University of Southampton

Faculty of Social Sciences

Statistical Methods for the Analysis of Ordinal Response Data

Altea Lorenzo-Arribas

Thesis for the degree of Doctor of Philosophy

June 20, 2019

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF SOCIAL SCIENCES

SCHOOL OF MATHEMATICS

Doctor of Philosophy

STATISTICAL METHODS FOR THE ANALYSIS OF ORDINAL RESPONSE DATA

Altea Lorenzo-Arribas

Ordinal response models and in particular cumulative link models are the most prevalent techniques for modelling ordered response data. This thesis examines the advantages of these models versus other approaches in the socio-economic literature and assesses the performance of available residual diagnostics measures by means of simulation studies and four case studies. Furthermore, it proposes solutions to specific issues of flexible versions of cumulative link models.

Contents

List of Figures

List of Tables

Author's Declaration

I, Altea Lorenzo-Arribas, declare that the thesis entitled

Statistical Methods for the Analysis of Ordered Response Data

and the work presented in the thesis are both my own, and have been generated by me as the result of my own original research. I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. None of this work has been published before submission.

Signed: .....

Date: .....

Acknowledgements

I would like to thank my supervisors Dr Antony M. Overstall and Dr Mark J. Brewer for their help and support, and Ame, Marce and Dom for their immense patience.

Chapter 1

Introduction

1.1 Ordinal data

Ordinal responses are common in social and economic research, where variables are often measured in the form of categories or by means of scales (e.g., Likert, 1932). Examples include measures of wellbeing and opinion surveys (Agresti, 1977). An ordinal or ordered scale, as defined by Guisan and Harrell (2000), is an ordering of values with no natural unit of measurement and only relative instead of quantitative differences between them. These ordered items convey information about the relationship between values (e.g., that one value is greater than another), but they do not indicate how much greater a value is. For instance, although 'excellent' is greater in value than 'very good', one cannot say with certainty that the distance between those two values is the same, less, or more than the distance between 'very good' and 'good'. Ordinal variables are typically obtained with ordinal scales that include closed-ended response categories in which the categories are labelled using words, numbers, or some combination of both. In order to make an informed decision on the type of model to be used for the analysis of this type of response data, further understanding of the way these data have been generated is required (Greenland, 1994, Lall et al., 2002). We summarise the most common types of ordinal data according to the subject providing the measure and the generating process in Table 1.1 as per definitions in Agresti (2010), Anderson (1984), Ishwaran and Gatsonis (2000), Lall et al. (2002), and O'Brien (1985).

Tailored methods exist to accurately model these ordinal data when they are response variables. The extent of application of different methods varies between disciplines. For instance, models incorporating grouped continuous variables in the response are very frequent in economic and market research data sets. In contrast, categories assessed by an observer are very common in medical research and require the use of particularly flexible models. Response outcomes measured in sequential processes require a model that takes the stages

Table 1.1: Types and provenance of ordinal data.

Agent evaluation		
Type of ordinal variable	**Description**	**Example**
Assessed variable	Values assessed by the subject of the study.	Self-reported quality of life.
Measured variable	Values directly measured by the subject of the study.	Walking frequency over the last month.
Reported variable	Values assessed by an observer.	Clinician assessing the density of a tumour on a patient's scan.
Process provenance		
Type of ordinal variable	**Description**	**Example**
Grouped or binned continuous variable	Observations ranked according to a criterion.	Cities classified according to population size.
Sequential variable	Values derived from stages in the sequential process.	Educational attainment.

into account and those that might incorporate a trend require specific modelling too (as we will see in Chapter 3). Different biases are also introduced by their specific characteristics. For instance, both assessed and reported variables are likely to present greater observer error and subjectivity than grouped continuous variables (Colombi et al., 2018).

1.2 Ordinal response data modelling

Methodological debate over how to treat variables measured on an ordinal scale is ongoing, much of it centred around whether (i) ordinal variables can be safely treated as continuous variables subject to ordinary linear modelling techniques as described above, (ii) whether ordinal variables require special statistical methods or (iii) whether ordinal variables should be replaced with truly continuous variables in causal models (Agresti, 2010, Kuzon et al., 1996). The linear model approach treats the categories' scores as real values, uses a standard regression to learn a real-valued function, and finally rounds to the closest label when generating predictions (Kramer et al., 2001). There is considerable support for this idea that methods for continuous variables should be used for ordinal variables because the power and flexibility gained from these methods offset any small biases incurred (Capuano, 2012, Knapp, 1990, Torra et al., 2006, Winship and Mare, 1984). However, this cannot be stated for those scale items with a small number of categories (fewer than five), because by as-

suming certain distances between the few categories, we would be likely to incur substantial errors (Rhemtulla et al., 2012). Additional arguments, such as lack of understanding or easier interpretation and tendency towards less resource intensive techniques, are also used in favour of continuous approaches (Long, 2014). Gaito (1980) goes further by sticking to Lord's principle that "the numbers don't know where they came from" and arguing that the nature of the data is irrelevant to the analysis. Moreover, some authors argue that the fact that we cannot make inferences about differences in the latent variable underlying a Likert scale, does not mean that it invalidates conclusions about the numbers (Norman, 2010).

According to Lall et al. (2002) the way the data is generated can be accommodated in a given ordinal response model providing more accurate and refined results. Furthermore, Hawkes (1971) and O'Brien (1982) suggest that the biases in using a continuous approach for ordinal variables are large and that special techniques for ordinal variables are required that reflect category-specific characteristics. Bias can occur in the estimation driven by the so-called 'ceiling and floor effect' that is associated with the range of values of the responses (Agresti, 2010, Heeren and D'Agostino, 1987). This effect becomes particularly prominent when the distribution of the ordinal data is highly skewed due to the fact that one of the potential outcomes is either rare or too frequent. In these cases, predictions of the 'continuous' regression can lie outside the range of categories of the outcome variable, which invalidates inference. It can also imply violation of the normality assumption. This limitation is overcome by ordinal models by 'naturally' bounding the predictions to lie within the response values range (Capuano, 2012). Other advantages of these models when compared to the above mentioned pragmatic approach include statistical power improvement (often described in the literature; e.g., Capuano et al., 2007, and Kosmidis, 2014). Specific benefits in terms of results (both significance and estimation) and in terms of interpretation (more accurate and based in thresholds, in the comparison of effects across categories) are also of great relevance (Capuano, 2012) and will be show in Chapter 4.

Other approaches in the literature outside of the scope of this thesis include: collapsing of categories (Bender and Grouven, 1998, MacCallum et al., 2002) - criticisms of its application (Agresti, 1977, Armstrong and Sloan, 1989); sliding dichotomy (dichotomy defined depending on the baseline condition of the patient in clinical trials; Ilodigwe et al., 2013, Murray et al., 2005); non-parametric versions (Agresti, 2010, Munzel and Langer, 2004) and their limitations (Singer et al., 2004); mixture models (CUB) (Iannario and Piccolo, 2012); and optimal numerical mapping (Torra et al., 2006). Finally, multinomial logit models are often considered attractive to model ordinal data because they do not assume normality, homoscedasticity, or linearity between the dependent and the independent variables (although they assume linearity between the logits and the independent variables). However, they

make no use of information about the ordering of the categories of the response variable and might "include a lot of extraneous and unnecessary parameters" (Williams, 2016).

1.2.1 The latent variable approach

The latent variable approach appropriately accounts for the ordinal variables' ordered nature by assuming a continuous underlying distribution (e.g., degree of happiness) which corresponds to a discontinuous ordinal-level observed distribution (typically measured on a Likert scale).

Suppose there are n responses and C categories where $Y_i \in \{1, \ldots, C\}$ and $Y_i^* \in \mathbb{R}$ are the response and latent response, respectively, from the ith unit. Let $\alpha_1, \ldots, \alpha_{C-1}$ be $C-1$ thresholds (or cut-points), the ordinal response variable is defined such that

$$Y_i = \begin{cases} 1 & \text{if } Y_i^* < \alpha_1; \\ 2 & \text{if } \alpha_1 \leq Y_i^* < \alpha_2; \\ \vdots & \vdots \\ \text{C} & \text{if } Y_i^* \geq \alpha_{C-1}. \end{cases} \tag{1.1}$$

Figure 1.1 shows an example with $C = 5$ response categories determined by 4 cut-points.

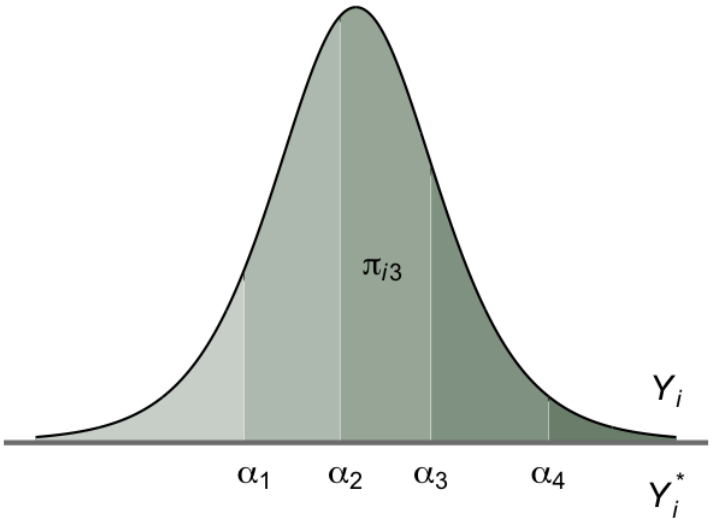

Figure 1.1: The latent variable approach to modelling an ordinal response Y_i with probability $\pi_{i3} = P(Y_i = 3)$ determined by the cut-points in the latent continuous variable Y_i^*, α_3 and α_2.

The existence of an underlying continuous variable is assumed both at the model fitting and interpretation stage, and by definition, its existence cannot be proved or disproved (Agresti, 2010, Fielding, 1999). However, some authors (Yule and Kendall, 1950) consider it as a misleading and artificial tool, and understand an ordinal model instead as "a nonlinear probability model without appealing to the idea of a latent variable" (Long and Freese, 2006). They also claim that there are serious issues associated with the measurement error derived from using ordinal measures of continuous underlying variables (O'Brien, 1985, Winship and Mare, 1984). Boes and Winkelmann (2006) further claim that in most applications of ordered response models the parameters of the latent model do not have direct interpretation *per se*. We argue that on the contrary the latent variable approach is the most meaningful and appropriate framework for the correct modelling of ordinal response data in socio-economic research. Researchers in this area routinely work with latent constructs as part of other techniques (e.g., structural equation modelling) and therefore find this approach to ordinal modelling and its interpretation particularly intuitive whereby they understand the response variable under study (e.g., wellbeing) as a combination of two parts (i.e., the underlying satisfaction level and the measurement process corresponding to the response categories; Bollen, 2002, Greene, 1993). Assessment of the impact of question formulation, rating scale definition and respondents understanding on the observed level of the response variable across the categories (Grilli and Rampichini, 2012) is beyond the scope of this thesis.

1.2.2 Uptake of ordinal models

Despite ordinal data being very common, this has not led to a widely adopted statistical modelling approach. A tendency to model ordinal response data as continuous data on the basis of an easier application and interpretation is noticeable in the area of socio-economics. We explore the reasons behind this low uptake and try to address these limitations as exemplified in the case studies described in Chapter 2.

Challenges in the interpretation of the results appear to have had an effect on the uptake of these models. The most natural way to interpret ordered response models (and discrete probability models in general) is to determine how a marginal change in one regressor changes the distribution of the outcome variable, i.e., all the outcome probabilities. As described in the previous section, difficulties arise in the interpretation of the latent variable which is often described as not "of intrinsic interest" (Greene and Hensher, 2010) as it cannot be observed and is purely artificial. A tendency in the literature to consider cut-points as "nuisance parameters of little interest" (Liu et al., 2009) has also been noted. We aim to emphasise the relevance of these thresholds and show the impact of the different distributions of these

cut-points on inference (see Chapter 4) and goodness of fit (see Chapter 5). The latter chapter also looks at the specific difficulties in interpreting the residual diagnostics for these models. The challenges above get magnified when the estimation produces a high number of parameters, which is the case for partial proportional odds models as described in Chapter 6.

We argue that further reasons for the low uptake in the area relate to the lack of software availability and accessibility. As we discuss in Section 3.6, despite more open software packages being available for ordinal response data models, none of them can handle the different specifications mentioned in this thesis altogether, and user-friendliness of these (both at the specification and output interpretation) is limited.

1.3 Contribution of thesis

A systematic review of the models for ordinal response variables in the wider literature as well as a exploratory background research trying to identify and gain a better conceptual grasp of the gaps in the socio-economic literature is presented in Chapter 3. This literature review is organised to cover methodology, applications and open software availability. We argue that an appropriate modelling framework is particularly required in the socio-economic area and provide the reasoning behind our claim in Chapter 4 by means of a couple of simulation studies and case studies (introduced in Chapter 2) highlighting inference related advantages of cumulative logit models. Chapter 5 presents the challenges found at the model checking stages and highlights the most appropriate residual diagnostics strategy when using cumulative logit models by performing a systematic review of three ordinal residuals. Chapter 6 explores issues with partial proportional odds models, specifically the prediction of negative probabilities, and proposes solutions to these as well as model selection strategies. Finally, Chapter 7 highlights areas for future work that will aim to fill the gaps and limitations in the currently available approaches. All the code used for the results and figures presented in this thesis is fully reproducible and can be found at: `https://github.com/altealo/Ordinal-response-data-modelling`.

Chapter 2

Case studies

2.1 Introduction

In this chapter we introduce the three case studies in the area of socio-economics that motivated this thesis. In addition, we introduce a data set that allows us to implement an extensive model selection exercise with a high number of covariates.

2.2 Connectedness to nature

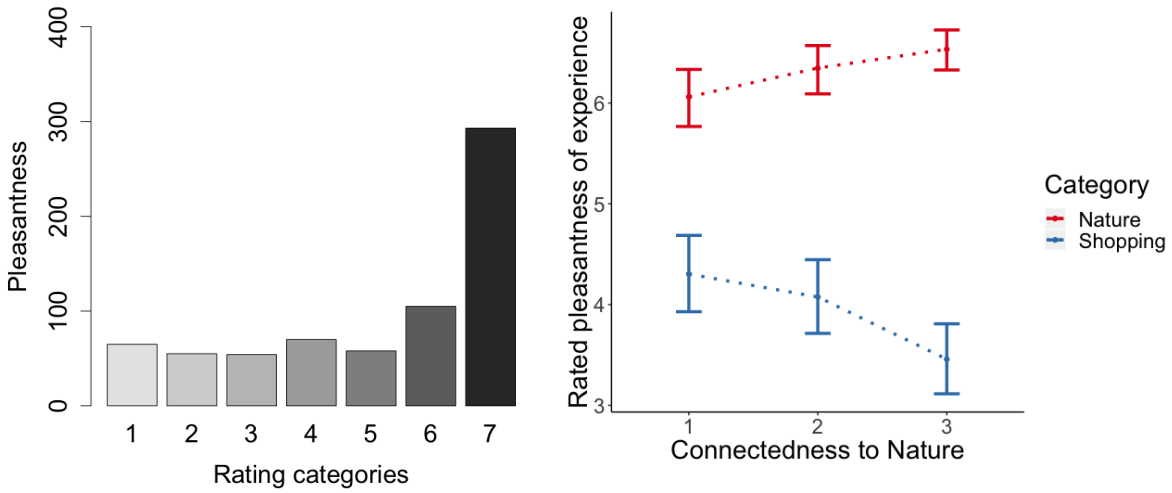

Figure 2.1: Connectedness to nature data set - *Pleasantness* ratings (left) and interaction of *experience* and *connectedness to nature* (right).

Data from Craig et al. (2018) evaluating shopping and nature experiences, and their association to other relevant cognitions and behaviours. $n = 357$ individuals were asked to describe a memory they had of an "everyday experience involving nature", as well as a "shopping experience". Both experiences were assessed by the respondents in relation to:

- rating of the *pleasantness* of the experience, 7-category semantic differential scale, ranging from unpleasant to pleasant, and
- another 6 ratings that will not be analysed in this thesis, e.g., *familiar-unfamiliar*, and *active-passive*.

In addition to these ratings, respondents reported their *connectedness to nature*, which in the original analysis is a binned 3-category variable derived from their responses to a 13-item scale.

Figure 2.1 (right) represents the median values of *pleasantness* for the two types of *experience* and the different levels of *connectedness to nature*. For the *nature* category, when the *CNS* level increases, the rating of *pleasantness* also increases, whereas the opposite is apparent for the *shopping* category, which might indicate the presence of an interaction between the two covariates. We analyse this case study in detail in Chapter 4.

2.3 Species desirability

Data from a survey conducted with members of the general public in study sites in 8 European countries, with sample sizes of around 300 per country (total $n = 2,378$) as reported in Fischer et al. (2011). Each respondent was asked to rate the desirability of an increase in population of three different species (plant, spider, large mammal) on a $C = 5$ semantic differential scale ranging from -2 (very undesirable) to 2 (very desirable), where 0 corresponds to neutral responses but also includes "don't know" responses following an *ad hoc* approach by Fischer et al. (2011) and therefore ignoring 'partial order' (Peyhardi et al., 2016; Zhang and Haksing Ip, 2010). The three different species were chosen such that respondents were likely to be familiar with them, given that they occurred in the study sites. For each study site, the authors selected (a) a large, charismatic, but potentially controversial mammal species - in most sites this was either red or roe deer, while in France the wolf was chosen, (b) a garden spider as an example for an invertebrate prevalent in all eight sites, distinctive enough to be known due to the 'cross' on its back, but not particularly charismatic, and (c) a non-native plant species, visible and widespread, but not necessarily known as non-native in the study site. For each species, respondents also provided their perceptions on whether the species:

- had *decreased or increased* in their country in the past 20 years,
- is *attractive or unattractive*,
- is *strong or vulnerable*,
- is *valuable or worthless*,
- is *common or rare*,
- is *harmful or harmless*, and
- is *foreign or native*,

all recorded on a semantic differential scale from -2 to +2. The beliefs were hypothesised to explain variation in the response variable, namely, the desirability of a species' population increase. The corresponding analysis explores the relationships between perceived desirability of the 3 species and the 6 species-related characteristics. We analyse this case study in Chapter 4.

2.4 Pro-environmental attitudes

Data from the "Scottish Environmental Attitudes and Behaviours Survey" (Ipsos MORI Scotland and Scottish Government, 2009) which evaluates respondents' awareness of environmental issues and their greener behaviour including; knowledge and attitudes towards climate change, travel behaviour, efficiency in the home, eco-friendly purchasing, and greenspace and wellbeing. The study found that high environmental engagement is more concentrated among certain groups in the population, with *educational attainment*, *social class*, and *age* being the strongest covariates of engagement.

Within this data set ($n = 3,054$) we model *educational attainment*, i.e., highest level of qualification obtained with $C = 5$ categories:

$$\text{none} < \text{O-level, standard} < \text{higher, A-level} < \text{HNC/HND} < \text{degree, professional}$$

versus *age* via a PPOM (see Figure 2.2 for a visualisation of the distribution of these categories and Figure 6.17 for a representation of the model's fitted probabilities). Although we acknowledge that for an appropriate analysis of the data, we would need to control for other covariates (e.g., *sex*), for the purposes of this methodological study, we look at one covariate only. We have inverted the original order of categories in educational attainment in order to have an intuitive interpretation (1 =none, 2=O-level, standard or equivalent, 3=

higher, A-level or equivalent, 4= HNC/HND or equivalent, 5= degree, professional) and have eliminated the 'missing/refused' category. We show in Chapter 6 that there are problems with this ordinal response model which lead to negative predictions of class probabilities.

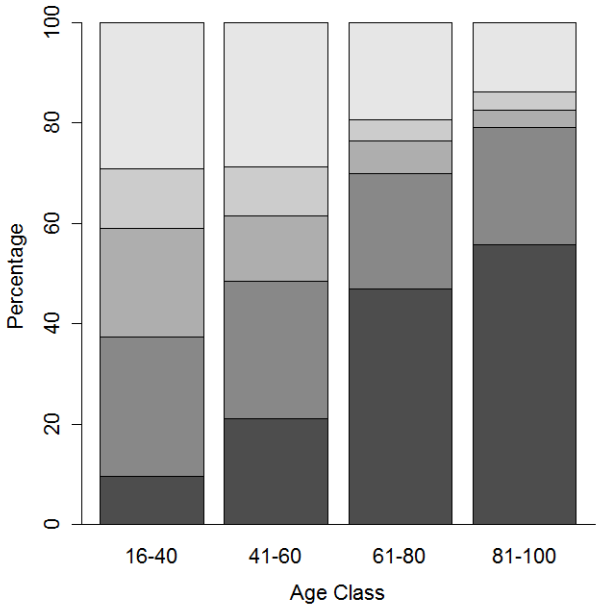

Figure 2.2: Pro-environmental attitudes data set - *educational attainment* category frequencies (percentage; different categories represented by shades of grey), for different *age* groups (created for this plot only).

2.5 Retinopathy

Data from the "Wisconsin Epidemiological Study of Diabetic Retinopathy" (Agresti, 2010, Archer et al., 2014) available in `R` package `ordinalgmifs`. The response under study is *severity of retinopathy* which was measured in the left and right eye of $n = 720$ subjects in $C = 4$ categories:

$$\text{none} < \text{mild} < \text{moderate} < \text{proliferative},$$

in order to assess the prevalence of diabetic retinopathy and determine the factors that influence the severity of the disease. Other variables included in the data set are

- *lerl*, *rerl*: left and right severity of retinopathy,
- *lme*, *rme*: left and right eye macular oedema,
- *lre*, *rre*: left and right eye refraction index,

- *liop*, *riop*: left and right eye intra-ocular eye pressure,
- *sbp*, *dbp*: systolic and diastolic blood pressure,
- *pr*: pulse rate,
- *prot*: presence of proteinuria,
- *gh*: glycosylated haemoglobin level,
- *diab*: duration of diabetes (in years), and
- patient characteristics including: *sex*, *age*, and *bmi* -body mass index.

Ignoring individual effects, we focus initially on the left eye measure and model it with respect to two covariates; *systolic blood pressure* and *left eye refraction index* (see variables' relationship represented in Figure 2.3) via two types of ordinal models in order to determine which one is a better fit.

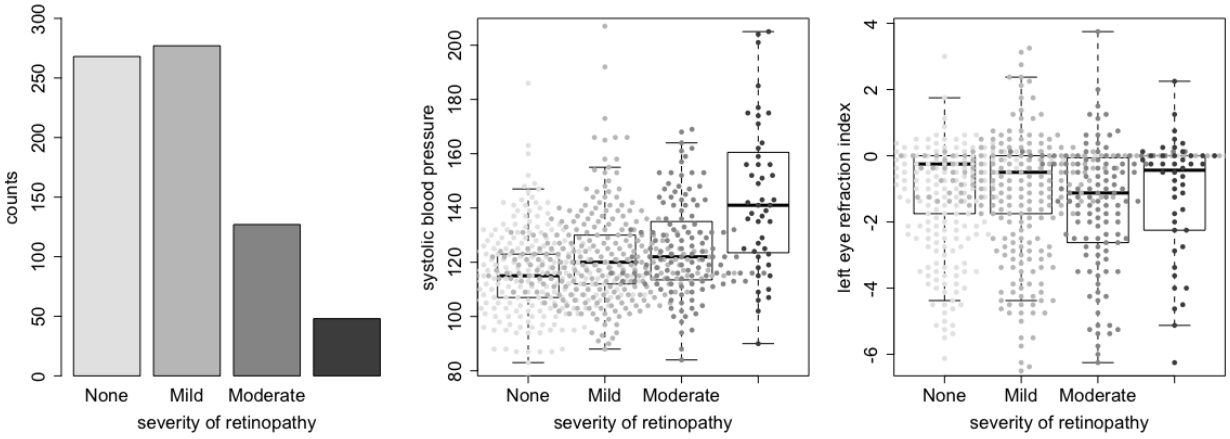

Figure 2.3: Retinopathy data set - *severity of retinopathy* category frequencies (left), and relationship with *systolic blood pressure* (middle) and *left eye refraction index* (right).

