**Inleiding regressie-analyse***Door: Danielle van Helvoirt | Docent: Manfred te Grotenhuis | Vakcode: SOW-MTB2045*

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

**Cursusomschrijving:** In de cursus SOW-MTB2045 "Inleiding Regressie-Analyse" wordt een overzicht gegeven van gangbare toepassingen van de lineaire en non-lineaire regressie-analyse binnen de sociale wetenschappen. In de hoorcolleges komen de theoretische en statistische grondbeginselen van deze analysetechniek ter sprake. In de werkcolleges ligt de nadruk op de praktische toepassing aan de hand van casussen uit het veld van de sociologie. De nadruk van de cursus ligt niet op de wiskundige achtergrond maar op de praktische toepasbaarheid. Om die reden leren studenten regressie-modellen toe te passen op concrete onderzoeksgegevens met behulp van het statistische programma SPSS.

**Doelstellingen:**

1. De student heeft inzicht in de theorie van de lineaire en logistische regressie-analyse met één of meer predictoren;
2. De student heeft inzicht in de theorie van de padanalyse;
3. De student heeft inzicht in het modelleren van interactie tussen twee predictoren;
4. Inzicht in collineariteit, assumpties t.a.v. residuen, kwadratische modellen en transformaties van variabelen;
5. Vaardigheid in het toepassen van de geleerde theoretische kennis binnen concreet onderzoek;
6. Vaardigheid in het hanteren van gangbare computerprogrammatuur (SPSS) met betrekking tot alle hierboven genoemde thema's;
7. Studenten moeten in staat zijn om bij elke thema de juiste spss-syntax te gebruiken en de relevante uitkomsten te interpreteren.

# College 1: herhaling toetsende statistiek en bivariate regressie

|  |  |  |
| --- | --- | --- |
| **Welke toets gebruik je wanneer?** | | |
|  | **X-en** | **Y** |
| **Simpele regressie** | Interval (1 X!) | Interval |
| **Multipele regressie** | Interval (meerdere) | Interval |
|  | Nominaal (1 of meer) | Interval |
|  | Nominaal/interval | Interval |
| **Logistische regressie** | Nominaal/interval | Dichotoom |

Aselecte steekproef 🡪 groep mensen afkomstig uit de populatie

E()= gemiddelde in populatie

σ ()= afhankelijk van steekproef 🡪 σ / √n

90% BI is 1,65 standaardfouten af van het gemiddelde. 95% BI is 2 standaardfouten van gemiddelde af. Berekening (95%) is gem +/- (2 \* stnddev).

s ≈ *σ* 🡪 *SE (**) = s / √ n.* Dit betekent dat s (standaarddeviatie in steekproef) gebruikt kan worden om de standaardfout te berekeken. (Als je niet snapt; kijk toetsende statistiek terug).

Standaardaanpak

1. Bereken gemiddelde steekproef
2. Bekeren bijbehorend SE
3. Bereken p-waarde
4. Vergelijk p-waarde met alfa
5. Bereken betrouwbaarheidsinterval

## Lineaire bivariate regressieanalyse

Lineaire regressie onderzoekt

* De relatie tussen Y en een of meerdere X variabelen
* Alle variabelen moeten een meetniveau hebben van minimaal interval
* Formule: Y=a + b\*X + e
  + a = intercept, b = richtingscoëfficiënt

Galton

* Kleinste kwadraten methode
* Neem de verschillen tussen geobserveerde lijn en regressielijn. Kwadrateer de verschillen en tel ze op. Hieruit volgt het kleinst mogelijke getal
* Residual Sum of Squares (dit staat in SPSS)
* Verklaarde variantie is 1- RSS/SSY
  + SSY is de totale variantie in Y

Samenvatting toetsen met regressieanalyse

1. Stel onderzoekshypothese op
2. Nulhypothese is doorgaans dat b-coëfficiënt 0 is
3. Schat de regressielijn
4. Vergelijk p-waarde met a en beslis of er significantie is
5. Bereken betrouwbaarheidsinterval

Intercept nuttig maken: Voorbeeld leeftijd. Stel je data is vanaf 18. Doe dan compute leeftijdnieuw = leeftijd – 18. Nu is het intercept dus eigenlijk de 18-jarigen.

