SOS3003

Anvendt statistisk dataanalyse i samfunnsvitenskap

Forelesingsnotat 08

Erling Berge
Institutt for sosiologi og statsvitenskap
NTNU

Forelesing VIII

- Manglande data

  Allison, Paul D 2002 “Missing Data”, Sage University Paper: QASS 136, London, Sage,

Det manglar Case i utvalet

- når ein person
  - nektar å svar,
  - Ikkje er heime,
  - er flytta,
  - Osv
- Problemet med frafall kjem inn under studiet av skeive utval. Generelt er dette eit meir alvorleg problem enn at vi manglar data på nokre variablar for nokre case (sjå Breen 1996 ”Regression Models: Censored, Sample Selected, or Truncated Data”, QASS Paper 111, London, Sage)
- Men problema er i slekt

Det manglar data på visse variablar når

- Personar nektar å svar på visse spørsmål
- Personar gløymer eller overser nokre spørsmål, eller intervjuar gjer det same
- Personar veit ikkje noko svar
- Spørsmålet er irrelevant
- I administrative register kan somme dokument ha gått tapt
- I forskingsdesign der vanskeleg målbare variablar berre vert målt for eit mindre tal personar i utvalet

Manglane data fører til problem

- Det er eit praktisk problem sidan alle statistiske prosedyrar føreset fullstendige datamatriser
- Det er eit analytisk problem sidan manglande data som regel gir skeive estimat av parametrane
- Det er eit viktig skilje mellom data som manglar av tilfeldige årsaker og dei som manglar av systematiske årsaker

Den enkle løysinga: fjern alle case med manglande data

- Listwise/ casewise fjerning av manglande data tyder at ein fjernar alle case som manglar data på ein eller fleire variablar inkludert i modellen
- Metoden har gode eigenskapar, men kan i somme høve ta ut av analysen mesteparten av casa
- Vanlege alternativ, som parvis (”pairwise”) fjerning, har vist seg å vere dårlegare
- Nyare metodar som ”maximum likelihood” og ”multiple imputation” har betre eigenskapar men er krevjande
- Det løner seg å gjere godt arbeid i datainnsamlinga

Typar av tilfeldig missing

- MCAR: missing completely at random
  - Tyder at mangel på data for ein person i på variabelen y ikkje er korrelert med verdien på y eller med verdien på nokon anna variabel i datasettet (dette hindrar ikkje at missing i seg sjølv kan korrelere internt case for case)
- MAR: missing at random
  - Tyder at mangel på data for ein person i på variabelen y ikkje er korrelert med verdien på y når ein kontrollerer for dei andre variablane i modellen
  - Meir formelt: Pr(Y=missing l Y,X) = Pr(Y=missing l X)

Prosessen som gir missing

- Kan ignorerast (ignorable)
  - Prosessen kan ignorerast dersom resultatet er MAR og parametrane som styrer missing prosessen ikkje er relatert til dei som skal estimerast
- Kan ikkje ignorerast (non-ignorable)
  - Prosessen kan ikkje ignorerast dersom resultatet ikkje er MAR. Modellestimering krev da ein eigen modell for missing-prosessen (sjå Breen 1996 ”Regression Models: Censored, Sample Selected, or Truncated Data”, QASS Paper 111, London, Sage)
- I det følgjande er det situasjonen med MAR data som vert drøfta

Konvensjonelle metodar

Vanlege metodar ved MAR data:

- Listevis utelating (Listwise deletion)
- Parvis utelating (Pairwise deletion)
- Dummy variabel korreksjon
- Innsetjing av verdi (Imputation)

Av dei vanleg brukte metodane er listevis utelating den beste

Listevis utelating (1)

- Kan alltid nyttast
- Dersom data er MCAR gir det eit enkelt tilfeldig utval av det opphavelege utvalet
- Mindre n gir sjølvsagt større variansestimat
- Også når data er MAR og missing på x-variablar er uavhengig av verdien på y vil listevis utelating gi forventingsrette estimat

Listevis utelating (2)

- I logistisk regresjon er listevis utelating problematisk berre dersom missing er relatert både til avhengig og uavhengige variablar
- Når missing berre er avhengig av den uavhengige variabelen sine eigne verdiar er listevis betre enn maximum likelihood og multiple imputation

Parvis utelating

- Tyder at alle utrekningar baserer seg på alt tilgjengeleg materiale sett parvis for alle par av variablar som inngår i analysen
- Dette fører til at ulike parametrar er rekna ut på grunnlag av ulike utval (variasjon i n frå observator til observator)
- Da er alle variansestimat og vanlege testobservatorar skeivt estimert
- Bruk ikkje parvis utelating!