We also consider an exercise of model selection with this case study in Chapter 6.

Chapter 3

Previous work

3.1 Introduction

There is a vast body of literature concerning the use of ordinal data. This chapter reviews publications that pose particular relevance to the modelling of ordinal response data, both from methodological and applied perspectives, with a focus on socio-economic applications. We mainly focus on four areas according to the specific needs of ongoing research; types of ordinal models, methods of estimation and inference, software, and topics for further research. The ultimate aims of this literature review are twofold: (i) to find gaps for further development and (ii) to determine the limitations of the existing software in terms of implementation of the proposed techniques. Models with ordered variables acting as covariates are not considered in this review.

Boes and Winkelmann (2006) claim that regression models for ordered responses have their origin in the biometrics and medical literature. Initially, the ordered probit model was proposed by Aitchison and Silvey (1957) "to analyse experiments in which the responses of the subjects to doses of stimulus" are classified into ordinal categories. Later, Snell (1964) suggests the use of the logistic instead of the normal distribution in the assessment of disease severity levels, an area in which these models have acquired great relevance (for a recent example of medical imaging ordinal analysis see Saffari et al., 2015) as they also have done in epidemiology (Armstrong and Sloan, 1989).

In the case of the social sciences, the first comprehensive treatment of ordered response models is attributed to McKelvey and Zavoina (1975) who extend the model of Aitchison and Silvey (1957) to more than one independent variable and apply it to the analysis of individual data as opposed to the central focus of applications within medical science, namely grouped data and the analysis of proportions. McCullagh (1980) independently develops the so-called 'cumulative link model' in the statistics literature.

In economics and market research the user rating for a new product (Resnick and Varian, 1997) or the reporting of the level and usage of insurance (Jeliazkov and Rahman, 2012) are both situations in which these models prove useful. There has been an increasing emphasis in medical research on the design and analysis of quality of life scales, which often imply the need for ordinal response modelling (Lall et al., 2002). Epidemiologists also use these models when estimating the risk of diseases, which have a natural ordering of severity or certainty (Ananth and Kleinbaum, 1997). These models are also present in cancer diagnostics, where ordered categories can be assigned to the stage of the spread of the illness. Cumulative link models are known in psychometrics as graded response models (Samejnima, 1969) or difference models (Thissen and Steinberg, 1986). Finally, ordinal regression models have often been neglected in past ecology studies (Guisan and Harrell, 2000). However, it has been widely recognised that examples such as ordinal abundance studies in plant distribution modelling are areas where the correct modelling approach offers significant benefits.

We highlight the most widely used models for ordinal response data modelling in the following sections and include an overall classification in Table 3.1 [1] based on (Agresti, 2013, Ananth and Kleinbaum, 1997, Armstrong and Sloan, 1989, Brant, 1990, Fullerton, 2009, Greenland, 1994, Peyhardi et al., 2015, Tutz, 2012, 1991).

3.2 Alternative ordinal response data models

Different families of ordinal models exist that are more appropriate for modelling specific types of ordinal data, according for instance to the generating process (as highlighted in Table 1.1 in Chapter 1) or even to the area of application. As a preamble to the next section where we describe in detail cumulative link models, we highlight polytomous, continuation ratio models, and trend odds models as particularly relevant.

3.2.1 Polytomous models

Polytomous models (PMs in Engel, 1988; also known as baseline-category logit models by Agresti, 2010) are extensions of the logistic regression model and are designed to analyse nominal scales where there are several categories. Given an ordered response variable Y_i with C categories and n observations, logits are formed in these models by comparing each

[1] With the only exception of the stereotype models, all the model specifications in the same row of the table have the same linear predictor.

response category j to an arbitrarily chosen baseline category b such that

$$\text{logit}(P(Y_i = j|Y_i = j \quad \text{or} \quad Y_i = b)) = \log\left(\frac{\pi_{ij}}{\pi_{ib}}\right) = \alpha_j + \sum_{k=1}^{p} \beta_{jk} x_{ik}, \\ i = 1, \ldots, n; j = 1, \ldots, C-1, \tag{3.1}$$

where π_{ij} and π_{ib} are the multinomial probabilities corresponding to a particular response category j and the baseline category b. By definition, they fail to reflect ordinality and lack parsimony as each covariate x_{ik} has $C-1$ parameters.

3.2.1.1 Stereotype model

The standard polytomous model can be modified to account for ordinality in the form of the stereotype model (SM; Anderson, 1984, Lunt, 2001). A stereotype model is a paired-category logit model defined such that

$$\text{logit}(P(Y_i = j|Y_i = j \quad \text{or} \quad Y_i = b)) = \log\left(\frac{\pi_{ij}}{\pi_{ib}}\right) = \alpha_j + \phi_j \sum_{k=1}^{p} \beta_k x_{ik}, \\ i = 1, \ldots, n; j = 1, \ldots, C-1, \tag{3.2}$$

where the coefficient $\phi_j \beta_k$ for the covariate x_{ik} represents the log odds ratio for categories j and b of Y with a unit increase in x_{ik}, and the ϕ_j parameters can be regarded as scores for the response categories with imposed constraints $1 = \phi_1 > \phi_2 > \cdots > \phi_{C-1} > \phi_C = 0$, for identifiability (Agresti, 2010). Given the scores ϕ_j, this model requires a single parameter β_k per covariate x_{ik} (similar to the definition of proportional odds assumption in Section 3.3). This makes it more parsimonious than the polytomous model while more flexible than the yet to be defined proportional odds models, and therefore more suited for ordinal data in medical contexts according to Bender and Grouven (1997).

3.2.1.2 Adjacent category model

The adjacent category model (ACM) is a specific form of generalised logit model for multinomial outcomes (Hosmer and Lemeshow, 2000) and it involves simultaneous estimates of the effects of covariates in pairs of adjacent categories, defined as

$$\text{logit}\left(P(Y_i = j|j \leq Y_i \leq j+1)\right) = \log\left(\frac{\pi_{ij}}{\pi_{i,j+1}}\right) = \alpha_j + \sum_{k=1}^{p} \beta_k x_{ik}, \\ i = 1, \ldots, n; j = 1, \ldots, C-1. \tag{3.3}$$

The effects of covariates can be constrained to be constant for comparisons of adjacent categories as it is done with the assumption of proportionality/parallelism of the cumulative link models (Section 3.3) and continuation ratio models (Section 3.2.2), respectively. Goodman (1979) refers to this assumption as uniform association (as reflected in Table 3.1). For those cases in which this assumption is not valid for one or some of the covariates represented by z_{ik}, we can also have a partial uniform version which can be defined as

$$\text{logit}\left(P(Y_i = j | j \leq Y_i \leq j+1)\right) = \log\left(\frac{\pi_{ij}}{\pi_{i,j+1}}\right) = \alpha_j + \sum_{k=1}^{p} \beta_k \boldsymbol{x}_{ik} + \sum_{k=1}^{q} \gamma_{jk} \boldsymbol{z}_{ik}, \quad i = 1, \ldots, n; j = 1, \ldots, C-1. \tag{3.4}$$

with $\gamma_{1k} = 0$ for all $k = 1, \ldots, q$ for identifiability.

3.2.2 Continuation ratio models

The continuation ratio model (CRM; Fienberg, 1980) is defined by

$$\text{logit}(P(Y_i = j | Y_i \geq j)) = \alpha_j + \sum_{k=1}^{p} \beta_k x_{ik}, \quad i = 1, \ldots, n; j = 1, \ldots, C-1. \tag{3.5}$$

CRMs are usually relevant when the ordinal variable may be thought of as a latent continuous variable progressing through various stages (Kosmidis, 2014). For instance, an example of outcome such that progression through the ordinal levels cannot be reversed is stage of cancer (Archer et al., 2014, Greenland, 1994). It has a strong connection to discrete survival analysis and the Cox proportional hazard model where we can understand the categories as referring to time intervals and we can define a discrete hazard function (Agresti, 2010, Harrell et al., 1998, Poßnecker and Tutz, 2016).

The above properties require using a link function in the model that can produce a non-symmetric response function, for example, the log-log function. This aims to reflect the fact that the outcome of this type of model depends on the direction chosen to model the variable (e.g., increasing or decreasing severity of cancer; Abreu et al., 2008).Other versions of this model are briefly described in Table 3.1.

3.2.2.1 Sequential model

The sequential model (SQM; Tutz, 1990) or unconstrained continuation ratio model is a specific case of the continuation ratio model where we alsodefined sequential logits such that

$$\text{logit}(P(Y_i = j|Y_i \geq j)) = \alpha_j + \sum_{k=1}^{p} \beta_{jk} x_{ik}, \quad i = 1, \ldots, n; j = 1, \ldots, C-1, \tag{3.6}$$

and where the effects are again the same for each logit. However, it provides greater flexibility than the (constrained) CRM by allowing the effect of each explanatory variable to depend on the response level (Fagerland and Hosmer, 2016).

The partial version of this model would be expressed as

$$\text{logit}(P(Y_i = j|Y_i \geq j)) = \alpha_j + \sum_{k=1}^{p} \beta_k x_{ik} + \sum_{k=1}^{q} \gamma_{jk} z_{ik}, \quad i = 1, \ldots, n; j = 1, \ldots, C-1, \tag{3.7}$$

with $\gamma_{1k} = 0$ for all $k = 1, \ldots, q$ for identifiability. By means of the parameters γ_{jk} we allow z_{ik} to contribute separate effects for each logit (Cole and Ananth, 2001). This model conveniently predicts valid category probabilities regardless of the range of values of the covariates (Agresti, 2010).

3.2.3 Trend odds models

The trend odds model (TOM) allows the odds parameter to increase or decrease in a monotonic manner across the cut-points (TO assumption; Capuano and Dawson, 2013) and is defined such that

$$\text{logit}(P(Y_i \leq j)) = \alpha_j + \sum_{k=1}^{p} (\beta_k + \gamma_k \tau_{jk}) x_{ik}, \quad i = 1, \ldots, n; j = 1, \ldots, C-1, \tag{3.8}$$

where for non-zero γ_k we have a trend in the log-odds ratio, with $\tau_{1k} = 0$ for all $k = 1, \ldots, q$, and $\tau_{jk} < \tau_{j+1,k}$ reflecting the monotonic structure of the model.

TOMs can also be understood as structured constrained cumulative link models (which will be described in the next section). According to Capuano (2012), the added complexity of these models provide improved power over partial proportional odds (that we will define next) when there are moderate to severe departures from proportionality. Since the model assesses the presence of a trend in the odds, it is consequently seen as particularly useful in the assessment of risk factors and disease etiology analyses (similarly to SQMs).

In summary, the choice of an appropriate family of ordinal models to fit our ordinal

response data will heavily depend on the type of data we are working with. SMs are recommended for their overall flexibility (as they are not dependent on the proportional odds assumption holding), ACMs are particularly useful when the interest lies in comparing adjacent categories rather than assessing differences with respect to a baseline, CRMs are specifically suited for cases where the individual under study can only continue to higher levels (categories) of the outcome if they have gone through lower levels and similarly, TOMs are recommended when there is a clear trend in the odds (therefore not requiring the proportional odds assumption to be met). In contrast to cumulative link models, ACMs and CRMs model conditional log-odds, i.e., the log-odds of being at a certain level given that they have been at other levels.

3.3 Cumulative logit models

In this thesis we will focus on cumulative link models that are based on the cumulative (or accumulated) response probabilities that denote the probability that a randomly selected observation falls in the jth category of a variable (Agresti, 1981). As such, they reflect the ordering of the values of the response variable (unlike PMs). Specific advantages, when compared to the alternative approaches mentioned above, include statistical power improvement (frequently highlighted in the literature; e.g., Capuano et al., 2007, Guzman-Castillo et al., 2015, and Kosmidis, 2014). Other benefits are; improved inference and a more straightforward interpretation (more accurate and based on the comparison of effects across categories). They are also more generic and flexible than process-specific models such as ACMs and CRMs (defined in previous section).

Consider an ordinal variable Y_i with observations $i = 1, \ldots, n$ and C categories, understood as a categorised version of a latent continuous variable $Y_i^* = \sum_{k=1}^{p} \beta_k x_{ik} + \varepsilon$ (as defined in (1.1)), with ε having a cumulative distribution function (cdf) of some standard form G. Suppose that the assumed model for Y is in a class of cumulative link models. Such model defined as

$$G^{-1}(P(Y_i \leq j)) = \alpha_j + \sum_{k=1}^{p} \beta_k x_{ik}, \tag{3.9}$$

links the cumulative probabilities to a linear predictor, where $j = 1, \ldots, C-1$ and G^{-1} is called the link function. We require that G^{-1} be monotonic and differentiable over the range of $E[Y_i] = \mu_i$. The threshold parameters are constrained such that $-\infty = \alpha_0 < \alpha_1 < \ldots < \alpha_{C-1} < \alpha_C = \infty$ and β_k are the covariate parameters to be estimated. This is basically reflecting the underlying latent distribution (as we first introduced in Section 1) such that

$$P(Y_i \leq j) = P(Y_i^* \leq \alpha_j) = G\big(\alpha_j - \sum_{k=1}^{p} \beta_k x_{ik}\big). \tag{3.10}$$

From a statistical perspective, the choice of link function (or latent variable distribution) is essentially a choice of scale on which the linear and additive effects operate. Typical choices of the link function for these models are logit, probit and complementary log-log. Other less known options include scobit (or skewed-logit) as proposed by Nagler (1994), which is shown to be appropriate where individuals with any initial probability of choosing either of two alternatives are most sensitive to changes in independent variables, but we will not be considering these models in further detail. We are aware of the fact that when using complementary log-log we would be dealing with proportional hazards, but this is also outside the scope of this thesis. A choice may often be made *a priori* on substantive grounds or for ease of interpretation (e.g., relative odds or hazards). The close agreement of the standard logistic and normal distributions over most of the range leads to the idea that logit or probit links will lead to similar results. Differences may be expected only where end category probabilities are small. Again, there is some theoretical consensus on this (e.g., Anderson and Philips, 1981a), and in much applied statistics literature, although Ameniya (1981) has expressed concerns (particularly in cases where data are heavily concentrated in the tails due to the characteristics of the problem under study or in multiple response or multivariate models). In general, logit is the most widely used link function because of its tractability, its computational convenience and its connection with odds ratios which make its interpretation intuitive (Fielding, 1999). Genter and Farewell (1985) show that the links may be discriminated for moderate sample sizes, although the different choices seem to be closely related to disciplines. Additionally, it has been shown that it ensures robust inference (Iannario et al., 2017). We therefore focus on ordinal response models with the logit link function, which will allow us to produce odds ratios by exponentiation of the model estimates. Thus, in the following subsections we define different cumulative logit models and associated concepts.

In cumulative logit models (CLMs hereafter) we have that G is the *cdf* of the standard logistic distribution such that $G(\varepsilon) = e^{\varepsilon}/(1 + e^{\varepsilon})$ and G^{-1} corresponds to the logit link function, therefore we can define the model as

$$\text{logit}(P(Y_i \leq j)) = \alpha_j - \sum_{k=1}^{p} \beta_k x_{ik}, \tag{3.11}$$

for a response variable Y_i with C ordered categories. The thresholds $\boldsymbol{\alpha}$ can be constrained to be either symmetric or equidistant based on the characteristics of the data (as we will see

in Chapter 4). Additionally, the cumulative logits are defined as

$$\text{logit}(P(Y_i \leq j)) = \log\left\{\sum_{k=1}^{j} \pi_{ik} / \sum_{k=j+1}^{C} \pi_{ik}\right\}, \quad i = 1, \ldots, n; j = 1, \ldots, C-1, \tag{3.12}$$

where $P(Y_i \leq j)$ is the cumulative probability that a given observation is less or equal than the jth level and $\pi_{ik} = P(Y_i = k)$, for $k = 1, \ldots, C$ (Agresti, 2010).

3.3.1 Proportional odds model

In the framework of the CLMs, a proportional odds model (POM hereafter) is defined as

$$\text{logit}(P(Y_i \leq j)) = \alpha_j + \sum_{k=1}^{p} \beta_k x_{ik}, \quad i = 1, \ldots, n; j = 1, \ldots, C-1, \tag{3.13}$$

where Y_i is interpreted in terms of the same covariate effect β [2] for each response category, independently of the underlying continuous latent variable Y_i^* thresholds $-\infty < \alpha_1 < \alpha_2 < \ldots < \alpha_{C-1} < \infty$ (Gill and Casella, 2009). Additionally, $Y_i = j$ if $\alpha_{j-1} < Y_i^* \leq \alpha_j$ (Agresti, 2010). The regression parameters β_k and α_j are unknown and therefore estimated.

The proportional odds (PO) assumption imposes equal slopes for the CLM (parallel regression lines; see Figure 3.1), or equivalently, the effects of the covariates β are assumed to be constant across the response variable categories.

Alternatively, we can rewrite (3.13) as

$$\text{logit}(P(Y_i \leq j)) = \alpha_j - \sum_{k=1}^{p} \beta_k x_{ik}, \quad i = 1, \ldots, n; j = 1, \ldots, C-1. \tag{3.14}$$

The negative sign before the linear predictor in the model implies that increasing a covariate x_{ik} with a positive β_k is associated with a shift towards the right-hand end of the response scale, namely, an increase in the probabilities of the higher categories (Grilli and Rampichini, 2012). This notation is often preferred in software because it makes the sign of each component of β have the usual interpretation in terms of whether the effect is positive or negative. However, we will use the positive notation in (3.13) hereafter.

The POM provides a single estimate of the log odds ratio over the cut-points which is not a weighted average of the cut-point-specific log odds ratios, but is the optimum estimate obtained using the maximum likelihood (ML hereafter) or weighted least squares methods

[2] The β coefficient is interpreted as the effect of a 1-unit change in x_{ik} on the log-odds of being in a lower category rather than a higher category of Y_i.

(see subsection 3.4). "The information that is contained in the ordering of the response categories is typically exploited by specifying a single parameter per covariate. The corresponding covariate therefore has a global effect that is not specific to the considered response category" (Poßnecker and Tutz, 2016). It is ideal in terms of the ease of interpretation of the data and in terms of model parsimony (Abreu et al., 2008). Other properties of this model include: invariability to direction of the dependent variable modelling (e.g., lower to higher or higher to lower; for logit, probit, and inverse Cauchy link functions; note that log-log models and complementary log-log are not as McCullagh (1980) states); and stochastic ordering (for logit, probit, log-log and complementary log-log).

While the proportionality in the odds ratios is often considered a property of the model and not necessarily a property of the 'real world' (Jones and Westerland, 2006), the PO assumption is still an important assumption to test just as it is to test normality and independence of residuals. The validity of this assumption can be checked by means of a Brant test (Brant, 1990), Rao's efficient score statistic (Rao, 1948), likelihood ratio tests, or the χ^2 test (Peterson and Harrell, 1990). Models violating this assumption should be used with care since they raise identification and interpretation issues (Agresti, 2010). Nonetheless, the literature neither shows a tendency towards testing the PO assumption nor reporting it (e.g., Jones and Sobel, 2000, Brumback et al., 2012, and Ramaswami and Sukumar, 2013). This could be due to the fact that PO is particularly restrictive (Ananth and Kleinbaum, 1997, Brant, 1990) as it imposes strong assumptions; i.e., neither coefficients nor thresholds differ across individuals. This is even more prominent when one considers more than one covariate, and in practice, it is rare that for all the covariates in the model the PO assumption holds (Lall et al., 2002), therefore it is frequently not satisfied. When the null assumption of PO is rejected, we need to know to what extent it is violated in a practical sense, i.e., we need to investigate if the assumption of PO holds for all or part of the covariates. In order to gain further information as to why the assumption has been rejected and to what extent this statistical significance effectively implies a practical significance, graphical methods can be used for assessing the PO assumption (Kim, 2003). This further allows us to control for the effect of large sample sizes, which is recommended for this strongly sample size dependent assumption. It is also worth highlighting that identification issues arise from the latent variable specification and while they are often overlooked, they affect the interpretation of model parameters. Most importantly, the effect of a predictor is not constant, since it depends on the value of the predictor itself and also on the values of the other predictors in the model. In particular, the effect of a continuous covariate is forced to be monotone on the two extreme categories, while it is not monotone on the central ones (Agresti, 2010).

Because of the PO assumption, POMs are not considered to be flexible enough (Anderson,

1984). The PO assumption can be relaxed by letting the variance of the residual in the underlying linear model depend on covariates (Cox, 1995) or, alternatively, to use a scaled link such as the scaled probit link proposed by Skrondal and Rabe-Hesketh (2004). An additional approach is to introduce latent classes (Breen and Luijkx, 2010). Amongst all, the most popular approach is the partial proportional odds model (Peterson and Harrell, 1990; PPOM hereafter) which is an extension of the POM that operates by setting the thresholds to depend on covariates or, alternatively, by allowing covariates to have category-specific parameters. It essentially allows some covariates to be modelled with the assumption of PO, whilst allowing others to have odds ratios which vary by cut-point (Lall et al., 2002). In this case, we would have, in addition to the global effect above mentioned, a "category-specific effect" (Poßnecker and Tutz, 2016) where we get one parameter per response category.

3.3.2 Partial proportional odds model

In the framework of the CLMs, an unconstrained PPOM for an ordinal response variable Y_i with C ordered categories, and two sets of covariates x_{ik} for which the PO assumption holds, and z_{ik} for which it does not hold, can be defined by the following expression

$$\begin{aligned} \text{logit}(P(Y_i \leq j)) &= \alpha_j + \sum_{k=1}^{p} \beta_k x_{ik} + \sum_{k=1}^{q} \gamma_{jk} z_{ik} = \\ &= \alpha_j + \sum_{k=1}^{p} \beta_k x_{ik} + \sum_{k=1}^{q} \gamma_k + u_{jk}) z_{ik}, \\ & i = 1, \ldots, n; j = 1, \ldots, C-1, \end{aligned} \tag{3.15}$$

where $-\infty < \alpha_1 < \alpha_2 < \ldots < \alpha_{C-1} < \infty$; $\boldsymbol{\beta}_k$ corresponds to the p covariates $\boldsymbol{x}_{ik}$ for which the PO holds -also referred to as 'global effects' - and $\boldsymbol{\gamma}_{jk} = \boldsymbol{\gamma}_k + \boldsymbol{u}_{jk}$ to the q covariates $\boldsymbol{z}_{ik}$ for which we relax the PO assumption, also known as 'category-specific effects' (Poßnecker and Tutz, 2016).The individual component $\boldsymbol{u}_{jk}$, represents the deviation of $\boldsymbol{\gamma}_{jk}$ from $\boldsymbol{\gamma}_k$ -fixed component-, the 'typical' value in the population for category j.

In addition to issues of potential over-parameterisation and lack of convergence, PPOMs can predict negative class probabilities in special circumstances (Williams, 2016). According to Hedeker et al. (2006), by having slightly different parameters, there could be cases in which the $C-1$ non-parallel regression lines would cross. For a binary factor (e.g., gender coded as 0 or 1), this crossing of regression lines occurs outside the range of admissible values (i.e., < 0 or > 1). However, in the case of continuous covariates this could happen within the range of the data and lead to negative fitted response probabilities.

