MISSING VALUES

Harry BG Ganzeboom
Opleiding Sociologie VUA
MA-Cursus Onderzoekslab
Maart 3-5 2020
Versie 2, 2 juli 2020

Citatie: Ganzeboom, Harry BG (2020). *Missing values. College-aantekeningen MA-cursus Onderzoekslab.* Amsterdam: Vrije Universiteit.
http://www.harryganzeboom.nl/Teaching/index.htm. Geraadpleegd: datum.

Inhoud

- Agenda
- Missing values
- Beschrijving missing values
- Pairwise-listwise / available-complete cases
- Missing values in Reliability
- Single value imputation
- Maximum likelihood
- Multiple imputation
- Paneldata
- Voorbeeldanalyses

AGENDA

- Kennismaking
- Inhoud en vorm van de cursus

Inhoud en vorm van de cursus

- Tijdschema
- Twee colleges, twee practica, twee practicumverslagen, een eindopdracht
- Literatuur: Allison
- Hoofdzaak: MV in SPSS
  - MV markering in SPSS
  - Listwise / pairwise
  - Enkelvoudige imputatie
  - Multiple Imputation
- Indien mogelijk: ML estimation in Stata SEM.

Wat moet je al weten

- SPSS: DESC, CORR, FACT, RELI, REGR.
- Gebruik SPSS syntax (met name COMPUTE en RECODE).
- Inferentiële statistiek: random sampling, steekproevenverdeling, SE, CI, H0, significantie testen.

Wat je moet leren

- Nadenken over de schade die MV mogelijk toebrengen aan je onderzoek: de twee belangrijkste leerpunten.
- Soorten Missingness
- Praktische omgang met MV (markeren) in SPSS.
- Handige manieren om MV enkelvoudig te imputeren.
- Multiple Imputation in SPSS.

MISSING VALUES

Missing values (MV)

- Missing values kun je in hoofdlijnen verdelen in twee soorten:
  - Een case is in zijn geheel missend (**case non-response**). Zelfs als die cases wel in de datamatrix staan, is er geen informatie.
  - Een waarde is missing binnen cases die verder wel geldige waarden heeft (**item non-response**).
- Hoewel het een radicaal onderscheid lijkt, is het toch vloeiend:
  - Of een case geheel of gedeeltelijk missend is, kan afhangen van je selectie van variabelen.
  - In panel data heb je vaak dat een case niet deelneemt aan een wave, maar daarvòòr en daarna wel.

MV in SPSS

- SPSS heeft twee manieren om missing values te behandelen
  - **System missing** [SYSMISS]: in de datamatrix zie je een punt (.).
  - **User-defined missings**: in de datamatrix zie je gewoon een waarde.
- User-defined missings kun je opnemen in kruistabellen (CROSS), system missings niet.
- User-defined missings kunnen value labels hebben, bv. “NA”, “Geen mening”, “Skipped”, etc.
- User-defined missing zijn best wel handig bij het beschrijven van data, maar bij analyse is het handiger om er sysmiss van te maken.
- Kies voor user-defined missings altijd een opvallende, bv. extreem negatieve waarde: -99.

SPSS syntax MV

```
MISSING VALUES v1 to v10 (-99 thru -1).
RECODE v1 to v10 (-99 = sysmiss).
```

Drie belangrijkste leerpunten

- Missing values leiden tot twee problemen:
  - (A) **Inefficiëntie**: je gebruikt minder data dan je eigenlijk hebt;
  - (B) **Bias**: je gebruikt selectieve data.

  Dit zijn VERSCHILLENDE problemen.
- Missing values analyse / treatment gaat NIET over de ontbrekende data, maar over hoe je data die je WEL hebt, kunt gebruiken.

Missing at random

- MCAR: Missing Completely at Random
- MAR: Missing at Random
- Het verschil zie je door listwise en pairwise correlaties te vergelijken: deze zijn (binnen toevalsgrenzen) aan elkaar gelijk als MCAR.
- Bij MCAR gaat het MV probleem alleen om de lagere N (inefficiëntie). Bij MAR gaat het ook om het biasprobleem.

