

Stappenplan Kwantitatieve analyse

Harry B.G. Ganzeboom

Deze versie: November 28 2013

Het onderstaande is een zo kort mogelijke samenvatting van stappen in een standaard kwantitatief onderzoek. Meer gedetailleerde informatie vind je in de volgende onderliggende documenten (op www.harryganzeboom.nl/teaching):

- Elementaire inleiding SPSS met syntax [link](#)
- Stappenplan dimensionaliteits- en betrouwbaarheidsanalyse van attitudeschalen [link](#)
- Guide to reading spss regression tables [link](#)
- Causale analyse – een handreiking [link](#)
- Do's and don't's van wetenschappelijk schrijven: [link](#)

Stap 1: Het voorbereiden van je bestand

- Data-invoer in SPSS
 - Label je data in het Engels
 - Variabelen namen moeten terugverwijzen naar de vragenlijst.
 - Missing values gedeclareerd en expliciet labelen.
 - Maak een datadocumentatie met een verwijzing naar de data.
 - Bewaar / archiveer je data. Doe hierin een verslag van het veldwerk: aard van de steekproef, responsproces.
 - Scheid identificerende persoonsgegevens (namen, adressen, telefoonnummers, emailadressen) van je kwantitatieve data en gooi ze naar verloop van tijd weg.
 - Codeer alle (andere) alphanumerieke gegevens (beroepen, anders dan) tot numerieke variabelen.
- Databewerking
 - Werk in syntax. Bewaar alleen je syntax, geen bewerkte variabelen of tussenbestanden.
 - Alle syntax moet beginnen met het oproepen van het databestand.
 - Commentaar in je syntax helpt, maar is niet zaligmakend: eerder moeten bewerkingen zelf goed leesbaar zijn.
 - Bekijk eerst alle frequenties om te zien of je iets raars of ongedocumenteerd ziet.
 - Bewerk nooit de oorspronkelijke gegevens – maak copieën van de variabelen en geef deze informatieve (voor zichzelf sprekende) namen.
- Variabelenconstructie

- Bedenk eerst of het meetniveau van een variabele nominaal dan wel ordinaal (of hoger) is. In de praktijk gaan we met ordinale gegevens om alsof ze metrisch zijn: bv. berekenen gemiddelde en standaarddeviaties.
- Nominale variabelen kun je (altijd) dichotomiseren. Dichotomieën kun je ook als metrische variabele beschouwen.

Stap 2: Index constructie

- Indexconstructie (van meerdere indicatoren naar een index)
 - Check meetniveaus en missing values. Maak een descriptives tabel.
 - Dimensionaliteitsanalyses:
 - Het meeste leer je van het kijken naar een correlatiematrix. Kijk naar:
 - Hoeveelheid cases per variabele: waar zitten de missing values?
 - Richting van de variabelen: zijn ze allemaal gelijk gepoold?
 - Zijn de correlaties homogeen (positief) of treden er groepjes van correlaties op: dat duidt erop dat je meerdere dingen aan het meten bent.
 - Een formele techniek om dimensionaliteit te onderzoeken is factoranalyse (in de praktijk meestal: principale componenten-analyse).
 - De knik-plot van eigenwaarden geeft meestal het best te hoeveelheid aangetroffen dimensies aan: knik-1.
 - Een helder beeld van de factorladingen (= hoe goed meten de indicatoren de onderliggende / bedoelde dimensie?) krijg je bij /ROTATE=OBLIMIN. Bekijk in dit geval de pattern matrix and de factor correlations (niet de structure matrix).
 - Als je besloten hebt dat er meerdere dimensies onder je indicatoren schuilgaan, is het verstandig de factoranalyse per dimensie opnieuw te doen.
 - Betrouwbaarheidsanalyse doe je *na* factoranalyse en alleen op de indicatoren die je wilt gebruiken om een onderliggende dimensie te meten. Betrouwbaarheidsanalyse is dus niet om dimensionaliteit te bepalen, maar om de kwaliteit van de resulterende index te berekenen / te optimaliseren.
 - Indexconstruct gaat met handigst via COMPUTE INDEX=mean(indicatoren). Gebruik vooral geen factorscores als je missing values hebt.

Stap 3: Causale analyse

- Uiteindelijk dienen je statistische analyse uit te monden in een analyse van oorzaak-gevolg relatie. Wat een oorzaak en wat een gevolg is, kun je niet leren uit statistische analyse, dat moet volgen uit je onderzoeksontwerp. Doorgaans doen we daarbij een beroep op tijdsvolgorde, maar ook andere argumenten worden gebruikt. Denk hierbij om:
 - Levensloop, bv. geslacht < opleiding < beroep.

