5.1

Causaliteit is een groot probleem in de statistiek en de filosofie. Een reden hiervoor is dat er niet een exacte definitie is van het woord causaliteit. Er wordt meestal aangenomen dat causaliteit kan worden vastgesteld als aan deze drie criteria is voldaan: associatie, richting van invloed en isolatie. We zullen deze drie in meer detail bespreken:

Associatie

Het eerste wat we leren is dat correlatie niet betekent dat er ook causaliteit is. Als er een sprake is van causaliteit dan betekent dit wel dat er ook een correlatie is. Wanneer twee variabelen een causaal verband hebben, moet een verandering in de één, een verandering in de andere veroorzaken.

Richting van causaliteit

Wanneer twee variabelen (A en B) geassocieerd zijn met elkaar kunnen er drie reden zijn:

- Het kan dat A B veroorzaakt
- Het kan zijn dat B A veroorzaakt
- Een andere variabele, C is de oorzaak van A en B.

Dus wanneer er een correlatie is tussen twee variabelen, weten we nog niet de richting van de causaliteit. Maar hoe kunnen we zeggen of A, B veroorzaakt of dat B, A veroorzaakt? Het antwoord is dat we altijd ervan uit gaan dat de oorzaak eerst komt en daarna pas het effect. Wanneer A veroorzaakt B, zal een verandering in A een verandering in B veroorzaken na een bepaalde tijd. Er moet dus onderzocht worden welke variabele als eerste verandert. Na een verandering in de onafhankelijke variabele moet na enige tijd een verandering worden geobserveerd in de afhankelijke variabele. Dit tijdsverschil tussen de oorzaak en het effect kan erg verschillen. Dit idee van temporele prioriteit is ook aanwezig in het design van een experiment, omdat de manipulatie van de onafhankelijke variabele altijd moet gebeuren voordat de afhankelijke variabele wordt gemeten. Dit kan kan alleen gedaan worden in experimenteel of cross-sectioneel onderzoek. In andere onderzoeksdesigns worden vaak alleen metingen gedaan op één tijdstip.

Isolatie

Om zeker te weten dat een onafhankelijke variabele, A de oorzaak is van verandering in de afhankelijke variabele B, moeten we afhankelijke variabele B isoleren om te zorgen dat deze niet door andere dingen wordt beïnvloed dan variabele A. Meestal kan deze isolatie niet helemaal worden bereikt, maar wel een zogenaamde *pseudo-isolatie*.
Experimentele controle kan worden gebruikt om de onafhankelijke variabelen te isoleren, in niet experimentele onderzoeken kan dit niet. Dit moet worden gedaan op een andere manier. We weten al dat de regressie hellingen het effect van de onafhankelijke variabele op de afhankelijke variabele laten zien, terwijl ze het effect van de andere onafhankelijke variabelen gelijk houden. Dit kan dus bestudeerd worden met behulp van regressie modellen.

Om deze verschillende criteria bij elkaar te voegen, moeten we niet alleen naar de statistische analyse kijken, maar ook naar de theorie. Wanneer psychologen data verzamelen willen ze zeker weten dat dit nauwkeurig gebeurt. Theorie heeft een belangrijke functie in het succesvol toepassen van een regressie analyse.
Een serie van verschillende bevindingen, die allemaal correleren kunnen bewijs zijn voor een causale relatie. Deze serie van verschillende bevindingen word de *signature* van dat proces genoemd.

5.2
We zullen nu gaan kijken naar het effect van de steekproefgrootte op de regressie analyse. Het idee is vooral dat hoe groter de steekproefgrootte des te beter. De standaardfout van het gemiddelde is gelijk aan:

$$se(\bar{x}) = \sqrt{\frac{sd^2}{n}} \qquad (50)$$

Uit deze formule kan je opmaken dat wanneer de steekproefgrootte groter wordt, de noemer ook groter wordt, en dus zal de standaardfout kleiner worden. Dit betekent dat de geschatte waarden voor de parameter preciezer zullen worden. Daarnaast zal een kleinere standaardfout de kans op het vinden van een significante waarde vergroten. Maar er zijn ook problemen met steekproeven die te groot zijn. Want het zoeken van meer gegevens kost veel tijd. Daarnaast zijn ook ethische besturen meer bezig met het onderzoeken van de steekproef groottes. Zij zeggen dat deelnemers hun tijd opgeven in de hoop dat er iets goeds mee wordt gedaan.
Er zijn twee manieren om een goede steekproefgrootte te bepalen. De eerste zijn vuistregels. Dit zijn makkelijke regels. De tweede is de power analyse. We zullen beide methoden bespreken.

