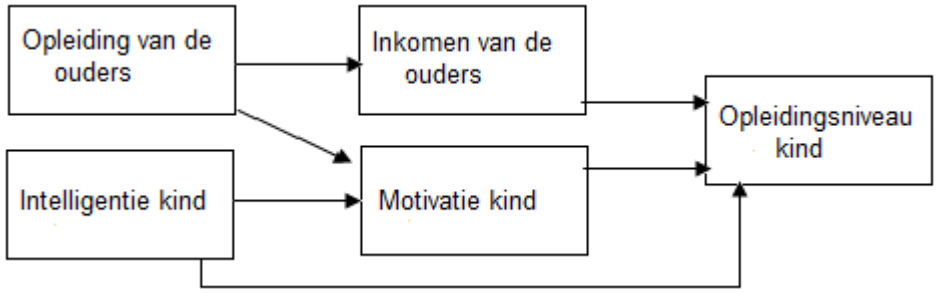
# 16. Padanalyse

Bii padanalyse gebruik je regressiemodellen om theorieën weer te geven over causale verbanden tussen variabelen. Bij regressie kun je echter maar een afhankelijke variabele invoeren, terwijl er in de werkelijkheid vaak meer afhankelijke variabelen in een model zitten. Neem bijvoorbeeld het volgende. Onze theorie stelt dat iemands hoogste opleidingsniveau afhankelijk is van een aantal factoren, te weten zijn/haar intelligentie (IQ), zijn motivatie, en het inkomen van de ouders. Maar deze verklarende variabelen zijn op zichzelf ook weer afhankelijke variabelen. Want de motivatie van een persoon is afhankelijk van die persoon's intelligentie en het opleidingsniveau van de ouders. Het inkomen van de ouders, (dat was een van de verklarende variabelen van het hoogste opleidingsniveau van het kind), is afhankelijk van het opleidingsniveau van de ouders. Omdat dit er zo vrij ingewikkeld uitziet, kunnen we het weergeven in een paddiagram:




In dit paddiagram woorden causale verbanden weergegeven aan de hand van de pijlen. Het is nu makkelijker te overzien wat er allemaal gebeurt.

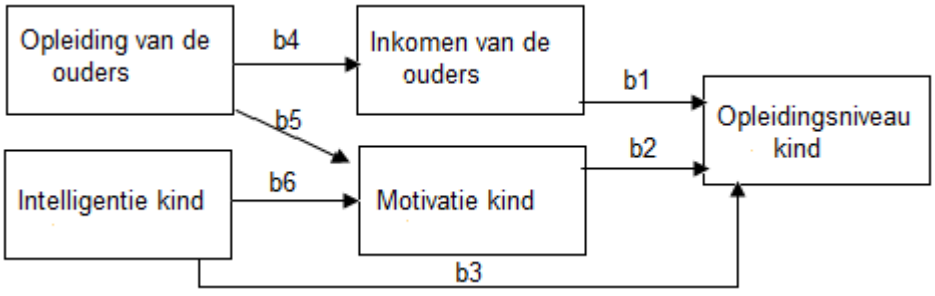
**Padcoëfficiënten**
Het paddiagram is een soort grafische weergave van je regressiemodel. Als we hier een regressiemodel van zouden willen maken, zouden we het opleidingsniveau van het kind als afhankelijke variabele nemen, en het inkomen van de ouders, de motivatie van het kind en de intelligentie van het kind als verklarende variabelen. Vervolgens zouden we nieuwe regressiemodellen moeten maken waarin de verklarende variabelen als afhankelijke variabelen worden gezien. Eerst begin je met datgene wat we willen voorspellen, namelijk het opleidingsniveau van het kind:

E(y = opleidingsniveau kind) = a + b1(inkomen ouders) + b2(motivatie kind) + b3(intelligentie kind)

Nu kunnen we nieuwe formules maken die deze variabelen weer kunnen schatten:

E(y = inkomen van de ouders) = a + b4(opleiding ouders)
E(y = motivatie kind) = a + b5(opleiding ouders) + b6(intelligentie kind)