Missing Not at Random

- Missing kunnen ook door onbekende, niet random (systematische) mechanismen optreden: MNAR.
- Belangrijke voorbeelden:
  - Wanneer de missings afhangen van de waarden van een variabele zelf (bv. hogere inkomensgroepen zijn geneigd geen opgave van inkomen te doen).
  - Non-response: de bereidheid om deel te nemen aan een volgende wave berust op onbekende (niet-gemeten) motivaties / verhuisgeneigdheid.
- MV-technieken bieden hier weinig soulaas – ze kunnen je alleen helpen als je wel informatie hebt over de het missing value mechanisme.
- Als je wel iets weet over de ontbrekende gevallen (bv. bij uitval in panels), kun je terugvallen op MAR.

BESCHRIJVING MISSING VALUES

Kijk naar de Frequencies

- Een goed begin van elke data-analyse is altijd de frequentie-verdelingen van alle variabelen te bekijken.
- Hier zie je dingen als extreme en onmogelijk waarden, of de MV correct gemarkeerd zijn.

Kijk naar de Descriptives

- DESC geeft niet alleen een overzicht van bereik (min, max), mean en SD, maar ook van het aantal *available cases* per variabele en het aantal *complete cases* (listwise).
- Een DESC tabel zou eigenlijk in elke analyse moeten worden gerapporteerd.
- Maar bepaalde gegevens ontbreken bij DESC: met name het totaal aantal cases en het totaal aantal cases met available information.
- Als je een batterij attitudevragen analyseert, is het handig ook summative index met de indicatoren op te nemen: dat geeft de *available N of Cases*.

Kijk naar Correlations

- Een heel goed uitgangspunt bij MV analyse is de (pearson) correlatiematrix.
- SPSS CORR geeft twee vormen: pairwise en listwise.
- (Je verkrijgt deze twee vormen op een iets fraaiere manier via REGR.)
- We kijken naar twee dingen: (A) de N of Cases (inefficiëntie) en (B) de correlaties zelf (bias).
- Eén ding krijg je ook hier niet (automatisch) te zien: de totale N of Cases.

N of Cases

- Je begint met je Total N of Cases – alle cases, of ze nu enige informatie bevatten of niet.
- Je bent vervolgens geïnteresseerd hoeveel cases überhaupt informatie bevatten.
- Er zijn twee belangrijke varianten van de available N of Cases:
  - De N of Cases die ten minste één geldige waarde bevatten.
  - De N of Cases die ten minste twee geldige waarden bevatten.
- De tweede variant is belangrijker als je geïnteresseerd bent in op correlaties berustende technieken zoals regressie en factor.
- Maar ook de eerste N of Cases is van belang, als M en SD van variabelen je interesseren. Denk aan vergelijkingen tussen waves in paneldata. Dan will je ook gebruik maken vam de cases die maar in één wave meedoen.

Syntax N of Cases

- Het is altijd verstandig al je missings als SYSMISS te declareren. Vervolgens kun je ze tellen:

```
Count NMiss = v1 to v10 (sysmiss).
```

- De frequentieverdeling van NMiss is het waard om bekeken te worden. Ze geeft het eerste volledige inzicht in de verschillende varianten van de N of Cases.

Patroon van MV

- Je wilt ook graag kijken naar het patroon van MV. Een eenvoudige syntax is:

```
Compute PatMiss=1.
Do repeat vvv=v1 to v10.
Compute Patmiss=Patmiss*10.
If (sysmiss(vvv)) PatMiss=Patmiss+1.
End repeat.
```

- De frequentie van PatMiss geeft je veel inzicht.
- **`Cross Patmiss by Nmis`** is ook een mooie om te bekijken. Herorden deze table op basis van de frequenties van PatMiss.