- Retrospectieve vragen, bv. eerste beroep < huidig beroep.
- Stabiel < veranderlijk, bv. inkomen < attitude.
- Algemeen < bijzonder, bv. politieke ideologie < stemgedrag.
- Gegeven een causale volgorde tussen variabelen, levert je causale modellering uiteindelijk een inzicht in de sterkte (evt: afwezigheid) van effecten, alsook sterkte van directe en indirecte effecten (mediatie), dan wel hoe effecten op elkaar inwerken (moderatie).
- In experimentele data heb je het gemakkelijk: de experimentele opzet leidt tot een eenduidige causale volgorde en confounding is per design uitgesloten.
- De meest algemeen gebruikte statistische techniek van causale modellering is OLS multiple regressie-analyse, equivalent aan variantie-analyse (ANOVA). Je bestudeert dan de lineaire relatie tussen een Y-variabele en meerdere X-variabelen. In SPSS gebruik je hiervoor REGRESSION of UNIANOVA (of GLM).
Voordelen van REGR boven ANOVA:
 - REGR kun je ook doen met *pairwise deletion of missing values*.
 - REGR kun je gemakkelijker stapsgewijs doen.
 - REGR biedt ook gestandaardiseerde coëfficiënten.

Voordelen van UNIANOVA (en GLM)

- UNIANOVA maakt automatisch dummy-variabelen en interacties met dummy-variabelen aan.
- UNIANOVA maakt plaatjes van interactie-effecten en groepsverschillen *adjusted for covariates*.
- Er vier soorten causale vraagstellingen die je met deze modellen kunt beantwoorden:
 - Gootsteenmodel: welke X-vars zijn allemaal van invloed op Y?
 - Confounding: is er effect van X op Y als ik confounders Z constant houd?
 - Mediation: in hoeverre is er effect van X op Y als ik mediators M constant houdt?
 - Moderation: in hoeverre hangt de sterkte van X op Y af van de condities Z?

Veruit de belangrijkste soort vraagstelling is die over confounding.

- Een regressiecoëfficiënt geeft weer hoeveel Y het resultaat is van 1 eenheid X, gegeven de andere (X) variabelen in het model. De interpretatie is partieel: je houdt de andere variabelen constant.
- Het kan leerzaam zijn om je multiple regressie-analyse te laten voorafgaan door een analyse van conditionele gemiddelden,
 - Bij meerwaardige X-variabelen krijg je zo een indruk van niet-lineariteit.
 - Bij tweewaardige X-variabelen is het aardig om het verschil in effect voor en na het constanthouden van confounders / mediators / moderatoren in een grafiek of tabel te laten zien.

- Als je een dichotome afhankelijke variabele hebt, kun je in de meeste gevallen toch met gewone OLS regressie uit de voeten. Formeel is logistische regressie vereist, maar dit voegt in de meeste gevallen weinig toe.
- Het gootsteenmodel is het in een stap uitvoeren van een regressie-analyse.
- Confounding breng je het helderst in beeld door je regressie in twee stappen uit te voeren: (A) eerst het ongecorrigeerde effect van X, (B) daarna wat daarvan wordt als je de confounders constant houdt.
- Mediation breng je als volgt in beeld:
 - Eerst de invloed van confounders constant houden om het totale effect van X op Y te berekenen.
 - Daarna mediators M een voor een toevoegen. Hoeveel van het effect van X blijft direct?
 - Daarna laten zien hoe de mediators M voorspeld worden uit de X-variabelen.
 - Indirecte effect kun je berekenen als het product van de betrokken directe effecten. Ook: indirect effect = totaal effect – overblijvend direct effect.
- Moderation (interactie) kun je op twee manieren in beeld brengen:
 - Splits je data in twee groepen en vergelijk de regressiemodellen.
 - Bereken een multiplicatieve interactieterm. Deze techniek is ook handzaam wanneer je interactievariabele meer dan twee groepen omvat of continu is. Interactiemodellen zijn gemakkelijker te interpreteren als je 0 als geldige waarde hebt: dat kan zijn als gemiddelde of als beginpunt.
- Belangrijke valkuilen en manieren om ze te vermijden:
 - Missing values: je moet altijd weten hoeveel cases je aan het analyseren bent. Vergelijk bij regressie-analyse steeds pairwise and listwise deletion.
 - Je houdt mediators constant terwijl je confounders constant wilt houden.
 - Bij interactiemodellen gebruik je een variabele zonder zinvol nulpunt (bv. leeftijd of – nog erger – jaartal. Je kunt dit gemakkelijk vermijden door lineaire transformatie (bv. trek 21 af van leeftijd).