STAPPENPLAN REGRESSIE-ANALYSE

Harry Ganzeboom

4 januari 2015

Regressie-analyse is een statistisch model om de lineaire effecten van een of meerdere onafhankelijke variabelen Xk op een enkele afhankelijke variabele Y te onderzoeken, volgens het additieve model:

$$Y^{\wedge} = B0 + B1*X1 + B2*X2 + \ldots + Bk*Xk$$

$$Y = B0 + B1*X1 + B2*X2 + \ldots + Bk*Xk + residu$$

Enkelvoudige regressie (*simple regression*, met slechts een X-variabele) wordt vooral gebruikt om een lineaire trend te modelleren. Door uitbreiding met dummy-variabelen of polynomische termen (X^2, X^3) kun je dit gemakkelijk uitbreiden naar niet-lineaire trends.

Meervoudige regressie (*multiple regression*, met meerdere X-variabelen) wordt gebruikt voor causale analyse. De interpretatie van de B-coëfficiënten is *partieel*: wat is het effect van een X op Y als je alle andere Xk constant houdt? Omdat je dit bij regressie-analyse met veel X-variabelen tegelijk kunt doen, is het hiervoor beter bruikbaar dan tabelsplitsingtechnieken.

Het regressie-model is veruit de meest gebruikte statistische techniek, binnen en buiten de sociale wetenschappen. Andere veel gebruikte technieken (zoals (partiele) correlatie, variantie-analyse, logistische regressie, general linear models en factoranalyse) kun je zien als een variant of uitbreiding van het regressie-model.

Enkelvoudige regressie kun je afbeelden als een rechte lijn in een plat vlak. Bij twee X-variabelen kun je je het regressiemodel voorstellen als een vlak in een driedimensionele kubus. Bij meer dan twee X-variabelen kun je spreken over een hyper-vlak, maar het visualiserende aspect gaat dan verloren.

De coëfficiënten in het model worden berekend door het minimaliseren van de residuen volgens het kleinste kwadraten criterium:

$$\text{SS-residuen} = \sum(Yi - Y^{\wedge})^2 \rightarrow \text{minimaal.}$$

Om deze reden wordt het regressiemodel ook wel aangeduid als het OLS (Ordinary Least Squares) model. Je hoeft je over het OLS minimaliseringsalgoritme geen zorgen te maken, SPSS rekent het voor je uit.

Belangrijke aannames bij het regressiemodel zijn:

- De (partiële) relatie tussen Y en X is lineair (lineariteit)

- De effecten van Xk op Y zijn additief (geen interactie).

Aan problemen met beide aannames kun je doorgaans tegemoet komen door uitbreidingen van het model. Verdere aannames hebben betrekking op de residuen:

- Residuen zijn homogeen rondom de regressielijn / het regressievlak: homoskedastiteit
- Residuen hebben een normale verdeling: normaliteit.
- Residuen zijn onafhankelijk van elkaar: onafhankelijkheid.

Schending van deze aannames heeft met name gevolgen voor de inferentiële statistiek (standard errors, t-waarden, significanties). De geschatte coëfficienten worden er doorgaans niet of nauwelijks door beïnvloed.

Stap 0: Beschrijven van de variabelen

Check de volgende zaken:

- Wat is het bereik (min, max), gemiddelde en spreiding (standaard deviatie) van de betrokken variabelen?
- Heeft de afhankelijke variabele Y minimaal ordinaal meetniveau? Je kunt regressie-analyse niet gebruiken voor een nominale afhankelijke variabele. Alternatieven zijn binomiale logistische regressie en multinomiale logistische regressie. Bij een ordinale afhankelijke variabele kan ordinale logistische regressie een alternatief zijn, maar meestal gebruiken we dan toch lineaire regressie.
- Verdeling van de afhankelijke variabele Y. Wees hierbij vooral beducht op uitbijters, die veel invloed op de schattingen kunnen hebben.
- Ook de X-variabelen dienen minimaal ordinaal meetniveau te hebben. Voor nominale variabelen heb je hier regressie met dummy-variabelen tot je beschikking.

Stap 1: Transformatie X-variabelen

- Een regressiemodel is veel gemakkelijker te interpreteren als de X-variabelen aan twee voorwaarden voldoen:
    - 0 is een geldig en interpreteerbare waarde
    - De eenheid van de variabele is gemakkelijke te interpreteren.
- Vier manieren om dit voor elkaar te krijgen zijn:
    - Herschalen van variabelen naar een 0..1 bereik.