Correlaties van MV

- Nog iets om naar te kijken is of de MV gecorreleerd optreden.

```
Recode V1 to V10 (sysmiss=1)(else=0)
 into MV1 to MV10.
Corr MV1 to MV10.
```

- Als je eigenlijk alleen maar nul-correlaties ziet, heb je bewijs voor MCAR.
- (Het kan ook MNAR zijn, maar dat kun je niet zien.)
- Let op dat correlaties gevoelig zijn voor scheve verdelingen: als MV heel zeldzaam zijn, zullen de correlaties ook zwak zijn.

PAIRWISE-LISTWISE AVAILABLE - COMPLETE CASES

SPSS: listwise

- Listwise == complete case analysis.
- Dit is in SPSS de standaard, behalve bij:
  - CORR
  - DESC.
- Let erop dat de “complete cases” bepaald worden aan de hand van de geselecteerde variabelen, niet aan de hand van de variabelen die actief zijn in de analyse.
- In Stata:
  - `pwcorr _all, obs`
  - `sum _all`

Complete cases analysis

- Bij complete cases analysis (SPSS: listwise) maak je alleen gebruik van de cases die op alle variabelen een geldige waarde hebben.
- Nadeel 1: Dit leidt tot de kleinst mogelijke effectieve steekproef.
- Nadeel 2: Als niet MCAR, is dit waarschijnlijk ook de meest vertekende steekproef.
- Er zijn ook voordelen:
  - Nooit rekenproblemen,
  - Geen sores met MV treatment,
  - ‘Conservatieve’ [voorzichtige] aanpak: je geeft de H0 veel kans; als de H0 verworpen wordt, zit je goed.

Available cases analysis

- Bij available cases [information] analysis (SPSS: pairwise) maak je gebruik van een pairwise berekende correlatie / covariantie matrix.
- Dit is voldoende voor twee veel gebruikte analysemodellen: lineaire (OLS) regressie en factoranalyse (en hun combinatie: SEM).
- Het werkt NIET voor: (binomiale, multinomiale en ordinale) logistische modellen.
- Het werkt ook niet voor RELIability, hoewel die techniek wel uitsluitend op correlaties berust.

Pairwise deletion in SPSS

- SPSS biedt de mogelijkheid van available information analysis in twee belangrijke programma's: REGR, FACTOR.
- De bijbehorende beschrijvende grootheden (M, SD, Corr, N) kun je bij deze programma's zelf opvragen, maar ook via DESC en CORR.
- Let op dat UNIANOVA – waarmee je heel handig OLS regressie kunt doen, geen mogelijkheid tot pairwise deletion biedt.
- Merk ook op dat de SPSS schattingen van de SE bij pairwise deletion NIET KLOPPEN.
- Dit probleem is belangrijker bij regressie-analyse dan bij factor-analyse.

Wat doet SPSS bij pairwise deletion of missing values?

- Dit is onbekend, maar het lijkt er het meest op dat bij de berekeningen een homogene N aan alle correlatieparen wordt toegekend.
- Die N kunt je aflezen in de ANOVA table (Total DF + 1). Hij lijkt hetzelfde te zijn als de minimale N in de pairwise correlatiematrix.
- ***M.a.w.: SPSS pairwise deletion is er niet gevoelig voor dat de ene variabele meer missings heeft dan een andere.***

Waarom het toch nuttig is om SPSS pairwise deletion te gebruiken

- Hoewel de SE berekeningen niet kloppen, zijn de coëfficiënten wel correct. Ze lijken heel veel op wat je zou krijgen als je heel ingewikkelde MV technieken (MI, ML) zou toepassen.
- Als je data niet MCAR zijn, zijn deze waarden vermoedelijk dichter in de buurt van de populatiewaarden dan de listwise berekeningen.
- Vergelijking tussen een listwise en een pairwise berekening vertelt je dus iets over het MV mechanisme, en met name welke vertekening complete cases analysis teweegbrengt.
- Naschrift HG: in het voorbeeld dat we in practicum bestudeerd hebben, bleek dit niet te kloppen: pairwise analyse gaf vreselijke fouten bij schatting van een model met interacties. Dit moet ik nog nader uitzoeken.