Vuistregels zijn meestal erg makkelijk. Green heeft een methode van vuistregels gemaakt om een minimale steekproefgrootte te bepalen. Hij zegt dat een minimale steekproefgrootte groter moet zijn dan 50 +8k, waarin k het aantal onafhankelijke variabelen is. Ook heeft hij gezegd wanneer je een significantie toets wil uitvoeren op de regressie hellingen, de steekproefgrootte groter moet zijn dan 104 + k. Het nadeel van deze regels is dat ze de verwachte effectgrootte of gewilde power van de toets niet meenemen. Dus deze regels missen generaliteit.

Om een power analyse te gebruiken hebben we de volgende informatie nodig:

- De waarde van alpha. Deze waarde is meestal 0.05 (5%). Wanneer alpha groter is wordt de kans dat we een significant effect vinden groter, maar tegelijkertijd wordt ook de kans op het vinden van een onecht resultaat groter. De kans op een type I fout is gelijk aan de waarde van alpha.
- De effectgrootte van de populatie waar we in geïnteresseerd zijn. De effectgrootte in een meervoudige regressie (multiple regression) is gelijk aan R^2. Hoe groter de effectgrootte, des te groter de kans om het te vinden. Maar wanneer de effectgrootte is heel klein, dan zal het vinden van het ook niet handig zijn. De effectgrootte kan bepaald worden op drie manieren:
    1. Effect gebaseerd op werkelijke kennis
    2. Baseer de schatting van de effectgrootte op voorgaand onderzoek

3. Gebruikt bepaalde regels on de verwachte effectgrootte te bepalen. Cohen heeft waarden van R^2 bepaald die de hoeveelheid van de effectgrootte aangeven. Wanneer R^2 =0.02 is er een klein effect, wanneer R^2=0.13 is er een gemiddeld effect en wanneer R^2 = 0.26 is er een groot effect.

- Een geschikt level van de power moet worden bepaald. De power is de kans op het vinden van een resultaat gegeven dat het effect bestaat in de populatie. Een regel is dat de power wordt gezet op 0.80 (80%). Dat betekent dat er een 80% kans is op het vinden van een significant resultaat wanneer er een effect is in de populatie. De kans op een type II fout is gelijk aan 1-power, dus 1-0.80=0.20 (20%).

Uit deze informatie kunnen we de benodigde aantal deelnemers bepalen. Programma's zoals G*Power kunnen worden gebruikt om dit te bereken. In G*Power worden grafieken gemaakt, welke verschillende aantallen van deelnemers laten zien. Uit deze grafieken kan je conclusies trekken, maar je kan niet een nauwgezette power berekening mee doen (zie de grafieken op pagina 122-125.)

5.3

We zullen nu gaan kijken naar collineariteit, ook wel multicollineariteit genoemd. Collineariteit refereert naar de grootte van de correlaties tussen de onafhankelijke variabelen in een regressie. Het komt voor omdat twee (of meer) onafhankelijke variabelen correleren. Dit betekent dat het moeilijk is voor de regressie berekening om te bepalen welke van de variabelen echt belangrijk is, eentje van de twee, of misschien beiden. Dus de standaardfouten zullen toenemen en ook de onnauwkeurigheid (helling coëfficiënten) zal toenemen. Dus we kunnen niet bepalen welke variabele belangrijk is voor het resultaat. De regressie berekeningen houden rekening met deze onzekerheid en hebben grotere standaardfouten. Dus meer formeel, wanneer decorrelatie tussen twee onafhankelijke variabelen gelijk is aan 1 (of dichtbij 1) of wanneer de multiple correlatie tussen elke onafhankelijke variabele 1 (of dichtbij 1) is, is er *perfecte of complete collineariteit*.