In our simulated example in Figure 3.1 (right panel) given that the regression line cor-

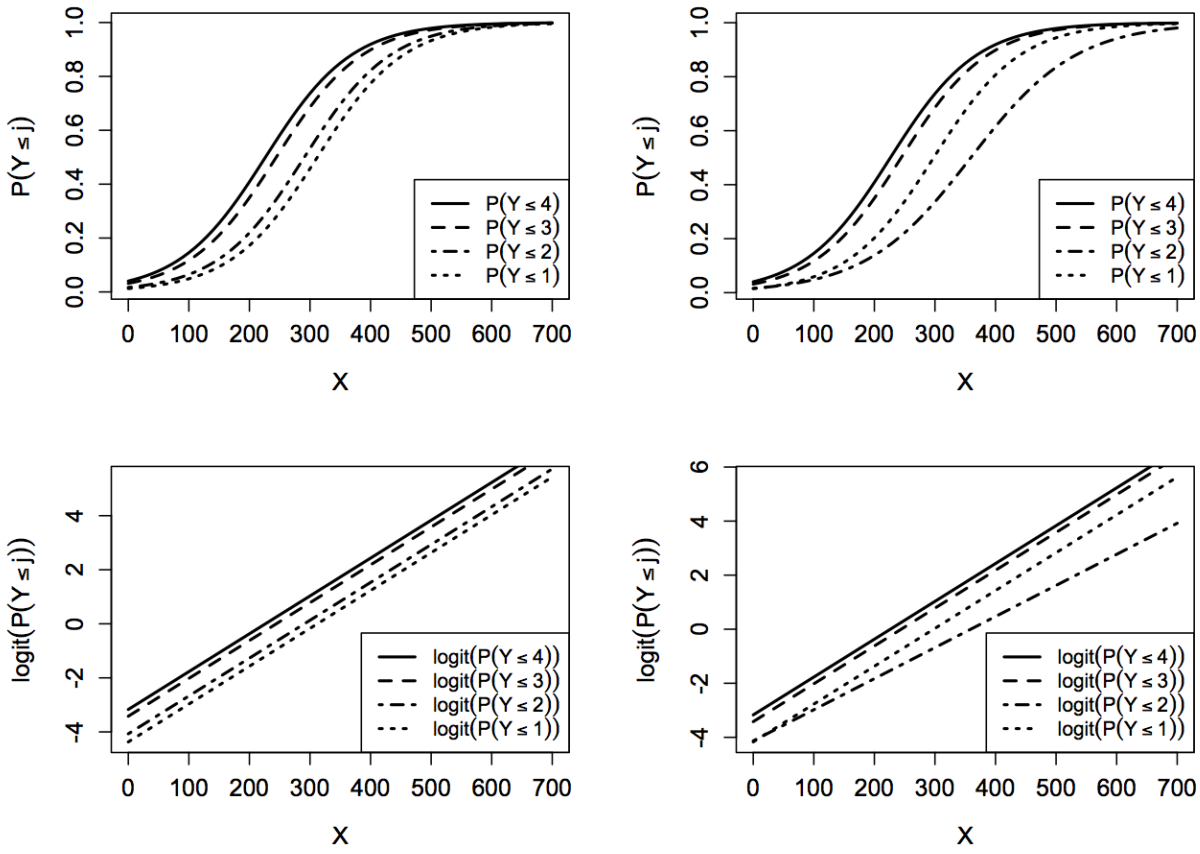

Figure 3.1: Modelled cumulative probabilities for different ordinal levels for a POM (left), and a PPOM (right) with $C = 5$. Note that for the POM, the corresponding regression lines on the logit scale (bottom) are parallel (thus have equal slopes) but this is not the case for the PPOM.

responding to $P(Y \leq 1)$ crosses the $P(Y \leq 2)$ regression line, we have that $P(Y = 2) < 0$ within the range of the covariate x.

We will discuss potential solutions to this issue in Chapter 6. Additionally, extra caution must be taken when using this highly parameterised PPOMs (Hedeker et al., 2006) to avoid overfitting. Despite these limitations, in most cases and particularly those with a small number of covariates or where simple structures for the thresholds can be imposed, PPOMs remain practical.

Table 3.1: Ordinal response models classification according to the approach to category comparison and validity of PO assumption or equivalent.

<table>
<tr><td></td><td colspan="3">Approach to category comparison</td></tr>
<tr><td>Validity of PO assumption (or equivalent)</td><td>A. Cumulative</td><td>B. Stage</td><td>C. Adjacent</td></tr>
<tr><td rowspan="3">1. For every independent variable</td><td>Proportional Odds (PO)</td><td>Continuation Ratio (CR)</td><td>Adjacent Category (AC)</td></tr>
<tr><td rowspan="2">- Effects described via categories' groupings or cumulative probabilities.
- Cumulative logits:
$\text{logit}(P(Y_i \leq j)) = \alpha_j + \sum_{k=1}^{p} \beta_k x_{ik}$</td><td>- Probability of progression across a sequence of stages or categories:
$\text{logit}(P(Y_i = j \mid Y_i \geq j)) = \alpha_j + \sum_{k=1}^{p} \beta_k x_{ik}$</td><td>- Uniform association assumption valid.
- Adjacent-categories logits:
$\text{logit}\left(P(Y_i = j \mid j \leq Y_i \leq j+1)\right) = \log\left(\frac{\pi_{ij}}{\pi_{i,j+1}}\right) = \alpha_j + \sum_{k=1}^{p} \beta_k x_{ik}$</td></tr>
<tr><td>Sequential (S)</td><td>Stereotype Ordered (SO)</td></tr>
<tr><td>2. For some independent variables</td><td>Partial Proportional Odds (PPO)</td><td>Partial Continuation Ratio (PCR)</td><td>Partial Adjacent Category (PAC)</td></tr>
<tr><td>- Cumulative logits:
$\text{logit}(P(Y_i \leq j)) = \alpha_j + \sum_{k=1}^{p} \beta_k x_{ik} + \sum_{k=1}^{q} \gamma_{jk} z_{ik} = = \alpha_j + \sum_{k=1}^{p} \beta_k x_{ik} + \sum_{k=1}^{q} (\gamma_k + u_{jk}) z_{ik}.$</td><td>- Relaxes the PO assumption for coefficients with significant variation across stages:
$\text{logit}\left(P(Y_i = j \mid Y_i \geq j)\right) = \alpha_j + \sum_{k=1}^{p} \beta_{jk} x_{ik}$</td><td>- Relaxes the PO assumption for coefficients with significant variation across logit equations:
$\text{logit}\left(P(Y_i = j \mid Y_i = j, j+1)\right) = \log(\frac{\pi_{ij}}{\pi_{i,j+1}}) = \alpha_j + \sum_{k=1}^{p} \beta_{jk} x_{ik}$</td></tr>
<tr><td rowspan="2">3. For some independent variables & additional constraints</td><td>Trend Odds (TO)</td><td>Continuation Ratio with Partial Proportionality Constraints (CRPPC)</td><td>Adjacent Category with Partial Proportionality Constraints (ACPPC)</td></tr>
<tr><td>- TO assumption valid.
- Odds parameters increase or decrease in a monotonic manner across the cutpoints:
$\text{logit}(P(Y_i \leq j)) = \alpha_j + \sum_{k=1}^{p} \left(\beta_k + \gamma_k \tau_{jk}\right) x_{ik}$</td><td>– Allows for a more flexible application of the PO assumption:
$\text{logit}(P(Y_i = j \mid Y_i \geq j)) = \alpha_j + \sum_{k=1}^{p} \beta_k x_{ik} + \tau_j \sum_{k=1}^{q} \gamma_k x_{ik}$</td><td>– Allows β_k to vary freely, by a common factor, or not at all across logit equations:
$\text{logit}\left(P(Y_i = j \mid Y_i = j, j+1)\right) = \alpha_j + \sum_{k=1}^{p} \beta_k x_{ik} + \tau_j \sum_{k=1}^{q} \gamma_k x_{ik}$</td></tr>
<tr><td>4. For no independent variables</td><td colspan="3">Polytomous (P)
- No ordering
- Baseline-category logits:
$\text{logit}\left(P(Y_i = j \mid Y_i = j \quad \text{or} \quad Y_i = b)\right) = \log\left(\frac{\pi_{ij}}{\pi_{ib}}\right) = \alpha_j + \sum_{k=1}^{p} \beta_{jk} x_{ik}$, with reference category b.</td></tr>
</table>

3.3.3 Cumulative logit mixed models

Ordinal data sets often incorporate multiple respondents' data, which can be analysed either by assuming that the item specific parameters are the same for all the respondents or otherwise, by accounting for the correlation between responses belonging to the same cluster and therefore incorporating clustering procedures (McParland and Gormley, 2011). In order to model these complex structures and to take into account additional patterns (for instance, individual response patterns as reported in Bauer and Sterba, 2011) proportional odds mixed models (POMMs hereafter) that incorporate random effects and account for both within- and between-respondent variance have been introduced (Grilli and Rampichini, 2012, Hedeker et al., 2006).

The simplest version of these models, the random intercept model, can be defined as a subject-specific version of the standard POM by introducing a random effect u_i, such that

$$\text{logit}(P(Y_{it} \leq j)) = u_i + \alpha_j + \sum_{k=1}^{p} \beta_k x_{kit} \quad i = 1, \ldots, n; j = 1, \ldots, C-1, \tag{3.16}$$

where Y_{it} denotes the response for observation t in cluster i (Agresti, 2010, 2015).

We can generalise (3.16) to incorporate multiple random effects such that

$$\text{logit}(P(Y_{it} \leq j)) = \alpha_j + \mathbf{u_i'w_{it}} + \boldsymbol{\beta}' \mathbf{x_{it}} \quad i = 1, \ldots, n; j = 1, \ldots, C-1, \tag{3.17}$$

where the multivariate random effects $\mathbf{u_i}$ have their own covariates (Agresti, 2010).

Partial proportional odds mixed model

Analogously to the expression for the random intercept POMM in (3.16), by introducing the subject term u_{g_i} we can express the random intercept partial proportional odds mixed model (PPOMM) as

$$\text{logit}(P(Y_{it} \leq j)) = u_i + \alpha_j + \sum_{k=1}^{p} \beta_k x_{kit} + \sum_{k=1}^{q} \gamma_{jk} z_{kit} \quad i = 1, \ldots, n; j = 1, \ldots, C-1, \tag{3.18}$$

which allows some of the coefficients to vary by category (here γ_{jk}, where $\gamma_{1k} = 0$ for all $k = 1, \ldots, q$ for some j).

We can generalise (3.18) to incorporate multiple random effects such that

$$\text{logit}(P(Y_{it} \leq j)) = \alpha_j + \mathbf{u_i'w_{it}} + \boldsymbol{\beta}'\mathbf{x_{it}} + \boldsymbol{\gamma_j'}\mathbf{z_{it}} \quad i = 1, \ldots, n; j = 1, \ldots, C-1, \tag{3.19}$$

Detailed interpretation of these models' results is out of the scope of this thesis but further guidance can be found in Agresti (2010) and Hedeker (2007).

3.3.4 Threshold structure specification

For ordinal responses with a high number of categories, many cut-points parameters need to be estimated, leading to the poor precision and likely computational instability that are often associated with models for repeated ordinal scores. However, the number of parameters required to be able to interpret the thresholds in cumulative models can be reduced by certain structure specifications. In particular for a response variable with C categories we can set the threshold structure to be:

- symmetric around zero, requiring $C/2$ parameters and symmetric, with $C/2 + 1$ parameters defined for a response variable with an odd number of categories as

$$\alpha_j = a + d_1[j - (j-1)] + d_2[(j-1) - (j-2)] + \ldots + d_{C/2-1}[(C-1) - (C-2)], \quad j = 1, \ldots, C-1, \tag{3.20}$$

  where the parameter a represents the central threshold, and d_1, d_2,..., are the distances between adjacent thresholds. Similarly for a response variable with an even number of categories such that

$$\alpha_j = a + b + d_1[j - (j-1)] + d_2[(j-1) - (j-2)] + \ldots + d_{C/2-1}[(C-1) - (C-2)], \quad j = 1, \ldots, C-1, \tag{3.21}$$

  where the parameters a and b represent the central thresholds, and d_1, d_2,..., are the distances between consecutive thresholds.

- equidistant, requiring 2 parameters a and d such that

$$\alpha_j = a + d(j-1), j = 1, \ldots, C-1, \tag{3.22}$$

  where a is the central threshold and d represents the distance between adjacent thresh-

olds.

These structures and the interpretation of the corresponding parameters are often ignored or belittled. For instance, Mayer and Foster, 2015 do not report them for their PPOM, and Kuesten et al., 2017 claim that they "have less practical meaning than the factors". For an intuitive graphical explanation, the specific case of 7 categories is illustrated in Appendix A. In the next chapter we illustrate the properties of some of the CLMs described above with different threshold structures using two sets of simulation studies and two case studies.

3.4 Estimation and inference

The most common estimation method for CLMs is maximum likelihood (ML) estimation, for which the log-likelihood function defined by Peterson and Harrell (1990)[3] is:

$$L = \sum_{i=1}^{n} \sum_{j=1}^{C} I_{ij} \log(P(Y_i = j)) = \sum_{i=1}^{n} \sum_{j=1}^{C} I_{ij} \log(\pi_{ij}), \tag{3.23}$$

where I_{ij} is an indicator variable for observation i such that:

$$I_{ij} = \begin{cases} 1 & if \quad Y_i = j \\ 0 & if \quad Y_i \neq j \end{cases} \tag{3.24}$$

and

$$\pi_{ij} = P(Y_i = j) = \begin{cases} \pi_{i1} = D_{i1} & if \quad Y_i = 1 \\ \pi_{ij} = D_{i,j} - D_{i,j-1} & if \quad 1 < Y_i \leq C-1 \\ \pi_{i,C} = 1 - D_{i,C} & if \quad Y_i = C \end{cases} \tag{3.25}$$

where $D_{ij} = P(Y_i \leq j)$ is the cumulative probability that a given observation is less or equal than the jth level (3.13). The above equations are valid for both POM and PPOM and they will be used in this thesis.

The use of the maximum likelihood estimator (MLE) in CLMs for ordinal data has been criticised mainly because it can be infinite which can have undesirable consequences (Kosmidis, 2014). The generalised empirical logistic transform (McCullagh, 1980) could be a solution to these issues as it has smaller asymptotic bias than the MLE and is guaranteed to give finite estimates of the difference in cumulative logits. However, its applicability is limited to CLMs, with no extension to more general cumulative link models.

Firth (1993) and Kosmidis and Firth (2009) propose a procedure to remove the leading term in the asymptotic bias of the MLE, known as the Firth correction or the reduced bias estimator. When the number of outcome categories is relatively large, the sample size is relatively small or some of the outcome categories are rare (e.g., cases of separation or quasi-complete separation), MLE can generate biased estimates of the regression parameters (Lipsitz et al., 2012). More importantly, there can be identifiability issues and the ML estimates might not exist or be infinite. According to Kosmidis (2014), when the researchers define an ordinal scale, the possible responses are largely determined by the end categories of that scale. Hence, the authors argue that the end categories should play a bigger role than the intermediate categories in the analysis, and a good estimation method should

[3]We have changed the original P_{ij} for π_{ij} for consistency of notation throughout this thesis.

not be as democratic as ML is in this respect. Provided that the ordinal scale is well defined, if an end category is not observed then it would be more appropriate to inflate its probability of occurrence slightly, instead of setting it to 0 as the MLE would do. Their proposed estimator would systematically address these issues, contrarily to potential *ad hoc* regularisation solutions such as ridge estimators. The main benefits of this reduced bias estimator according to Kosmidis (2014) are that; it is finite and it reduces asymptotic bias while still respecting the invariance properties of the CLMs.

Generalised estimating equations (GEE) are also recommended in the ordinal regression framework (Nooraee et al., 2014, Parsons et al., 2006, Toledano and Gatsonis, 1996). Although they require the user to specify the correlation structure (Lumley, 1996), which can be seen in certain cases as a drawback, one highlighted advantage is that parameter estimation in data sets with unbalanced missingness is much easier than with MLE.

Notable studies of the predictions from ordinal response models include Anderson and Philips (1981b), Campbel et al. (1991), Campbel and Donner (1989), Rudolfer et al. (1995), and Tutz and Hechenbichler (2005). Given that the assumptions underlying MLE are met, inference can be mostly based on usual methods such as the standard Wald and likelihood ratio tests. Nonetheless, there are slight differences: firstly, although some authors claim inference about the threshold parameters is meaningless, and is generally not carried out (e.g., Greene and Hensher, 2010), we believe it needs to be carefully assessed (read Sasidharan and Menendez, 2014 for an example where it matters); secondly, techniques used for group comparisons in standard linear regression (e.g., adding interaction effects) can be highly problematic in ordinal regression (Allison, 1999, Kuha and Milss, 2018). As Hoetker (2004) notes, even for small contributions to residual variation, comparisons of coefficients across groups can highlight or hide significant differences that are not necessarily consistent with reality (Allison, 1999, Williams, 2009, 2010). When dealing with non-linear models, instead of looking at the parameters, it is advisable to report probabilities and marginal effects on probabilities, since these have a clearer interpretation and are not prone to recalling issues (Long and Freese, 2006).

Statistical inference for CLMs is based on standard concepts used in linear regression, such as standard errors and Wald confidence intervals. For small sample sizes or extreme categories, likelihood ratio tests, and standard errors and confidence intervals based on the profile likelihood are recommended instead (Murphy et al., 1997).

For an easy interpretation of the model effects, Agresti (2010) suggests the comparison of response category probabilities at specific values of the model covariates. He also recommends the use of the confidence intervals associated to those probabilities as a measure of precision of the estimates.

The comparison of parameters from models with different covariates and/or different link functions is out of the scope of this thesis but has been addressed in the literature (Fielding, 2004, Bauer, 2009, Mood, 2010).

In the `R` package `ordinal` incorporates confidence intervals in the predictions through the `predict` function. For models fitted using `VGAM`, the generic function `confint` can be used instead. ML estimation is incorporated in all of the ordinal model fitting functions.

3.5 Goodness of fit and residual diagnostics

Model assessment in ordinal response models also entails many practical issues (Genter and Farewell, 1985). Fielding (1999) states that traditional goodness of fit measures that are available for linear fixed and random-effects models and diagnostic plots of individual residuals are generally unavailable and difficult to implement for these ordinal models. The main problems are due to the discrete nature of the data. Cox and Snell (1968) and Pregibon (1981) review the discreteness issues for single binary variables and Landwehr et al. (1984) focus on the interpretation difficulties of graphical methods due to this discreteness.

Extensions to the Hosmer-Lemeshow goodness of fit statistic (which was originally proposed for binary logistic regression by Hosmer and Lemeshow, 1980) to be used in an ordinal framework are proposed by Fagerland and Hosmer (2012), Lipsitz et al. (1996), and Pulkstenis and Robinson (2004). This test requires assigning ordinal scores to the estimated probabilities π_{ij} which we then arrange into G groups [4]. Let us denote the sums of the observed and estimated frequencies in each group for each of the C response categories by O_{jk} and E_{jk}, respectively. The statistic is given by

$$S_{HL} = \sum_{j=1}^{C} \sum_{k=1}^{G} \frac{(O_{jk} - E_{jk})^2}{E_{jk}}. \tag{3.26}$$

However, limitations have been found for these approaches. Fagerland and Hosmer (2012) argue that the method by Lipsitz et al. (1996) is not always computable for small sample sizes and that the approach by Pulkstenis and Robinson (2004) requires continuous and categorical covariates, and can underperform when many categorical covariates are present. They propose a set of alternative tests derived from Fagerland et al. (2008) multinomial logistic regression extension, which showed a higher sensitivity to the lack of fit when compared to

[4] The authors suggest a choice of the number of groups satistying that $G > p+1$, with p being the number of covariates in the model. The results will be very sensitive to this choice, with a small value of G reducing the chances to detect misspecification, and a large value making it difficult to decide whether differences are due to chance or misspecification.

the others in a set of six situations but did not show an improvement in power. Lin and Chen (2008), in contrast, assess goodness of fit of the PO model by a nonparametric local linear smoothing technique. Their method would also suffer from high dimensionality when dealing with many covariates, but outperforms Pulkstenis and Robinson (2004) in terms of power. Graphical diagnostic tools have also been proposed. Liu et al. (2009) evaluate model misspecification of the functional form of specific covariates and the misspecification of the link function for the PO model. Other attempts include the use of ROC curves for all possible collapsing of categories (Rajan and Zhou, 2012, Toledano and Gatsonis, 1996) and graphical comparisons of relative risks (Ananth and Kleinbaum, 1997). Overall, goodness of fit tests are used to evaluate model fitting and provide limited information in the form of a single p-value. Residual assessement on the contrary, "enables us to examine a given model from different angles, focus on each component one at a time, visualize the practical deviation (rather than merely statistical significance), and advise model improvement" (Liu and Zhang, 2018).

Residual diagnostic techniques for models where the outcomes are ordinal are generally accepted as not well developed (O'Connell and Liu, 2011). Successful approaches for ordered category response models are unavoidably closely linked to the latent variable approach. From earlier ideas on residuals (Cox and Snell, 1968), the work of Chesher and Irish (1987) introduces a range of graphical methods and specification tests for those cases for which cut-points are known. Machin and Stewart (1990) extend these methods to models with 'unknown' thresholds. Hosmer and Lemeshow (2000) suggest extensions to the residual strategies originally developed for binary logistic models (Pregibon, 1981) while Bender and Benner (2000) propose graphical strategies to examine the feasibility of the PO assumption. However, the interpretation of these diagnostics is not simple (Di Iorio and Iannario, 2012). O'Connell and Liu (2011) argue that the two most straightforward approaches would be (i) the one considering the ordinal response variable as continuous and applying ordinary least squares residual analysis, and one (ii) where the ordinal response model is decomposed into underlying binary logistic models and where the residuals are assessed separately. The uptake of the second approach is high among researchers (Harrell, 2001, Hosmer and Lemeshow, 2000, Long and Freese, 2006), but it seems contradictory to advocate for an appropriate ordinal modelling and assess the fitting via non-ordinal methodologies.

Finally, two of the most popular Bayesian approaches to residuals' diagnostics in ordinal models are described below:

Posterior predictive checks

Simulated values from the joint posterior predictive distribution of replicated data are drawn and these samples from the simulation are compared to the observed data. Any systematic differences between the simulations and the data indicate potential failings of the model. They allow for measurement of the "statistical significance" of the lack of fit. They have the limitation that the distribution of Bayesian p-values is not uniform under the null hypothesis that the model fits (Robins et al., 2000).

Bayesian latent residuals

Gelman et al. (2000) suggest a way to avoid the discreteness of the residuals is by working with residuals defined in an underlying continuous model in terms of latent continuous variables, as suggested by Albert and Chib (1993). This also seems more consistent with the modelling approach which also uses the latent approach. Supposing a latent variable $(Y_i^*|\alpha, b_i) \sim N(x_{ik}\beta_k, 1)$, we define the latent studentised Pearson residual such that

$$r_{ij}^*(\theta, b_i) = (Y_i^* - E(Y_i^*|, b_i))/(Var(Y_i^*|\alpha, b_i)). \tag{3.27}$$

However, according to Gelman et al. (2000), the plots based on latent continuous residuals can be too noisy to be useful. The difficulty can be that the discrete data provide very little information about the individual latent variables, and so any patterns in the latent residuals disappear in the posterior uncertainty.

For the reasons explained above and given implementation practicalities, we limit our study to the residuals introduced by Dunn and Smyth (1996) (D-S hereafter), Li and Shepherd (2012) (L-S hereafter), and Liu and Zhang (2018) (i.e., surrogate residuals) that we will explore in detail via a simulation study in Chapter 5.

3.6 Software availability

Until recently, the limited application of ordinal response models was linked to restrictions on the availability of software (Fielding, 1999, Lall et al., 2002). Now, freely available software makes the process easier. We focus on pieces of software specifically designed to fit models with ordinal responses by means of cumulative link models.

There are multiple libraries in the open software platform (R Core Team, 2018) which can implement CLMs, including; `vcrpart`, `polr`, `VGAM`, `MCMCglmm` -only with probit link function-, `rms`, `mixcat`, `DPpackage`, `lcmm`, and `GMMBoost`). We have initially favoured `ordinal` (Chris-

tensen, 2015) in this thesis because it includes a substantially more user-friendly and complete implementation of ordinal response models via ML than some of the other mentioned packages, and because of its flexibility to model different types of thresholds. However, `VGAM` allows further flexibility when working with PPOMs (e.g., more than one PPO variable allowed).

Packages including functions implementing residual diagnostics for ordinal response data models include; `VGAM` (function `resid` with relevant `type` options `response` and `pearson` as described in Appendix C), `rms` (function `lrm`), `vcrpart` (function `resid.olmm`) and `PResiduals` (function `presid`, which extracts Li-Shepherd residuals but do not support models fitted via `clm` (`ordinal`). A function to calculate the Dunn-Smyth residuals for the `clm` function of the `ordinal` package is available at `http://web.maths.unsw.edu.au/~loicthibaut/ordinalLab.html`) .