Waarom pairwise deletion in kwade reuk staat

- Veel auteurs staan wantrouwend tegen pairwise deletion MV.
- Het meest genoemde concrete bezwaar is dat de pairwise correlatie matrix op verschillende subsamples berust en tot inconsistente resultaten kan leiden.
- In uiterste gevallen lopen de berekeningen vast; naar mijn ervaring is het zeldzaam en komt eigenlijk alleen voor als een pairwise N (bijna) nul is.
- Dat lijkt mij allemaal geen reden om listwise deletion te doen, integendeel.

MV IN RELIABILITY

MV in SPSS Reliability

- Cronbach's Alpha is een functie van de **gemiddelde** correlatie tussen de indicatoren van een meetinstrument.
- Het is daarom enigszins verrassend dat SPSS RELI geen optie biedt voor pairwise berekening van deze correlaties.
- En dat kan heel vervelend zijn: als MV heel gespreid zijn over veel indicatoren, hou je listwise soms heel weinig over.
- Het is daarom goed bij RELI analyse de N of Cases in de gaten te houden.
- Dit is juist belangrijk omdat de RELI procedure erop gericht is variabelen uit je analyse weg te laten. Het weglaten van een variabele heeft niet alleen consequenties voor de geschatte betrouwbaarheid, maar ook voor de effectieve N of Cases!!

Zelf betrouwbaarheidsanalyse doen

- Een belangrijk uitkomst van RELI is de kolom “corrected item-total” correlation.
- De kolom is de correlatie tussen de indicator en het gemiddelde van de overige indicatoren: de item-restcorrelatie.
- Een heel simpel, maar effectief idee om te kijken of elke indicator goed bij de rest hoort.
- Je kunt deze kolom heel gemakkelijk nadoen door zelf je item-restcorrelaties uit te rekenen.

Syntax item-rest correlaties

```
Compute Rest4 = mean(V1,V2,V3).
Compute Rest3 = mean(V1,V2,V4).
Compute Rest2 = mean(V1,V3,V4).
Compute Rest1 = mean(V2,V3,V4).
Corr V1 V2 V3 V4 with Rest1 Rest2 Rest3
Rest4.
```

- Op deze manier heb je niet zo'n last van de wisselende N.
- De informatie die je hieruit krijgt is overigens niet zo veel anders dan wanneer je gewoon de pairwise correlaties bekijkt.

SINGLE VALUE IMPUTATION

Wat kun je doen?

- Single value imputation:
  - Logical substitution
  - Mean substitution
  - Predicted value substitution

  (Deze aanpakken hebben met elkaar gemeen dat ze betrekkelijk eenvoudig zijn toe te passen en over algemeen tot incorrecte resultaten leiden; vaak zijn ze wel handig.)
- Maximum Likelihood: MLMV in Stata.
- Multiple Imputation (in SPSS of Stata).

  (ML en MI lijken erg te verschillen, maar zijn toch asymptotisch aan elkaar gelijk.)

Logical imputation

- Van logical imputation is sprake wanneer je de werkelijke waarde van een missing kunt afleiden uit de logica van een vragenlijst.
- Belangrijk voorbeeld: arbeidsinkomen is 0, wanneer iemand gezegd heeft geen baan te hebben (en de vraag naar inkomen niet gesteld is).
- Hoewel er in dit voorbeeld geen probleem lijkt te zijn, is het goed je te realiseren dat de imputation toch fout kan zijn: het antwoord op de filtervraag naar de baan kan fout zijn gegeven.

