



Universiteit
Leiden
The Netherlands

Distance models for analysis of multivariate binary data

Worku, H.M.

Citation

Worku, H. M. (2018, December 20). *Distance models for analysis of multivariate binary data*. Retrieved from <https://hdl.handle.net/1887/67140>

Version: Not Applicable (or Unknown)

License: [Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden](#)

Downloaded from: <https://hdl.handle.net/1887/67140>

Note: To cite this publication please use the final published version (if applicable).

Cover Page



Universiteit Leiden



The following handle holds various files of this Leiden University dissertation:

<http://hdl.handle.net/1887/67140>

Author: Worku, H.M.

Title: Distance models for analysis of multivariate binary data

Issue Date: 2018-12-20

Samenvatting

In dit proefschrift beschrijven we recent ontwikkelde statistische tools voor het analyseren van multivariate binaire data. Multivariate binaire data, gedefinieerd als verzamelde gegevens van meerdere binaire afhankelijke variabelen en één of meer onafhankelijke variabelen, komen voor in allerlei onderzoeksdisciplines. Neem bijvoorbeeld de Indonesische Kinderen Studie (ICS). In deze studie is er data verzameld van meer dan drieduizend kinderen die medisch onderzocht zijn op luchtweginfectie, diarree-infectie, en xeroftalmie. Het doel van de ICS was om te achterhalen of kinderen met een deficiëntie in Vitamine A een verhoogd risico lopen op luchtweg- en diarree-infectie.

Een ander voorbeeld waarbij multivariate binaire wordt gebruikt is de Nederlandse Studie naar Depressie en Angst (NESDA). De gegevens die door NESDA verzamelt worden dienen ten doel om de interactie tussen persoonlijkheidseigenschappen enerzijds en de comorbiditeit van depressie- en angststoornissen anderzijds te kunnen onderzoeken. In dit onderzoeksgebied van psychologische stoornissen zijn psychologen en epidemiologen veelal genteresseerd in comorbiditeit en hoe comorbiditeit gerelateerd kan worden aan risicofactoren zoals persoonlijkheidseigenschappen en achtergrondkenmerken.

Er zijn talloze statistische methoden beschikbaar voor het analyseren van multivariate continue afhankelijke variabelen doordat er goed gebruik gemaakt kan worden van de multivariate normale kansverdeling. De multivariate regressie en de multivariate variantie analyse (MANOVA), om er maar een paar te noemen, behoren tot de populaire statistische methoden die hier worden toegepast. Echter, voor de multivariate categorische data is

het aanbod van methoden en technieken gering. De huidige beschikbare methoden en technieken bouwen voort op assumpties die niet gecontroleerd kunnen worden (zoals het bestaan van de latente variabelen in latent variable models en structural equation models), of komen met vereisten dat de onafhankelijke variabelen gecategoriseerd dienen te worden (zoals de GEE2 methode voor marginale modellen). Met behulp van een Monte Carlo simulatie studie laten in hoofdstuk 2 we zien dat het toepassen van een latente variable model op multivariate binaire data tot gebrekkige resultaten leidt wanneer er slechts twee of drie indicatoren per latente variabele zijn.

In dit proefschrift presenteren we een aangepaste versie van het ideal point classification (IPC) model waarmee multivariate binaire gegevens geanalyseerd kunnen worden. Het IPC model is een probabilistisch multidimensional “unfolding” model en veel lijkend Ideal Point Discriminant Analysis (IPDA). Hoofdstuk 3 begint eerst met een studie van de eigenschappen van het IPC model voor het analyseren van bivariate binaire gegevens. Door gebruik te maken van een kader gebaseerd op de bivariate logistische regressie, kunnen de afhankelijke variabelen worden gerepresenteerd in een drie-dimensionale Euclidische ruimte. In deze drie-dimensionale ruimte heeft de eerste dimensie betrekking op de prevalentie van de eerste afhankelijke variabele; de tweede heeft betrekking op de prevalentie van de tweede variabele; en, de derde dimensie heeft betrekking op de samenhang tussen de twee afhankelijke variabelen. Op basis van een simulatie studie kunnen we aantonen dat met het IPC model het niet volledig mogelijk is om de daadwerkelijke parameters van de binaire data te achterhalen, dat wil zeggen, de twee marginale prevalentie parameters en de parameter voor de associatie tussen de twee afhankelijke variabelen. In hoofdstuk 3 laten we vervolgens zien dat met een re-parameterisatie van het IPC model het wel mogelijk is om deze parameters terug te vinden. Dit aangepaste model noemen we het Bivariate IPC (BIPC) model.

