**Grootschalig veldonderzoek**Door: Marleen Hasselt | Docenten: Rob Eisinga & Peter van Groenestijn | Vakcode: SOW-MTB2004

***Deze samenvatting is afkomstig uit het collegejaar 2017-2018. Het kan zijn dat sommige onderdelen iets verschillen van de huidige tentamenstof. Let hier op!***

**Cursusomschrijving:** De cursus grootschalig veldonderzoek bestaat uit drie onderdelen:  
  
a. Het eerste deel is gericht op het verkrijgen van inzicht in het construeren van vragenlijsten, het verzamelen van data met name via telefonische interviews en het zorgvuldig documenteren van de data zodanig dat deze door derden te gebruiken zijn. De volgende onderwerpen komen hierbij aan bod: vraagformulering en vragenlijst constructie, typen steekproeven en steekproeftrekking, nonresponse, representativiteit en weging, typen van missing values en hoe daarmee om te gaan, problemen/moeilijkheden die typerend zijn voor een telefonisch survey onderzoek.  
  
b. Het tweede deel bestaat uit het toepassen van de verworven kennis. Studenten voeren twee opdrachten uit m.b.t. steekproeftrekking, representativiteit en weging; zij verrichten literatuuronderzoek naar bestaande vragenlijsten en formuleren indien nodig nieuwe vragen; zij evalueren vragen gesteld in voorgaande leerproject-surveys en verwijderen/herformuleren vragen voor het lopende survey indien nodig. Tenslotte worden de eigen data verzameld met behulp van een CATI (computer-assisted telephone interview) systeem.  
  
c. Het derde deel bestaat uit het toepassen van de in deel a opgedane kennis aan de hand van de in deel b verzamelde data. De datadocumentatie vindt plaats. Studenten maken van het telefonisch survey een data documentatie rapport met daarin veldwerkverslag (waaronder steekproefkader en -trekking, nonresponse, representativiteit), geannoteerde variabelenlijst, coderingen, schriftelijke vragenlijst, literatuur.

**Doelstellingen:**Het verwerven van kennis over en opdoen van praktische ervaring in de uitvoering van een (telefonisch) survey. In het bijzonder gaat het over kennis en de toepassing daarvan inzake:  
1. typen steekproeven en steekproeftrekking;  
2. weging en representativiteit;  
3. oorzaken van en omgaan met nonresponse;  
4. typen van missing values en hoe daarmee om te gaan;  
5. het formuleren van vragen en het construeren van vragenlijsten;  
6. literatuuronderzoek naar bestaande vragen(lijsten) van andere surveys;  
7. dataverzameling via telefonisch interviewen;  
8. het opzetten van de datadocumentatie in een wetenschappelijk rapport.  
Aan het einde van de cursus is de student in staat zelfstandig een eenvoudig survey-onderzoek uit te voeren.

**Hoorcollege 1 Grootschalig veldonderzoek, hoe en wat?**

**De empirische cyclus**

1. Onderzoeksvragen
2. Conceptueel model
3. *Onderzoeksontwerp*
4. *Selectie onderzoekseenheden*
5. *Selectie dataverzameling technieken*
6. *Data-ordening*
7. *Analyse*
8. Conclusie en interpretatie
9. Verslaglegging
10. Antwoord op onderzoeksvragen

Grootschalig veldonderzoek = beschrijven of verklaren van verschijnselen via ondervraging van een relatief groot aantal mensen (respondenten) op een relatief groot aantal kenmerken.   
Het doel van dit type onderzoek is

* Beschrijven van verschijnselen
* Verklaren (eventueel voorspellen) van verschijnselen
* Toetsen van hypothesen over verschijnselen

Voorbeelden van grootschalig veldonderzoek: opinie-, markt-, kiezersonderzoek

**Enquêtes**

Er zijn verschillende manieren waarop je grootschalig veldonderzoek kan doen. Een daarvan is de enquête. Er zijn verschillende enquête varianten:

* Mondeling interview  
  + Beste respons  
  + Beste kwaliteit  
  - Duur en tijdrovend  
  - Meer last van sociale wenselijkheid en interviewer effecten
* Telefonisch  
  + Snel  
  + Relatief goedkoop  
  + Goede kwaliteit  
  + Minder last van sociale wenselijkheid  
  - Minder goede respons  
  - Korte interviews  
  - Geen visuele of schriftelijke hulpmiddelen
* Schriftelijk interview  
  + Goedkoop  
  + Minder last van sociale wenselijkheid  
  - Respons is minder goed
* Web-interview (CAPI, CATI, CAWI)  
  + Goedkoop  
  + Minder last van sociale wenselijkheid  
  - Respons is minder goed  
  - Afhankelijk van internet

**Steekproeven trekken**Doelpopulatie: groep die je wilt onderzoeken en waarover je uitspraken wilt doen.   
Steekproef: selectie uit de populatie ten behoeve van je onderzoek.