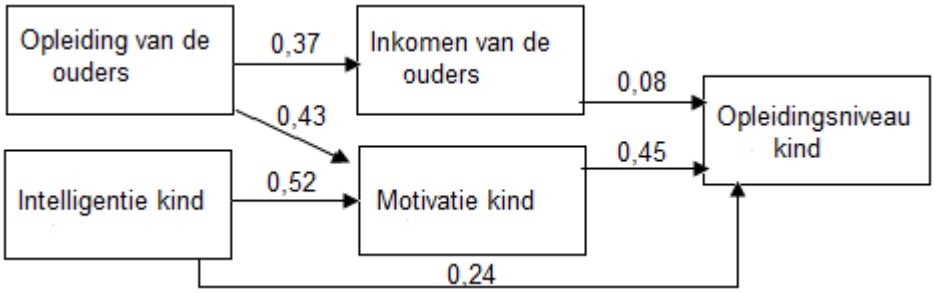
Als we deze regressiemodellen zouden laten runnen in SPSS dan krijg je de coëfficiënten. Deze zouden we dan kunnen invullen in het paddiagram, zodat het duidelijker wordt wat het effect is van de variabelen op elkaar. Eerst is het makkelijker om aan te geven welke b-coëfficiënt bij welke pijl hoort:




**Directe en indirecte effecten**
We kunnen nu makkelijker directe en indirecte effecten waarnemen. Een indirect effect is wanneer een variabele effect heeft op de afhankelijke variabele, via een tussenliggende variabele. In dit

model zijn er drie indirecte relaties vast te stellen, namelijk 1) de opleiding van ouders, via het inkomen van de ouders op het opleidingsniveau van het kind, 2) de opleiding van de ouders, via de motivatie van het kind op het opleidingsniveau van het kind en 3) de intelligentie van het kind, via de motivatie van het kind op het opleidingsniveau van het kind. Een direct effect is wanneer er dus geen variabele tussenligt. Er zijn hier ook drie directe effecten, namelijk 1) het inkomen van de ouders op de opleiding van het kind, 2) de motivatie van het kind op de opleiding van het kind en 3) de intelligentie van het kind op de opleiding van het kind.

*Sterkte van directe en indirecte verbanden*
Stel dat we in de regressieanalyse de coëfficiënten hebben opgevraagd van deze verbanden. Je kunt ze dan invullen in het schema, dat is het makkelijkst.

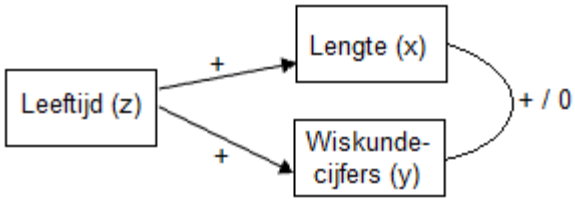



We zien nu dat voor elke punt dat inkomen omhoog gaat, gaat het opleidingsniveau van het kind met 0,08 omhoog. Voor elke punt dat de motivatie van het kind omhoog gaat, gaat het opleidingsniveau van het kind met 0,45 omhoog. Etc.

Voor het berekenen van het indirecte effect, bijvoorbeeld het indirecte effect van de intelligentie van het kind via de motivatie van het kind, moeten we de coëfficiënten vermenigvuldigen. In dit geval is dat b6 * b2 = 0,52 * 0,45 = 0,234. Dat is het *indirecte effect* van intelligentie. Het *directe effect* van intelligentie is 0,24. Het *totale effect* van de intelligentie van het kind is dan 0,234 + 0,24 = 0,474.
Wanneer we nu relatief gaan kijken naar dit model, dan kunnen we zien dat 0,234/0,474 * 100 = 49,4% van het effect van het intelligentie van het kind een indirect effect is. Dat is natuurlijk een heel groot indirect effect.

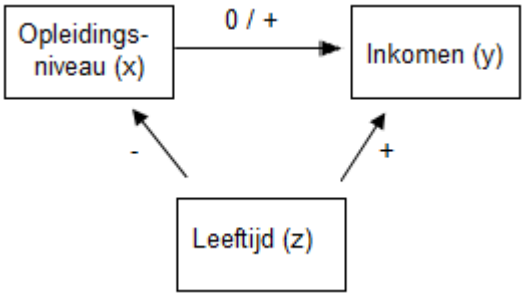
**Schijnverband, onderdrukt verband en Simpsons paradox**
In de statistiek kom je vaak nog andere verbanden tegen. We bespreken er hier drie.