Dummy variabel korreksjon

Dersom data manglar på den uavhengige variabelen x

- Sett x* = x dersom x ikkje er missing
  og   x* = c (ein vilkårleg konstant) når x er missing
- Definer D=1 hvis x er missing, 0 elles
- Bruk x* og D i regresjonen i staden for x
- I nominalskalavariable kan missing få sin eigen dummy

Studiar viser at sjølv med MCAR data er parameter-estimata skeive

Bruk ikkje dummy-variabel korreksjon!

Innsetjing av verdi (imputasjon)

- Målet her er å erstatte missing verdiar med rimelege gjettingar på kva verdien kunne vere før ein gjennomfører analysen som om dette var verkelege verdiar, t.d.
  - Gjennomsnitt av valide verdiar
  - Regresjonsestimat basert på mange variablar og case med gyldige observasjonar
- Parameterestimata er konsistente, men variansestimata er skeive (systematisk for små) og testobservatorar er for store
- Unngå om mogeleg å nytte enkel imputasjon

Oppsummering om konvensjonelle metodar for manglande data

- Vanlege metodar for korreksjon av manglande data gjer problema verre
- Ver nøye med datainnsamlinga slik at det er eit minimum av manglande data
- Prøv å samle inn data som kan hjelpe til med å modellere prosessen som fører til missing
- Der data manglar **bruk listevis utelating** dersom ikkje maximum likelihood eller multiple imputasjon er tilgjengeleg

Nye metodar for ignorerbare manglande data (MAR data): Maximum Likelihood

- Konklusjonar
  - Baserer seg på sannsynet for å observere nett dei variabelverdiane vi har funne i utvalet
  - ML gir optimale parameterestimat i store utval når data er MAR
  - Men ML krev ein modell for den felles fordelinga av alle variablane i utvalet som manglar data, og den er vanskeleg å bruke for mange typar modellar

ML-metoden: eksempel (1)

- Observerer y og x for 200 case
- 150 er fordelt som vist
- For 19 case med Y=1 er x missing og for 31 case med Y=2 er x missing
- Vi ønskjer å finne sannsyna p_{ij} i populasjonen

	Y=1	Y=2
X=1	52	21
X=2	34	43

	Y=1	Y=2
X=1	p_{11}	p_{12}
X=2	p_{21}	p_{22}

ML-metoden: eksempel (2)

- I ein tabell med I rekkjer og J kolonnar, fullstendig informasjon om alle case og med n_{ij} case i celle ij er Likelihooden

$$\mathcal{L} = \prod_{i,j} \left(p_{ij} \right)^{n_{ij}}$$

Dvs produktet av alle sannsyn for kvar tabellcelle opphøgd med cellefrekvensen som potens

ML-metoden: eksempel (3)

For ein firefeltstabell vert Likelihooden:

$$\mathcal{L} = (p_{11})^{n_{11}} (p_{12})^{n_{12}} (p_{21})^{n_{21}} (p_{22})^{n_{22}}$$

For dei 150 casa i tabellen ovanfor der vi har alle observasjonane vert den

$$\mathcal{L} = (p_{11})^{52} (p_{12})^{21} (p_{21})^{34} (p_{22})^{43}$$

ML-metoden: eksempel (4)

- For tabellar er ML estimatoren for $p_{ij} = n_{ij}/n$
- Dette gir oss gode estimat i den tabellen der vi ikkje har manglande data (listevis utelating av case)
- Korleis kan ein ta omsyn til det vi veit om y for dei 50 som manglar data på x?
- Sidan vi har MAR må dei 50 ekstra casa med kjent Y følgje marginalfordelinga til y
- Pr(Y=1) = ($p_{11} + p_{21}$) og Pr(Y=2) = ($p_{12} + p_{22}$)