Een aanname behorende bij het regressie model is dat er geen perfecte collineariteit mag optreden, en wanneer dit wel zo is, zullen de meeste statistische computerprogramma's stoppen en een fout laten zien. Maar perfecte collineariteit komt maar weinig voor in echte data. Wanneer dit wel gebeurd, is het hoogstwaarschijnlijk dat er twee of meer onafhankelijke variabelen bij elkaar zijn gevoegd om zo een extra variabele te creëren. Meestal wanneer er collineariteit is, is deze hoog genoeg om problemen te veroorzaken, maar niet hoog genoeg om de aanname officieel te schenden.

Wanneer je een regressie analyse hebt waarbij de regressie coëfficiënten niet significant zijn, maar de totale regressie wel significant is, kan collineariteit hier een rol hebben gespeeld. Ook als uit de regressie analyse blijkt dat een groot deel van de variantie verklaard kan worden is belangrijk om naar de grootte van de geschatte parameters te kijken Er zijn verschillende manieren waarop onderzocht kan worden in hoeverre de collineariteit tot problemen leidt.

Een manier is om de correlatiematrix te bekijken. Hoge correlaties tussen onafhankelijke variabelen kunnen duiden op multicollineariteit maar een lage correlatie betekent niet automatisch dat er geen probleem is. De correlatie laat ons zien hoe veel variantie de twee variabelen delen. We zijn geïnteresseerd in het deel van de variantie van elke onafhankelijke variabele dat gedeeld wordt met alle andere onafhankelijke variabelen.
We kunnen dit vinden door een meervoudige regressie analyse uit te voeren. De meeste statistische software geeft ook andere waarden die hierbij gebruikt kunnen worden, zoals de *tolerance en de variance inflation factor (VIF).* Tolerance is een uitbreiding van de R^2. De tolerance van een onafhankelijke variabele is de hoeveelheid van de onafhankelijke variabele die niet kan worden voorspeld door de andere onafhankelijke variabele in de regressie analyse. De waarde van R (in meervoudige correlatie) is gelijk aan de waarde van *r* (bivariate correlatie), en de tolerance kan daarom via *r* worden berekend. De waarde van tolerance ligt tussen 0-1. Een tolerance van 0 voor een variabele betekent dat deze variabele helemaal voorspeld wordt door de andere onafhankelijke variabele. Dit is perfecte collineariteit. Wanneer een variabele een tolerance heeft van 1, betekent dit dat de variabele helemaal niet correleert met de andere onafhankelijke variabele.
De variance inflation factor (VIF) ligt dichtbij tolerance. Wanneer er meer dan twee onafhankelijke variabelen zijn berekenen we de VIF als volgt:

$$VIF = 1/tolerance \qquad (110)$$

Deze waarde laat zien hoeveel de standaardfout van de variabele is toegenomen door collineariteit. De toename in de standaardfout is gelijk aan de wortel van de VIF.

Maar wat moet je doen wanneer je collineariteit in je data vindt? Het beste is, wanneer collineariteit een groot probleem is, om de data niet te gebruiken en nieuwedata te verzamelen. Dit is veel werk, maar er zijn ook andere opties:

1. Meer gegevens verzamelen
   Het is niet een grote verbetering. Collineariteit zorgt voor een toename in de standaard fouten. Een grotere steekproef heeft kleinere standaardfouten, dit zal een beetje helpen tegen de effecten van collineariteit maar kan een collineariteit van 1 niet voldoende verbeteren.
2. Verwijder of combineer variabelen
   Wanneer variabelen een hoge correlatie hebben betekent dit dat ze dezelfde dingen meten, en dat de informatie van deze variabelen (deels) overbodig is. Wanneer je veel onafhankelijke variabelen hebt, kan je dit aantal verkleinen door *principal components analysis (PCA).* Dit lijkt op een factor analyse, beide technieken maken groepen van de originele variabelen. Dit zorgt voor minder gecorreleerde variabelen.
3. Stepwise entry
   Dit is een vorm van hiërarchische regressie. Je kan dit toepassen wanneer de collineariteit een probleem geeft in het selecteren van de variabelen voor de analyse.

4. Ridge regressie
   Wanneer de collineariteit zo groot is dat de regressie procedure niet kan worden uitgevoerd, is een mogelijke oplossing een ridge regressie. Dit is een complexe methode die moeilijk te interpreteren is en daarom zelden wordt gebruikt.