Een beperking van het Bivariate IPC model is dat het niet toegankelijk is om uit te breiden naar multivariate binaire gegevens (meer dan twee binaire afhankelijke variabe-

len). Door deze beperking van het BIPC model, wordt in hoofdstuk 4 voorgesteld om het Multivariate Logistische Afstands (MLD) model te gebruiken voor het analyseren van multivariate binaire data. Het MLD model is een vereniging van twee soorten domeinen van statistische methoden: het domein van de Multidimensional Scaling (MDS) en het domein van het Generalized Linear Model (GLM). Het MLD-model kan tegelijkertijd gebruikt worden voor zowel het beoordelen van de dimensionale structuur van de data als het schatten van het effect van de onafhankelijke variabelen op de afhankelijke variabelen. Zo biedt het MLD-model de mogelijkheid om op NESDA data tegelijkertijd de dimensionale structuur van psychologische stoornissen te onderzoeken als het effect van persoonlijkheidseigenschappen en achtergrondkenmerken op de prevalentie van psychologische stoornissen.

Voor ondersteuning van interpretatie doeleinden lenen de resultaten de MLD analyse zich goed voor de grafische weergave in een biplot. Een ander voordeel van het MLD-model ten opzichte van marginale modellen is dat MLD-model toegepast kan worden in combinatie met dimensie reductie, waarmee de complexiteit van het standaard multivariate GLM wordt vereenvoudigd door minder parameters te hoeven schatten. Met deze dimensie-reductie methode wordt de deur geopend naar verder onderzoek.

Wanneer de afstanden tussen de twee categorieën op elke afhankelijke variabele eenzelfde waarde krijgen toegewezen, dan kan het MLD-model geschat worden door gebruik te maken van de GEE methode. Onder deze restrictie van 'gelijke afstanden' is het dan ook mogelijk om het MLD-model te schatten met behulp van bestaande statische software pakketten zoals de **genmod** procedure in SAS, of het **geepack**-pakket in R. Wanneer er geen gebruik wordt gemaakt van de gelijke afstanden restrictie, dan is het MLD-model een op zichzelf staand marginaal model. In hoofdstuk 5 presenteren we het **mldm**-pakket dat is ontwikkeld in R om het MLD-model op data te kunnen toepassen. De belangrijkste functie in dit pakket is `mldm.fit`, hiermee kunnen we het MLD-model schatten. Vervolgens kan met de functie `mldm.fit` het geschatte model grafisch worden weergegeven in een biplot. De functie `mldm.fit` heeft ook als output een object genaamd `QIC`, met dit

object kunnen verschillende kandidaat-modellen worden vergeleken. Het **mldm**-pakket is publiek toegankelijk en beschikbaar op het online database-systeem GitHub, te vinden via het URL adres: <https://github.com/workuhm1/mldm-package-github>.

Ten slotte raden we onderzoekers aan om voorzichtig te zijn met het toepassen van latent variable models of structural equation models op multivariate binaire gegevens. De prestatie van statistische methoden gebaseerd op deze modellen is ondermaats met slechts enkele indicatoren per latente variabele (d.w.z. 2 of 3). Een alternatief statisch model dat minder assumpties vereist is mogelijk beter toepasbaar, bijvoorbeeld het multivariaat logistische afstands model.