Er kan op verschillende manieren een steekproef getrokken worden:

* Kanssteekproef: elk element in de populatie heeft een bepaalde kans om in de steekproef te worden opgenomen, die kans kan berekend worden.
  + Systematische steekproef: steekproef trekken uit een lijst (gemeentelijke basis administratie, etc.).
  + Gestratificeerde steekproef: eerst deelpopulaties onderscheiden, dan per deelpopulatie een steekproef trekken (onderverdeling in allochtonen en autochtonen, etc.).
  + Clustersteekproef: eerst clusters selecteren, dan per cluster alle eenheden onderzoeken (steekproef van scholen, dan van klassen en daarin vervolgens alle leerlingen onderzoeken).
  + Getrapte steekproef: via meerdere stappen trekken (gewogen steekproef van gemeente, steekproef van inwoners van gemeente).
* Niet-kanssteekproef (monster): elke andere soort steekproef (vrienden, mensen die zichzelf aanmelden).

Wil je generaliserende uitspraken doen, kies dan altijd voor de kanssteekproef.

**Doelpopulatie en steekproefkaders**Steekproeven trekken we uit een steekproefkader.   
Steekproefkader: bevolkingsregisters, direct mail bestanden, andere adressenbestanden, verzameling geldige telefoonnummers.   
Onderdekking: onderzoekselementen doelpopulatie komen niet in steekproefkader voor (bijv. geen telefoon, mobiel, geheim nummer).   
Overdekking: onderzoekselementen steekproefkader horen niet tot de doelpopulatie (bijv. bedrijven in onderzoek naar particuliere personen).

**Vragenlijst constructie en vraagformulering**

1. Inventariseer altijd eerste bestaande operationalisaties  
   + Betrouwbaarheid en validiteit dikwijls bekend/ onderzocht  
   + Vergelijkbaarheid met andere data en onderzoeken
2. Bestudeer bestaande operationalisaties kritisch
   1. Zijn er veel ongeldige antwoorden -> duidelijkere vragen maken
   2. Is de verdeling erg scheef? -> antwoordmogelijkheden aanpassen
   3. Veel items voor een concept? -> items schrappen
3. Als je zelf vragen gaat formuleren: voorkom formuleringsfouten (zie onderstaande paragraaf)
4. Logische opbouw van de vragenlijst:
   1. Begin met een paar eenvoudige interessante vragen.
   2. Val met de deur in huis. Begin met inhoudelijke en relevante vragen.
   3. Plaats vragen die bij elkaar horen in een blok bijeen.
   4. Logische volgorde, logisch voor de respondent
   5. Moeilijke en gevoelige vragen: voorzichtig inleiding, in het midden van de vragenlijst.
   6. Achtergrondkenmerken achterin vragenlijst
   7. Vragen met risico op non-respons (als inkomen) achterin plaatsen.
5. Uittesten van de vragenlijst met proefinterviews is aanbevolen.

**Vragen formuleren**

Je hebt verschillende vraagvormen:

* Openvragen: + meer informatie + suggereert geen antwoorden + doorvragen – lastig kwantificeerbaar – tijdrovend.
* Gesloten vragen: + kwantitatief + snel + grootschalig – beperkte informatie – opties soms onduidelijk of niet uitputtend – respondenten interpreteren vraag verschillend.
* Combinatie: opsomming met optie -> anders namelijk…

Vragen moeten begrijpelijk en relevant zijn voor respondenten.

