KM 6(1982)
pag. 148-152

Formules voor semipartiële correlaties in de sociaal-wetenschappelijke literatuur
(of: wat er mis kan gaan als je je niet om afleidingen bekommert).

G.H. Maassen *)

Samenvatting:
Aanleiding tot het schrijven van deze notitie is de schemer die hangt rond het begrip semipartiële correlatie, door sommige auteurs ook wel part correlation genoemd. In verschillende sociaal-wetenschappelijke boeken kan men dit begrip vinden, samen met een formule voor zijn eenvoudigste vorm (de eerste orde), zie bijv. McNemar (1969) of Nunnally (1978). Bij Nie e.a. (1975) vindt men tevens algemenere formules die echter dwingen tot het eerst uitrekenen van multipele correlaties. De schrijver van deze notitie is één enkel boek bekend, dat een formule geeft voor een hogere orde semipartiële correlatie uitgedrukt in eenvoudiger grootheden. Deze formule ziet er overtuigend uit, maar is fout. Merkwaardig is dat deze fout zelfs na herhaaldelijk toepassen van de formule in het boek niet wordt ontdekt.

Laten we op vertrouwd terrein beginnen. Ruime bekendheid bij onderzoekers geniet het begrip partiële correlatie. In zijn eenvoudigste vorm (eerste orde) is het de correlatie tussen twee variabelen 2 en 3, met uitschakeling van de invloed van een derde, zeg 1. Uitschakeling van de invloed van 1 op de scores van 2 en 3 kan men zich als volgt voorstellen:

$$x_{2.1} = x_2 - r_{12}\frac{S_2}{S_1}x_1 \quad \text{en} \quad x_{3.1} = x_3 - r_{13}\frac{S_3}{S_1}x_1$$

(We maken in deze notitie gebruik van deviatiescores, dus met gemiddelde gelijk aan 0.) De partiële correlatie, meestal aangeduid met $r_{23.1}$, is nu gelijk aan de correlatie tussen de $x_{2.1}$-scores en de $x_{3.1}$-scores (dus $r_{23.1} = r_{(2.1)(3.1)}$). Uitwerken leidt tot de formule:

$$r_{23.1} = \frac{r_{23} - r_{12}\,r_{13}}{\sqrt{1-r_{12}^2}\sqrt{1-r_{13}^2}} \qquad (1)$$

In de onderzoeksliteratuur vindt men talloze toepassingen van deze formule. Vanzelfsprekend minder vaak treft men de partiële correlatie van de tweede orde

*) Instituut voor pedagogische en andragogische wetenschappen
Heidelberglaan 1, Utrecht

aan, dat is de correlatie tussen de variabelen 3 en 4, met uitschakeling van de invloed van de variabelen 1 en 2. Als men zich de uitschakeling van 1 en 2 uit 3 en 4 als volgt voorstelt:

$$x_{3\cdot12} = x_{3\cdot1} - r_{(2\cdot1)(3\cdot1)} \frac{S_{3\cdot1}}{S_{2\cdot1}} x_{2\cdot1} \text{ en } x_{4\cdot12} = x_{4\cdot1} - r_{(4\cdot1)(2\cdot1)} \frac{S_{4\cdot1}}{S_{2\cdot1}} x_{2\cdot1}$$

dan is de partiële correlatie van de tweede orde $r_{34\cdot12}$ gelijk aan de correlatie tussen de $x_{3\cdot12}$-scores en de $x_{4\cdot12}$-scores. Uitwerken hiervan levert:

$$r_{34\cdot12} = \frac{r_{34\cdot1} - r_{24\cdot1}\, r_{23\cdot1}}{\sqrt{1-r^2_{23\cdot1}}\ \sqrt{1-r^2_{24\cdot1}}} \qquad (2)$$

Tot zover de partiële correlaties.

Semipartiële correlaties wijken in die zin van partiële correlaties af, dat de invloed van een of meer variabelen wordt uitgeschakeld bij slechts één van de variabelen waarover de correlatie wordt berekend. Dus de semipartiële van een variabele y met een variabele 2, uit welke laatste de invloed van 1 is uitgeschakeld, is - om in bovenstaande terminologie te spreken - de correlatie tussen y-scores en $x_{2\cdot1}$-scores. Deze correlatie wordt meestal met $r_{y(2\cdot1)}$ aangeduid en de formule hiervoor is:

$$r_{y(2\cdot1)} = \frac{r_{y2} - r_{y1}\, r_{12}}{\sqrt{1-r^2_{12}}} \qquad (3)$$

Gezien de aard van deze correlatie zal men niet gemakkelijk een voorbeeld vinden van een situatie waarin de onderzoeker in deze correlatie is geïnteresseerd. Maar in de theorie achter de multipele regressie rekening spelen de semipartiële correlaties een belangrijke rol.

