Académie Universitaire
Wallonie-Europe

UNIVERSITÉ DE LIÈGE
FACULTÉ DES SCIENCES
Département de Mathématique

Contribution to the statistical analysis of incomplete longitudinal ordinal data

Anne-Françoise DONNEAU

Dissertation présentée
en vue de l'obtention du grade de
Docteur en Sciences

2013

UNIVERSITY OF LIEGE
Faculty of Science
Department of Mathematics

Contribution to the statistical analysis of incomplete longitudinal ordinal data

Anne-Françoise DONNEAU

Supervisors

Professor A. Albert
Professor G. Haesbroeck

February 2013

Acknowledgements

At the end of my doctoral thesis I would like to thank all those who one way or another made its completion possible.

First and foremost, I would like to express my deepest gratitude to my main supervisor, Professor Adelin Albert. Soon after my master degree in biostatistics in 2003, Professor Albert welcomed me in his department of medical informatics and biostatistics at the Faculty of Medicine. As an assistant, I worked for almost 10 years under his guidance and longstanding experience. Over the years I gained teaching expertise, research abilities and statistical practice skills. He originally suggested the topic of my thesis considering that ordinal variables play a major role in medical sciences. His valuable suggestions and careful review of the several draft versions of my thesis were highly appreciated. I am much indebted for his continuous support and enthusiasm throughout my stay in his department. Next, I am deeply grateful to Professor Gentiane Haesbroeck, my copromotor, for her particular attention to my research activities and her sustained encouragements. I would also like to thank Professor Philippe Lambert, President of the jury, as well as the other members of the jury, in particular Professors Geert Molenberghs, Emmanuel Lesaffre and Bernard Vrijens, for their interest in this work.

I take the opportunity to express my gratitude to the EORTC Belgium members who generously shared some of their quality of life datasets to apply our methods but also gave me the opportunity to present at several occasions the progresses of my work at their ‘Stats Club Meeting’. In particular, I would to thank Murielle Mauer for the numerous stimulating exchanges we had about the analysis of ordi-

nal data.

During all these years I cannot forget my current and previous colleagues of the Department of Public Health and the Department of Mathematics who made my working environment so stimulating and enjoyable. In particular, I would like to thank Sophie Vanbelle and Sylvie Streel who successively shared my office and created a deep friendship, Lixin Zhang who helped me improving my computer programing skills at the time she was in our department, Nadia Dardenne and Laurence Seidel for their daily agreeable company and for ’lending’ me their computers during the week-end and the holidays to run my unending simulations! My special thanks go also to Eddy Husson, Nicolas Gillain, Anna Marchetta, Claire Lepère, Vanessa Strauven and particularly Professor Michèle Guillaume for introducing me to the field of Public Health and for boosting me during the last steps of my thesis.

I would like to address my heartfelt acknowledgments to my parents and family who never failed to encourage me and who were supportive in all my career choices. Finally, my deepest appreciation goes to my companion Xavier for supporting my ups and downs and for postponing some projects during the realization of this thesis.

Summary

Instead of viewing ordinal variables as fully-fledged variables, some researchers consider them as being either nominal or quantitative. Therefore, when analysing data, they generally apply methods that are inappropriate because they ignore the ordinal feature of the variables. Methods for analysing ordinal outcome variables, in particular in longitudinal settings with missing data, are the main focus of this thesis.

In the first part of this work, we consider the analysis of ordinal outcome data in situations where assessments are made only once for each subject. After providing a definition of an ordinal variable, the different ways to assess its relationship with a set of covariates are investigated. In this perspective, we focus on the well-known proportional odds regression model and on its strict underlying assumption. Some authors developed numerical and graphical methods to test the proportional odds assumption. When this assumption fails for some or for all the covariates, fitting a more general model, such as the partial proportional odds model or the non-proportional odds model, can be the solution.

Next, our investigations on the analysis of ordinal outcome variables move to the longitudinal setting and the associated unavoidable problem of missing data. In this context, the marginal model framework was considered with the generalized estimating equations (GEE), popular for the analysis of non-Gaussian correlated data. After an adaptation of the GEE to the ordinal outcome framework, we proposed to handle the presence of missing data by considering multiple imputation (MI) prior to the data analysis. Even if not strictly speaking appropriate for ordinal data, it is a common practice for researchers to impute ordinal missing data using Markov Chain Monte Carlo (MCMC) methods designed for continuous ones, for example the multivariate normal imputation (MNI) method. There is a need however to use more appropriate MI approaches specifically designed for ordinal data. In this context, we introduced the ordinal multiple imputation method (OIM) based on the proportional odds model.