*Schijnverband*
Er is sprake van een schijnverband wanneer je een verband vaststelt tussen x en y, maar dat verband verdwijnt wanneer je controleert voor z. Deze variabele z heeft dan invloed op zowel x als y, waardoor het lijkt alsof er een verband bestaat tussen x en y. We noemen de variabele ‘z’ dan een intermediaire variabele.
Een voorbeeld is wanneer je in je data vaststelt dat er een verband is tussen iemands lengte (x) en zijn wiskundecijfers (y). Het is een positief verband, hoe groter iemand is, hoe hoger de wiskundecijfers. Dit lijkt een verband te zijn, totdat je controleert voor leeftijd. Dan bestaat het verband niet meer. Dat komt omdat leeftijd een positief verband heeft met lengte, en ook een positief verband met wiskundecijfers. Hoe ouder je wordt, hoe langer je wordt, en hoe beter je wordt in wiskunde. Er dan feitelijk geen verband tussen lengte en wiskundecijfers. Een schijnverband ziet er grafisch zo uit:




*Onderdrukt verband*

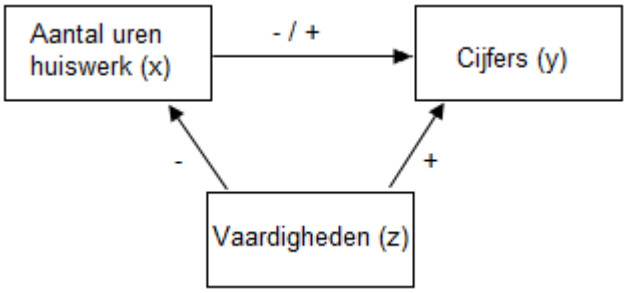
Er is sprake van een onderdrukt verband, wanneer er eerst geen verband lijkt te zijn, maar na het invoeren van controlevariabelen wel een verband. Bijvoorbeeld wanneer je kijkt naar opleidingsniveau en inkomen, met een controlevariabele voor leeftijd. Stel dat we deze controlevariabele er eerst uitlaten, dan zul je zien dat er zowel hoog als laagopgeleiden zijn, die zowel een hoog als een laag inkomen hebben. Er lijkt dan helemaal geen verband te zijn tussen opleiding en inkomen. Totdat je controleert voor leeftijd. Want dan blijkt dat leeftijd positief gecorreleerd is met inkomen (hoe ouder, hoe hoger je inkomen), en negatief geassocieerd met opleiding (hoe ouder, hoe minder opleiding). Na deze controle zul je zien dat er wel een positief verband is tussen opleiding en inkomen (hoe hoger opgeleid, hoe hoger inkomen). Grafisch ziet dat er zo uit:




Dus zonder controlevariabele wordt er niet gecontroleerd voor het feit dat oudere mensen meer verdienen, en lager opgeleid zijn. Je ziet dan alleen dat er laagopgeleiden zijn, die veel verdienen. En je ziet ook dat er hoogopgeleiden zijn, die maar weinig verdienen. Dat dit komt door hun leeftijd dat zie je dan niet, totdat je er voor controleert.

*Simpsons Paradox*

Simpsons Paradox is een specifiek geval van een onderdrukt verband, waarbij de richting van het verband omkeert na het invoeren van een controlevariabele. Dus een negatief verband verandert in een positief verband, of andersom. Een voorbeeld hiervan is het aantal uren besteed aan huiswerk en de behaalde cijfers. De controlevariabele is hier de vaardigheden van het kind. Je kan in je data vinden dat kinderen die weinig uren besteden aan hun huiswerk, hoge cijfers halen, en dat kinderen die veel tijd besteden aan hun huiswerk juist lagere cijfers halen. Je vindt dan een negatief verband tussen het aantal uren besteed aan huiswerk en de behaalde cijfers. Dat lijkt natuurlijk wel heel raar. Totdat je controleert voor de intelligentie of de vaardigheden van het kind. Want kinderen met meer vaardigheden zijn minder lang bezig met hun huiswerk, en halen toch dezelfde cijfers als kinderen die weinig vaardigheden hebben en wel lang bezig zijn met hun huiswerk. Wanneer je controleert voor deze vaardigheden, dan zul je wel een positief verband vinden tussen huiswerk en cijfers. Grafische weergave:




Dus eerst vind je een negatief verband tussen het aantal uren huiswerk en cijfers, maar wanneer je controleert voor de vaardigheden, bestaat er een positief verband.