Several packages are available to apply penalties. However, the most popular packages "for penalized ordinal regression, such as `penalized` and `glmnet` do not currently fit ordinal models" (Wurm et al., 2017). "`glmpath` and `glmnet` fit models along the entire regularisation path, whereas `lasso2` requires the user to select *a priori* the penalty parameter. Other packages, such as `ncvreg` and `SIS` fit a regularisation path but do not include options for estimating coefficients that are not included in the penalty term, which is needed for estimation of the α_k terms" (Archer et al., 2014). `ordinalgmifs` (Archer et al., 2014; no longer updated in CRAN), allows fitting of CLMs for high-dimensional data settings. Finally, `ordinalNet` (Wurm et al., 2017) fits a broad class of ordinal regression models with an elastic net penalty via a coordinate descent algorithm.

In SAS (SAS Institute, 2008), `PROC LOGISTIC` and `PROC CATMOD` allow modelling of ordered response data, and in Stata: `gologit2` (Williams, 2006) includes POMs and PPOMs; `oglm` (Williams, 2009) estimates generalised ordered logit models that deal well with issues associated with the PO assumption; and `meologit` (Stata Corp, 2015) estimates mixed model versions of PO models. However, as far as we are aware, none of these functions allow for different threshold structures.

In sum, although there has been great development of specific libraries, none of the above mentioned include a comprehensive user-friendly treatment of CLMs and their extensions, which further justifies the aims of this thesis.

3.7 Retinopathy

We start with the retinopathy case study described in Chapter 2. We focus on the left eye measure of *retinopathy severity* and model it via a PPOM with two covariates; *systolic blood*

pressure (for which we relax the PO assumption) and *left eye refraction index*. As we can see from Figure 3.2, there is prominent crossing of predicted probabilities for this PPOM.

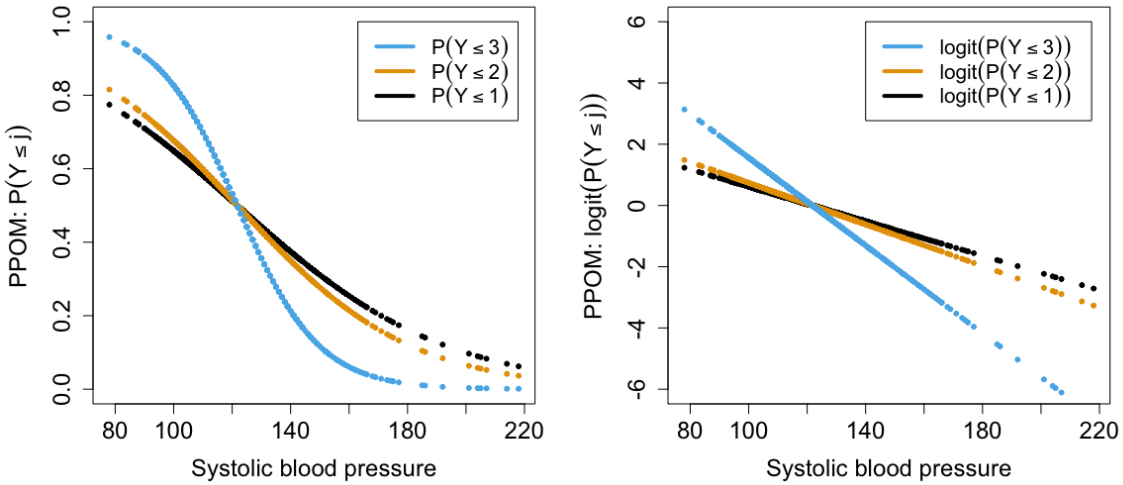

Figure 3.2: PPOM of *left eye retinopathy severity* vs *systolic blood pressure*. Predicted regression lines on the probability scale (left) and on the logit scale (right).

3.8 Conclusions

Amongst the different models for ordinal response data, CLMs are prevalent in the socio-economic literature. Prominently, POMs are widely used despite the restrictive nature of the assumption, which is often not tested. Still, plenty of papers ignore the ordered nature of the data (e.g., Fischer et al., 2011). Findings from the literature and software review indicate three areas for further research that are the main focus of this thesis: (i) additional justification for the use of ordinal models (particularly CLMs) in terms of effects on inference, (ii) model diagnostics to ensure accurate model specifications, and (iii) solutions to issues in PPOMs.

Chapter 4

Ordinal models for ordinal responses - an evaluation

In this chapter we perform a comparative assessment of ordinal models by comparing them to other standard practices when modelling ordinal response data, and evaluating the consequences of using inappropriate modelling approaches in terms of inference and practicalities.

4.1 Introduction

When dealing with ordered data, issues arise from the descriptive through to the modelling stage for both independent covariates (Van der Jagt et al., 2014) and response variables (Orme and Combs-Orme, 2009). This was initially highlighted by Stevens (1946) and measurement theory has since been a fruitful topic of research. In particular, ordinal response variable modelling is often tackled in a simplified manner by applying standard linear regression to equidistant numeric scores of the response categories (Capuano, 2012, Norman, 2010, Torra et al., 2006). This approach inherently considers the ordinal response variable as continuous but it is only likely to be accurate when the number of categories is high, when the distance between successive response categories is truly equidistant, and when there are a limited number of observations in the end categories as this can help avoid problems with variance heterogeneity (Christensen, 2015). In addition to the already mentioned power and performance benefits from a correct (ordinal) approach (Bauer and Sterba, 2011, Capuano, 2012), Agresti (2010) highlights the benefits derived from the estimation of probabilities for the response categories at fixed settings of the covariates and alerts us of the biases caused by spurious effects associated with potential 'ceiling and floor effects' (asymmetries caused by predominant lowest or highest category responses often derived from the truncation of the scales). However, despite further widespread discussions on the appropriateness of such

models (Greene and Hensher, 2010, Winship and Mare, 1984), certain specific effects on inferences associated with this modelling approach have not yet been explicitly described in the literature.

This chapter addresses issues that have been previously discussed (Agresti, 2010) but contributes further to the literature by focusing on the application of CLMs and the advantages derived from their application, both at the numerical and graphical simulation level and for the three different potential types of thresholds. It also exemplifies these advantages by means of two case studies. Additionally, these examples allow us to highlight advantages in terms of interpretability. Finally, it shows the strengths and limitations of the available open software and points towards potential improvements both at the implementation and interpretation level.

4.2 Simulation study

A simulation study has been developed to compare the two main approaches for modelling ordinal response data (i.e., numeric vs ordinal) with the latent model, all based on an interpretation of an underlying latent response (as described in subsection 1.2.1 and defined in (1.1)), in order to understand when a POM is truly necessary - e.g., to avoid making the wrong inference with a numeric model.

4.2.1 Numerical simulation

In order to consider the effect of modelling choice on inference, we run the following simulation study:

1. We simulate an ordinal response variable Y with $C = 6$ categories which is a categorised version of a latent continuous variable $Y^* = \boldsymbol{\alpha} + \varepsilon$ with $\varepsilon \sim N(0, 100)$ by means of 5 known cut-points $\alpha_1, \ldots, \alpha_5$. 4 different data sets are generated according to threshold structures

   - unconstrained example 1 with $\boldsymbol{\alpha} = (1, 2, 3, 8, 9)$,
   - unconstrained example 2 with $\boldsymbol{\alpha} = (3, 7, 8, 9, 10)$,
   - symmetric with $\boldsymbol{\alpha} = (1, 2, 5, 8, 9)$,
   - equidistant with $\boldsymbol{\alpha} = (2, 4, 6, 8, 10)$.

2. Three sample sizes are considered; $n = 100$, $n = 1,000$, and $n = 5,000$.

3. We repeat the simulation 1,000 times.

4. We model these data separately against a covariate x by means of 3 different approaches:

   (i) the original latent response via OLS linear regression (i.e. the 'correct' model)

   $$Y^* = \alpha + \beta x + \varepsilon, \tag{4.1}$$

   as a benchmark since Y^* is not observable. If we rely on the observed Y, whatever the model we specify, it is impossible to recover the 'true' parameters of model (4.1), since the scale of Y^* is not identifiable from any model on Y.

   (ii) the ordinal response as if it were continuous via OLS linear regression

   $$Y = \alpha + \beta x + \varepsilon, \tag{4.2}$$

   where we are arbitrarily assigning the values 1 to 6 to the categories [1], and

   (iii) the ordinal response via a POM with unconstrained thresholds,

   $$\text{logit}(P(Y \leq j)) = \alpha_j + \beta x. \tag{4.3}$$

5. We study the proportion of occurrences of statistically significant and non-significant results (at both the 5% and 1% significance levels) in a Wald χ^2 test with null hypothesis $H_0 : \beta = 0$.

Results from the simulations are plotted in Figure 4.1 and show that there is a closer agreement between the results of (iii) and (i) than between (ii) and (i). Except for the unconstrained thresholds example 2, for a sample size of $n = 5,000$[2] in around 30% of the simulated data sets, model (iii) alone produces the same inferences as the latent response model (the 'correct' model) for 5% significance (slightly fewer for 1%). All three models give the same results of the significance test around 50% of the time for 5% significance and 70% for 1%. Only rarely (3% of the time for both 5% and 1% significance) do the 'ordinal as continuous' results alone match the correct model's results. Results for the equidistant thresholds data set are particularly interesting because one might have expected the proportion of time in which the three models resulted in the same decisions to be substantially higher than in the case of the other threshold data sets. However, differences in the variance

[1] Modelling Y by means of the linear model has some drawbacks but is easy to estimate and the results can be easily interpreted, so it can be useful in a preliminary analysis, especially when Y has many categories and the distribution is not too skewed.

[2] We argue that due to uncertainty, for lower sample sizes results do not show the unbiased behaviour of the models.

of the estimates are likely to have caused this proportion to be only slightly greater than for the other cases. The unconstrained example 2 shows the highest percentages (76% for 5% significance and 87% for 1%) of consistent statistically significant results for all three models.

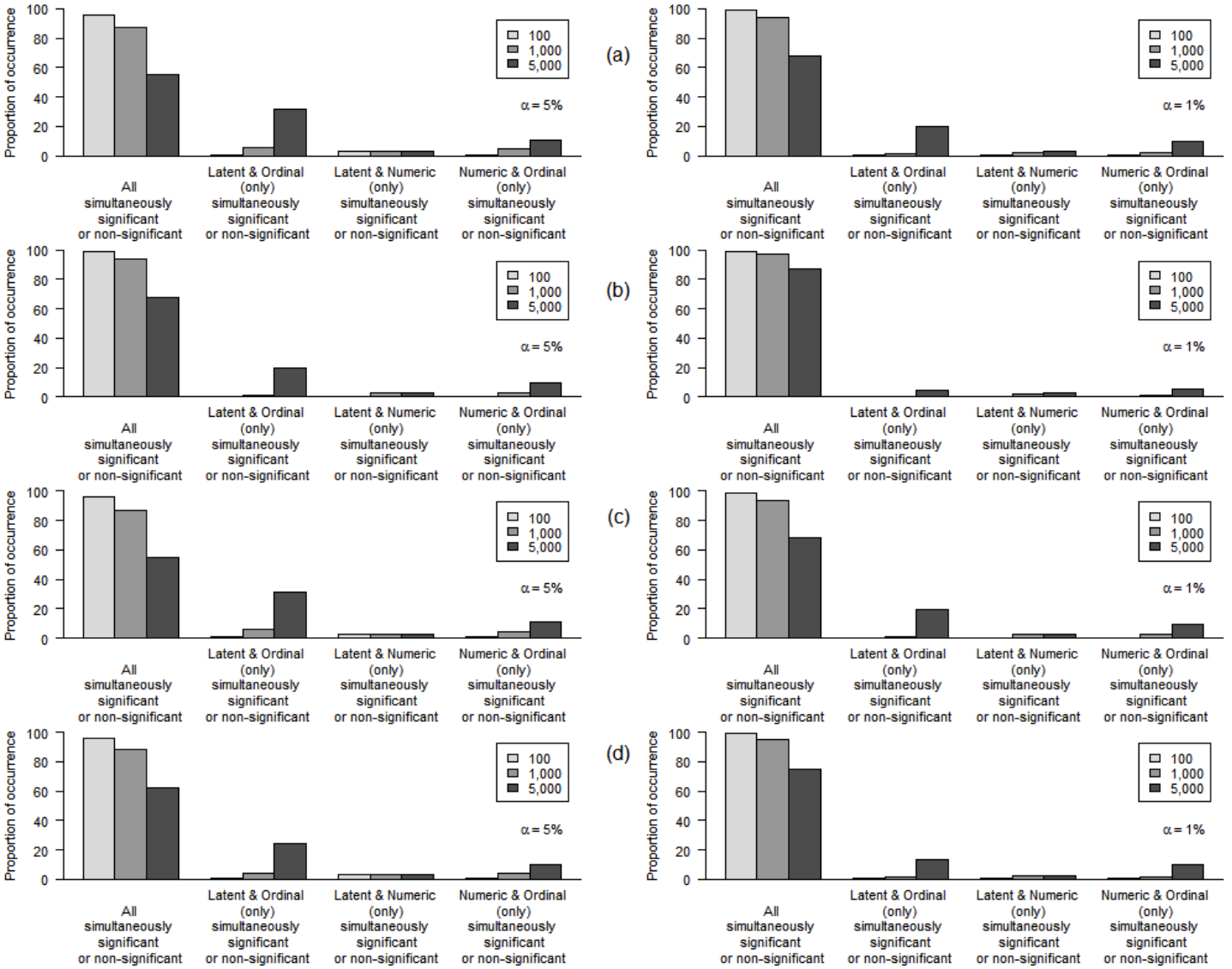

Figure 4.1: Proportion of occurrence of significant/non-significant results for simulated data sets: (a) unconstrained example 1 with $\boldsymbol{\alpha} = (1, 2, 3, 8, 9)$, (b) unconstrained example 2 with $\boldsymbol{\alpha} = (3, 7, 8, 9, 10)$, (c) symmetric with $\boldsymbol{\alpha} = (1, 2, 5, 8, 9)$, and (d) equidistant with $\boldsymbol{\alpha} = (2, 4, 6, 8, 10)$, at 5% (left panel) and 1% significance levels (right panel). Simulation was run 1,000 times and the sample sizes are shown in the legend.

4.2.2 Graphical simulation

A set of alternative simulations is performed to show the frequent lack of consistency between analysing ordinal scales as continuous and analysing the underlying (simulated) latent variable:

1. We consider an ordinal response variable Y with $C = 5$ categories which we assume to be a categorised version of a latent continuous variable $Y^* = 2 + 2x + \varepsilon$ with $\varepsilon \sim N(0, 0.1)$ and defined such that for a category j, $Y = j$ if $\alpha_{j-1} < Y^* \leq \alpha_j$ (Agresti, 2010), therefore the continuous scale is considered as naturally divided into 5 regions by 4 known cut-points $\alpha_1, \ldots, \alpha_4$ as shown in Figure 4.2 (shown on the right y-axis of the left panels and represented by the horizontal, dotted lines), and $\beta = 2$. We compare 4 examples of data sets with different threshold structures:

   - unconstrained example 1 with $\boldsymbol{\alpha} = (0.2, 0.4, 0.6, 2.8)$,
   - unconstrained example 2 with $\boldsymbol{\alpha} = (1, 2.5, 3.1, 3.6)$,
   - symmetric with $\boldsymbol{\alpha} = (0.2, 1.2, 2.8, 3.8)$,
   - equidistant with $\boldsymbol{\alpha} = (0.5, 1.5, 2.5, 3.5)$.

2. We suppose the existence of data on an underlying continuous latent response Y^* with normal errors, and show simulated points ($n = 500$) plotted (in grey) against a covariate x with a uniform distribution $U(1, 7)$. The true relationship is represented as a straight line, drawn in black (Figure 4.2) such that

$$Y^* = \alpha + \beta x + \varepsilon, \tag{4.4}$$

   where $\varepsilon \sim N(0, 0.2)$.

3. Then we arbitrarily assign the values 1 to 5 to the categories instead (which is essentially what we are doing when dealing with the variable as continuous with parametric techniques; left y-axis) and plot the resulting regression line in red

$$Y = \alpha + \beta x. \tag{4.5}$$

4. Finally, we model the response variable Y using a POM with unconstrained thresholds, drawn in blue (Figure 4.2)

$$\text{logit}(P(Y \leq j)) = \alpha_j + \beta x. \tag{4.6}$$

The plots on the left-hand side of Figure 4.2 highlight the consequences of using two different approaches (numeric and ordinal) for modelling ordinal responses for the four different data sets, and compare the results to the latent (true) model of the underlying continuous variable. Note that subsequently treating the ordinal response as numeric introduced a 'warping' of the true scale, indicated by the values on the right-hand y-axes (where midpoints of the classes are shown). Treating the derived ordinal scale as numeric and using linear regression produced the red lines, which missed the true relationships owing to the bias caused by the (unknown) warping of the real scale. Using a POM, the scale was not warped in this way, and back-calculating the dotted lines onto the numeric scale produced fits (shown in blue) which were hard to distinguish from the true relationships. In the data set where thresholds were equidistant (d), as expected, no bias was apparent as shown by the overlapping lines in (d). The bias was also negligible for the unconstrained spacing in (b). For the unconstrained example in (a) and the symmetric threshold spacing in (c), the bias appeared in the estimated slope. This bias has an effect on significance tests, as shown in the first part of the simulation study.

The plots on the right-hand side of Figure 4.2 graphically compare response category frequencies for the three models in the four cases of simulated thresholds. Overall, response categories for the ordinal and latent (true) models have very similar frequencies, whereas the numeric model differs in frequency counts for most cases, except for (d) equidistant thresholds where all the response categories present approximately the same frequencies for all models. Even worse, response categories 1 and 2 in the numeric model in (a) and response category 5 in the numeric model in (c) present 0 frequencies. Finally, it is also worth mentioning that in (b) we initially had predicted numeric responses which rounded to 0, outside of the set of chosen arbitrary categories; these were collapsed into category 1; also category 5 has a frequency of 1 for the numeric model, and it is only due to scaling that it is not obvious in the plot.

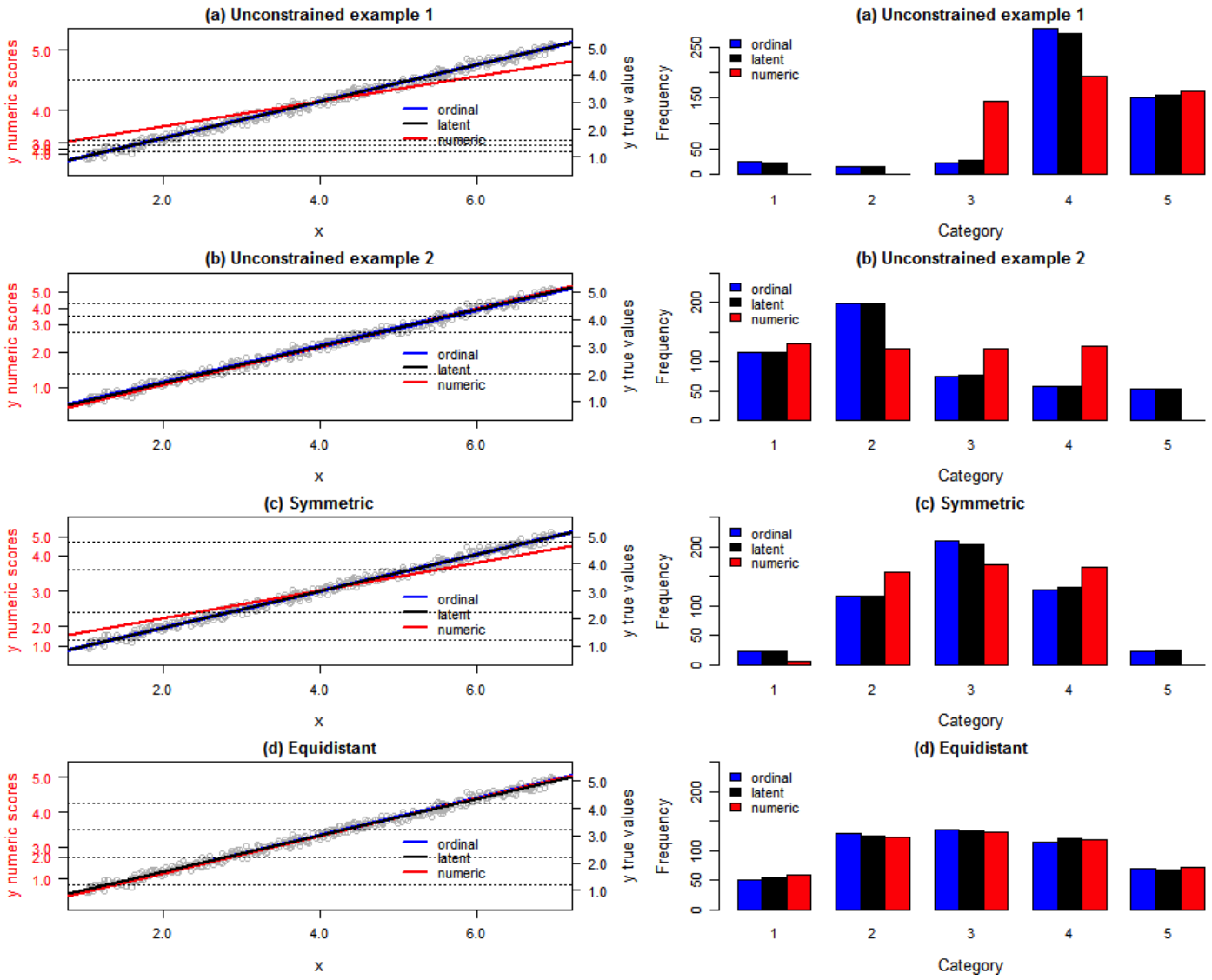

Figure 4.2: Model fitting (left) and predictions (right) comparison for simulated thresholds: (a) unconstrained example 1 with $\boldsymbol{\alpha} = (0.2, 0.4, 0.6, 2.8)$, (b) unconstrained example 2 with $\boldsymbol{\alpha} = (1, 2.5, 3.1, 3.6)$, (c) symmetric with $\boldsymbol{\alpha} = (0.2, 1.2, 2.8, 3.8)$, and (d) equidistant with $\boldsymbol{\alpha} = (0.5, 1.5, 2.5, 3.5)$.

4.3 Case studies

We now show two case studies based on real data where the effects shown above with simulated data have a practical impact (both case studies have been described in Chapter 2). We firstly revisit the analysis in (Craig et al., 2018; Case study 1) evaluating shopping and nature experiences, and their association to connectedness to nature. Secondly, we review the analysis in Fischer et al. (2011) (Case study 2), exploring the relationships between perceived desirability of 3 species and 6 attributes.

Here for both cases (Case studies 1 and 2), potential differences in results between numeric and ordinal models were assessed. In Case study 1 we compared a PPOM with symmetric thresholds to a linear model and in Case study 2 we focused on the comparison between a POMM with unconstrained thresholds and a linear mixed model.

All ordinal models were fitted using `R` package `ordinal` (Christensen, 2015) except from the PPOMM in Case study 2 which was fitted with `vcrpart` (Bürgin and Ritschard, 2017).

4.3.1 Connectedness to nature

Our first case study deals with the results from a survey assessing the effects of the interaction of connectedness to nature and experiences in nature and shopping. Respondents ($n = 357$) were asked to provide ratings of *pleasantness* (as a response with $C = 7$ categories where 7 is most and 1 is least pleasant) for 2 categories of the covariate *experience* (*nature* and *shopping*) given their reported measurement for a second covariate, connectedness to nature (*CNS*), which is a 3-level ordinal variable ranging from 1 being low to 3 being high levels of *CNS*.

We initially modelled the data using a linear model including the interaction spotted in Figure 2.1 between *CNS* and *experience* (results are shown in Table 4.1), and we modelled the same data via an analogous ordinal model such that

$$\text{logit}(P(Y_i \leq j)) = \alpha_j + \sum_{k=1}^{p} \beta_k x_{ik} + \sum_{k=1}^{q} (\gamma_{jk} z_{ik} + \delta_{jk} x_{ik} \times z_{ik}) \tag{4.7}$$

(results are shown in Table 4.2) to assess the differences in the results. When we accounted for variability at the respondent level via mixed model versions of both models, the likelihood ratio test on the variance component showed no statistically significant differences and the effects estimates were approximately equal and therefore the most parsimonious (non-mixed) models were reported.