Then, we conducted a comprehensive simulation experiment in which we investigated the effect of several factors on the estimation of the parameters of the proportional odds model. These factors included the number of categories of the ordinal outcome, the sample size, the number of time points, the rate of missingness, the type of missingness (monotone or non-monotone) and the form of the ordinal data distribution (well-balanced or skewed). Our work shows that, whatever the pattern of missingness, the estimates derived under the MNI are highly biased, while those obtained under the OIM are almost unbiased even for datasets with a high proportion of missing data. In fact, we found that the MNI approach markedly modifies the underlying distribution of the ordinal data as opposed to the OIM.

In the final part of our work, we addressed the problem of testing the proportional odds assumption for incomplete longitudinal ordinal data. Under the MNI, the type I error rate was significantly higher than the 5% nominal level while the power was markedly increased as compared to that of the full dataset (absence of missing data). Under OIM, the opposite picture occurred with lowered type I error and loss of power.

As a general conclusion, our findings demonstrate that the widely used MI method based on multivariate normality in the analysis of incomplete longitudinal ordinal data does severely bias estimates of the proportional odds regression coefficients. Furthermore, multiple imputation methods, MNI or OIM, are inadequate to test the proportional assumption since they modify the distribution of the data in favor or against this popular assumption. The test based on the so-called "complete-case" dataset, i.e. eliminating all subjects with missing observations, did actually perform well in spite of a loss of power, at least when the rate of missingness was moderate. Throughout this work, methods were applied to real life datasets and more particularly to quality of life data of an EORTC cancer clinical trial which motivated the present research work.

Résumé

Au lieu de considérer les variables ordinales comme des variables à part entière, certains chercheurs et utilisateurs les considèrent comme nominales ou quantitatives. Par conséquent, lors de l'analyse statistique de telles données, ils utilisent généralement des méthodes inappropriées puisqu'ils ignorent le caractère ordinal des variables traitées. Cette thèse s'est intéressée aux méthodes statistiques pour variables ordinales, en particulier dans le cadre d'études longitudinales avec données manquantes.

Dans la première partie du travail, nous abordons l'analyse de données ordinales non répétées obtenues sur un échantillon de sujets. Après avoir donné la définition d'une variable ordinale, les différentes façons d'évaluer la relation entre une variable ordinale et un ensemble de covariables sont étudiées. Dans cette perspective, nous nous concentrons sur le célèbre modèle de régression dit des "cotes proportionnelles" (ou "proportional odds") ainsi que sur son hypothèse sous-jacente. Plusieurs méthodes numériques et graphiques ont d'ailleurs été développées par différents auteurs pour tester cette hypothèse. Lorsque celle-ci n'est pas vérifiée pour toutes ou pour une partie des covariables, l'utilisation d'un modèle plus général, comme le modèle à "cotes proportionnelles partielles" ou le modèle à "cotes non-proportionnelles", peut être envisagé.

Ensuite, nous considérons l'analyse statistique de données ordinales longitudinales, souvent compliquée par la présence d'observations manquantes. Dans ce contexte, la méthode des équations d'estimation généralisée (GEE), populaire pour l'analyse des données non gaussiennes corrélées, s'avère intéressante. Après adaptation du modèle GEE aux variables ordinales, nous proposons de gérer le problème des observations manquantes en appliquant la technique d'imputation multiple (MI) avant l'analyse proprement-dite des données. Même si l'approche n'est pas spécialement appropriée aux données ordinales, les utilisateurs ont pris l'habitude d'imputer les données manquantes ordinales à l'aide des méthodes MCMC (Monte-Carlo Markov Chain) conçues pour données continues ; c'est le cas de la méthode d'imputation normale multivariée (MNI). Il est toutefois nécessaire d'utiliser des techniques d'imputation mieux adaptées aux données ordinales. Pour ce faire, nous avons introduit la méthode d'imputation multiple de régression ordinale (OIM) basée sur le modèle des cotes proportionnelles. Dans une autre partie du travail, nous avons réalisé une vaste étude de simulations afin d'étudier l'effet de différents facteurs sur l'estimation des paramètres du modèle des cotes proportionnelles. Ces facteurs comprennent le nombre de catégories de la variable ordinale, la taille de l'échantillon, le nombre de mesures répétées pour chaque sujet, le pourcentage de données manquantes, le profil de données manquantes (monotone ou non-monotone) ainsi que la forme de la distribution des données ordinales (homogène ou asymétrique). Notre étude montre que, quel que soit le profil de données manquantes, les estimations obtenues après application de la méthode MNI sont fortement biaisées, tandis que celles obtenues avec la méthode OIM sont pratiquement fidèles, même dans le cas d'échantillons comportant un pourcentage élevé de données manquantes. En fait, nous avons constaté que, contrairement à la méthode OIM, l'approche MNI modifie substantiellement la distribution sous-jacente de l'échantillon des données ordinales.