- Hercoderen van variabelen naar een 0/1 dichotomie.
- Variabelen transformeren naar een percentiel (rang) score. Deze hebben minimum 0 en maximum 100 (of 1).
- Variabelen standaardiseren naar Z-waarden. Deze hebben gemiddelde 0 en de standaarddeviatie als eenheid.

Stap 2: Onderzoek lineariteit

- Het is soms de moeite waard om de lineariteitsaanname nader te onderzoeken. Een directe lineariteitstest biedt **`Means /stat=anova.`**
- Aan andere manier is de onafhankelijke variabele op te delen naar dummy-variabelen en te kijken of de verklaarde variantie significant toeneemt in vergelijking met een gewoon lineaire model. Deze methode is algemener dan de lineariteitstest in **`Means`**, en kan bv. ook worden toegepast als er nog meer X-vars zijn.
- Bij beide voorgaande methoden is het belangrijk dat je X-variabelen een beperkt aantal categorieën heeft. Bij een continue X (bv. leeftijd) is handzame truc om de variabelen op de delen in stukjes en die te scoren naar het klassemidden. Zie bijlage B.
- Als je niet wilt categoiseren, kun je niet-lineariteit onderzoeken door het toevoegen van polynomische termen (machten van X). Op die manier laat je kromlijnige verband toe.

Stap 3: Stapsgewijze schatting van het regressiemodel[1]

- Een regressiemodel is het gemakkelijkst te interpreteren als je het stapsgewijs opbouwt, namelijk door het een voor een of bloksgewijs toevoegen van X-variabelen. Door het vergelijken van coëfficiënten tussen de modellen, leer je hoe de partiële interpretatie van de relaties in elkaar zit.
- In SPSS kun je stapsgewijze schatting gemakkelijk doen via een meervoudige specificatie van het **`/ENTER`** statement.
- Volg bij het toevoegen van variabelen altijd de volgorde:
  - Hoofdeffect, confounders, mediatoren, interactietermen

  Zowel toevoeging van confounders als mediatoren kan tot gevolg hebben dat het hoofdeffect verandert, maar de causale interpretatie ervan is sterk verschillend.

[1] Pallant (2013: 155) noemt deze method “hierarchical regression”, maar deze term wordt in de literatuur ook wel gebruikt voor regressie met meerdere niveaus, bv. leerlingen binnen scholen *(multi-level regression)*.

- SPSS zet de verschillende modellen onder elkaar. Als je je eigen tabel maakt, zet je ze naast elkaar. Via bewerking van de output in Excel gaat dit tamelijk gemakkelijk.

Stap 4: Interpretatie van de regressie-coëfficiënten

- Het in elke stap geschatte regressiemodel vind je links onderaan, onder "Unstandardized Coëfficiënts B". De "(Constant)" [de intercept B0] geeft aan wat de verwachte Y is als alle X-variabelen op 0 staan. De overige B-coëfficiënten geven aan hoe Y verandert als de X een eenheid verandert. Als de X-vars verschillende eenheden hebben (en dat hebben ze vaak), kun je B-coëfficiënten alleen *tussen modellen* en *NIET tussen variabelen* vergelijken.

- Onder "Standardized Coëfficiënts Beta" vind je dezelfde informatie in gestandaardiseerde termen. Net zoals correlaties variëren deze getallen tussen -1.00 en +1.00. Hun interpretatie is: hoeveel meer standaarddeviatie van Y krijg je als X met 1 standaarddeviatie verandert? Deze coëfficiënten kun je *tussen variabelen* vergelijken. Het geeft je reden om te concluderen dat het ene effect sterker is dan het andere.

- De standardized coëfficiënts hebben geen intercept, maar zo je wil is deze 0 (nl. het gemiddelde van de gestandaardiseerde Y-var).

Stap 5: Interpretatie van de inferentiële statistiek

- De kolom Standard Error [SE] geeft de geschatte steekproevenvariatie aan van de regressie-coëfficiënten. De SE is de geschatte standaarddeviatie van de steekproevenverdeling. T is de geschatte toetsgrootheid (T = B / SE) en de Sig-kolom geeft de overschrijdingskans van het steekproefresultaat onder de H0 aan, en zou daarom beter P genoemd kunnen worden. Als Sig < 0.05, dan noemen we de betreffende coëfficiënt statistisch significant. De toetsing gaat altijd over de H0: B=0. Voor de intercept is dit (B0=0) meestal een zinloze hypothese, voor de andere coëfficiënten betekent het: geen effect.