Table 4.1: Case study 1: summary statistics for the linear model ($n = 357$). Effect estimates of *CNS* and *experience* are not absolute, but in relation to contrasting categories (*CNS1* and *nature* respectively). The t value corresponds to the test statistic which has a t-distribution with *n-p* degrees of freedom under the null hypothesis that the effect size is 0.

	Estimate	**Std. error**	**t value**	**P(>| t |)**
CNS2	0.287	0.230	1.246	0.214
CNS3	0.474	0.226	2.097	0.037
Shopping	-1.758	0.241	-7.289	<0.001
CNS2*Shopping	-0.514	0.326	-1.578	0.116
CNS3*Shopping	-1.319	0.320	-4.128	<0.001

Table 4.2: Case study 1: summary statistics for PPOM (with the PO assumption relaxed for the variable *experience*) with symmetric thresholds. Effect estimates of *CNS* and *experience* are not absolute, but in relation to contrasting categories (*CNS1* and *nature* respectively). The z value corresponds to the test statistic which has an asymptotic standard normal distribution under the null hypothesis that the effect size is 0. Threshold coefficients interpretation can be found in Appendix A.

	Estimate	**Std. error**	**z value**	**Pr(>| z |)**
CNS2	0.714	0.271	2.636	0.008
CNS3	1.157	0.284	4.070	<0.001
CNS2*Shopping	-0.907	0.360	-2.523	0.012
CNS3*Shopping	-1.910	0.369	-5.171	<0.001
Threshold coefficients				
Intercepts				
Central 1	-2.080	0.256	-8.112	<0.001
Central 2	-1.591	0.215	-7.394	<0.001
Spacing 1	0.383	0.087	4.421	<0.001
Spacing 2	1.419	0.143	9.895	<0.001
Shopping (Nature as reference)				
Central 1	1.472	0.314	4.695	<0.001
Central 2	1.685	0.279	6.029	<0.001
Spacing 1	0.224	0.104	2.150	0.032
Spacing 2	-0.022	0.171	-0.127	0.899

Both the effects of *CNS2* and the interaction of *CNS2* with *shopping* (in both cases with

respect to the corresponding terms for the reference category *CNS1*) appeared not to be statistically significant in the linear model, whereas, in the results from the ordinal model, they were statistically significant (at 1% and 5% significance levels respectively).

We initially tested the PO assumption, which did not hold for our data and then set the structure of the thresholds as informed by the design of the survey and patterns in the responses. The final model was a PPOM with symmetric thresholds relaxing the PO assumption for the variable *experience*; results are shown in Table 4.2. This model was chosen because it was statistically different from the POM and it did not show signs of overfitting or lack of precision in the estimation (in contrast to the example with the non-proportional odds interaction of *CNS* and *experience* which had far higher values in the standard errors).

When we compared Table 4.1 and Table 4.2, results were found to be qualitatively different in terms of parameter statistical significance. The main differences between the results corresponded to the effect of *CNS2*. When compared to the reference level *CNS1* for the ordinal model, it became statistically significant at the 5% significance level. Similarly the interaction of *CNS2* and *shopping* became significant at the same significance level. This would essentially mean that in addition to the statistically significant differences between a low and high level of *CNS* effect on the response, there are also significant differences between the medium and high levels. For the interaction of *CNS* and *experience*, similar conclusions were reached therefore we can state that there is an overall effect of both *CNS* and the interaction of *CNS* and *experience*. Using a model appropriate to the response data type thus produced different inference compared to a simpler model (and in addition, in this example did not compromise the residual diagnostics).

Table 4.3: Case study 1: Pleasantness predicted category probabilities from PPOM for *nature* (upper) and *shopping* (lower) experiences for the different levels of *CNS*.

Nature	**Categories**						
Level	**1**	**2**	**3**	**4**	**5**	**6**	**7**
CNS1	0.029	0.049	0.033	0.058	0.061	0.227	0.543
CNS2	0.015	0.025	0.018	0.033	0.037	0.164	0.708
CNS3	0.009	0.017	0.012	0.022	0.026	0.123	0.791
Shopping	**Categories**						
Level	**1**	**2**	**3**	**4**	**5**	**6**	**7**
CNS1	0.119	0.110	0.124	0.171	0.145	0.148	0.184
CNS2	0.140	0.124	0.133	0.173	0.138	0.134	0.157
CNS3	0.222	0.164	0.150	0.164	0.111	0.094	0.096

Additionally, another important advantage of the ordinal approach is that threshold coefficients can be translated into cumulative odds and ultimately into predicted category probabilities (Agresti, 2010, p5; Simonoff, 2003). As shown in Table 4.3, the probabilities of the higher values of the response categories (5, 6 and 7) corresponding to the *nature* experience were considerably higher (range: 0.026-0.791) than the ones for the lower values (1, 2 and 3; range: 0.009-0.049). Additionally, as expected, the higher the level of *CNS*, the higher the probability is for the *nature* experience rating 7. In contrast, for the *shopping* experience all the response categories presented a low probability ranging from 0.110 to 0.184 (see Appendix A for details of the calculations).

Figure B.1 (Appendix B) compares these predicted probabilities to the relative frequency of each of the 7 response categories and for each combination of the factors *CNS* and *experience*. A very close fit of both predicted probabilities and observed relative frequencies is clear, except for rating 6 in *CNS1* and nature experience. Besides the differences in probabilities highlighted from the previous tables, the plots showcase the subtle but noticeable interaction between *CNS* and *experience* which further justifies the choice of this PPOM.

4.3.2 Species desirability

For the second case study, we further modelled the data published in Fischer et al. (2011) as described in Chapter 2. In order to explore whether desirability of a 'population increase of certain species' can be understood at a level that transcends individual (types of) species, a linear mixed model was fitted with subject as a random effect and an interaction *country* by *species*. Results from this model are shown in Table 4.4. All the covariates appeared to have statistically significant effects at 1% significance level (only 5% for *foreign-native*), and only the contrasts of *spider* with respect to *plant* were significant at 10% significance level.

To determine whether differences can be found when using the appropriate ordinal model, we fitted a PPOMM that ensured straightforward comparability (results are shown in Table 4.5). Small differences were found in the p-values for *spider* between Tables 4.4 and 4.5, which was significant at the 10% significance level in the linear mixed model but non-significant in the ordinal mixed model.

Moreover, results in Table B.2 (Appendix B) allow further interpretation of the data on the basis of the threshold positions and ultimately in the form of 24 sets of category probabilities, for specific cases of interest to the researchers. Figure B.3 compares these predicted probabilities to the observed relative frequencies for each of the 5 response categories and for each combination of the factors *country* and *species*. In general, a relatively close fit of both predicted probabilities and observed relative frequencies is visible, with very close fits for

the *large mammal* in Flanders, France, Hungary, Romania and Slovakia, and for the *plant* in Flanders and Hungary. The rating patterns (both actual and predicted) were particularly similar across countries for the *spider*.

Figure B.2 (Appendix B) represents overall relative frequencies for each of the species and sheds further light on the different patterns in the responses. Overall, an increase in the large mammal population seems to present the highest probability of being considered desirable (ratings 1 and 2). Changes in the spider species provoked the highest number of indifferent responses but also the lowest number of 2 ratings. The pattern in the desirability responses to an increase in the plant species can be considered approximately constant, with slightly higher probabilities for the higher ratings.

Table 4.4: Case study 2: summary statistics for linear mixed model ($n = 7,134$). Effect estimates of species are not absolute, but in relation to a contrasting category (*plant*). The t value corresponds to the test statistic which has a t-distribution with n-p degrees of freedom under the null hypothesis that the effect size is 0. Differences with respect to results in Fischer et al. (2011) are caused by different sample sizes. For *country* and the interaction *country*species* we would have 8 and 8×3 values respectively, therefore, for reasons of simplicity the individual level estimates and standard errors are not shown; the F value corresponds to Type III ANOVA.

	Estimate	**Std. error**	**t value**	**P(>| t |)**
Attractive-unattractive	-0.140	0.014	-10.033	<0.001
Common-rare	0.056	0.013	4.455	<0.001
Decrease-increase	-0.273	0.012	-22.922	<0.001
Foreign-native	0.025	0.012	2.084	0.037
Harmful-harmless	0.097	0.011	8.707	<0.001
Strong-vulnerable	0.038	0.011	3.344	<0.001
Valuable-worthless	-0.164	0.015	-11.069	<0.001
Species (non-native plant contrast)				
Large mammal	0.070	0.082	0.852	0.395
Spider	0.150	0.084	1.781	0.075
			F value	**p-value**
Country	-	-	4.503	<0.001
Country*species	-	-	5.971	<0.001

Table 4.5: Case study 2: summary statistics for PPOMM (with the PO assumption relaxed for the variable *attractive-unattractive* and *harmful-harmless*) with unconstrained thresholds. Effect estimates of species are not absolute, but in relation to a contrasting category (*plant*). The z value corresponds to the test statistic which has an asymptotic standard normal distribution under the null hypothesis that the effect size is 0. For the interaction *country*species* we would have 8×3 combinations therefore for simplicity reasons the individual level estimates and standard errors are not shown. Threshold coefficients interpretation can be found in Appendix A.

	Estimate	**Std. error**	**z value**	**P(>| z |)**
Common-rare	0.145	0.028	5.190	<0.001
Decrease-increase	-0.696	0.024	-28.595	<0.001
Foreign-native	0.072	0.027	2.618	0.009
Strong-vulnerable	0.079	0.025	3.144	0.002
Valuable-worthless	-0.391	0.033	12.025	<0.001
Species (non-native plant contrast)				
Large mammal	0.010	0.206	0.050	0.960
Spider	0.262	0.212	1.236	0.216
Country+Country*species	-	-	143.44	<0.001
Threshold coefficients				
Intercepts				
-2/1	2.473	0.163	15.185	<0.001
-1/0	1.202	0.156	7.705	<0.001
0/1	-1.449	0.157	-9.237	<0.001
1/2	-3.292	0.165	-19.962	<0.001
Slopes (Attractive-unattractive)				
-2/1	-0.281	0.044	-6.428	<0.001
-1/0	-0.285	0.039	-7.345	<0.001
0/1	-0.415	0.037	-11.080	<0.001
1/2	-0.325	0.042	-7.685	<0.001
Slopes (Harmful-harmless)				
-2/-1	0.349	0.039	9.054	<0.001
-1/0	0.312	0.033	9.421	<0.001
0/1	0.144	0.032	4.491	<0.001
1/2	0.100	0.037	2.715	0.006

4.4 Conclusions

In this chapter we have assessed how susceptible models are to the way in which we treat variables, in particular in the case of ordinal variables. The results of our simulation studies have shown the advantages of an appropriate ordinal modelling of ordered response data, by highlighting potential differences both at the intercept and slope level. The advantages concern the results both in significance and parameters estimates. Furthermore, interpretation of these results is more accurate and based on thresholds, which in practice may have broad implications. In addition, through two case studies from the psychology literature, we have provided practical evidence of these advantages.

The consequences of an inappropriate numeric modelling approach are twofold: firstly, it does not take the ordered nature (and not necessarily equally spaced categories) of the data into consideration; secondly, it may not provide accurate inference, which from an applied perspective, can lead to the wrong conclusions and ultimately might invalidate research evidence in certain cases.

We have found by means of numerical simulations a considerably higher percentage of simultaneous statistically significant results for the ordinal and the latent variable modelling particularly for large sample sizes, and with different percentages for different threshold patterns. Although ordinal models are trickier to fit, gains from fitting the correct model will become increasingly apparent for larger sample sizes. These differences have also been shown in graphical representations which portray biases in slope and intercept estimation, with substantial differences for the symmetric threshold structure.

The case studies showed that, although it was not clear *a priori* whether treating the response variable as ordinal was going to make a difference, there were differences in the output that were worth taking into account. In the first case study, results from the reported CLM showed noticeable differences when compared to an initial linear model. In contrast, in the second case study, a linear mixed model was fitted which accounted for individual differences between and within respondents, but did not take into account the ordinal nature of the response variable. After fitting the same mixed model in an ordinal framework, the differences between both results were not as extreme as in the first case study but were still noteworthy. Additionally, ordinal response models have allowed us to provide a thorough interpretation that included the comparison of effects across categories through predicted probabilities derived from the threshold coefficients.

Chapter 5

Goodness of fit and residual diagnostics

5.1 Introduction

The available global goodness of fit tests for ordinal models (e.g., Hosmer-Lemeshow as defined in (3.26)) are useful to evaluate model fit. However, we argue that residual diagnostics allow the researcher to assess the model at specific levels (e.g., for each component of the model) and for different types of lack of fit (e.g., violation of the PO assumption), and provide an evaluation of practical significance (and therefore hints at potential model improvements) in contrast to the statistical significance measures derived from the statistical tests.

Within the area of residual diagnostics, there is a lack of consensus on which residuals to use for detecting lack of fit in ordinal response models and little clarity on what the behaviour of these residuals should ideally be (Abreu et al., 2009). This chapter assesses by means of, firstly a series of examples and simulations, and secondly, two case studies, the main residual diagnostics available for these models (which to the best of our knowledge are those defined by Dunn and Smyth, 1996, Li and Shepherd, 2012, and Liu and Zhang, 2018), both graphically and numerically, and highlights advantages and disadvantages of each of these residuals.

5.2 Ordinal residuals

There are several possible ways to measure the contrast between observed and fitted values, and therefore several associated definitions of residuals for ordinal models (Hosmer and Lemeshow, 2000). According to Li and Shepherd (2012), these residuals must present the

following features:

1. *Single value per subject.* Result in only one value per subject irrespective of the number of categories of the ordinal outcome.

2. *Overall direction.* Reflect the overall direction of the observed value compared with the fitted value.

3. *Monotonicity.* Be monotonic with respect to the observed value for those ordinal responses with common covariates. If we observe $x_k = x_j$ and $Y_k < Y_j$, then $r_k < r_j$ almost surely (Liu and Zhang, 2018).

4. *Symmetry around zero.* Have a range of possible values that is symmetric about zero (symmetry of the residual distribution). This leads to the residuals having *zero expectation* $E(r) = 0$. Otherwise, we would have biased predictions as the model would not be fitting the data well.

5. *Order preservation.* Preserve order without assigning arbitrary scores to the categories (i.e., it does not require the assignment of arbitrary numbers to categories).

They also argue that some traditional residuals (see Appendix C for definitions) that could be adapted to ordinal response models (e.g., Pearson and deviance residuals) do not satisfy properties $1-4$. Property 5 does not hold either for Pearson residuals because they are defined as the numeric difference between the fitted and observed values and therefore they are directly affected by the arbitrary choice of scores (Liu and Zhang, 2018). Additionally, visual inspection of these traditional residuals in scatterplots of residuals vs fitted show highly structured artefacts caused by the ordinal nature of the response, which means they provide "very limited meaningful information for model diagnosis" (Feng et al., 2017). For these reasons, we will not assess these residuals designed for continuous response variables. We instead focus on three types of residual diagnostics appropriate for CLMs and describe how they respect the five properties above. To ensure consistency of notation across the three types of residuals, we have slightly modified the definitions from the original references.

5.2.1 Li-Shepherd residuals

For a set of ordered categories $j = 1, \ldots, C$ with order $1 < \ldots < C$, and a fitted distribution for categories in the model F, L-S residuals (Li and Shepherd, 2012) are defined as the mean of the sign contrast of the random response variable Y_i, and the specific observed value j. In other words, the difference between the probability of the ordinal response being less and

being greater than the observed value under the model. We thus define the L-S residual r_i as

$$r_i = E[sign(Y_i, j)] = P(Y_i < j) - P(Y_i > j), \tag{5.1}$$

which for a POM with parameters $\theta = (\boldsymbol{\alpha}, \beta)$ and log-likelihood L_i for observation i as defined in (3.23) can be estimated such that

$$\hat{r}_i = -\sum_{j=1}^{C-1} \frac{\partial L_i}{\partial \alpha_j}\Big|_{\hat{\theta}} \tag{5.2}$$

where $\hat{\theta}$ is the model's ML estimate.

L-S residuals are closely related to other 'traditional' residuals defined for continuous models:

- Score residuals u_i as defined for CLMs (Li and Shepherd, 2012, Therneau et al., 1990) in Appendix C as
$$u_i = \frac{\partial L_i}{\partial \theta}. \tag{5.3}$$
are directly linked to L-S residuals as defined in (5.2). For this reason, they are also known as partially aggregated score residuals.

- For a binary response variable Y with categories 0 and 1, we can define L-S residuals as
$$r_i = Y_i - P(Y_i = 1). \tag{5.4}$$
which corresponds to the definition of the unscaled Pearson residual for binary outcomes (Hosmer and Lemeshow, 2000; see Appendix C).

- L-S residuals also capture information similar to that of latent residuals (also described in Appendix C) and are defined on the probability scale of the fitted distribution irrespective of the choice of link function.

These residuals have the "ideal" features proposed in Li and Shepherd (2012):

- *Single value per subject, overall direction, and order preservation.* True by definition, irrespective of the number of categories of the ordinal response variable.

- *Monotonicity.* For L-S residuals, where we follow the notation in (3.25) for cumulative probabilities $P_{ij} = P(Y_i \leq j)$, we have that since $r_i(j+1, F) - r_i(j, F) = P_{i,j+1} -$

$P_{i,j-1} \geq 0$ [1] then $r_i(j+1,F) \geq r_i(j,F)$, and since $P_{i1} \geq 0$ and $P_{i,C-1} \leq 1$, then $r_i(1,F) = P_{i1} - 1 \geq -1$ and $r_i(C,F) = P_{i,C-1} \leq 1$. We can finally conclude that: $-1 \leq r_i(1,F) \leq \ldots \leq r_i(C,F) \leq 1$.

- *Symmetry around zero with zero expectation.* The fact that its range of possible values is symmetric about zero in $[-1,1]$ is directly derived from $-1 \leq r_i(1,F) \leq \ldots \leq r_i(C,F) \leq 1$. Additionally, given the random response variable Y_i with fitted distribution $F = (p_1, \ldots, p_s)$, then we can consider our residuals as a random variable such that $E(r_i) = \sum_{j=1}^{C} p_j \{P_{i,j-1} - (1 - P_{ij})\} = \sum_{j=1}^{C} p_j P_{i,j-1} - \sum_{j=1}^{C} p_j (1 - P_{ij}) = \sum_{j=1}^{C} p_j (\sum_{k<j} p_k) - \sum_{j=1}^{C} p_j (\sum_{k>j} p_k) = \sum_{j_1<j_2} p_{j_1} p_{j_2} = 0$.

In addition to the above, L-S residuals satisfy the following property that also holds for traditional 'linear' residuals:

- *Zero sum of residuals* such that

$$\sum_{i=1}^{n} \hat{r}_i = -\sum_{i=1}^{n} \sum_{j=1}^{C-1} \frac{\partial L_i}{\partial \alpha_j}\bigg|_{\hat{\theta}} = -\sum_{j=1}^{C-1} \frac{\partial L}{\partial \alpha_j}\bigg|_{\hat{\theta}} = 0. \tag{5.5}$$

We calculate L-S residuals in `R` using the package `PResiduals` (Dupont et al., 2017).

An advantage of these residuals is that they are not randomly generated, which contributes to their user-friendliness. For model diagnostics purposes rather than checking for normality of these residuals, we compare them by definition to a uniform distribution $U(-1,1)$. This stems from the asymptotic behaviour of the residual's variance (Shepherd et al., 2016). Additionally, (Shepherd et al., 2016) state that they are likely to reflect heteroscedasticity when present.

On the negative side, the discreteness of L-S residuals (reflected in the form of banded patterns), and their dependence on the covariates x_i, limit their usefulness in model diagnostics. Some authors (e.g., Arbogast and Lin, 2005) argue that "this approach is highly subjective: it is difficult to determine whether a seemingly unusual residual pattern reflects a faulty functional form [of covariates] or natural variation." Another limitation is that given the fact that they are bounded between -1 and 1, they do not appropriately detect outliers. Finally, they are not useful for checking the PO assumption (Shepherd et al., 2016).

We study further pros and cons in our systematic simulation study in Section 5.3.

5.2.2 Dunn-Smyth residuals

Dunn and Smyth (1996) introduced randomised quantile residuals (D-S residuals hereafter)

[1] $r_i(j+1,F) - r_i(j,F) = P_{ij} + P_{i,j+1} - 1 - (P_{i,j-1} + P_{ij} - 1) = P_{i,j+1} - P_{i,j-1}$.

based on quantile residuals (Brillinger and Preisler, 1983, Machado and Santos-Silva, 2005, and Hong and He, 2010) and the "crude residuals" proposed by Cox and Snell (1968) for binomial data (see Appendix C for further description).

In order to understand these residuals we define first quantile residuals as generated by inverting the *cdf* for each response observation and finding the corresponding standard normal quantile. In consequence, D-S residuals are exactly standard normal under the true model (when the parameters are known; Feng et al., 2017).

We assume that Y_i with $i = 1, \ldots, n$ are independent responses indexed by parameters $\mu_i = E(y_i)$ and $\boldsymbol{\alpha}$. If the *cdf* of Y_i, $F(Y; \mu; \boldsymbol{\alpha})$, is continuous, quantile residuals can be defined as

$$r_i = \Phi^{-1}(F(Y_i; \hat{\mu}_i, \hat{\boldsymbol{\alpha}})), \tag{5.6}$$

where Φ is the cumulative distribution of the standard normal.

Randomisation is introduced to produce continuous normal residuals (Smyth et al., 2017) instead of resulting in a set of parallel curves corresponding to distinct response values which would not be easily interpretable. However, this introduced randomisation means that multiple realisations of the residuals need to be produced to ensure that a pattern is consistent across the different runs.

If $F(Y; \mu; \boldsymbol{\alpha})$ is not continuous, D-S residuals for Y_i can be defined as

$$r_i = \Phi^{-1}(U_i), \tag{5.7}$$

where Φ^{-1} is the quantile function of the standard normal distribution and U_i is a uniform random variable on the interval $(a_i, b_i]$ with $a_i = \lim_{Y \to Y_i} F(Y; \hat{\mu}_i, \hat{\boldsymbol{\alpha}})$ and $b_i = F(Y_i; \hat{\mu}_i, \hat{\boldsymbol{\alpha}})$. To calculate these residuals in `R` we follow the approach for CLMs proposed by Christensen (2012) and the associated function `clm.residuals`.

The main advantages of D-S residuals over L-S residuals are as follows:

- Under the true model D-S residuals follow standard normal distributions when the true parameters are known (Feng et al., 2017). Therefore, we can examine the adequacy of a fitted model by checking if these residuals follow a standard normal distribution, which visually can be done by means of Q-Q plots. This ensures user-friendliness of graphical interpretation, as it is one of the distributions people are used to interpret.

- Additionally, D-S residuals are continuous which ensures that the wide range of tools applicable to continuous variables also apply here. They can be useful for examining linearity, checking for outliers, and measuring residual correlation.

As a limitation, the fact that randomisation is used to ensure that the residuals are con-

tinuous means that from one run to another we will have different plots for the same data set and fitted model. Dunn and Smyth (1996) suggest four runs each time to ensure consistency throughout. In addition to that, jittering is incorporated for visualisation purposes. As an alternative, we use bootstrapping as suggested by Greenwell et al. (2018) for surrogate residuals (described in the next Subsection 5.2.3).

5.2.3 Surrogate residuals

A surrogate approach to ordinal residuals is suggested by Liu and Zhang (2018). The idea is to transform the problem of checking the distribution of an ordinal outcome Y to that of checking the distribution of a continuous outcome S (simulated variable on the latent scale), which they call a surrogate variable. The authors define the continuous variable S_i as a surrogate of Y_i and then obtain residuals r_i based on S_i such that

$$r_i = S_i - E[S_i]. \tag{5.8}$$

S_i is defined by sampling conditionally on the observed ordinal outcomes $Y = (y_1, ..., y_n)$ according to a hypothetical probability model that is coherent with the assumed model for Y_i, and $E[S_i]$ is the expectation calculated under the null hypothesis. In summary, they propose conditional sampling so that the continuous space of the simulated data can be the area of work (i.e., S_i is the simulated value on the latent scale).