Voorkom formuleringsfouten:

* Taalgebruik: geen moeilijke woorden -> gebruik een inleiding/ omschrijving
* Geen ontkennende vragen en geen dubbele ontkenning
* Voorkom vragen of antwoorden met vage telwoorden *(Hoe vaak bent u slachtoffer geweest van fietsendiefstal? Vaak, soms, nooit?)*
* Geen lange vragen, geen motivering in vragen
* Geen dubbelzinnige vragen
* Geen dubbele vragen: niet meer dan een onderwerp tegelijk -> opsplitsen in meerdere vragen.
* Geen vragen met feiten waarover discussie mogelijk is
* Geen vragen met mogelijk verschillende normen
* Geen suggestieve vragen (*Vindt u ook niet dat de politie haar werk goed doet?)*
* Geen veronderstellende vragen
* Vermijd items met verschillende interpretatiemogelijkheden
* Antwoorden moeten aansluiten bij de vragen
* Antwoorden moeten uitputtend en wederzijds uitsluitend zijn.

**Hoorcollege 2 Non respons, representativiteit en weging**

**Missing data**

Je hebt verschillende soorten missing data:

* Missing cases (unit nonrespons)
  + Respondent is niet bereidwillig (weigering)
  + Niet bereikbaar (niet thuis/ aanwezig)
  + Niet in staat deel te nemen (ziekte, taalproblemen)

Weigering is de laatste decennia sterk toegenomen

* Missing values (item nonrespons)
  + Ontbrekende observaties
  + In survey onderzoek is dit vaak: weet niet/ geen antwoord/ niet van toepassing.

**Toevallige en systematische non-respons**

Non-respons kan toevallig of systematisch zijn. Bij toevallige non-respons:

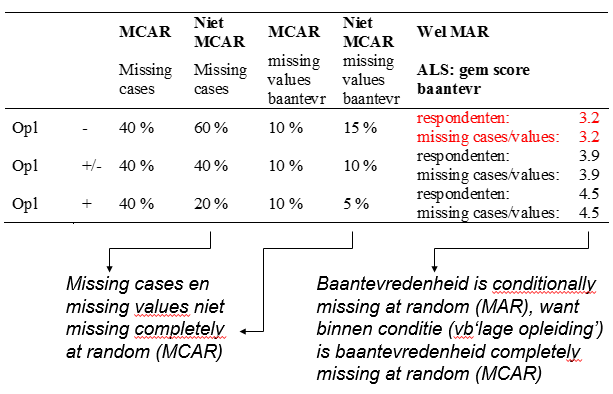
* Wijken cases (non-respondenten) niet systematisch af van cases (respondenten).
* Respondenten met missing values wijken niet systematisch af van respondenten zonder missing values.

Er zijn twee soorten toevallige non-respons:

* Missing Completely At Random (MCAR) -> waarden ontbreken volledig willekeurig (zuiver toeval). Kans dat Y missing is, is onafhankelijk van de waarde van Y, en van andere variabelen.
* (conditionally) Missing At Random (MAR) -> (tussenvorm) gecontroleerd voor andere variabelen X zijn de missings op Y onafhankelijk van score op Y.

Er is een soort systematische/ selectieve non-response:

* Not Missing At Random (NMAR) -> Missing op Y zijn afhankelijk van score op Y.

Voorbeeld (conditionally) Missing At Random (MAR):

**Gevolgen van non-respons**

1. Minder informatie (kleinere N): dit betekend dat je minder precieze uitspraken kan doen en dat je een grotere kans op het niet-significante resultaten.
2. Wellicht vertekende resultaten: selectieve non-respons -> indien cases met missing values systematisch afwijken van cases zonder missing values, zijn resultaten vertekend/ biased.
3. Hoe groot is het probleem? Afhankelijk van: de omvang van nonrespons, in combinatie met de mate waarin nonrespons systematisch is.

De gevolgen van non-respons zijn ook afhankelijk van de onderzoeksvraag: nonrespons heeft meestal minder gevolgen voor het vaststellen van samenhangen dan voor vaststellen van gemiddelden of frequenties.

Voorbeeld 1: *Hoe tevreden is men in Nederland over hun baan?* Non-respons leidt tot overschatting, door hogere non-respons onder lager opgeleiden (die immers lager scoren op tevredenheid).

Voorbeeld 2: *Samenhang opleiding en baantevredenheid?* Non-respons hoeft niet tot vertekening van samenhang te leiden.

