikevaihto_1	187965.91
21	max_ir2_liikevaihto_1	197022.02
22	max_iv1_liikevaihto_1	211740.00
23	max_iv2_liikevaihto_1	239055.00

Imputointivertailu yksikkötasolla 7

The MEANS Procedure

Variable	Mean	Minimum	10th Pctl	90th Pctl	95th Pctl	Maximum
metodi_i	24647.25	170.0174216	12729.98	26604.98	35150.02	185645.02
metodi_ir	43056.87	81.2411725	8737.53	86751.20	87486.20	185998.07
metodi_i2	1208.15	0.6031560	27.9848232	2932.90	5769.14	24524.32
metodi_ir2	2793.85	0.0915178	20.4572202	5748.28	11849.66	98742.02
metodi_iv1	1330.04	0	30.0000000	3915.00	6690.00	24915.00
metodi_iv2	1575.38	0	30.0000000	3705.00	9525.00	26730.00

6. Selosta moni-imputoinnin perusidea ml. estimointi (sekä piste-estimaatti että väliestimaatti). Väliestimaatin eräs muoto eli varianssiestimaatti on luennoissa esitetty kahdessa muodossa, Rubinin ja Björnstadin. Sano jotain näiden kahden kaavan erosta.

7. Miten voit imputoida jos sinulla ei ole varsinaista aputietoa (tietysti datassa on merkitty jokainen tilastoyksikkö olemassa ja merkitty jollakin koodilla)?
APUTIETO: kaksi selkeästi erilaista vaihtoehtoa olen esittänyt luentomateriaalissa.

8. Milloin voit saada suhdeasteikon muuttujan imputoinnissa negatiivisia arvoja? Mitä teet korjataksesi tällaisen epäkelvon tilanteen?

9. Imputointimalli on tärkeä osa imputoinnissa. Minkälaisia malleja voi olla ja miten malli estimoidaan jos estimoidaan? Kerro myös miten malli saadaan jos sitä ei estimoida. Esimerkki voi olla ihan riittävä osoitus asian ymmärtämisestä.

10. Imputointitoiminto voidaan imputointimallin ollessa konkreettisena olemassa (joko estimoituna tai ei) toteuttaa joko vastaajaluovuttaja- tai malliluovuttajamenetelmällä. Kerro kummankin vaihtoehdon hyvistä vs huonoista puolista.

11. Vastaa seuraaviin väitteisiin joko kyllä (=1) tai ei (=0); voit laittaa perustelunkin:
(a) Imputoitua muuttujaa ei saa laittaa selittäjäksi imputointimallissa.
(b) Vastaajaluovuttajamenetelmä on eräänlainen painotusmenetelmä.
© Vastaajaluovuttajamenetelmä ei voi olla stokastinen.
(d) Jos puuttuvat tiedot korvataan imputointisoluissa havaittujen arvojen keskiarvoilla, on kyseessä deterministinen malliluovuttajamenetelmä.

12. Surveyn mikrodatassa voi olla monenlaisia puuttuvuuksia. Olen kurssimateriaalin alkupuolella esittänyt näistä yhteenvedon joka sisältää viisi eri tilannetta. Esitä näistä ainakin neljä ja pohdiskele imputointia samalla ja siis myös imputoinnin motivointia.

Helsingin yliopisto
Imputointimenetelmät

Loppukoe Seppo Laaksonen
19.5.2009

Merkitse jokaiseen vastauspaperiin nimesi ja opiskelijanumerosi. Vastaa lyhyesti kahdeksaan (8) kysymykseen. Tulokset kerron sähköpostitse. Tarpeen mukaan selostan tehtäviä erillispalavereissa.

1. Alempi väritetty teksti on luennoista. Sieltä puuttuu sanoja joiden tilalle on asetettu **ISO KIRJAIN**. Imputoi ne. Anna vastaus tyyliin: **A**='imputointi.'

Imputointi on prosessi jonka tässä katson koostuvan seuraavista 6 osatehtävästä:

(i) Datan **A** jolloin päätetään myös mitä imputoidaan

(ii) **B** hankinta ja huolto

(iii) **C** suunnittelu ja rakentaminen ml. imputointisolujen mahdollinen luonti **D** varten.