Stap 6: Interpretatie van de ANOVA tabel

- De ANOVA tabel geeft de variantie analyse van het geschatte model, met de fundamentele som: SS-totaal = SS-regression + SS-residual. De bijbehorende hoeveelheid vrijheidsgraden geeft de totale hoeveelheid observaties (N-1) aan, en de hoeveelheid geschatte effecten, waarbij geldt: DF-total = DF-regression + DF-residual.

- De ANOVA tabel levert uiteindelijk een F-test op, die de significantie van het gehele model toetst. Meestal is deze niet zo interessant, omdat deze grootheid altijd significant is, als ten minste een van de T-toetsen significant is.

Stap 7: Interpretatie van de verklaarde variantie

- De Model Summary tabel geeft per model de proportie verklaarde variantie aan (=SS-regression / SS-total) en de daaruit berekende Multiple R. Meestal rapporteren we de Adjusted R-square, deze geeft een (vaak kleine) correctie voor de hoeveelheid geschatte effecten.

- De Standard Error of the Estimate is eigenlijk geen standard error (het gaat namelijk niet over sampling variability), en zou beter heten: Root Mean Squared Residual. Het is de gemiddelde absolute afwijking van het regressievlak. Wordt kleiner als de verklaarde variantie toeneemt.

- Bij stapsgewijze regresie en **`/stat=change`** lees je in de Model Summary of de uitbreiding van het model in opeenvolgende stappen een statistisch significante verbetering oplevert.

Stap 8: Check de invloed van missing values

- Regressie kun je zowel met *listwise* als *pairwise deletion of missing values* berekend worden. Je kunt met **`/DESC = DEF N`** heel mooi te zien krijgen waar de missing values optreden.

- Probeer je bij missing values eerst af te vragen hoe ze ontstaan zijn. Wanneer het patroon door puur toeval ontstaan is, zullen pairwise en listwise correlatie matrix veel op elkaar lijken (nl. binnen toevalsgrenzen hetzelfde zijn) en zijn de pairwise schattingen preciezer dan de listwise schattingen. Als er een systematische afwijking tussen de twee situaties optreedt (correlaties verschillen, en de geschatte modellen verschillen ook), dan zijn je listwise schattingen vertekend. Het behandelen van zulke missing values kan het beste via Multiple Imputation. Een tussengeval is wanneer de correlaties wel behoorlijk verschillen tussen listwise en pairwise, maar het geschatte model niet. Ook in dat geval zijn de pairwise schattingen precieser.

- Uiteindelijk draait het bij missing values meer om de inferentiele statistiek (de standard errors, significantie) dan de beschrijvende statistiek (het regressiemodel zelf). Wees dus bijzonder alert op situaties waarin de significantie van effecten verandert.

Stap 9: Multicollineariteit?

- Pallant (2013: 163-164) wil graag dat we multi-collineariteit testen. ‘Inzicht’ hierin zou je krijgen via de Tolerance en VIF coëfficiënten die kunt opvragen door **`/STAT=DEF TOL`** te specificeren. De door haar genoemde criteria zijn: Tol < .10 en VIF > 10 (omdat VIF = 1/Tol, is dit hetzelfde).

- We spreken van multi-collineariteit als de X-variabelen zodanig sterk met elkaar samenhangen dat het niet mogelijk is om uitspraken over hun partiële effecten te doen: als

X1 en X2 sterk met elkaar samenhangen, kun je nu eenmaal niet de een variëren, terwijl je de andere constant houdt.

- Maak je geen zorgen, multicollineariteit komt in de praktijk nauwelijks voor. In de gevallen waarin dat wel zo lijkt te zijn, gaat het meestal om modellen met polynomische of interactietermen, waarop de hele redenering van multicollineariteit niet opgaat.

Stap 10: Andere regression diagnostics

- SPSS biedt een uitgebreid aantal mogelijkheden om andere assumpties van regressie-analyse (anders dan lineariteit en collineariteit) te onderzoeken. Pallant (2013) noemt: *outliers, normality, homoskedasticity, independence*. In het meeste praktische werk is dit allemaal niet zo belangrijk. Het speelt echter een grote rol wanneer onze analyses betrekking hebben op een beperkt aantal macro-eenheden, zoals landen of organisaties.

Bijlage A: Minimale SPSS syntax (met standaardwaarden onderlijnd)

```
REGRession
  /MISSing = LISTwise / PAIRwise
  /DESCriptives = DEFault N
  /STATistics = DEFault TOLerance
  /DEPendent = Y / ENTER = X1 /ENTER = X2.
```