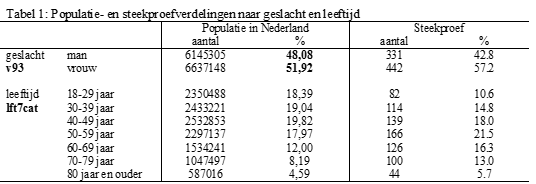
**Omgaan met missing cases** (=respondenten die niet mee doen)

Vooraf:

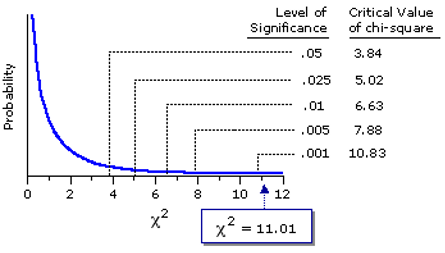
* Onderzoeksopzet: probeer missing cases te vermijden door middel van strategie: en gericht op het vergroten van de respons -> aankondigen enquête, overhaalpogingen interviewer, beloning respondenten, respondent methode laten kiezen.
* Probeer hierbij vooral de bereikbaarheid en bereidwilligheid van groepen NMAR te vergroten.

Achteraf:

* Controleer of missing cases toevallig zijn met behulp van aanvullende informatie:
  + Dataverzameling: aantekeningen interviewer, contactformulier, non-respons vragenlijst.
  + Data-analyse: controleer (toets) representativiteit op basis van hulp variabelen
* Weging van de data indien nodig.

Toets representativiteit met behulp van hulpvariabelen.

1. Op basis van de populatieverdeling zijn de *verwachte* aantallen bij een steekproefomvang van n=773:  
   Man: 773 \* 0,4808 = 372 respondenten  
   Vrouw: 773 \* 0,5192 = 401 respondenten
2. De *geobserveerde* aantallen in de steekproef van n=773 zijn:  
   Man = 331 respondenten  
   Vrouw = 442 respondenten
3. De verschillende tussen *geobserveerde* en *verwachte* aantallen zijn:   
   Man: -41  
   Vrouw 41
4. Zijn deze verschillen significant?  
   H0: geobserveerde en verwachte frequentieverdling komen met elkaar overeen. Voor toetsing van deze nul-hypothese gebruiken we Chi-kwadraat toets:   
   
5. Bij Chi-kwadraat van 8,567 en df= 1 hoort een p-waarde van 0,003. Omdat p kleiner is dan 0,05 kun je H0 verwerpen. Dit betekend dat de steekproef significant afwijkt van de populatie.

Df = aantal vrijheidsgraden = 1. Aantal klassen of categorieën (hier van geslacht, dus 2) - 1).

Weging van data

1. Check de representativiteit met betrekking tot hulpvariabelen:   
    populatie steekproef  
   man 48% 43%

vrouw 52% 57%

1. Data wegen naar geslacht:  
   Weegfactor = proportie in populatie / proportie in steekproef  
   - Weegfactor voor mannen = 0,48/0,43= 1,12  
   - Weegfactor voor vrouwen = 0,52 / 0,57 = 0,91

**Omgaan met missing values** (=respondenten beantwoorden sommige vragen niet)

Zie college 5

**Hoorcollege 3: Samenvoegen datafiles en variabele constructie**

Bij het samenvoegen van databestanden zijn er vier scenario’s mogelijk:

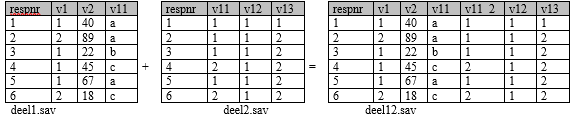
1. Parallel samenvoegen
2. Niet parallel samenvoegen
3. Niet parallel samenvoegen context gegevens
4. Onder elkaar plakken van bestanden

Voordat je databestanden gaat samenvoegen moeten er aan drie randvoorwaarden voldaan worden.

1. Naamgeving van de sleutelvariabele (de variabele waarop gekoppeld wordt) identiek in de bestanden die je gaat samenvoegen.
2. Formaat van de sleutelvariabele moet identiek zijn in de bestanden
3. De samen te voegen files moeten gesorteerd zijn op identieke wijze (oplopend/ aflopend) op de sleutelvariabele.