(iv) Imputointitehtävä tai imputointitoiminto

(v) Estimointi sisältäen piste-estimoinnin, **E** ja **F** sekä näiden pohjalta kokonaisvarianssin (ja keskivirheen)

2. Toteuta imputointi vastaavalla tavalla kuin tehtävässä 1 tekstille:

Usein imputoinnissa on tavoitteena menestyä hyvin myös muuttujien välisten yhteyksien kuvaamisessa. Tämä onnistuu jos kuhunkin muuttujaan sovellettu imputointi onnistuu hyvin. Monet menetelmät (ja huono aineisto) eivät kuitenkaan takaa tätä. Seuraavanlaisia ratkaisuja on sovellettu:

*(i) Ei **H** ollenkaan vaan puuttuvat tiedot jätetään analyysistä pois. Tässä on ongelmana havaintomäärän **I** ja tästä johtuva tarkkuuden heikkeneminen sekä muussa kuin MCAR-tilanteessa myös tulosten vääristyminen.*
(ii) Käytetään analyysimenetelmää, jossa **J** on mukana (ehkä selittäjänä, sen teho ei aina vahva).
(iii) Puuttuvuudesta johtuva harha oikaistaan **K** (mainittu edelläkin).
(iv) Sovelletaan vastaajaluovuttaja -menetelmää siten, että samalta luovuttajalta otetaan tiedot **L** tai joukolle kiinnostuksen kohteena olevia muuttujia.

(v) Sovelletaan ns. **M** imputointia (mainittu jo edellä) jossa ensin imputoidaan **N** muuttuja, sitten **P** läheinen käyttäen **Q** imputoitua arvoa apumuuttujana, ja niin edespäin.

3. Alempana on SAS-ohjelma. Selosta mitä siinä tehdään, tietystikin kurssilla esitettyjen asioiden näkökulmasta. Olisiko olemassa muukin imputointivaihtoehto samaan tilanteeseen?

```
proc genmod data=alku descending; class c1;
model y= x1 x2 c1 /link=probit type1;
output out=jatko p=pred; /*weight w;*/
data jatko2; set jatko;
if pred ne.;
u=ranuni(776);
if y ne . then yimp=y;
else if pred>u then yimp=1; else yimp=0;
run;
```

4. Sinulla on mikrodata jossa on puuttuvaa tietoa. Mitä teet ennen kuin mahdollisesti aloitat imputoinnit? Mistä syystä lähdet imputoimaan?

5. Seuraavassa on kaksi tulostetta joissa metodeilla i, ir, i2, ir2, iv1 ja iv2 on imputoitu muuttujaa liikevaihto_1. Tulokset koskevat vain imputoituja tilastoyksikköjä. Oikeat arvot on tässä tiedetty. Ensimmäisessä tulosteessa kohdassa obs=3 on oikea keskiarvo ja alla vastaavat eri imputointitulokset, sitten kohdassa 10 oikea suhteellinen keskihajonta ja kohdassa 17 oikea maksimiarvo. Jälkimmäistä tulostetta varten on muodostettu uusi muuttuja kullekin metodille siten että muuttujan arvo on absoluuttinen yksilötason poikkeama oikeasta arvosta. Sitten on laskettu tästä mainitut tunnusluvut.

Tehtävänäsi on eri näkökohtiin huomiota kiinnittäen valita paras metodi. Tee se sulkemalla ensin pois huonoin tai huonoimmat ja sitten valitse lopulta joku parhaaksi vaikkei ratkaisusi olisikaan kiistaton. Eli siis kerro valintasi syyt ajatellen asiakastasi jolle imputoinnin olet tehnyt.