Dans la dernière partie de notre travail, nous avons abordé le problème du test d'hypothèse des "cotes proportionnelles" à partir d'un échantillon de données ordinales longitudinales incomplètes. Avec la méthode MNI, le taux d'erreur de type I est significativement plus élevé que le niveau nominal classique de 5% ; de la même manière la puissance est nettement augmentée par rapport à celle obtenue à partir de l'échantillon complet (c'est-à-dire sans données manquantes). La situation est inversée lorsqu'on applique la méthode OIM; en effet, celle-ci diminue l'erreur de type I et conduit à une perte de puissance du test sous l'hypothèse alternative.

En conclusion générale, nos résultats démontrent que la méthode d'imputation multiple MCMC basée sur la normalité multivariée, largement utilisée en pratique pour l'analyse des données longitudinales ordinales incomplètes, biaise nettement les coefficients du modèle de régression de cotes proportionnelles. De plus, nous montrons que, quelle que soit la méthode d'imputation multiple utilisée, MNI ou

OIM, il est particulièrement hasardeux de tester l'hypothèse des cotes proportionnelles car la distribution des données de l'échantillon s'en trouve fortement modifiée en faveur ou en défaveur de cette hypothèse. Par ailleurs, nous avons remarqué que le test basé sur les "cas complets", c'est-à-dire sur l'ensemble des sujets de l'étude ne présentant pas de données manquantes, fournit de bons résultats malgré une perte de puissance, du moins lorsque le pourcentage de données manquantes n'est pas trop élevé. Tout au long de ce travail, les méthodes ont été appliquées à des exemples concrets et plus particulièrement à une étude de qualité de vie de patients cancereux provenant d'un essai clinique de l'EORTC, qui fut d'ailleurs à l'origine de notre travail de recherche.

Samenvatting

In plaats van ordinale gegevens als een op zichzelf stand type te beschouwen, is het niet ongebruikelijk ze voor ofwel nominaal ofwel metrisch te aanzien. Dit heeft voor gevolg dat minder gepaste methodologie gebruikt wordt bij het analyseren van dergelijke gegevens; immers, de ordinale structuur wordt op die manier over het hoofd gezien. Methodologie voor de analyse van ordinale gegevens, in het bijzonder in het kader van longitudinale studies met ontbrekende gegevens, vormt het onderwerp van dit proefschrift.

In het eerste deel van dit werk beschouwen we ordinale respons in een univariate context, waarbij slechts één meting per studiesubject wordt verzameld. We definiëren ordinale gegevens en exploreren de verschillende manieren om het verband tussen de ordinale respons enerzijds en een reeks covariabelen anderzijds te onderzoeken. We leggen de klemtoon op het welbekende proportional odds regressiemodel, alsmede op de onderliggende, strikte veronderstellingen. Bepaalde auteurs ontwikkelden numerieke en grafische methoden om de proportional odds veronderstelling te toetsen. Wanneer de veronderstelling niet voldaan is voor sommige of voor geen enkele van de covariabelen, kan men bijvoorbeeld het partiële proportional odds model of het niet-proportional odds model beschouwen.