Mean imputation

- Een ruwe manier om de cases die een missing value hebben, er weer bij te betrekken is om in plaats van de missing het gemiddelde in te vullen.
- SPSS FACTOR biedt deze mogelijkheid bij berekening van factorscores.
- Het gemiddelde (of een andere centrale) waarde heeft als voordeel dat het weinig invloed op covarianties heeft.
- In regressie-analyse wordt mean-substitution verbeterd door een dummy (0/1) variabelen toe te voegen die aangeeft of er al dan niet sprake is van substitutie.
  Naschrift HG: dit is twijfelachtig. Allison raadt het af.

Mean substitution - problemen

- Mean substitutie (met controle dummy) lijkt heel onschuldig, maar kan tot vertekeningen leiden.
- Door het gemiddelde in te vullen, verklein je de spreiding (SD) van een variabele, en dus ook de correlaties.
- Ook beïnvloedt mean substitution verklaarde variantie, en daarmee alle inferentiële statistiek.
- Voor de hand liggend probleem is verder dat kruistabellen tamelijk grote onzin kunnen vertonen.
- Hoe schadelijk het allemaal is, lijkt me erg af te hangen van de hoeveelheid gesubstitueerde missings en de rol daarvan in een causaal model.

Predicted (single) value substitution

- Een voor de hand liggende verbetering is om de MV niet te vervangen door het gemiddelde maar door een best passende waarde.
- Regression-based: voorspel Y uit een stel X-vars, en gebruik het model om de missende Y-waarden te scchatten.
- De handige praktijk om bij indexconstructie alleen gebruik te maken van de geldige indicatoren, is eigenlijk een sluwe vorm van predicted value substitution.

Predicted value substitution - problemen

- Predicted value substitution vermindert mogelijk de problemen van mean substitution, maar lost ze niet op:
- Je vermindert nog steeds de variantie van variabelen, en beinvloedt (verzwakt) dus correlaties.
- Door een voorspelde Y te gebruiken, verhoog je de verklaarde variantie en vertekent dus de inferentiële statistiek.
- In kruistabellen blijf je problemen hebben: de geïmputeerde Y liggen in de buurt van gemiddelde Y.

De verkeerde afslag

- Mean / predicted value substitution hebben met elkaar gemeen dat ze de verkeerde afslag nemen in MV treatment.
- Ze zijn erop uit om zo goed mogelijk de gegevens die je NIET hebt, te benaderen.
- Maar je analyse moet erop gericht zijn om gebruik te maken van de gegeven die wel hebt en de invloed van de gesubstitueerde waarden te neutraliseren (== geen invloed op schattingen (bias) en efficiëntie (SE's).
- Hot deck imputation gaat wel in de goede richting.

Hot deck – nearest neigbour imputation

- Een klassieke aanpak uit het ponskaart-sorteer tijdperk.
- Sorteer je datamatrix op basis de verwachte waarde van het regressie-model waarmee je Y voorspelt (dit heet ook wel de propensity score).
- Imputeer nu voor de missende waarde de dichtstbijzijnde geldige waarde (vorige case).
- Dit heeft de volgende voordelen:
  - De verdeling van de geïmputeerde missende waarden zal binnen toevalsgrenzen die van de geldige waarden zijn.
  - De correlaties van de geïmputeerde waarden zullen veel lijken op die van niet-geïmputeerde waarden. Hoeveel, zal afhangen van de mate waarin Y van de X-en afhangt.
- Hot deck neutraliseert dus de invloed van de imputatie op de geschatte effecten (minder bias); je heb er niet veel aan om correcte SE's te vinden.

Available information bij batterij-vragen

- Als je een multiple-indicator schaal aan het analyseren bent, is een handige manier om van MV af te komen:
  - **Compute index = mean(V1 to V6).**
- Dit statement (ik weet er geen goede naam voor) middelt (ongewogen) alleen de geldige waarden in V1 to V6. Het levert dus de available N van schaal.
- Mijn ervaringen hiermee zijn dat het beter werkt dan factorscores en zeker dan factorscores met mean-substitution, die SPSS aanbiedt.