An alternative to the generic expression for a CLM expressed in (3.11) is

$$\text{logit}(P(Y_i \leq j)) = \alpha_j + f(x_i, \beta). \tag{5.9}$$

where $i = 1, \ldots, n$ and $j = 1, \ldots, C-1$.

The concept of latent variable induces a joint distribution of Y and a hypothetical variable $Z = -f(x, \beta) + \varepsilon$ where ε follows a certain distribution G. The joint distribution is determined by setting $Y = j$ if $\alpha_{j-1} < Z \leq \alpha_j$, where $j = 1, \ldots, C-1$. Then, the marginal distribution of Y is the same as the distribution specified by the assumed model (5.9) and S follows a truncated normal distribution $S \sim TN(\alpha_{j-1}, \alpha_j)$.

Therefore, we define our residual variable as $r_i = S_i - E[S_i] = S - E[Z_i] = S_i + f(x, \beta) - \int_{-\infty}^{\infty} udG(u)$, where $\int_{-\infty}^{\infty} udG(u)$ is the mean distribution. In practice, given the data (x_i, y_i) and a fitted model, we estimate the conditional distribution $Z_i|Y_i = y_i$ by plugging in the parameter estimates $\hat{\beta}$ and $\hat{\alpha}_j$'s. From the distribution $\hat{f}_\alpha(z|y_i)$, we randomly draw a sample

s_i. Then, the ith residual is

$$\hat{r}_i = s_i + f(x_i, \hat{\beta}) - \int_{-\infty}^{\infty} u dG(u). \tag{5.10}$$

In short, the surrogate approach pursues conditional sampling so that we can work on the continuous space of the simulated data, rather than the discrete space of the original data.

Several advantages over L-S residuals are suggested by the authors:

- Surrogate residuals have a continuous nature, which allows for the use of the more extensive generic diagnostic tools for continuous outcomes. L-S residuals on the contrary are of a categorical nature and therefore can result in "strips" in graphic plots and make visual examination difficult (as we have found out too and is described in Section 5.3).

- The null distribution of the surrogate residual is independent of the covariates x_i which simplifies visual checks as stated by Liu and Zhang (2018), whereas L-S residuals have a null distribution and variance that depend on x and vary across the values of it which limits their utility.

- Surrogate residuals have an explicit null distribution related to the link function G^{-1} such that $G(c + \int u dG(u))$, that is, $P(r \leq c|x) = P(r \leq c) = G(c + \int u dG(u))$.

- Surrogate residuals provide information on the components of the model that are misspecified (heteroscedasticity for instance) and therefore might help improving the model fit.

The surrogate approach is overall also described by the authors as 'broader' than those using latent variables. This relates to the fact that it is not restricted to the logit, but would apply to other distributions too. Additionally, the authors claim that the jittering technique used in D-S residuals (Hong and He, 2010, Machado and Santos-Silva, 2005, Stevens, 1950), where an independent noise variable is added to "smooth" the discrete outcome is a special case of the surrogate method. We aim to make this comparison explicit and assess whether surrogate residuals are in practice the same as D-S residuals for the specific case where we have a CLM (where the G distribution in the surrogate approach would be the normal distribution for D-S). That is to say that when G is normal, then the surrogate residuals would be equivalent to the D-S residuals for any model (including CLMs).

We calculate these residuals in `R` using the package `sure` (Greenwell et al., 2018) that generates a bootstrap sample B of n_B surrogate residuals r_i; the residual scatterplots include

all the n_B residuals per plot, and the Q-Q plots represent the median of the B bootstrap distribution (Greenwell et al., 2018, Liu and Zhang, 2018).

5.3 Simulation study

We use a simulation study to systematically compare the performance across a range of scenarios of the different ordinal diagnostic methods (L-S, D-S, and surrogate residuals in particular) and to ultimately answer the following questions:

(i) Do the ideal properties hold for the three types of residuals in practice?

(ii) Do these residuals accurately reflect lack of fit or model misspecification for all the scenarios? Are all the model misspecifications (e.g., mean structure misspecification, link misspecification, odds proportionality, heteroscedasticity, and missing covariates and interaction terms) reflected in the residual patterns?

(iii) Are Q-Q plots sensible for all the residuals? Do they detect the different patterns of violation of distributional assumptions? Are there any other suitable plots?

(iv) Do D-S and surrogate residuals provide ostensibly the same results?

We address (i) in Subsection 5.3.1, and (ii) in 5.3.2.1, 5.3.2.2, and 5.3.2.3. We look at (iii) in detail in 5.3.2.4, and (iv) is discussed in 5.3.2.5. In addition to the above aims, we assess throughout the ease of interpretation of each of the residual types in search of potential user-friendly solutions that promote uptake of the corresponding ordinal models.

5.3.1 Graphical assessment of residual properties

We start by visually assessing the ideal properties of the three ordinal residuals in practice using a modified version of one of the examples in Liu and Zhang (2018) where we simulate an ordinal response Y with $C = 4$ categories determined by equidistant thresholds $\boldsymbol{\alpha} = (0, 4, 8)$ from a latent continuous variable $Y^* = 16 - 8x + \varepsilon$, with a random covariate $x \sim N(3, 2)$ and random errors $\varepsilon \sim N(0, 1)$. We fit a CLM to these data (the authors use a probit link function instead) such that

$$\text{logit}(P(Y_i \leq j)) = \alpha_j + \beta x_i, \quad i = 1, \ldots, 2000; j = 1, 2, 3, \tag{5.11}$$

with estimates $\hat{\boldsymbol{\alpha}} = (-8.6, -6.5, -4.5)$ and $\hat{\beta} = 4.4$. We determine each of the three types of residuals for these simulated data, while checking whether the ideal features hold for all three. The three residual classes have expectation zero, preserve monotonicity and show an overall upward direction and order with the response categories (most noticeable for surrogate residuals) for both POM and PPOM as shown in Figure 5.1. We find L-S, D-S and surrogate residuals to be approximately symmetric around zero for both POM and PPOM (see Figure 5.2). This is true for multiple runs of the random D-S residuals.

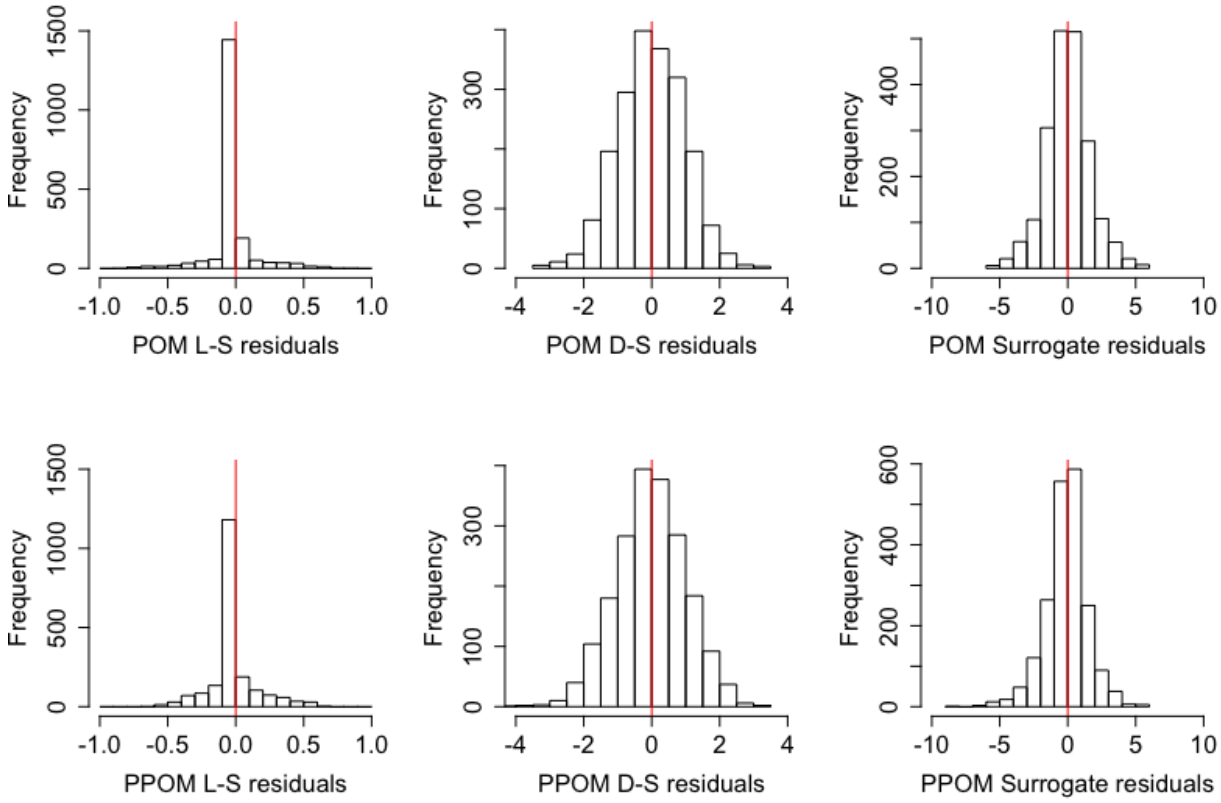

Figure 5.2: Histograms of L-S (left), D-S (middle) and surrogate residuals (right) for a POM (top) and a PPOM (bottom).

5.3.2 Assessment of effects of model misspecification

We continue by simulating data from the specific ordinal models stated below and assessing the residuals from different forms of model misspecification as described in the following scenarios:

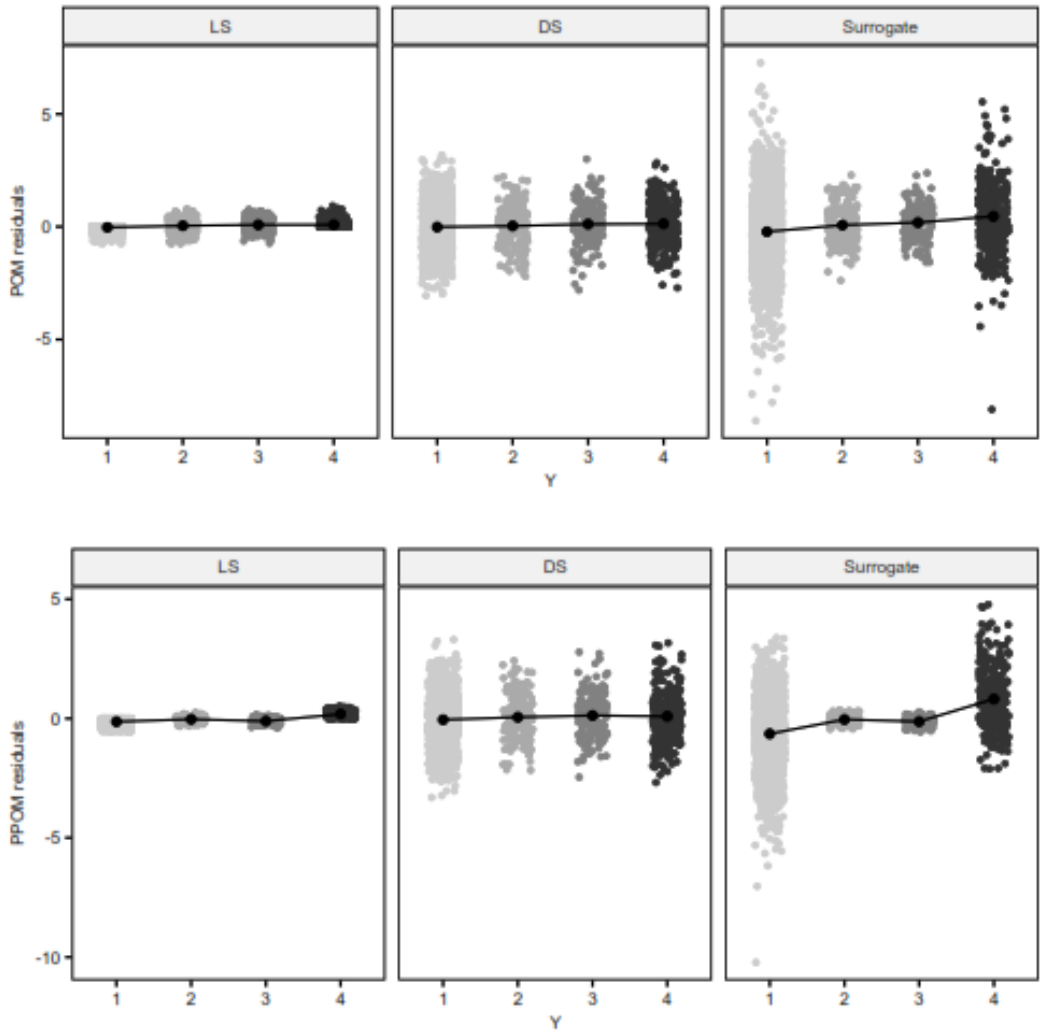

Figure 5.1: Mean values of L-S (left), D-S (middle) and surrogate residuals (right) for a POM (top) and a PPOM (bottom) according to the response categories, for a single simulated dataset.

Scenario 1. PO misspecification

In this first scenario, we assess the difference in effects of misspecification and imposing the proportionality of odds in our CLMs. We fit a POM such that

$$\text{logit}(P(Y_i \leq j)) = \alpha_j + \beta x_i, \quad i = 1, \ldots, n; j = 1, \ldots, C-1, \tag{5.12}$$

to data generated from a PPOM defined as

$$\text{logit}(P(Y_i \leq j)) = \alpha_j + \beta_j x_i, \quad i = 1, \ldots, n; j = 1, \ldots, C-1, \tag{5.13}$$

with $x \sim N(0, 1)$, $n = 500$, $C = 6$, and threshold structures:

- symmetric with $\boldsymbol{\alpha} = (-2.5, -1.2, 0.0, 1.2, 2.5)$,
- equidistant with $\boldsymbol{\alpha} = (-3.0, -1.5, 0.0, 1.5, 3.0)$,
- unconstrained with $\boldsymbol{\alpha} = (-2.5, -1.0, 0.0, 0.5, 3.0)$.

Shepherd et al. (2016) argue that L-S residuals are not sensitive to this misspecification and we aim to assess whether this is also true for the other residuals.

Scenario 2. Link misspecification

In this scenario we misspecify the link function in our model as logit (and probit) alternatively for ordinal data simulated from a cumulative Cauchit model where the latent variable is defined according to (1.1) and includes a quadratic term in x (Liu and Zhang, 2018) such that

$$Y^* = 16 - 8x + x^2 + \varepsilon, \tag{5.14}$$

with $\varepsilon \sim Cauchy(n, 0, 1)$, $x \sim U(1, 7)$, $n = 500$, $C = 5$ and *vice versa*, we fit a cumulative Cauchit model such that

$$\text{Cauchit}(P(Y_i \leq j)) = \alpha_j + \beta x_i, \quad i = 1, \ldots, n; j = 1, \ldots, C-1, \tag{5.15}$$

for data generated from a CLM (and cumulative probit model)

$$\text{logit}(P(Y_i \leq j)) = \alpha_j + \beta x_i, \quad i = 1, \ldots, n; j = 1, \ldots, C-1, \tag{5.16}$$

with the following threshold structures (as shown in Figure 5.3)

- symmetric with $\boldsymbol{\alpha} = (-36, -6, 34, 64)$,

- equidistant example 1 with $\boldsymbol{\alpha} = (-16, -12, -8, -4)$,
- equidistant example 2 with $\boldsymbol{\alpha} = (0, 4, 8, 12)$,
- unconstrained with $\boldsymbol{\alpha} = (-1.5, 0, 1, 3)$.

We anticipate heavy tails and kurtotic patterns to be reflected in the residuals as a consequence of this misspecification. Both the symmetric and the equidistant example 1 threshold structures result in a skewed distribution of the data, with sparse categories (other than the third category for symmetric and the fifth for the equidistant). We expect these to have an impact on the corresponding residuals too.

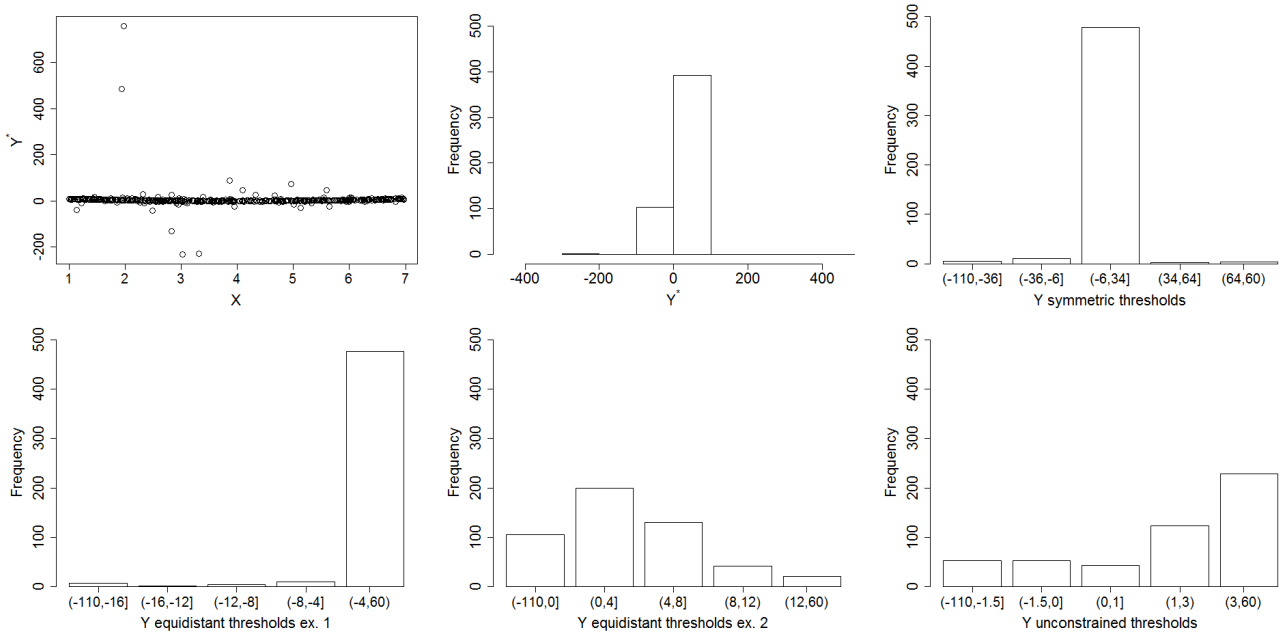

Figure 5.3: Latent response Y^* and threshold structures defining the ordinal response Y (*Scenario 2*).

We acknowledge as a limitation of this scenario that for non-linear models in practice, the model-fitting process automatically compensates link misspecifications by attenuating regression coefficients (Neuhaus 1999). As noted by Liu and Zhang (2018): "As a result, the assumed model may provide an adequate approximation to the true $P(Y = 1)$ (Neuhaus, 1999), and diagnostics could be very difficult."

Scenario 3. Missing quadratic term

In this scenario we consider the case of model misspecification where a straight-line model is fitted to a data set which exhibits a quadratic relationship with $x \sim U(1, 7)$ and

$\varepsilon \sim N(0,1)$, such that

$$Y^* = 16 - 8x + x^2 + \varepsilon, \tag{5.17}$$

(whose relationship with Y is determined by (1.1)).

This quadratic relationship has not been reflected in the POM

$$\text{logit}(P(Y_i \leq j)) = \alpha_j + \beta x_i, \quad i = 1, \ldots, n; j = 1, \ldots, C-1, \tag{5.18}$$

and *vice versa*, where $n = 500$ and $C = 4$. To further assess the patterns in the residuals, we also fit a linear and a quadratic model to the residuals such that

$$r_i = \alpha + \beta x_i + \delta x_i^2 + \varepsilon, \quad i = 1, \ldots, n; \varepsilon \sim N(0,1). \tag{5.19}$$

We generate the data by means of the following threshold structures (as shown in Figure 5.4)

- symmetric with $\boldsymbol{\alpha} = (-2, 0, 2)$,
- equidistant with $\boldsymbol{\alpha} = (0, 4, 8)$,
- unconstrained with $\boldsymbol{\alpha} = (-1, 2, 9)$.

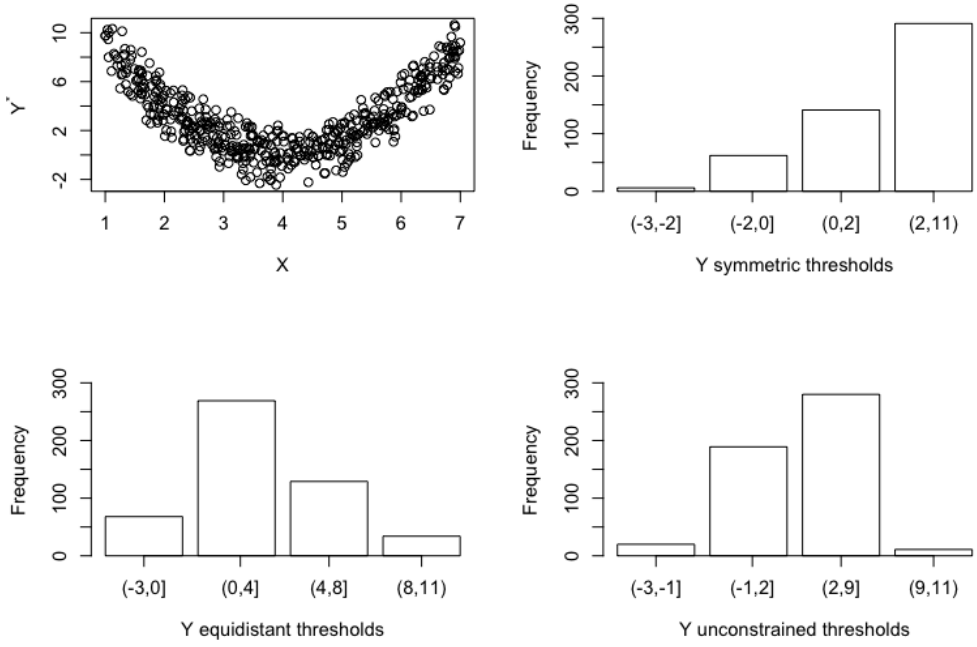

Figure 5.4: Latent response Y^* and threshold structures defining the ordinal response Y (*Scenario 3*).

We consider a range of related mismatches through three levels of comparison: generating model, fitted model, and a third fitted model for the residuals. We consider only 5 combinations of special interest, where 'quadratic' refers to an ordinal model including both the linear (x) and the quadratic term (x^2) such that $r = \beta_1 x + \beta_2 x^2$, and 'linear' refers to a model with the linear term only (x) such that $r = \beta_1 x$.

Table 5.1: Subscenarios for the quadratic model scenario.

	Generating model	Fitted model	Residuals fitted model
1	Quadratic	Quadratic	Quadratic
2	Quadratic	Quadratic	Linear
3	Linear	Linear	Linear
4	Quadratic	Linear	Quadratic
5	Quadratic	Linear	Linear

Out of the scope of this simulation study but potentially a useful tool to identify this kind of misspecification is the so-called link test proposed by Pregibon (1979).