Vervolgens verleggen we de aandacht naar de longitudinale situatie en het er onlosmakelijk mee verbonden probleem van onvolledige gegevens. In deze context leggen we de nadruk op veralgemeende schattingsvergelijkingen (generalized estimating equations, GEE), een veelgebruikte methode voor de analyse van niet-Gaussische gecorreleerde gegevens. Na aanpassing van GEE aan het ordinale kader, pakken

we het probleem van ontbrekende gegevens aan via multiple imputation (MI) als stap voorafgaand aan de eigenlijke analyse van de gegevens. Zelfs indien het strikt genomen niet aangewezen is voor ordinale gegevens, is het toch gebruikelijk om dergelijke gegevens te imputeren via Monte Carlo Markov Chain (MCMC) methoden, die eigenlijk ontwikkeld zijn voor continue gegevens. Een typisch voorbeeld is multivariaat normale imputatie (MNI). Het is duidelijk nodig van meer adequate MI methodologie te ontwikkelen voor ordinale gegevens. We stellen daarom de ordinale multiple imputation methode (OIM) voor, gebaseerd op het proportional odds model.

Hierop verder bouwend hebben we een uitgebreid simulatie-experiment uitgevoerd, waar we het effect van verscheidene factoren op het schatten van parameters in het proportional odds model onderzoeken. De factoren omvatten het aantal categorieën in de ordinale respons, de steekproefgrootte, het aantal tijdspunten, de proportie ontbrekende gegevens, het type van non-respons (monotoon of niet-monotoon) en de vorm van de verdeling van de ordinale gegevens (zo goed als gebalanceerd versus scheef verdeeld). Ons werk toont aan dat, onafhankelijk van het patroon van ontbrekende gegevens, de schatter onder MNI zeer vertekend is, terwijl deze gestoeld op OIM bijna onvertekend is, zelfs bij gegevens met een grote proportie uitval. We konden aantonen dat MNI de onderliggende verdeling van de ordinale uitkomst in belangrijke mate wijzigt, waar dat bij OIM niet het geval is.

In het laatste deel van dit werk hebben we het probleem bestudeerd van het toetsen voor de proportional odds veronderstelling in het kader van onvolledige longitudinale gegevens. Bij MNI steeg de kans op een type I fout gevoelig boven de 5%, terwijl het statistisch vermogen in belangrijke mate toenam in vergelijking met gegevens zonder ontbekende waarden. OIM zorgt voor het omgekeerde plaatje, d.w.z. met een verlaagde kans op een type I fout en een verlies aan statistisch vermogen.

Samengevat hebben we aangetoond dat de veelgebruikte MI methodologie, gebaseerd op de multivariaat normale verdeling voor de analyse van onvolledige ordinale longitudinale gegevens zorgt voor ernstig vertekende schattingen van de proportional odds regressiecoëfficiënten. Bovendien zijn MI methoden, MNI en OIM, ongeschikt om de proportional odds veronderstelling te toetsen omdat ze de verdeling wijzingen, ofwel in het voordeel ofwel in het nadeel van de assumptie. Een toets gebaseerd op de zogenaamde complete case gegevens, d.w.z. de dataset die ontstaat na het weglaten van alle subjecten met uitval, deed het echter wel goed, ondanks een verlies aan statistisch vermogen; dit geldt ten minste in situaties waar een lichte tot matige uitval wordt geregistreerd.

Doorheen het ganse werk werden de voorgestelde methoden toegepast op werkelijk bestaande gegevens, in het bijzonder op gegevens over levenskwaliteit in een klinische studie van het EORTC betreffende kankerpatiënten. Deze gegevens vormden ook een belangrijke motivatie om dit werk aan te vatten.

Contents

Glossary

CC	Complete case analysis
CI	Confidence Interval
EORTC	European Organisation for Research and Treatment of Cancer
GEE	Generalized estimating equations
GEE1	First order generalized estimating equations
GEE2	Second order generalized estimating equations
GLM	Generalized linear models
LOCF	Last observation carried forward
MAR	Missing at random
MCAR	Missing completely at random
MCMC	Markov Chain Monte Carlo
MI	Multiple imputation
ML	Maximum Likelihood

MNAR	Missing not at random
MNI	Multivariate normal imputation
MSE	Mean square error
OIM	Ordinal imputation method
OR	Odds ratio
POM	Proportional odds model
PPOM	Partial proportional odds model
QoL	Quality of life
RB	Relative bias
SD	Standard deviation
SE	Standard error
WGEE	Weighted generalized estimating equations

General introduction

Ordinal outcome variables occur in many domains of science, such as psychology, sociology or medicine. For instance, in a clinical context, they may be used to evaluate the efficacy of a treatment (low, medium, high) or the evolution of a patient's disease (deterioration, stabilization, improvement) and in social sciences to assess people's opinion on some topic (strongly disagree, disagree, neither disagree nor agree, agree, strongly agree).