Imputointivertailu estimaattitasolla 2

Obs	_NAME_	COL1
2	_FREQ_	278.00
3	Mean_liikevaihto_1	10700.50
4	mean_i_liikevaihto_1	26679.98
5	mean_ir_liikevaihto_1	25326.45
6	mean_i2_liikevaihto_1	10121.79
7	mean_ir2_liikevaihto_1	11271.34
8	mean_iv1_liikevaihto_1	10905.55
9	mean_iv2_liikevaihto_1	10830.59
10	cv_liikevaihto_1	220.08
11	cv_i_liikevaihto_1	0.00
12	cv_ir_liikevaihto_1	172.50
13	cv_i2_liikevaihto_1	215.50
14	cv_ir2_liikevaihto_1	233.73
15	cv_iv1_liikevaihto_1	225.16
16	cv_iv2_liikevaihto_1	233.48
17	max_liikevaihto_1	212325.00
18	max_i_liikevaihto_1	26679.98
19	max_ir_liikevaihto_1	87726.20
20	max_i2_liikevaihto_1	187965.91
21	max_ir2_liikevaihto_1	197022.02
22	max_iv1_liikevaihto_1	211740.00
23	max_iv2_liikevaihto_1	239055.00

Imputointivertailu yksikkötasolla 7

The MEANS Procedure

Variable	Mean	Minimum	10th Pctl	90th Pctl	95th Pctl	Maximum
metodi_i	24647.25	170.0174216	12729.98	26604.98	35150.02	185645.02
metodi_ir	43056.87	81.2411725	8737.53	86751.20	87486.20	185998.07
metodi_i2	1208.15	0.6031560	27.9848232	2932.90	5769.14	24524.32
metodi_ir2	2793.85	0.0915178	20.4572202	5748.28	11849.66	98742.02
metodi_iv1	1330.04	0	30.0000000	3915.00	6690.00	24915.00
metodi_iv2	1575.38	0	30.0000000	3705.00	9525.00	26730.00

6. Selosta moni-imputoinnin perusidea ml. estimointi (sekä piste-estimaatti että väliestimaatti). Väliestimaatin eräs muoto eli varianssiestimaatti on luennoissa esitetty kahdessa muodossa, Rubinin ja Björnstadin. Sano jotain näiden kahden kaavan erosta.

7. Miten voit imputoida jos sinulla ei ole varsinaista aputietoa (tietysti datassa on merkitty jokainen tilastoyksikkö olemassa ja merkitty jollakin koodilla)? APUTIETO: kaksi selkeästi vaihtoehtoa olen esittänyt luentomateriaalissa.

8. Milloin voit saada suhdeasteikon muuttujan imputoinnissa negatiivisia arvoja? Mitä teet korjataksesi tällaisen epäkelvon tilanteen?

9. Imputointimalli on tärkeä osa imputoinnissa. Minkälaisia malleja voi olla ja miten malli estimoidaan jos estimoidaan? Kerro myös miten malli saadaan jos sitä ei estimoida. Esimerkki voi olla ihan riittävä osoitus asian ymmärtämisestä.

10. Imputointitoiminto voidaan imputointimallin ollessa konkreettisena olemassa (joko estimoituna tai ei) toteuttaa joko vastaajaluovuttaja- tai malliluovuttajamenetelmällä. Kerro kummankin vaihtoehdon hyvistä vs huonoista puolista.

11. Vastaa seuraaviin väitteisiin joko kyllä (=1) tai ei (=0); voit laittaa perustelunkin:
(a) Imputoitua muuttujaa ei saa laittaa selittäjäksi imputointimallissa.
(b) Vastaajaluovuttajamenetelmä on eräänlainen painotusmenetelmä.
© Vastaajaluovuttajamenetelmä ei voi olla stokastinen.
(d) Jos puuttuvat tiedot korvataan imputointisoluissa havaittujen arvojen keskiarvoilla, on kyseessä deterministinen malliluovuttajamenetelmä.

12. Binäärinen regressiomalli (logit ja probit) on käytetty moneen tarkoitukseen, myös imputointiin. Imputoinneissa sitä voi käyttää sekä (i) vastaajaluovuttaja- että malliluovuttajapuolella. Malliluovuttajapuolella on kurssilla ollut esillä sekä (ii) deterministinen että stokastinen vaihtoehto. Anna konkreettinen esimerkki joko kohdasta (i) tai kohdasta (ii), eli siis miten kumpikin vaihtoehto toteutetaan.