Scenario 4. Missing covariate

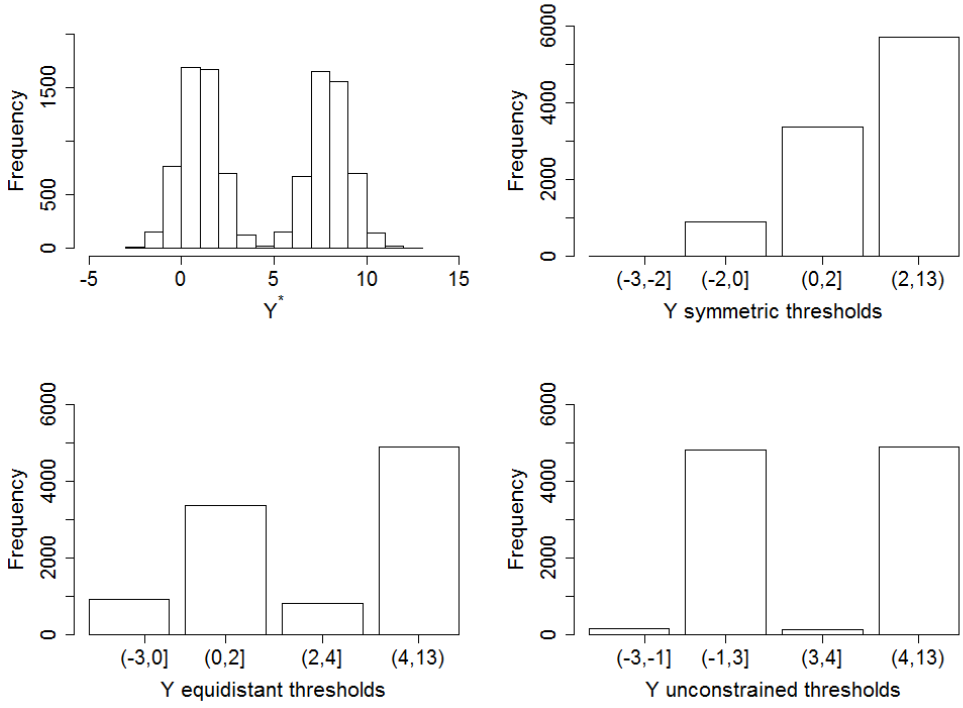

Figure 5.5: Latent response Y^* and threshold structures defining the ordinal response Y (*Scenario 4*).

We now consider the performance of the three candidate residual definitions for data generated from a model with an extra covariate z

$$Y^* = 2 - x + 7z + \varepsilon, \tag{5.20}$$

we are not accounting for in the fitted model

$$\text{logit}(P(Y_i \leq j)) = \alpha_j + \beta x_i, \quad i = 1, \ldots, n; j = 1, \ldots, C-1, \tag{5.21}$$

where $n = 10,000$ and $C = 4$. We consider $x \sim N(1, 0.3)$, $z \sim B(10000, 0.5)$, and $\varepsilon \sim N(0, 1)$ defining the latent variable Y^* (whose relationship with Y is determined by (1.1)), and the following threshold structures delimiting the categories of the ordinal response variables Y_i for our models (as shown in Figure 5.5).

- symmetric with $\boldsymbol{\alpha} = (-2, 0, 2)$,
- equidistant with $\boldsymbol{\alpha} = (0, 2, 4)$,
- unconstrained with $\boldsymbol{\alpha} = (-1, 1, 2)$.

This is called "omitted variable bias" and can lead to biased estimates of the model coefficients.

Scenario 5. Heteroscedastictiy

In this final scenario we assess the effect of model heteroscedasticity in the residuals under study. We consider a non-constant error variance dependent on the covariate σ_x:

$$\text{logit}(P(Y_i \leq j)) = \big(\alpha_j + \beta x_i\big)/\sigma_x, \quad i = 1, \ldots, n; j = 1, \ldots, C-1, \tag{5.22}$$

which is not accounted for in the POM:

$$\text{logit}(P(Y_i \leq j)) = \big(\alpha_j + \beta x_i\big), \quad i = 1, \ldots, n; j = 1, \ldots, C-1. \tag{5.23}$$

This misspecification is expected to result in 'funnel' patterns, and according to Shepherd et al. (2016), these should be reflected in L-S residuals plots.

In this specific case we fit the POM to data simulated from a probit model with heteroscedasticity $\sigma_x = x^2$, an ordered response variable Y with $C = 5$ response categories, $n = 2000$, symmetric thresholds $\boldsymbol{\alpha} = (-36, -6, 34, 64)$, $\beta = 4$, and $x \sim U(2, 7)$ (data set `df2` available in `R` package `sure`).

The above five scenarios represent some of the most relevant potential model misspecifications and where available, model parameters have been chosen following those in Liu and Zhang (2018). While some of these have also been assessed in Greenwell et al. (2018), we expand their work by also looking at CLMs (rather than probit) and by considering slightly modified scenarios and additional threshold structures. We assess the three types of residuals (in contrast to Liu and Zhang (2018) which only cover L-S and surrogate residuals) of the misspecified models and compare them to those for the true model (in a similar fashion to Arbogast and Lin, 2005). We study individually each of these cases as shown in Appendix D and summarise them in the following subsections.

5.3.2.1 Sensitivity to PO misspecification

While some traditional residuals (e.g., score and partial residuals; Harrell, 1996) are supposed to be very sensitive to violations of the PO assumption, Shepherd et al. (2016) state that L-S residuals "may provide little or no information on the adequacy of [...] the PO assumption". We have dedicated Scenario 1 to study this for both traditional and ordinal residuals and for the three types of thresholds.

Scenario 1. PO misspecification.

In addition to replicating the work of Harrell (2018) who only compares score, binary score, partial residuals, and L-S residuals using the package `rms`, we also check the sensitivity of D-S and surrogate residuals to a PO misspecification, by simulating an ordered response variable Y with $C = 6$ categories ($n = 500$) from a POM and a PPOM respectively versus a normal random covariate $x \sim N(0, 1)$ by means of two sets of POMs. We intend to reach wider conclusions applying to both traditional and ordinal-specific residuals, and different threshold structures.

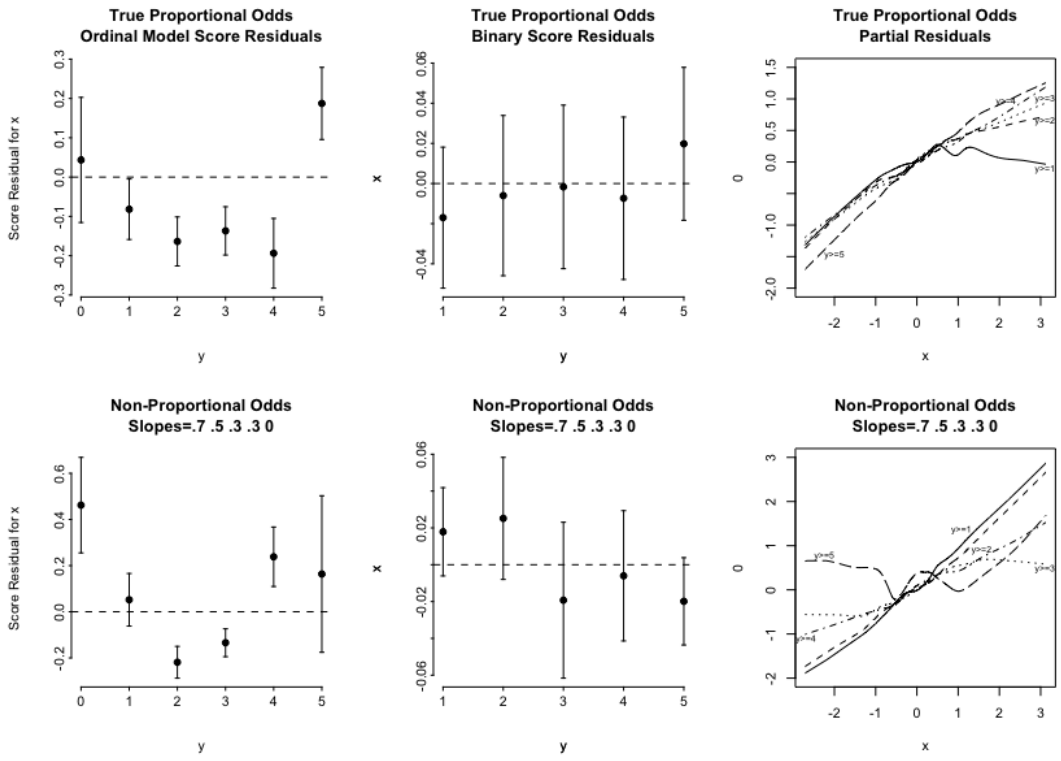

Figure 5.6: Score residuals (left), binary score residuals (middle), and partial residuals (right) for POMs fitted to data generated from a POM with symmetric thresholds $\boldsymbol{\alpha} = (-2.5, -1.2, 0.0, 1.2, 2.5)$ (top) and data generated from a PPOM with the same threshold structure (bottom). 95% confidence bands are provided for score and binary score residuals.

We start by assessing the case where we impose *symmetric thresholds* $\boldsymbol{\alpha} = (-2.5, -1.2, 0.0, 1.2, 2.5)$. We first plot the resulting traditional residuals for a POM of the data generated from the above mentioned POM and PPOM (see Figure 5.6).

We continue studying the model's fit by representing the ordinal residuals vs covariate for the same models (Figure 5.7), and the corresponding Q-Q plots (Figure 5.8) to further assess the effect of the proportionality misspecification on the residual distribution assumptions.

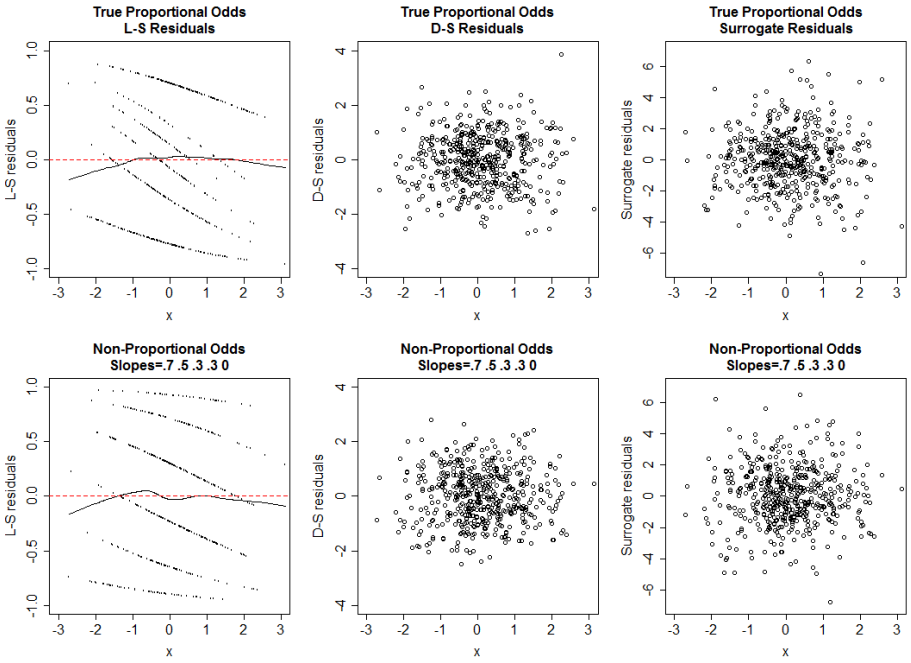

Figure 5.7: Residuals vs covariate plots for POM for data generated from a POM (top) and data generated from a PPOM (bottom) with symmetric thresholds $\boldsymbol{\alpha} = (-2.5, -1.2, 0.0, 1.2, 2.5)$. Left: L-S residuals, Middle: D-S residuals, Right: surrogate residuals.

From the plots we can infer that the different types of residuals reflect differently misspecifications with respect to the PO assumption. In the hypothetical case when the PO assumption holds and the covariate behaves linearly, we would expect in Figure 5.6 a U-shaped pattern in the score residuals plots (a), a horizontal pattern in the binary score plots (b) and parallel curves in the partial residual plots (c). Score residuals do not seem to reflect the misspecification for our example while the binary score residuals are probably the ones reflecting the misspecification better when we model the PPOM generated data as POM. Both Figures 5.7 and 5.8 show again very similar patterns for the correctly specified POM and the wrongly modelled PPOM in the cases of L-S, D-S and surrogate residuals. Both L-S residuals versus covariate plots show parallel curves that are difficult to interpret, while the corresponding Q-Q plots comparing the residual distribution to a uniform $U(-1, 1)$ reflect that the distribution assumptions hold for both the POM and the PPOM. In the cases of D-S and surrogate residuals, both the scatter clouds for the residuals versus covariate plots and the good fit to the normal distribution shown in the Q-Q plots, show no evidence of poor fit irrespectively of the POM or PPOM specification. However, both Q-Q plots for L-S and surrogate residuals seem to present worse patterns when the correct POM model is fitted (Figure 5.8).

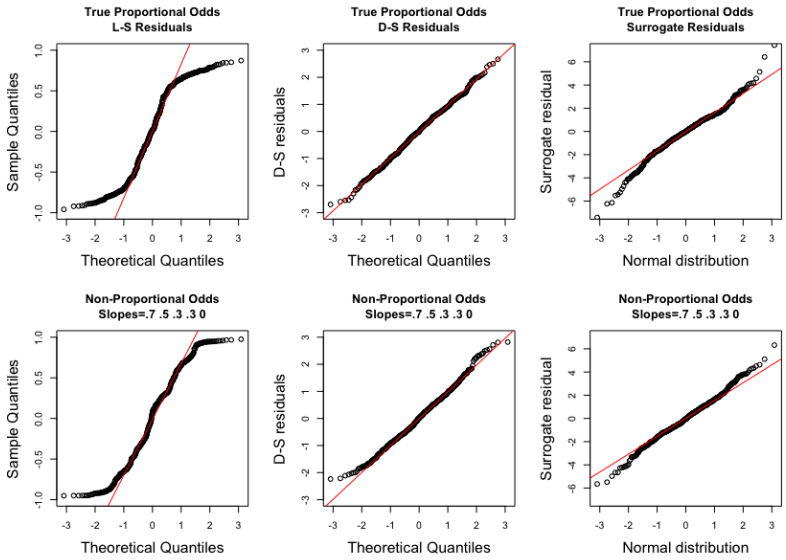

Figure 5.8: Q-Q plots for POM for data generated from a POM (top) and data generated from a PPOM (bottom) with symmetric thresholds $\boldsymbol{\alpha} = (-2.5, -1.2, 0.0, 1.2, 2.5)$: Left: L-S residuals, Middle: D-S residuals, Right: surrogate residuals. Theoretical quantiles for L-S residuals follow a uniform distribution $U(-1, 1)$ while for D-S residuals and surrogate residuals they follow a standard normal distribution $N(0, 1)$.

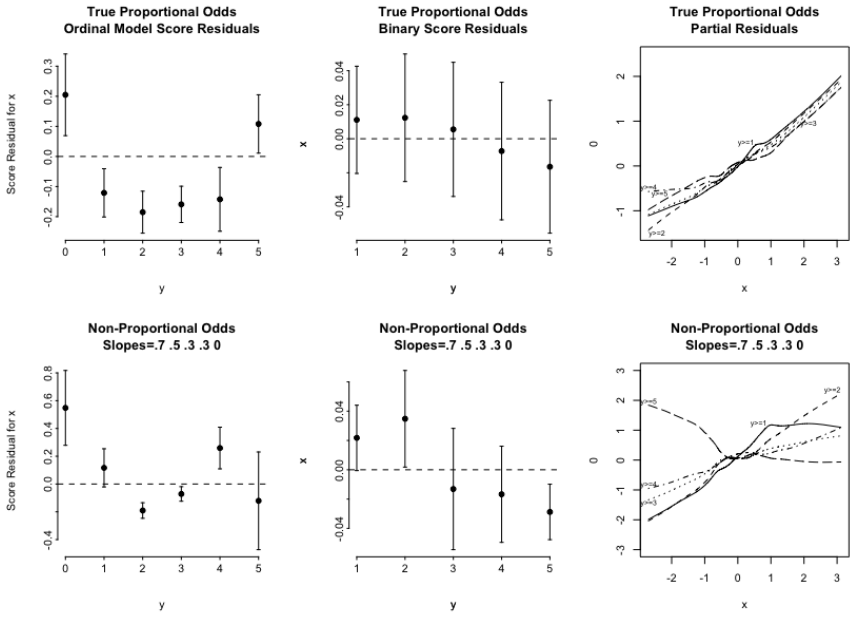

Figure 5.9: Score residuals (left), binary score residuals (middle), and partial residuals (right) for data generated from a POM with equidistant thresholds $\boldsymbol{\alpha} = (-3.0, -1.5, 0.0, 1.5, 3.0)$ (top) and data generated from a PPOM with the same threshold structure (bottom). 95% confidence bands are provided for score and binary score residuals.

For the same POM with *equidistant thresholds* $\boldsymbol{\alpha} = (-3.0, -1.5, 0.0, 1.5, 3.0)$ (which are in fact also symmetric at the same time), there are clear differences in the residuals patterns as a result of the misspecification (see Figure 5.9). In particular, the binary score residuals shown in the bottom plot are further apart from the horizontal pattern, and the partial residuals clearly show non-parallel cumulative curves. However, the differences are not apparent for L-S, D-S, and surrogate residuals (see Figures D.1 and D.2 in Appendix D).

For a version of the POM with *unconstrained thresholds* $\boldsymbol{\alpha} = (-2.5, -1.0, 0, 0.5, 3.0)$, we find again that the partial residuals (Figure 5.10) reflect the differences the most (compared to L-S, D-S, and surrogate residuals as shown in Figures D.3 and D.4)

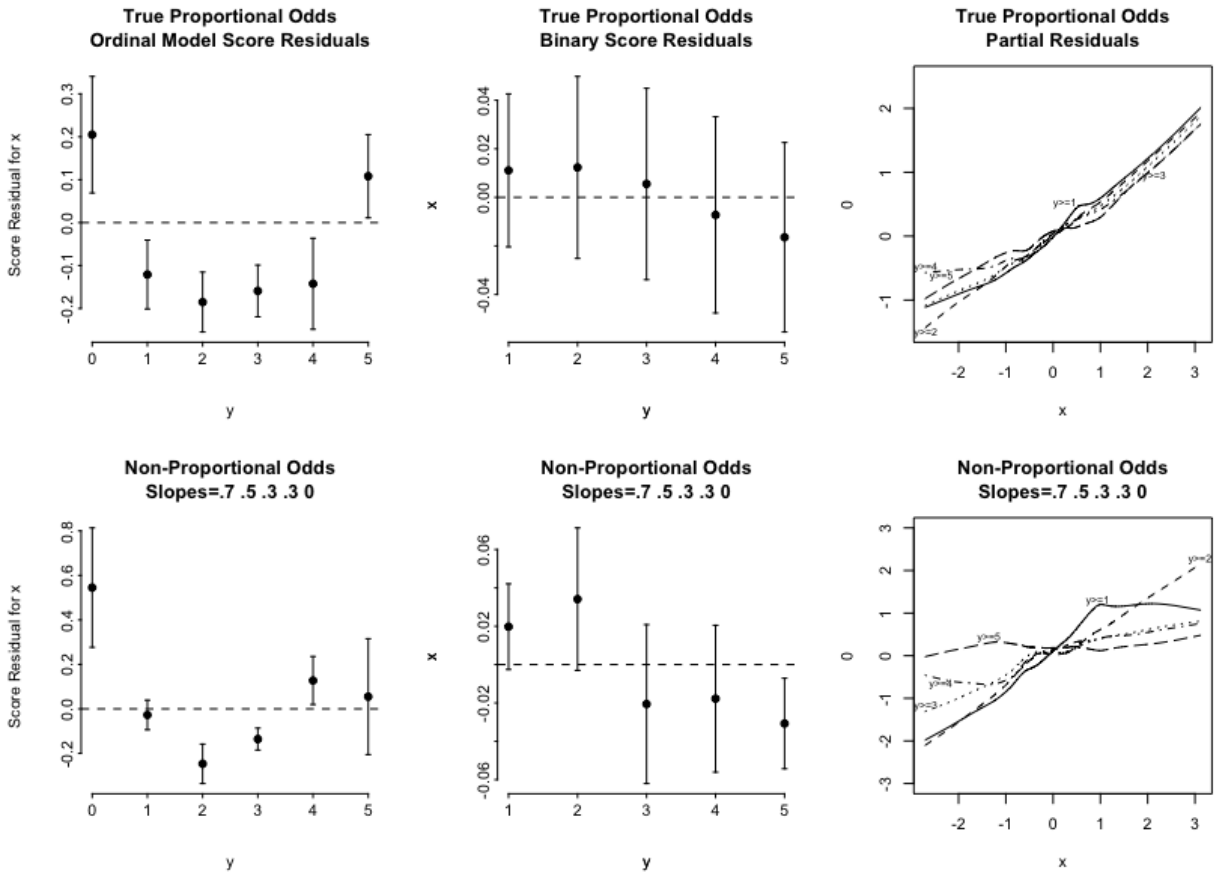

Figure 5.10: Score residuals (left), binary score residuals (middle), and partial residuals (right) for data generated from a POM with unconstrained thresholds $\boldsymbol{\alpha} = (-2.5, -1.0, 0.0, 0.5, 3.0)$ (top) and data generated from a PPOM (bottom). 95% confidence bands are provided for score and binary score residuals.

In summary, while binary score and partial residuals seem to show different patterns for odds proportionality misspecification with equidistant and unconstrained thresholds struc-

tures they are known to not perform as well for other misspecifications and will not be the focus of out chapter. In addition, we have seen that the residuals of interest -L-S, D-S, and surrogate residuals- do not visually reflect this misspecification. For this reason and for practicality, in our next example we ignore potential PPOMs and model the generated ordinal data only as POM.

5.3.2.2 Sensitivity to threshold structures

As we have seen in the previous subsection, different thresholds categorising the latent variable have an effect on the residuals. This is also most apparent in the results from Scenarios 2 and 4.

Scenario 2. Link misspecification.

In this scenario we find that the residual plots corresponding to *equidistant* and *symmetric threshold* structures with $\boldsymbol{\alpha} = (-16, -12, -8, -4)$ and $\boldsymbol{\alpha} = (-36, -6, 34, 64)$ respectively, fail to reflect the misspecification (see Figure 5.11 representing the empirical cumulative distribution function (*ecdf*) of the D-S residuals for the equidistant case for instance), while the unconstrained structure accurately shows the misspecification for logit and probit cumulative link models. Shapiro-Wilk tests of the D-S residuals for the cumulative logit and probit models in the equidistant subscenario are statistically significant at the 5% significance level in 97.4% of the 10,000 runs of the simulation, while the percentage is only slightly higher (97.5%) for the cumulative Cauchit models. Given the results of the corresponding Kolmogorov-Smirnov tests for the comparisons logit-Cauchit and probit-Cauchit in which 96.1% of the p-values are statistically significant at the 5% significance level, we can also justify the failure of these two subscenarios to reflect the link misspecification.

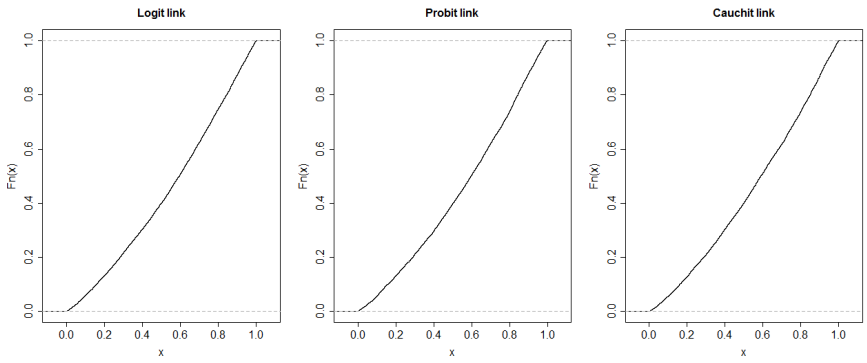

Figure 5.11: *ecdf* of p-values from Shapiro-Wilk tests for D-S residuals of cumulative link models with: (a) logit, (b) probit, and (c) Cauchit link functions, and equidistant thresholds $\boldsymbol{\alpha} = (-16, -12, -8, -4)$ for data generated from a Cauchy random distribution.

Scenario 4. Missing covariate.

In this scenario we find that the *equidistant thresholds* specification $\boldsymbol{\alpha} = (0, 2, 4)$ seems to be overall the one for which a bigger noticeable difference is present between the wrong and right model for D-S residuals (as shown in Figures 5.13 and 5.15; corresponding plots for a cumulative probit can be found in Appendix D). We appreciate unusual patterns for the L-S residuals which are most prominent for *unconstrained thresholds* $\boldsymbol{\alpha} = (-1, 1, 2)$ (see Figure 5.14), but we believe these are an artefact related to the discrete nature of these residuals (similarly to the banding in residuals vs covariate plots).