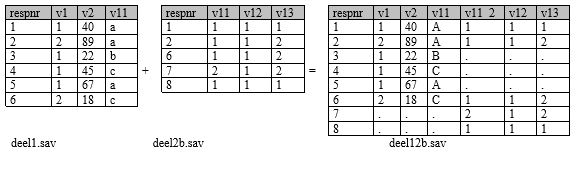
**Scenario 1: Parallel samenvoegen van databestanden**

Bij parallel samenvoegen wil je verschillende variabele van dezelfde cases (bijvoorbeeld respondenten) aan elkaar plakken, de verschillende bestanden bevatten identieke cases/ respondenten maar verschillende variabelen (vaak op uitzondering van het respondentnummer). Wil je de variabele uit beide bestanden overnemen die dezelfde naam hebben, dan moet je die in een van de twee bestanden van een andere naam voorzien of de variabele hernoemen bij het samenvoegen.

Deze parallel samenvoeging gebruik je bijvoorbeeld in gevallen waarbij je gegevens van beide gegevens per respondent naast elkaar wilt plakken voor onderlinge vergelijking. Je hebt dan te make met het aan elkaar plakken van matrices door het toevoegen van variabelen via overeenkomstige respondentnummers. Denk er aan dat beide matrices gesorteerd moeten staan op de sleutelvariabele.   


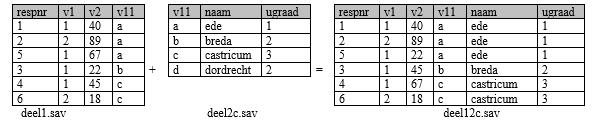
**Scenario 2: Niet parallel samenvoegen**

Niet parallel samenvoegen houdt in dat de bestanden naast dezelfde ook verschillende cases/ respondenten bevatten (bijvoorbeeld panels waarbij personen in een meting ontbreken) of één van de bestanden bevat van één vase/ respondent meer records (zoals gemeenten versus respondenten). Bij het koppelen van ebstanden met verschillende cases ontstaan ‘gaten’ op plekken van variabelen uit het ene bestand en cases uit het andere. SPSS plaats hier zogenaamde system-missings.



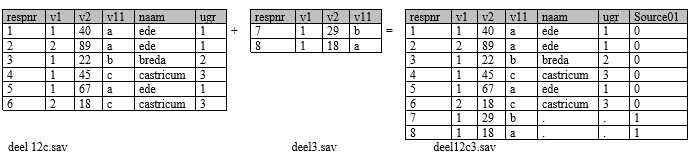
**Scenario 3: niet parallel samenvoegen context gegevens**

Het principe van dit scenario is niet anders dan bij scenario 2. Alleen voeg je hier onderzoeksdata samen met contextgegevens. Contextgegevens bijvoorbeeld gemeentegegevens in (SOCON). Er is bijvoorbeeld één sleutelvariabele *Z* die in beide files voorkomt. In beide files moet dan op dezelfde wijzen op *Z* gesorteerd worden. Er zijn waarschijnlijk mensen die in dezelfde gemeente wonen, dus moeten die dezelfde contextgegevens krijgen toegevoegd.



**Scenario 4: aan elkaar plakken van bestanden**

Bestanden bevatten (grotendeels) dezelfde variabelen maar verschillende cases/ respondenten. Dit toevoegen van cases zal normaal gesproken alleen gebeuren als het gaat om enkele verschillende cases, bijvoorbeeld wanneer je gegevens uit verschillende panels waarvan de respondenten niet overlappen wilt combineren (denk aan een onderzoek met controlegroepen), of wanneer je respondenten uit één enquête in verschillende bestanden hebt opgeslagen. Het is hierbij wel belangrijk dat identieke variabele namen in beide bestanden ook variabelen zijn die inhoudelijke hetzelfde zijn.



SPSS-begrip verklaringen:

* Non active dataset is keyed table: de uitgangsfile is het startpunt. Dat wil zeggen dat de respondenten die in deze file zitten, na samenvoeging ook in de totale file terugkomen. Met andere woorden: de data van de niet active file worden indien mogelijk (dat wil zeggen overeenkomstige respondentnummers) toegevoegd aan de uitgangsfile.
* Active set is keyed table: nu is het omgekeerde het geval. Slechts de respondenten die in de niet-actieve file voorkomen, worden in de totale file meegenomen.
* Both file provide cases: als je wilt dat alle respondenten worden opgenomen in het eindresultaat, ongeacht of ze in beide of slecht in één datafile voorkomen. Deze optie wordt het meest gebruikt.