MAXIMUM LIKELIHOOD

Maximum Likelihood

- Hoewel ML niet binnen SPSS beschikbaar is, is inzicht erin van groot belang.
- ML is beschikbaar (als MLMV) in Stata SEM en eigenlijk heel gemakkelijk uit te voeren. Je hoeft niet veel Stata of SEM te kunnen.
- Ook (en al lang) beschikbaar in LISREL en MPLUS, en heet daar FIML (Full Information Maximum Likelihood).

Het opdelen van de correlatiematrix

- ML begint bij het opdelen van de correlatiematrix over subsamples die gedefinieerd zijn door de PatMiss.
- We krijgen dan de listwise matrix en een boel gedeeltelijk gevulde matrices.
- Het is goed je op dit moment af te vragen hoe je aan deze matrices MCAR kunt zien.

Regressie- en factoranalyse als algebra

- Je kunt elke regressie- / factor-analyse opschrijven als een eenvoudig stelsel algebraïsche vergelijkingen, waarin de correlaties de bekenden zijn en de coëfficiënten de onbekenden.
- SEM lost dit vergelijkingenstelsel voor je op (als het oplosbaar is).
- Bij de oplossing wordt de N van de correlaties als weging gebruikt: als een grotere N betrokken is, weegt deze matrix meer mee in de uiteindelijke (gewogen) oplossing.

ML in Stata

- In Stata SEM is ML voor missing values beschikbaar als een optie op een SEM model:

  `sem (…) (…) , method(mlmv)`

- Het gevolg is dat het model er nu wat langer over doet om te convergeren. Een enkele keer convergeert het niet terwijl het complete cases model wel convergeert.
- Ik geef een demonstratie. Het is verder om andere redenen heel erg de moeite waard om SEM te leren.
- De resultaten lijken sprekend op de verschillen tussen listwise / pairwise bij SPSS REGR en FACT. En dat is het ook.

ML voor- en nadelen

- ML heeft de volgende voordelen:
  - Het is eigenlijk heel gemakkelijk uit te voeren en zeer intuïtief.
  - Het is (asymptotisch) equivalent aan MI (zie verderop), maar is niet van toevalscomponenten afhankelijk.
  - Gemakkelijk om structurele en meetmodellen te verweven.
- Nadeel is dat het lastig is om hulpvariabelen te gebruiken. (Het kan wel, maar het is lastig.)

Mijn puzzel

- Er is heel veel literatuur over Multiple Imputation. ML wordt vaak maar zijdelings of helemaal niet besproken.
- Toch is ML in veel gevallen veel gemakkelijker uit te voeren.
- Ligt dit aan de ontoegankelijkheid van SEM programmatuur? Of is er meer aan de hand?
- ML werkt goed bij lineaire modellen, niet bij logistische modellen. Is dit een reden dat ze zo weinig gebruikt worden?

MULTIPLE IMPUTATION

Multiple Imputation (MI)

- MI is de meest gebruikte manier om missing values te behandelen.
- Het zit in SPSS en is niet zo moeilijk uit te voeren.
- MI is niet zo intuïtief, het leidt vaak tot een misverstand, nl. dat je bezig bent data te verzinnen. Dat is niet zo, maar het misverstand is onuitroeibaar.
- ***Ook bij MI gaat het erom de data te gebruiken die je wel hebt, niet om de data die je niet hebt.***
- Een nadeel is verder dat de uitkomst een toevalscomponent heeft – en die kun je nog manipuleren ook en dus verleidelijk voor bedrog.

Wat gebeurt er bij MI?

MI valt uiteen in twee stappen: (A) Imputatie en (B) Schatting:

- Stap A.1: Zoals bij single value imputation voorspel je missende waarde, bv met een regressiemodel.
- Stap A.2: aan de voorspelde waarde wordt een toevalscomponent toegevoegd, random getrokken uit de residuen van het voorspellende model.
- Stap A.3: dit doe je meerdere (multiple) keren, bv. 25 keer (hoe vaker hoe beter).
- Stap B.1: zo verkrijg je meerdere (bv. 25) datamatrices. In elk hiervan reken je het model je interesse uit. De coëfficiënten variëren door toeval. Hun gemiddeld is je puntschatting.
- Stap B.2: door toepassing van Rubin's Rules leid je de SE van de geschatte coëfficiënten af – hiermee doe je significantietesten en confidence intervals.