Usually, explanatory variables or other outcome variables are collected at the same time as the ordinal outcome variable. It is then of interest to assess the association between the ordinal outcome variable and these covariates. Methods to assess such association can be classified into two broad families: the non model-based and the model-based methods. While the non model-based methods generally restrict the association between the ordinal response variable with only one covariate (or another outcome variable), model-based methods allow for multivariate or adjusted analysis by considering the association between the ordinal response variable and a set of covariates. The most popular model, among the model-based methods, is the proportional odds model (POM) proposed by McCullagh (1980). This model presents interesting properties under the strict condition of identical cumulative odds ratios across the cut-offs of the ordinal outcome. A more general model, derived by Peterson and Harrel (1990), relaxes this assumption by allowing non-proportional odds for all or a subset of the covariates. To fit the adequate model, it is necessary to verify the assumption of proportional odds. This can be done in three different ways. First, a likelihood ratio test that compares the likelihood of the two models can be considered. The drawback of such an approach is that both models need to be fitted. Next, there are approaches where only one of the two models has to be fitted; these are based on Wald or score tests. Other less popular models for ordinal response variables include the continuation ratio model, the adjacent-category and the stereotype model.

In more general situations, however, ordinal variables are assessed at several occasions, as in classical longitudinal data analysis. Longitudinal ordinal data arise naturally in many clinical settings. For example, in randomized clinical trials, the regular assessment of the patient's quality of life (QoL) by means of a Likert-type scale has become popular. In presence of repeated observations for each subject, the observed ordinal responses are dependent. Thus, methods that accounts for the correlations between the repeated observations have to be considered. Extensions of generalized linear models (GLMs) to the longitudinal setting can by classified into three families, namely, the marginal, random-effects, or conditional model family. The choice of one of the three model families is guided by the objective of the study. Marginal models are appropriate when interest lies on population average; when some patients are suspected to present different behaviors, random-effects models are more appropriate; conditional model family are considered when interest lies in the effect of previous ordinal outcomes on the current ordinal response. A quite popular marginal model among the non-likelihood framework for the analysis of non-Gaussian correlated data is the generalized estimating equations (GEE) (Liang and Zeger, 1986). The GEE approach was extended to ordinal variables by considering a marginal proportional (or non-proportional) odds model to relate the response to the covariates. With the increased application of the GEE for repeated ordinal data, methods to assess the adequacy of the fitted model have become necessary. Two broad classes of goodness of fit statistics have been developed, namely those comparing observed versus predicted values (i.e. using residuals) and those measuring how well ordinal responses are predicted from the GEE approach.

Most longitudinal studies suffer from another major problem, namely missingness; subjects may prematurely drop out from the study or miss one or more follow-up assessments. As missing data usually occur for reasons outside of the control of the investigators and may be related to the outcome measurement of interest, the mechanism generating the missing values has to be considered when analysing such data. For this purpose, a classification of the missingness mechanism was introduced by Rubin (1976). In his terminology, three broad classes are defined:

missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). This classification is valid under frequentist, likelihood, or Bayesian framework, and simplification can be derived within the likelihood framework with the ignorability concept. However, as the causes of missingness are varied, the classification of the mechanisms generating missing values is difficult as it may rely on untestable assumptions, especially in complex situations such as the MNAR context. In addition, as distinction between MAR and MNAR process is questionable, the possibility that missing data followed an MNAR process should not be rejected. This can be handled by realizing sensitivity analysis. The latter however is out of scope of the present work. Various methods to handle missing data have been proposed. The most simple approach consists in discarding subjects with missing data. Furthermore, more efficient methods based on either weighting process (e.g., weighting generalized estimation equations), simple imputation methods (mean imputation, last observation carried forward) or multiple imputation methods should be preferred. Since several years, the multiple imputation (MI) methods have become a reference solution for missing data problem. In the presence of non-Gaussian data, multiple imputation based on GEE (MI-GEE) has been adopted to ensure valid results under MAR assumption. Although appropriate MI approaches are available for ordinal data, it is a common practice for researchers to impute ordinal data using MI based on the multivariate normal distribution.