**Hoorcollege 5 Omgaan met missing data**

**Omgaan met missing cases**Check de representativiteit op basis van hulpvariabelen. Indien het niet representatief is kun je de data wegen door middel van weegfactoren. Zie college 3.

**Omgaan met missing values**

Je kan op verschillende manieren omgaan met missing values in de data analyse:

1. Ontbrekende gegevens verwijderen
2. Ontbrekende gegevens als aparte categorie of aparte variabele opnemen in de analyse
3. Schatting maken voor de missing values + schattingen imputeren.

Ontbrekende gegevens verwijderen

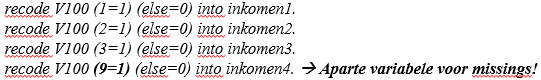
In dit geval laten we cases/ respondenten met missing values in de analyse buiten beschouwing. Het nadeel hiervan is, is dat je respondenten verliest en je N dus kleiner wordt. Ontbrekende gegevens verwijderen kan op twee manieren:

1. Listwise deletion: heeft de case één missing? Case wordt geheel verwijderd uit het bestand. + rekentechnisch is dit eenvoudig. – verlies respondenten. Dit kun je dan ook alleen toepassen bij relatief gering aantal missings.
2. Pairwise deletion: Dit houdt in dat SPSS alleen de missing values laat vallen en de rest van de case behoudt. Bijvoorbeeld, van een respondent die haar leeftijd niet heeft ingevuld, neem je wel ingevulde velden mee. + minder verlies van respondenten. – steekproefsamenstelling (N) verschilt per berekening. Daardoor krijg je soms vreemde uitkomsten of zelfs onvergelijkbare statistics/ rekenproblemen.

Ontbrekende gegevens als aparte variabele opnemen in de analyse

Categorische variabelen: cases met ontbrekende gegevens niet verwijderen maar als aparte categorie in de analyse opnemen. Dit heeft als voordeel ten opzichte van listwise deletion dat er een groter aantal cases (N) en dus meer informatie mee wordt genomen in het onderzoek.

Voorbeeld: inkomen met 3 categorieën, missing=9.  
Dummificeren: voor elke categorie – inclusief de categorie van respondenten met missing – wordt een afzonderlijke variabele (dummie-variabele) gemaakt.

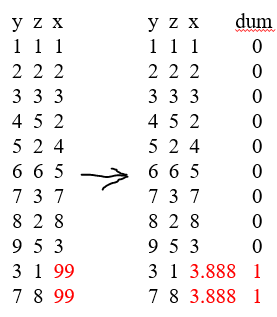


De 4 inkomensdummies (waarvan 1 referentiecategorie) worden allemaal opgenomen in de analyse.

Methode van Cohen: dummy variabele adjustment method -> alleen toepassen als waarde missing is, omdat valide score eenvoudigweg niet kan bestaan. Bijvoorbeeld als je iemand vraagt die enig kind is thuis wat de opleiding van zijn boer/ zus is. Ander voorbeeld: vragen over opvoeding van kinderen aan respondenten zonder kinderen.

De procedure van Cohen in stappen:

1. Maak een dummie variabele met code 1, als variabele X missing, anders code 0.
2. Als X missing, neem dan voor X het gemiddelde van X
3. Neem naast X ook dummie variabele op in het regressiemodel.



Schatting maken voor de missing values + schatting imputeren (d.w.z. invullen)

Missing values worden vervangen door een geschatte (geprediceerde) waarde. Deze ontbrekende waarden worden voorspeld door:

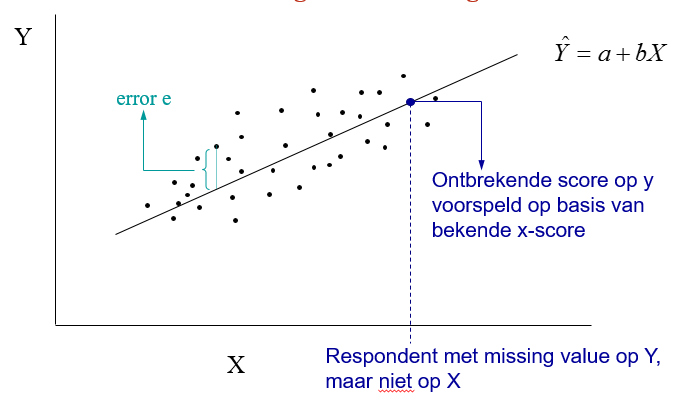
1. Constante waarde (bijvoorbeeld: gemiddelde, mediaan, etc.)
2. Hoogst correlerende item uit de meetschaal
3. Regressie-schatting
4. Regressie-schatting met variantie-correctie
5. Meervoudige substitutie

*Constante waarden*Missing value vervangen door vaste waarde per variabele (gemiddelde, mediaan, modus). Maar is dit wel realistisch? Want bij een hoger inkomen zijn er vaka meer missings bij de vraag naar inkomen. Een andere nadeel is dat door schatting van (co)variantie, correlaties vertekend raken.

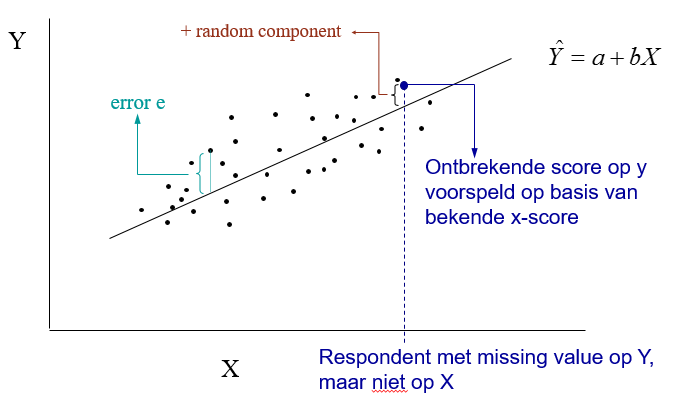
*Hoogst correlerende item uit de meetschaal*

Construeer een meetschaal, waarbij respondenten met missing value buiten beschouwing worden gelaten (d.m.v. listwise deletion). Indien de set van items een meetschaal vorm met een hoge betrouwbaarheid kun je de missing value op een item vervangen door een geldige score van de respondent op het hoogst correlerende andere item van die meetschaal.   
Bijvoorbeeld: onveiligheidsgevoel -> 6 uitspraken/ items met elk 5 antwoordmogelijkheden. Op 1 uitspraak ontbreekt het antwoord; vervang missing door valide score op de hoogst correlerende uitspraak.

*Regressieschatting*

Missing value op Y vervangen door geschatte waarde op basis van regressie-vergelijking: Y=a+bX+e. Bereken de regressie=coëffecient op basis van respondenten zonder missings op Y en X. Voordeel: maximaal gebruik van beschikbare informatie. Nadeel: je hebt goede predictoren nodig, varianties Y zijn te klein.

*Regressie-schatting met variantie correctie*

Regressie schatting inclusief random component. Je berekdn de regressiecoëfficient o.b.v. respondenten zonder missings op X en Y. Vervand missing op Y door geschatte waarde inclusief een random component.

Voordeel: variantie in y blijft vrijwel gelijk.  
Nadeel: toevalsfluctuatie -> goede predictoren nodig.

*Meervoudige substitutie*

Niet één maar 3,4 of meer substituties. Dit betekend voor elke substituie een aparte data analyse: vervang de missing value door gesubstitueerde waarde, analyseer de complete data. De gemiddelde uitkomst geeft de gewenste schatting. De spreiding in de uitkomsten geeft onzekerheid. Voordeel: alle beschikbare informatie wordt gebruikt. Nadelen: bij sommige substitutie-methoden wordt de variantie onderschat -> eerder significante resultaten. En de structuur in de gegevens wordt overschat -> resultaten worden mooier dan ze in werkelijkheid zijn.

**Kortom, tot slot over missing values**

Conventionele methoden zijn:

* Listwise deletion
* Pairwaise deletion
* Substitutie door gemiddelde
* Of foor score op hoogst correlerende schaal-item
* Opnemen aparte missing categorie
* Zoals bij methode van Cohen.

Geavanceerde methode is:

* Regressie, met en zonder variantie-correcte

State of the art:

* Meervoudige imputatie