SPSS doet al deze stappen in een keer voor je en levert je de gepoolde schatting.

Voor- en nadelen van MI

- Voordelen van MI:
  - Je kunt bij de imputatiestap gebruik maken van andere (en meer) variabelen dan bij berekenen van je eigenlijke model: de hulpvariabelen.
  - Als je eenmaal een MI-datamatrix hebt, doe je eigenlijk verder hetzelfde als bij gewone analyses.
- Nadelen van MI:
  - Imputatiestap kost veel tijd, vooral als je veel data hebt.
  - Geen berekening van Beta (kun je ondervangen door data zelf te standaardiseren).
  - Geen berekening van verklaarde variantie.
  - Er komt telkens iets anders uit, als je het random proces op een ander punt begint (SET SEED zet dit vast).

Rubin's Rules

- Stel je hebt k MI replicates en je bent geïnteresseerd in de SE van een bepaalde B.
- Bereken de gemiddelde within-variantie van deze B's: mean(SEk*SEk).
- Bereken de gemiddelde between-variantie van de B's: mean(B – Bk)**2.
- De variantie van B is het gewogen gemiddelde van deze beide varianties.
- Waarom dit zo is, heeft Donald Rubin bedacht. Hij heeft er geen Nobelprijs voor gekregen, maar dat zou nog kunnen gebeuren.

Praktische tips bij MI

- Probeer bij de voorspellende variabelen in de imputatiestap alles erbij te betrekken dat ermee te maken heeft.
- Ook de afhankelijke variabele!
- Het werkt het beste als je (ook) met gestandaardiseerde data werkt.
- Meer imputaties kosten meer tijd, maar maken het resultaat wel stabieler.
- Maak bij regressie gebruik van /missing=pairwise, dat is de relevante vergelijking tussen het Original Model en het Pooled Model.
- Let bij interpretatie meer op de SE's dan op de B's.

Wat kun je doen bij MNAR?

- MI en ML werken als op de een of andere manier het MV mechanisme random is.
- Maar dat hoeft niet zo te zijn: MNAR; je kunt echter niet weten of data MNAR zijn.
- Het probleem is principieel: als er geen informatie is over het MV mechanisme, weet je niks.
- Oplossingen draaien allemaal om het idee om er toch een MAR situatie van te maken:
  - Opnieuw data te verzamelen onder de missings.
  - Beredeneren dat de missende cases veel lijken op de cases die laat binnenkwamen.

Maakt MCAR en MAR verschil?

- Het verschil tussen MCAR en MAR wordt in elke MV tekst benadrukt.
- Toch heeft het verschil geen consequentie hoe je te werk gaat in MI (of ML).
- Het verschil zit eigenlijk alleen maar in de soort problemen:
  - MCAR leidt alleen tot een efficiëntieprobleem
  - MAR leidt ook tot een biasprobleem.

De afhankelijke variabele Y

- Het is bij MI van groot belang dat je in de imputatiestap ook de Y-var gebruikt.
- In de schattingsstap laat je deze geïmputeerde waarden Y juist weer weg.
- Als je alleen maar missing values in Y hebt en niet in de X-vars, hebben MI en ML geen zin.