As with only one observation per subject, the validity of the proportional odds assumption has to be assessed in longitudinal setting. For this purpose, the dependence among repeated observations over subject as well as the possible presence of missing data have to be accounting for.

The thesis is structured in seven chapters. Chapter 1 provides a definition of an ordinal variable. Based on the way they are collected, two broad classifications for ordinal variables are presented, namely the “grouped continuous” ordinal variables and the as “judged” or “assessed” ordinal variables. Next, some datasets that are used as illustration throughout the thesis are introduced. To conclude this first chapter, classical non model-based tests that study the relationship between an ordinal variable and another variable are summarized for different types of variables. Each situation is illustrated by an application on a real dataset.

The well-known proportional odds model is introduced in Chapter 2. Its definition and properties are first outlined. The different steps of the model fitting and associated computational issues are detailed. Then, the various ways to interpret the derived results are exposed with a focus on the cumulative odds ratio. Next, the different options, with their advantages and drawbacks, to assess the proportional odds assumption are summarized. The use of the proportional odds model is then illustrated when considering different types of covariates. Alternative solutions when the proportional odds assumption is not respected are also briefly introduced. This model, referred to as partial proportional odds model, is first described and then compared to the proportional odds model on a dataset.

Chapter 3 provides an introduction to the longitudinal setting. After a presentation of the notation, the classification of the longitudinal models within the three broad families of the marginal, random-effect and conditional model are exposed. Within the marginal models, a brief review of the GEE methods is done. Afterward, extension of the GEE to the context of ordinal outcome variable are presented. In this way, the proportional odds model as well as the partial proportional odds model are expanded and illustrated within the longitudinal setting. A GEE2 version based on the use of the global odds ratio as a measure of association is detailed and illustrated by an example. To close this chapter, hypothesis testing for the proportional odds assumption are proposed and applied on the same GEE illustrative examples.

The next chapters constitutes the original part of our work by attempting to assess the problem of missingness in ordinal longitudinal datasets when using the MI-GEE approach. Chapter 4 focuses on the problem of missingness. In this perspective, the distinction between the processes responsible for the missingness proposed by Little and Rubin is outlined. The different common issues for handling missing data are briefly summarized. Then, after a review of the theoretical background of the MI concept, we expose two MI methods commonly used with ordinal variables. The first method based on the multivariate Gaussian distribution is named as the Multivariate Normal Imputation (MNI). The second MI method, designed for ordinal variable, is built on the proportional odds model. It is referred to as the Ordinal Imputation Model (OIM) and has to be adapted to the missingness patterns present in the dataset.

The performance of both MI methods are investigated in Chapter 5 for monotone missingness patterns. In this context, a large simulation study based on the estimation of the parameters of a longitudinal proportional odds model was conducted. After a presentation of the simulation experimental plan, results derived under MNI and OIM methods are given. This chapter concludes with some recommendations regarding the imputation of longitudinal ordinal data.

The same investigation within the non-monotone missingness setting is realized in Chapter 6. The adjustments of the previous simulation plan to the non-monotone context are exposed. The results are then presented and an application of both MI approaches are also illustrated on a QoL dataset. Some advices about multiple imputation of longitudinal ordinal data in the presence of non-monotone setting close this chapter.

Finally, Chapter 7 investigates the proportional odds assumption within the MI-GEE approach. The way to combine results of the proportional odds assumption test issued from multiply imputed dataset are introduced. Then, the chapter studies the impact of both MI methods on the type I and power of the proposed test.

In summary, this thesis focuses on ordinal variables and on statistical methods used in the analysis of longitudinal ordinal datasets with missing outcome values. Emphasis is placed on the proportional odds model and the MI-GEE method. The impact of considering two different MI methods is investigated for the estimation of the parameters of a proportional odds model but also for the power of a proportional odds assumption test. Based on the results of these simulation studies, we provide some recommendations on the best way to impute ordinal data prone to missingness when using MI-GEE method.