# College 2: Multivariate lineaire regressieanalyse

Twee manieren

1. Stepwise
   * Voor exploratief onderzoek
2. Enter
   * Als je gerichte hypothesen wilt toetsen

## Stepwise

Wat doet stepwise?

1. Selecteert de x-variabele die de sterkste Pearson correlatie heeft met Y
2. Selecteert vervolgens de tweede x-variabele die onder controle van de eerste variabele de grootste bijdrage levert aan de verklaarde variantie. Controle of de eerste variabele nog significant is
3. Herhaling vorige stap met nieuwe variabele
4. Programma stopt als er geen variabelen meer zijn die nog iets kunnen toevoegen aan de verklaarde variantie

### Pin & pout

* PIN bepaalt wanneer een variabele wordt opgenomen. Standaard met een p-waarde kleiner of gelijk aan 0,05 DUBBELZIJDIG. (Beter om in 0,10 en 0,20 te zetten)
* POUT bepaalt wanneer een variabele wordt verwijderd. Standaard met een p-waarde van groter of gelijk aan 0,10 DUBBELZIJDIG.
* Voor enkelzijdig toetsen met a=5% moet PIN op 0,10 en POUT op 0,20.
* Beter om te kijken naaar adjusted R-square dan de gewone. Deze controleert voor het aantal predictoren.

Beta is de unieke verklaring van een variabele, dus zonder de gemeenschappelijke verklaring.

Als stepwise klaar is, ga je met enter verder. Je maakt een test voor wat er gebeurt als je toch de vervallen variabelen meeneemt. Je kijkt naar de opname van de variabelen tegelijk. Focus op de significantie en verandering in R-square. Als ze echt niet meemogen laat je ze weg en ga je toch een enter uitvoeren. Je verwijderd de variabelen als de f-toets significant groter is dan 0,05.

Stel: tegen je verwachting in is een variabele niet geselecteerd?

1. Gerelateerde predictoren zijn wél geselecteerd (kerkgang/religiositeit)
2. Door het bestaan van een causale keten (dus alleen indirect effect)
3. Doordat de variabele er gewoon niet toe doet

Deze drie zijn iet problematisch, deze heironder wel

1. Door suppressie
2. Door interactie

# College 3: Het invoeren van nominale en ordinale variabelen

Hoe in te voeren in een regressieanalyse?

* Maak er een ratio variabele van aan de hand van dummies.
  + Ratio want nulpunt, gelijke afstanden, eenduidige rangorde
* Kies een referentiecategorie
  + Deze wordt niet opgenomen in de analyse
* Er wordt nu dus ten opzichte van de referentiecategorie geanalyseerd. De intercept is dus ook de referentiecategorie
* Toetsen op significantie van de gehele oude variabele gaat via test (zoals bij stepwise). Je voegt dan de dummies in (zonder de referentiecategorie)
* Maak een SHEAF coëfficiënt aan. Neem de b-coëfficienten en vermenigvuldig deze met de dummies. Tel dit op.
* Voeg de sheaf toe aan de regressieanalyse en kijk naar de volgorde van belangrijkheid. Dit is te zien aan de hand van de Beta

# College 4: Interactie

* Maak een interactievariabele aan met compute
* Voeg hem toe in regressieanalyse
* Een interactie is een verschil tussen effecten
* Het intercept is voor alle 0 scores, dus ook voor de interactie
* De b-coëfficiënt is dan het verschil met het intercept

# College 5: Logistische regressieanalyse

## Bivariaat

Logistische regressie gebruik je wanneer je een dichotome afhankelijke variabele hebt.