PANELDATA

MV bij paneldata

- MV spelen een enorme rol bij analyse van multi-wave panel data. Hierin is vrijwel altijd sprake van uitval (‘attrition’) en dus zit je volop in de MV problemen: (A) je hebt steeds minder complete cases (inefficiëntie) en (B) er is gevaar voor selectieve uitval (bias).
- Toch is de veronderstelling van MAR ook bij paneldata helemaal niet gek: juist doordat je eerdere of latere metingen wel hebt, weet je relatief veel over mogelijke oorzaken van de uitval.
- Panel data lijken ook in veel opzichten op (randomized) split ballot designs, waar je groepen respondenten met opzet (overlappende) delen van een vragenlijst voorlegt. Ook hierin kom je met de complete cases nergens.
- Je kunt dit type incomplete data te lijf met zowel MI als ML, maar het gebruik van MI lijkt me hier nog moeizamer dan bij gewone cross-secties. ML is veel voor de hand liggender.

Bekende misverstanden over paneldesigns

- Door panel attrition neemt geleidelijk de N of Cases af (en wordt mogelijk selectiever). Je schiet niet op door waves weg te laten: je krijgt niet meer data door data weg te laten.
- Bij respondentbenadering in panels moet je iedereen telkens opnieuw benaderen. Het is voor goede analyse erg waardevol als je (ook) deelnemers hebt die waves hebben overgeslagen en opnieuw zijn ingestroomd.
- Er is volgens mij weinig tot geen voordeel om mensen meer dan drie keer mee te laten doen. Je kunt beter nieuwe respondenten toevoegen dan achter oude respondenten aan jagen. Als je lange termijn onderzoeks-vragen hebt, kun je beter je waves met onregelmatige tussenafstand uitzetten.
- Met paneldata kun je geen sociale veranderingen in kaart brengen, panels gaan over individuele veranderingen. Sociale (historische) veranderingen zie je alleen in herhaald cross-sectie onderzoek. In panel onderzoek betekent dat: refreshing panels – telkens nieuwe respondenten toevoegen.

VOORBEELDANALYSES

Incomplete attituden

- In onderzoek met attitudenvragen treden vaak MV op:
  - Weigering of niet weten (in dit geval erg verwant)
  - Vermoeidheidseffects (respondent fatigue) in batterijvragen.
- Er is vaak een markant verschil tussen “complete N of Cases” en “available N of Cases”.
- Scheepers, Peer, Manfred Te Grotenhuis, and Frans Van der Slik. 2002. “Education, Religiosity and Moral Attitudes: Explaining Cross-National Effect Differences.” *Sociology of Religion* 63 (2): 157. https://doi.org/10.2307/3712563.

Intergenerationele statusoverdracht

- Bij onderzoek naar statusoverdracht tussen ouders en kinderen (mannen en vrouwen) zijn we vaak geïnteresseerd in opleidingen en beroepen van vaders en moeders.
- Kenmerken van ouders worden vaak geplaagd door MV
  - Weigering
  - Niet weten
  - Moeders hebben vaak geen beroep gehad.
- Hoe kun je nu een compleet beeld van intergenerationele statusoverdracht geven, waarbij niet alleen de complete ouders in de anayse zijn?
- Vries, Jannes De, and Harry BG Ganzeboom. 2008. “Hoe Meet Ik Beroep? Open en Gesloten Vragen Naar Beroep Toegepast in Statusverwervingsonderzoek.” *Mens En Maatschappij* 83 (1/2): 71-95 / 190-191.

Inkomens in Hongarije

- MV treden veelvuldig op in inkomensmetingen
  - Weigering
  - Niet weten
- MV problemen worden groter als je heel nauwkeurig naar inkomen informeert door naar verschillende inkomenscomponenten te vragen.
- Jansen, Wim, Willem-Jan Verhoeven, Peter Robert, and Jos AG Dessens. 2013. “The Long and Short of Asking Questions about Income: A Comparison Using Data from Hungary.” *Quality & Quantity* 47 (4): 1957–69. https://doi.org/10.1007/s11135-011-9636-5.
- Ganzeboom, Harry BG, and Derk C Sparendam. 2019. “About the Long and Short of Asking Income Questions. A Reconstruction and Extension.” *[Working Paper, Last Revised 2019/07/15]*.