ML-metoden: eksempel (5)

- Når vi tar omsyn til alt vi veit om dei 200 casa blir Likelihooden

$$\mathcal{L} = (p_{11})^{52} (p_{12})^{21} (p_{21})^{34} (p_{22})^{43} (p_{11}+p_{21})^{19} (p_{11}+p_{21})^{31}$$

- ML-estimatorane vil no vere

$$\widehat{p}_{ij} = \widehat{p}(x = i \mid y = j)\,\widehat{p}(y = j)$$

Fall 2004 © Erling Berge 2004 21

ML-metoden: eksempel (6)

- Tar vi omsyn til informasjonen vi har om case med manglande data får vi andre estimat av parametrane

Estimat av	Missing utelatt	Missing med
p_{11}	0.346	0.317
p_{21}	0.227	0.208
P_{12}	0.140	0.156
p_{22}	0.287	0.319

Fall 2004 © Erling Berge 2004 22

ML-metoden

- I det generelle tilfellet av manglande data finst det to tilnærmingar
  - EM metoden, ein tostegsmetode der ein startar med ein forventa verdi på dei manglande data som vert nytta til å estimere parametrar som igjen vert nytta til å gi betre gjetting på forventa verdi som igjen …..

    (metoden gir skeive estimat av standardfeil)
  - Direkte ML estimat er betre (men er tilgjengeleg berre for lineære og log-lineære modellar)

Nye metodar for ignorerbare manglande data (MAR data): Multippel Imputasjon

- Konklusjonar
  - Baserer seg på ein tilfeldig komponent som vert lagt til estimat av dei einskilde manglande opplysningane
  - Har like gode eigenskapar som ML og er enklare å implementere for alle slags modellar.
  - Men den gir ulike resultat for kvar gong den blir brukt

Multiple Imputasjon (1)

- MI har dei same optimale eigenskapane som ML, kan brukast på alle slags data og med alle slags modellar, og kan i prinsippet utførast med vanleg analyseverktøy
- Bruken av MI kan vere temmeleg krokete slik at det er lett å gjere feil. Og sjølv om det vert gjort rett vil ein aldri få same resultat to gonger på grunn av bruken av ein tilfeldig komponent i gjettinga (imputasjonen)

Multiple Imputasjon (2)

- Bruk av data frå enkel imputasjon (med eller utan ein tilfeldig komponent) vil underestimere variansane til parametrane. Konvensjonelle teknikkar klarer ikkje å justere for at data faktisk er generert ved imputasjon
- Løysinga for imputasjon med tilfeldig komponent er å gjenta prosessen mange gonger og bruke den observerte variasjonen i parameterestimat til å justere estimata av variansane
- Allison, side 30-31 forklarer korleis dette kan gjerast

Multiple Imputasjon (3)

- MI krev ein modell som kan nyttast til å gjette på manglande data. Som regel er det føresetnad om normalfordelte variablar og lineære samband. Men modellar kan lagast særskilt for kvart problem
- MI kan ikkje handtere interaksjon
- MI modellen bør ha med alle variablane i analysemodellen (også avhengig variabel)
- MI fungerer berre for måleskalavariable. Tar ein med nominalskalavariable trengst spesiell programvare
- Testing av fleire koeffisientar under eitt er meir komplisert

Fall 2004 © Erling Berge 2004 27

Data som manglar systematisk

- Krev som regel ein modell av korleis fråfallet oppstår
- ML og MI tilnærmingane kan framleis nyttast, men med mye strengare restriksjonar og resultata er svært sensitive for brot på føresetnadene

Fall 2004 © Erling Berge 2004 28

Oppsummering

- Dersom nok data vert igjen er listevis utelating den enklaste løysinga
- Dersom listevis utelating ikkje fungerer bør ein freiste med multippel imputasjon
- Dersom ein har mistanke om at data ikkje er MAR må ein lage ein modell for prosessen som skaper missing. Denne kan eventuelt nyttast saman med ML eller MI. Gode resultat krev at modellen for missing er korrekt