UniAnova als alternatief voor Regression

Harry BG Ganzeboom

M&T lunch Afdeling Sociologie

26/27 mei 2020 / herzien 27 juni 2021

Dit document citeren als: Ganzeboom, Harry BG (2021). “UniAnova als alternatief voor Regression”. http://www.harryganzeboom.nl/Teaching/index.htm. [last accessed: *date*]

Conclusies

- Regressie en variantie-analyse zijn dezelfde statistische modellen.
- Of je experimentele of observatiedata bestudeert maakt niet uit voor de keuze van de procedure; in beginsel kan alles met zowel Regression als UniAnova / GenLin.
- De keuze tussen procedures hangt meer af van bijkomstige voor- en nadelen en de manier waarop je het aan studenten kunt uitleggen.
- SPSS Regression heeft een groot aantal handige voordelen die je bij UniAnova niet hebt.
- UniAnova heeft een aantal handige voordelen die je bij Regression niet hebt (en ook niet bij GenLin).

SPSS Regression

- Syntax

```
REGR /dep=Y /enter=X1 /enter=X2 /enter=X3.
```

- Dit staat bekend als stapsgewijze (‘stepwise’) regressie. Eigenlijk is dit een verkeerde aanduiding, want je kunt ook backward elimination, en stepwise (op en neer) doen, maar dat wordt in de praktijk weinig gebruikt.
- Nuttige extra syntax:

```
REGR /dep=Y /enter=X1 /enter=X2 /enter=X3
/missing=pair /des=def corr N.
```

SPSS Regression: de verdiensten

- Stepwise (== forward)
  - Alle coëfficiënten overzichtelijk in een tabel (onder elkaar).
  - F-test Change (op toename R2)
- B en Beta in een tabel.
- Het enige SPSS programma dan gebruik kan maken van correlaties / covarianties met pairwise deletion of missing values
  - Dit geeft een paar belangrijke tools om MV op te sporen: N of cases in univariate en bivariate data.
  - Een vergelijking tussen een pairwise en een listwise schatting kan je inzicht geven of de listwise data selectief zijn (toets op MAR).
- Zeer eenvoudige syntax.
- Bootstrapped SE available
- Collinearity diagnostics available: TOL en VIF.

SPSS Regression: de grootste verdienste

- Stapsgewijze regressie laat je nadenken over causale modellen: welke variabelen zijn oorzaak, welke gevolg, mediatoren, confounders?
- (Natuurlijk is dit een bijkomstigheid. Je kunt ook over causaliteit nadenken zonder stapsgewijze regressie, en je kunt stapsgewijze modellen ook berekenen in andere programma's dan SPSS Regression. Niettemin: de stappen zetten mensen aan het denken.
- Voorkeursvolgorde: (1) centrale X, (2) confounders (→ totale effect van X), (3) mediatoren (→ partiële effect van X)

SPSS Regression: de nadelen

- In de Descriptives ontbreken de min-max.
- Je moet even doorhebben waar je de effectieve N of cases moet zoeken.
  - En bij pairwise deletion is het niet duidelijk hoe deze tot stand komt.
- Pairwise estimation van SE is niet correct (bootstrap helpt niet). Voor de goede schatting moet je naar SEM met MLMV / FIML.
- Je moet dummy-variabelen en interactie-termen zelf definiëren
  - Kan heel veel syntax opleveren, vooral als je niet weet hoe DO REPEAT werkt
  - Error-prone, in het bijzonder bij missing values

UNIANOVA

- Oorsprong: variantieanalyse voor experimentele designs
- Regressie-analyse heet daar “analysis of covariance, ANCOVA”.
- Je vindt het bij “General Linear Model”; maar het is geen GLM, daarvoor moet je naar Generalized Linear Model → GENLIN
- UniAnova biedt hetzelfde model als Regression, maar het perspectief is soms wat anders:
  - Je krijgt niet standaard de B-coefficienten van het model te zien!
  - Geen pairwise deletion of missing values
  - Geen stapsgewijze opbouw van het model

UNIANOVA - nadelen

- Iets ingewikkelder syntax: je moet eerst je variabelen list opgeven, en vervolgens je model:

  unianova Y with X1 by X2 /print=parameter /design=X1 X2*X1.

- Het woord “design” is ook niet helemaal triviaal
- Variantie-analyse tabel heeft extra regel: ‘corrected total’.
- Geen beta-coëfficiënten. Je kunt deze verkrijgen door de betrokken variabelen zelf te standaardiseren – hetgeen eigenlijk een veel beter idee is dan de beta in Regress, met name wanneer je categorische variabelen en interacties hebt.

UNIANOVA - voordelen

- Automatisch dummy variabelen
  - Vooral handig als je er veel hebt
  - Werkt zelfs goed als je categorische variabele een string is (bv een landnaam).
- Automatische interacties
  - Via symbolische manipulaties (geen multiplicatieve interactietermen aanmaken).
- Meer statistics:
  - EMMEANS; expected value of Y for categorical variables at means of controls.
  - Effect size
  - Observed power

UNIANOVA: de grootste verdienste

- De grote verdienste van UNIANOVA is de heldere manier waarop referentiecategorieën en referentie-effecten worden aangegeven, namelijk als B=0*, met als noot: *aliased.
- Deze manier van noteren is veruit de helderste en altijd aan te bevelen, ook al kost het een extra regel in je tabel.
- UniAnova neemt altijd de laatste categorie als referentie, en dat is niet altijd de verstandigste keuze. Regression neemt standaard de meest omvangrijke categorie als referentie en dat is doorgaans een verstandiger keuze.
- Je kunt de keuze gemakkelijk manipuleren door je variabelen zo het hercoderen dat je eigen keuze de laatste categorie wordt.

UNIANOVA – waarom is het handig?

- Het grote voordeel van het gebruik van Unianova zit in de automatisch aanmaak van dummy en interactie-variabelen
- Zeer eenvoudige en vanzelfsprekende syntax
  - Categorisch variabelen staan achter BY en continue variabelen achter WITH.
  - (Het is soms handig om zowel een categorische als continue variant van een variabelen te hebben (bv YEAR en YR).
- Zeer aansprekende parameters: referentiecategorie duidelijk gemarkeerd in het model, als een gefixeerde 0.

GENLIN

- De echte GLM's staan in SPSS achter Generalized Linear Models.
- GLM betrekt zich op twee elementen:
  - De link-functie: hoe is Y verbonden met de data? Bij OLS via Identity, maar andere mogelijk link functies zijn: (ordinal) logit, (ordinal) probit, count, en nog meer.
  - Welke statistische verdeling hoort bij de evaluatie van je model: normaal, logistisch, poisson.
- Model specificatie is hetzelfde als bij Unianova, alleen de keywords zijn anders:

  unianova Y with X1 by X2 /print=parameter /design=X1 X2*X1.

  genlin Y with X1 by X2 /print=solution /model=X1 X2*X1.
- Minder statistics dan bij UNIANOVA.