Odds ratio’s

* p1/p0 = odds
* Als odds onder de 0 komt is er meer kans op niet dan op wel
* Odds ratio = odds groep 1 / odds groep 2
* In spss:
  + Onder de B-coëfficiënt staan de *logitparameters*. Deze geven aan of er een toename in een kans is of niet. Positief betekent dat de kans op wel toeneemt
  + Onder de Exp(B) staan de odds en oddsratio’s. De constante geeft de odds aan voor de met 0 gecodeerde groep. Bij de variabele staat de odds ratio. (de met 0 gecodeerde groep is groep 2)

Wald

* Is (B/SE)2
* Komt vaak overeen met de Chi-kwadraatscore
* Toets op significantie van de invloed van variabelen

-2 Loglikelihood

* Brengt tot uitdrukking hoeveel de geobserveerde waarde afwijkt van de voorspelde waarde
* -2 LL kan niet groter worden dan dit getal + de chikwadraat
* Er zijn verschillende maten voor de verklaarde variantie, namelijk McFadden R2, Cox & Snell R2 en de Nagelkerke R2

## Multivariaat

* Soms is het beter om catagorische variabelen aan te maken (dummies). Dit is te testen door naar de ‘*Omnibus tests of model coefficients*’ te kijken. Als de chi groter is dan het aantal vrijheidsgraden is het waarschijnlijk significant beter

B-step

* Door te kijken naar de Change in -2 LogLikelihood is te zien welke variabele het belangrijkst is in het model
* Gebruik ruime PIN en POUT (.90 en .99)

Kansen uitrekenen

* p1 = e a + b \* X / ( e a + b \* X + 1)

# College 6: Padanalyse

3 voorwaarden voor causaliteit

1. Samenhang
2. Tijdsvolgorde
3. Samenhang niet het resultaat van derde variabele

Voor causale verbanden kijken we naar de Beta’s. Dit zijn gestandaardiseerde scores en daarom is het makkelijker hiermee te analyseren. De B-coëfficiënten veranderen wanneer de variabele groter of kleiner wordt.

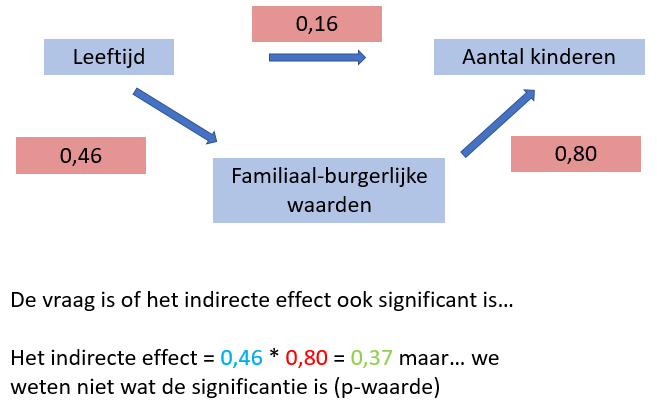
* Bepaal XZ via Pearson
* Bepaal XY via regressie (beta)
* Bepaal ZY via regressie (beta)

Modellen:

1. Interpretatie
   * X🡪Z🡪Y
   * Na controle voor Z is XY niet meer significant
2. Schijnrelatie
   * X🡨Z🡪Y
   * Z verklaart allebei de variabelen
3. Hybride
   * Interpretatie
     + Na controle voor Z blijft XY ook significant
   * Schijn
     + Na controle voor Z blijft XY ook significant
4. Suppressie
   * Na toevoeging van Z wordt de relatie tussen XY sterker
5. Interactie (zie college 4)

## Macro PROCESS

Wanneer er een indirect effect is kan worden getest of dit significant is.

Indirecte effect voorbeeld:

In SPSS 23 is dit te toetsen via:

ANALYSE🡪REGRESSION🡪PROCESS v3.0 by Andrew F. Hayes

Let op dat er model 4 staat! Vink ‘effect size’ aan onder options.

Kijk hierna naar de onderste rij. Het laat de 95% BI zien. Kun je veranderen onder de models. Als de 0 niet in de BI voorkomt dan is het significant!