Interpreteren we een gekwadrateerde multipele correlatie coëfficient als de proportie variantie in de afhankelijke variabele (noemen we hier y) die maximaal door een lineaire combinatie van onafhankelijke variabelen kan worden verklaard, dan kunnen we deze proportie opgebouwd denken uit proporties variantie die achtereenvolgens worden toegevoegd met elke onafhankelijke variabele die in de regressieanalyse wordt opgenomen (en waaruit dus de invloed van de variabelen die al eerder in analyse zijn opgenomen is uitgeschakeld). In formulevorm (voor de vier variabelen situatie):

$$R^2_{y\cdot123} = r^2_{y1} + r^2_{y(2\cdot1)} + r^2_{y(3\cdot12)} \qquad (4)$$

De laatste term is een semipartiële correlatie van de tweede orde. Het enige sociaal-wetenschappelijke boek waarin schrijver van dit artikeltje een formule

hiervoor heeft aangetroffen is Kerlinger & Pedhazur's 'Multiple Regression in Behavioral Research' (1973, pg. 96):

$$r_{y(3\cdot12)} = \frac{r_{y(3\cdot1)} - r_{y(2\cdot1)}\, r_{3(2\cdot1)}}{\sqrt{1 - r^2_{3(2\cdot1)}}}$$

We nummeren deze formule niet, want de formule is fout (hoe overtuigend haar uiterlijk ook is, gezien de analogieën die er tussen partiële en semipartiële correlaties bestaan).

We vermoeden dat de fout is ontstaan doordat de schrijvers zoals zoveel auteurs van statistiekboeken voor sociale wetenschappers zich niet bekommeren om afleidingen.

Tot nu toe hebben we de formules ook zonder afleidingen gegeven, omdat deze gemakkelijk in bestaande literatuur zijn terug te vinden. We zullen de juiste formule voor de semipartiële correlatie presenteren, maar nu wel voorafgegaan door een afleiding. Gebruikmakend van de hierboven ingevoerde terminologie kunnen we stellen dat $r_{y(3\cdot12)}$ gelijk is aan de correlatie tussen y en $x_{3.12}$, dus:

$$r_{y(3\cdot12)} = \frac{\Sigma y\left(x_{3.1} - r_{23.1}\,\frac{S_{3.1}}{S_{2.1}}\,x_{2.1}\right)}{N\, S_y\, S_{3\cdot12}}$$

$S_{3.12}$ is de standaardschattingsfout bij de regressie van $x_{3.1}$ op $x_{2.1}$ en volgens de bekende formule $S_{y.x} = S_y\sqrt{1-r^2_{xy}}$ geldt dus:

$$S_{3.12} = S_{3.1}\sqrt{1-r^2_{23.1}}$$

$$r_{y(3.12)} = \frac{\Sigma yx_{3.1} - r_{23.1}\,\frac{S_{3.1}}{S_{2.1}}\,\Sigma yx_{2.1}}{NS_y\, S_{3.1}\sqrt{1-r^2_{23.1}}}$$

Vervangen we de voorkomende covarianties door respectievelijk $S_yS_{3.1}r_{y(3\cdot1)}$ en $S_yS_{2.1}r_{y(2\cdot1)}$, dan ontstaat na vereenvoudiging uiteindelijk:

$$r_{y(3\cdot12)} = \frac{r_{y(3.1)} - r_{y(2.1)}\, r_{23.1}}{\sqrt{1-r^2_{23.1}}} \qquad (5)$$

Is er sprake van een drukfout bij Kerlinger & Pedhazur ($r_{3(2\cdot1)}$ i.p.v. $r_{32\cdot1}$)? Allerminst. Het unieke geval doet zich voor dat zij hun foutieve formule enkele malen toepassen (bijv. pg. 96 en pg. 288), de resultaten vergelijken met de uitkomsten verkregen via $r^2_{y(3.12)} = R^2_{y.123} - R^2_{y.21}$ (zie formule (4)) en de inderdaad kleine verschillen toeschrijven aan afrondingsfouten. Dat zij de fout niet

ontdekken wordt veroorzaakt door de toevallige situaties waarin zij hun berekeningen uitvoerden. De schrijver van dit artikeltje is al enkele malen gestuit op situaties waarin Kerlinger's formule leidt tot zeer grote afwijkingen.
Nog een pikant detail. Kerlinger en Pedhazur zetten hun lezers (die semipartiële correlaties willen berekenen zonder de grote computers te willen inschakelen voor het bepalen van multipele correlaties) via een verwijzing op het spoor van een algoritme dat geschikt is voor berekening met de hand: Nunnally (1978), pg. 182. Ook Nunnally geeft geen formule, maar narekenen leert dat dit algoritme gebaseerd is op:

$$r_{y(3\cdot12)} = \frac{r_{y3} - r_{y1}\, r_{31} - r_{3(2\cdot1)}\, r_{y(2\cdot1)}}{\sqrt{1-r^2_{13} - r^2_{3(2\cdot1)}}} \qquad (6)$$

De lezer verifieert gemakkelijk dat (6) overgaat in (5) door teller en noemer door $\sqrt{1-r^2_{13}}$ te delen. Ook aan de achtergronden van dit algoritme hebben Kerlinger en Pedhazur kennelijk geen aandacht besteed.

Hierboven zijn de formules gegeven, uitgaande van bekende begrippen uit de regressieanalyse: de nulde-orde correlatiecoëfficient en de standaardschattingsfout. Zuiver formeel bezien kan men ook formule (3) als uitgangspunt nemen[1].
De semipartiële correlatie van de tweede orde die centraal staat in dit verhaal volgt direct door substitutie van $x_{3\cdot1}$ in de plaats van x_2 en $x_{2\cdot1}$ in de plaats van x_1 in formule (3). We moeten daarbij bedenken dat $x_{3.12} = x_{(3.1).(2.1)}$ (zie ook de definitie van $x_{3.12}$ voorafgaande aan formule (2)).
De semipartiële correlatie van de derde orde ontstaat door substitutie van $x_{4\cdot12}$ in de plaats van x_2 en $x_{3.12}$ in de plaats van x_1.
Omdat $x_{4.123} = x_{(4.12).(3.12)}$ volgt dan:

$$r_{y(4.123)} = \frac{r_{y(4.12)} - r_{y(3.12)}\, r_{34.12}}{\sqrt{1-r^2_{34.12}}} \qquad (7)$$

Maar ook de partiële correlaties volgen uit (3). Substitutie van $x_{3\cdot1}$ in de plaats van y en de eigenschap $r_{23.1} = r_{(3.1)(2.1)}$ hebben bijvoorbeeld tot gevolg:

$$r_{23.1} = \frac{r_{2(3.1)} - r_{1(3.1)}\, r_{12}}{\sqrt{1-r^2_{12}}},$$

waarin $r_{1(3.1)} = 0$. Dus

1 Albert Verbeek maakte me hierop attent.

$$r_{23.1} = \frac{r_{2(3.1)}}{\sqrt{1-r_{12}^2}} = \frac{r_{23} - r_{12}r_{13}}{\sqrt{1-r_{12}^2}\sqrt{1-r_{13}^2}} .$$

Tenslotte nog een opmerking over het verband tussen partiële en semipartiële correlaties.
Voert men bijv. met SPSS een multipele regressie analyse uit volgens de FORWARD (= default)-methode dan wordt bij elke stap die variabele opgenomen die de hoogste partiële correlatie heeft met de afhankelijke variabele (onder uitpartialiseren van de variabelen die al in de analyse zijn opgenomen). Het waarom is niet helemaal duidelijk, want op grond van (4) zou men de semipartiële correlatie als criterium verwachten. En in de lijstjes op de computeroutput onder de kop 'variables not in the equation' zou men liever semipartiële dan partiële correlaties zien. Overigens wordt het resultaat hierdoor niet beïnvloed. Dat blijkt als men bedenkt dat de volgende relaties tussen semipartiële en partiële correlaties gelden;
voor de eerste orde:

$$r_{y(2.1)} = \sqrt{1-r_{y1}^2}\, r_{y2.1}$$

en voor de tweede orde:

$$r_{y(3.12)} = \sqrt{(1-r_{y1}^2)(1-r_{y2.1}^2)}\, r_{y3.12}$$

Zo blijkt dat voor alle 'variables not in the equation' geldt, dat partiële en semipartiële correlatie een constante factor verschillen. Bij elke stap kan men dus keuze van een variabele even goed op partiële als op semipartiële correlaties baseren.

Literatuur:

Kerlinger, F.N. and Pedhazur, E.J., Multiple Regression in Behavioral Research, Holt Rinehart and Winston (1973).

McNemar, Q., Psychological Statistics. Wiley (1969), pg. 182-186.

Nie, N.H. e.a., Statistical Package for the Social Sciences, McGraw-Hill (1975), pg. 332-345.

Nunnally, J.C., Psychometric theory, McGraw-Hill (1978), pg. 168-185.