Of particular relevance is the clear positive effect of the threshold specification in the model fitting using the `ordinal` function `clm`. We have found that by specifying the equidistant structure, the corresponding D-S residuals (the only ones we can use with `clm`) accurately reflect the unexplained variance misspecification (see Figure 5.15). That is, D-S residual plots only reflect the omitted variable bias if the correct threshold structure is specified.

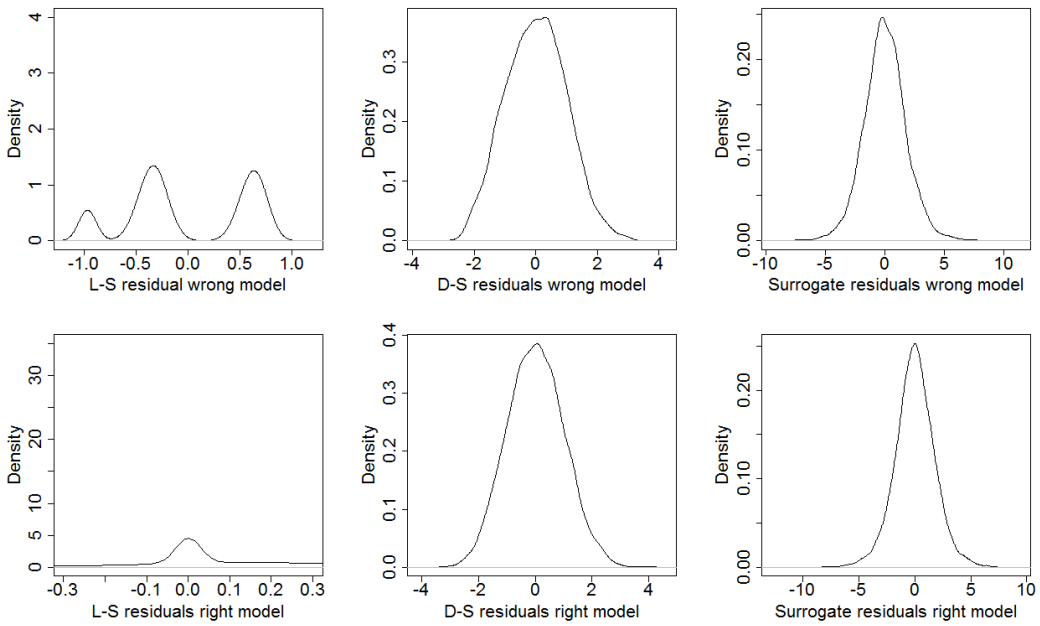

Figure 5.12: Comparison of density plots for ordinal residuals with symmetric thresholds $\boldsymbol{\alpha} = (-2, 0, 2)$ for the wrong CLM missing one of the covariates (top) versus right model with both covariates (bottom). Left: L-S residuals, middle: D-S residuals, right: surrogate residuals.

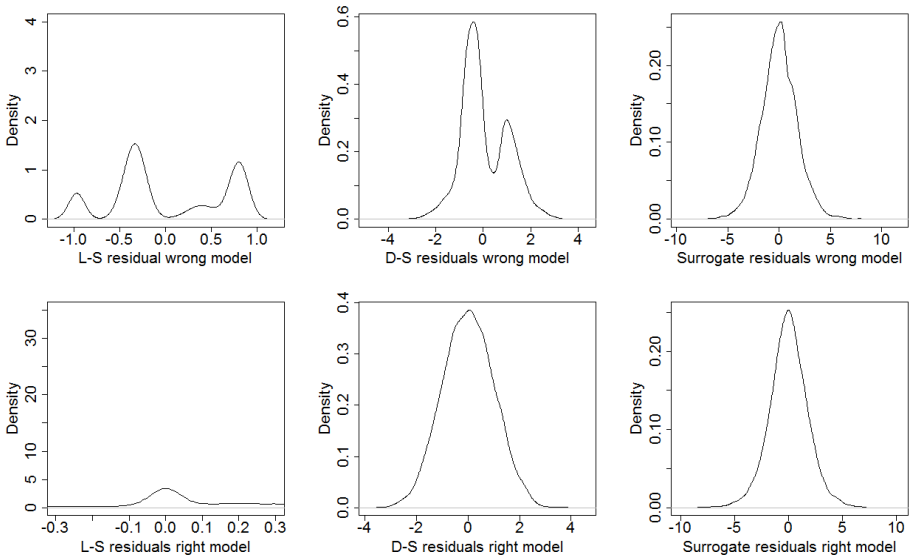

Figure 5.13: Comparison of density plots for ordinal residuals with equidistant thresholds $\boldsymbol{\alpha} = (0, 2, 4)$ for the wrong CLM missing one of the covariates (top) versus right model with both covariates (bottom). Left: L-S residuals, middle: D-S residuals, right: surrogate residuals.

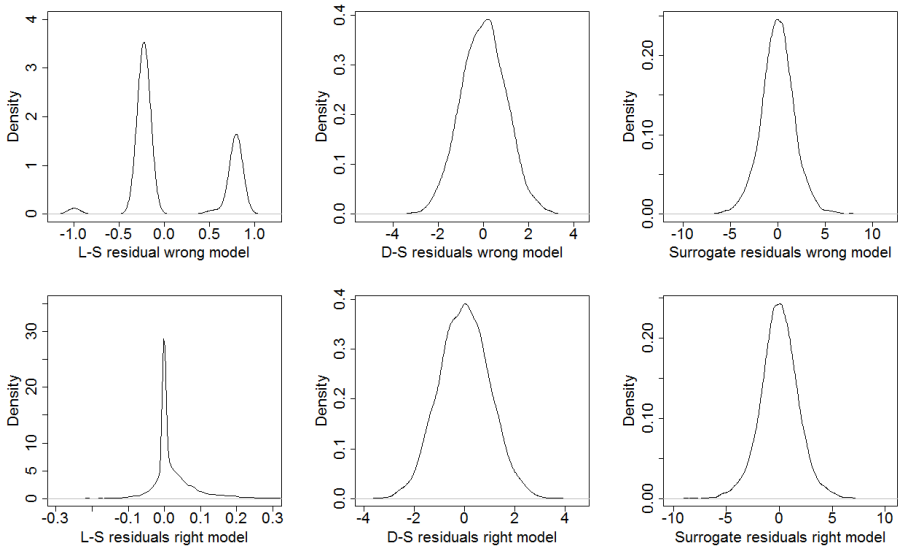

Figure 5.14: Comparison of density plots for ordinal residuals with unconstrained thresholds $\boldsymbol{\alpha} = (-1, 1, 2)$ for the wrong CLM missing one of the covariates (top) versus right model with both covariates (bottom). Left: L-S residuals, middle: D-S residuals, right: surrogate residuals.

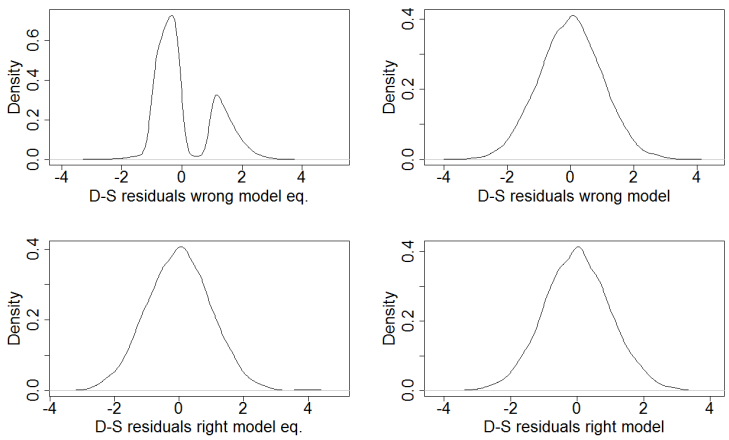

Figure 5.15: Comparison of density plots for ordinal residuals with equidistant thresholds $\boldsymbol{\alpha} = (0, 2, 4)$ and equidistant thresholds specification in `clm` (left) function for wrong CLM missing one of the covariates (top) versus right model with both covariates (bottom).

5.3.2.3 Limitations of L-S residuals

Given the semi-discrete patterns or banding in most of L-S residuals, we have concluded that L-S residuals are not particularly helpful for ordinal models residual diagnostics. This has been mainly reflected in residuals vs covariate plots and Q-Q plots specifically for Scenario 3.

Scenario 3. Missing quadratic term.

For instance, following Liu and Zhang (2018) and as described in (5.17) we generate the model of the ordinal response variable Y as having a linear term in $x \sim U(1, 7)$ and a quadratic term x^2, and fit the data via a CLM reflecting both the linear and the quadratic term too (*Subscenario 3.1. Quadratic-Quadratic-Quadratic* and *Subscenario 3.2. Quadratic-Quadratic-Linear*). We consider the latent variable as an ordinal variable with categories defined by the *equidistant thresholds* $\boldsymbol{\alpha} = (0, 4, 8)$. As shown in Figure 5.16 for one of the runs, while the surrogate and D-S residuals show normality of the residuals, L-S residuals do not follow a uniform distribution and present unusual patterns when plotted versus the covariates. These patterns are consistent throughout all the runs in our simulation for this specific case and the other subscenarios and for the probit link function too (see Appendix D).

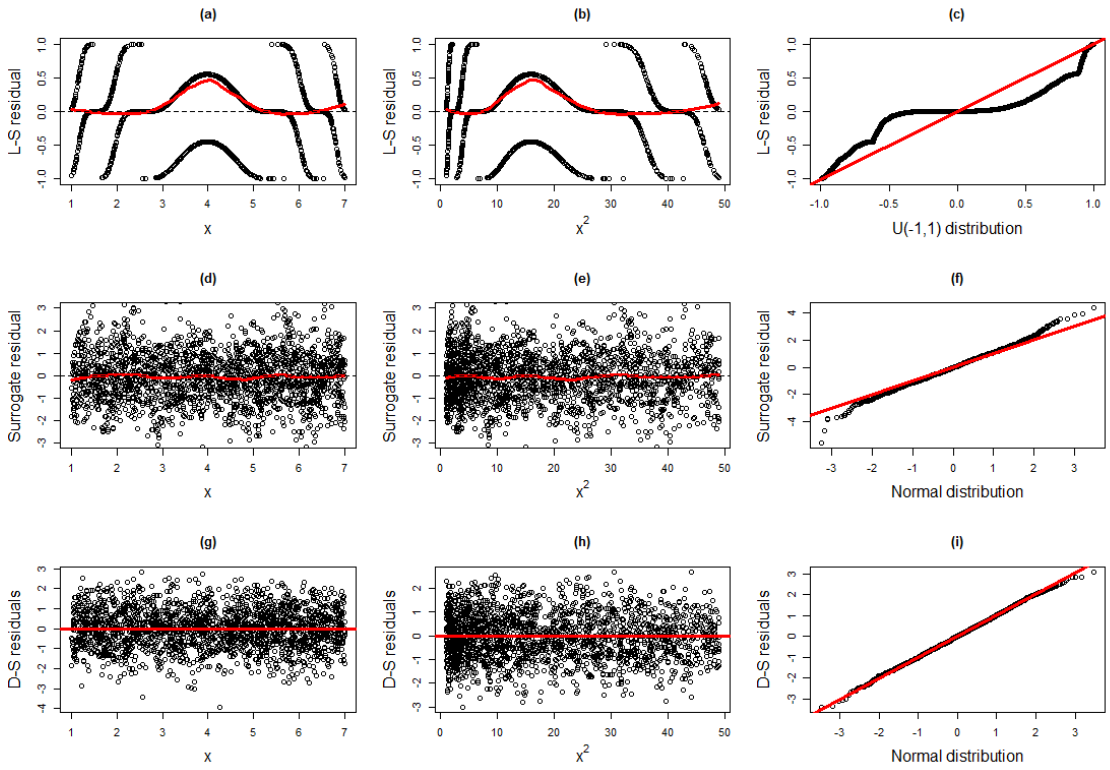

Figure 5.16: Comparison of L-S (top), surrogate (middle) and D-S (bottom) residuals for the true CLM including a quadratic term (*Quadratic-Quadratic subscenario*), and an equidistant threshold structure $\boldsymbol{\alpha} = (0, 4, 8)$.

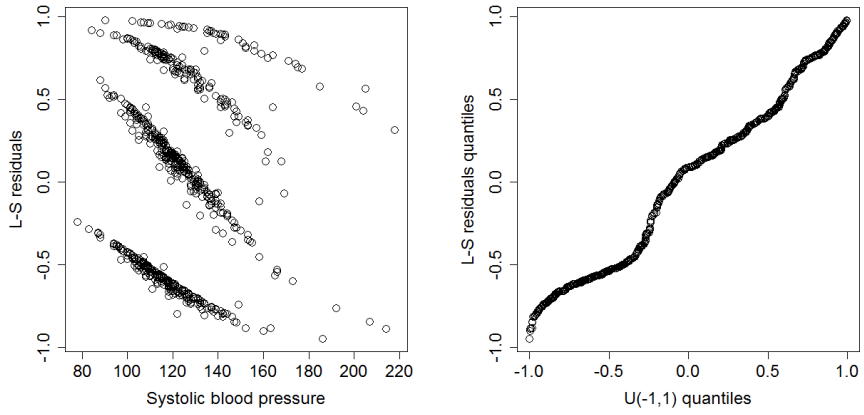

Figure 5.17: POM L-S residuals for eye retinopathy data set.

In the retinopathy example (introduced in Chapter 2), we assess the L-S residuals for the POM[2] (see Figure 5.17). The discrete "striped" pattern in the plot on the left show the

[2]L-S residuals cannot be calculated for PPOMs with the currently available software.

difficulty when trying to assess these residuals for ordinal response models.

5.3.2.4 Comparison of D-S and surrogate residuals

D-S and surrogate residuals, overall, seem to provide more consistent information. In particular, they provide very similar results and we have found that D-S and surrogate behave in similar fashion to different misspecifications. Most noticeable patterns appear in Scenarios 3 and 5.

Scenario 3. Missing quadratic term.

We can see for example that for the *Subscenario 3.2. Quadratic-Quadratic-Linear*, there is a high correlation between both residuals for *unconstrained thresholds* as shown in Figure 5.18.

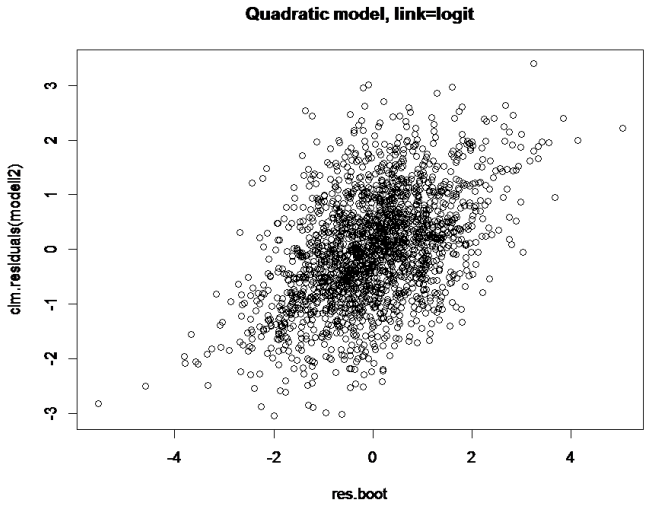

Figure 5.18: Comparison of D-S and surrogate residuals for the *Quadratic-Quadratic-Linear subscenario.*

Additionally, surrogate residuals appear to be more sensitive to quadratic misspecifications, and fail to follow the expected normal distribution for *3.2. Quadratic-Quadratic-Linear*, *3.4. Quadratic-Linear-Quadratic*, and *Subscenarios 3.5. Quadratic-Linear-Linear* for the wrong CLMs (as expected; see Appendix D).

Scenario 5. Heteroscedasticity.

Surrogate residuals also appear to be more sensitive to heteroscedasticity, as the 'funnel' patterns in Figure 5.19 show for a CLM with *unconstrained thresholds* $\boldsymbol{\alpha} = (-36, -6, 34, 64)$, $\beta = 4$, $x \sim U(2, 7)$, and $\sigma_x = x^2$.

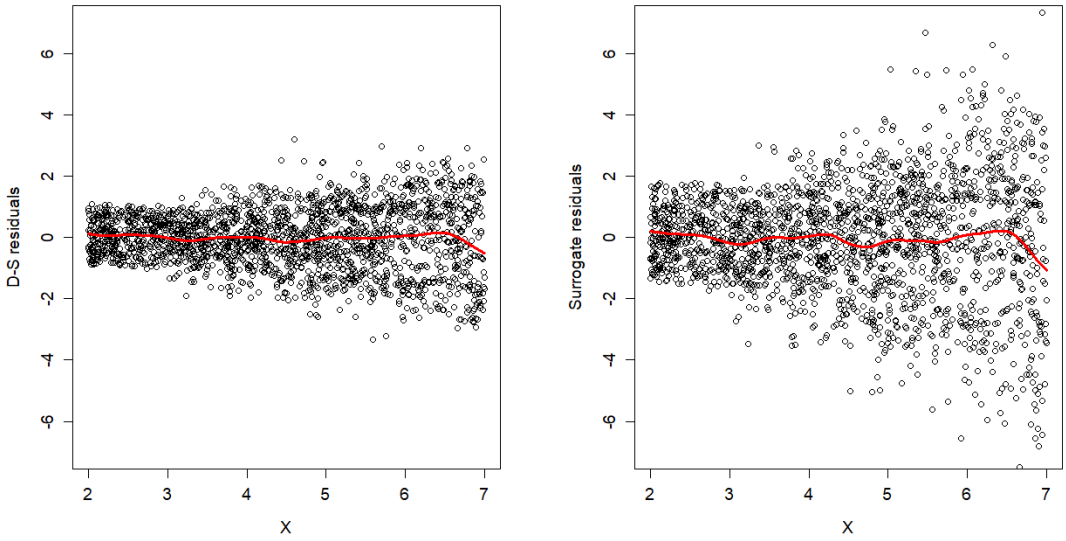

Figure 5.19: Comparison of D-S and surrogate residuals under heteroscedasticity for an unconstrained threshold structure $\boldsymbol{\alpha} = (-36, -6, 34, 64)$.

In the retinopathy example, we find that the two types of residuals produce different patterns for the PPOM. In particular, both the D-S residual vs covariate and the corresponding Q-Q plot show the model assumptions hold (as expected; see Figure 5.20) while for the surrogate residuals, the residual vs covariate plot does not show a perfect scatter, and the Q-Q plot shows a slight deviation from normality (see Figure 5.21).

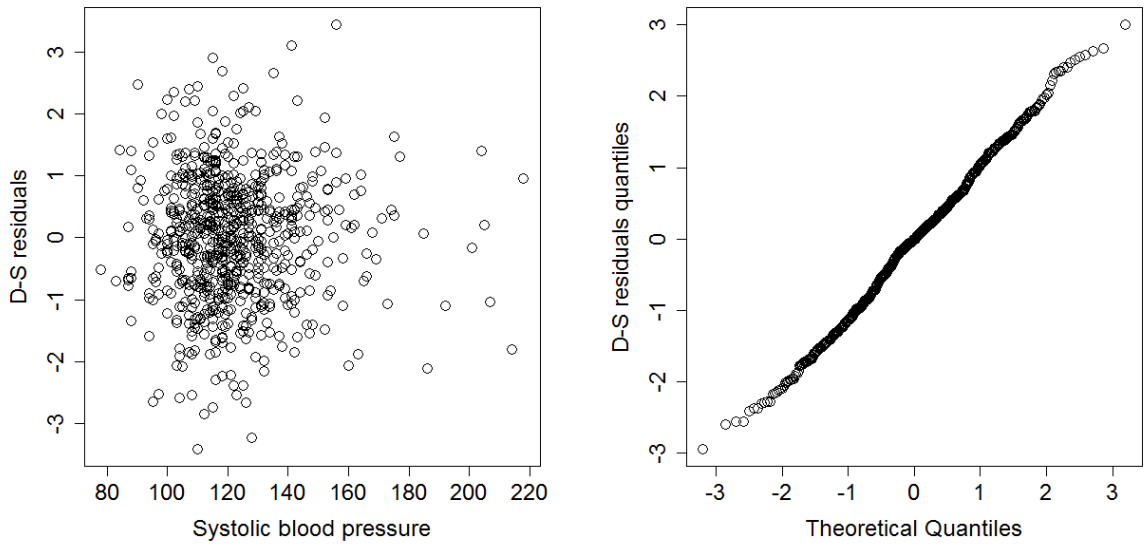

Figure 5.20: PPOM D-S residuals for eye retinopathy data set.

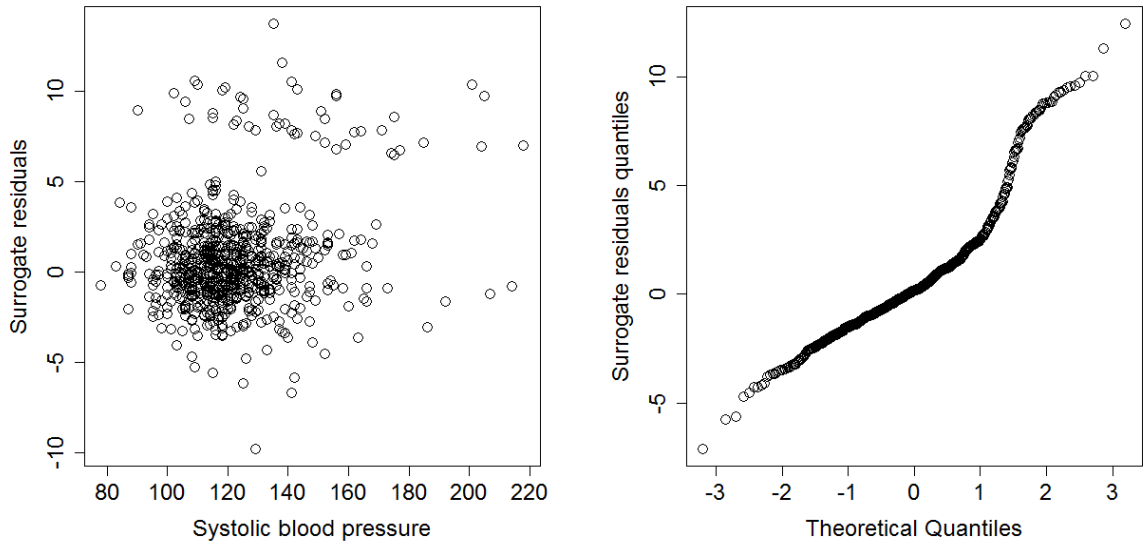

Figure 5.21: PPOM surrogate residuals for eye retinopathy data set.

Having considered the comparison of D-S residuals with alternatives both in the real data and simulated examples, both here and in the Appendix D, it appears that there is practical evidence that the D-S residuals provide a more accurate representation of this model fit and the residuals' distribution within the specific scenarios explored in this thesis. These results are in contrast to those from Liu and Zhang (2018) and future simulation work will aim to systematically assess this behaviour.

5.3.2.5 Assessment of residual plots

Although Q-Q plots are in general a very useful tool to investigate potential deviations from the normality assumptions, when we adapt them to compare the L-S residuals distribution to the reference $U(-1, 1)$ distribution they do not seem to perform as well apart from some exceptions. Most relevant to this issue are Scenarios 3 and 4.

Scenario 3. Missing quadratic term.

For instance, for the *Subscenario 3.5. Quadratic-Linear-Linear* with *equidistant thresholds* $\boldsymbol{\alpha} = (0, 4, 8)$ in which we generate the data from a quadratic model, fit it linearly, and then linearly model the residuals, all of the residuals versus covariate plots detect the quadratic pattern that is wrongly excluded from the model, while the Q-Q plots for D-S and surrogate fail to detect the misspecification, in fact they appear to fit well. Q-Q plots for L-S residuals (see the case for run 10,000 in Figure 5.22 (c)) do however reflect a non-uniform unusual banded pattern which does not relate to lack of fit but is rather an artefact due to

the way L-S residuals are defined.

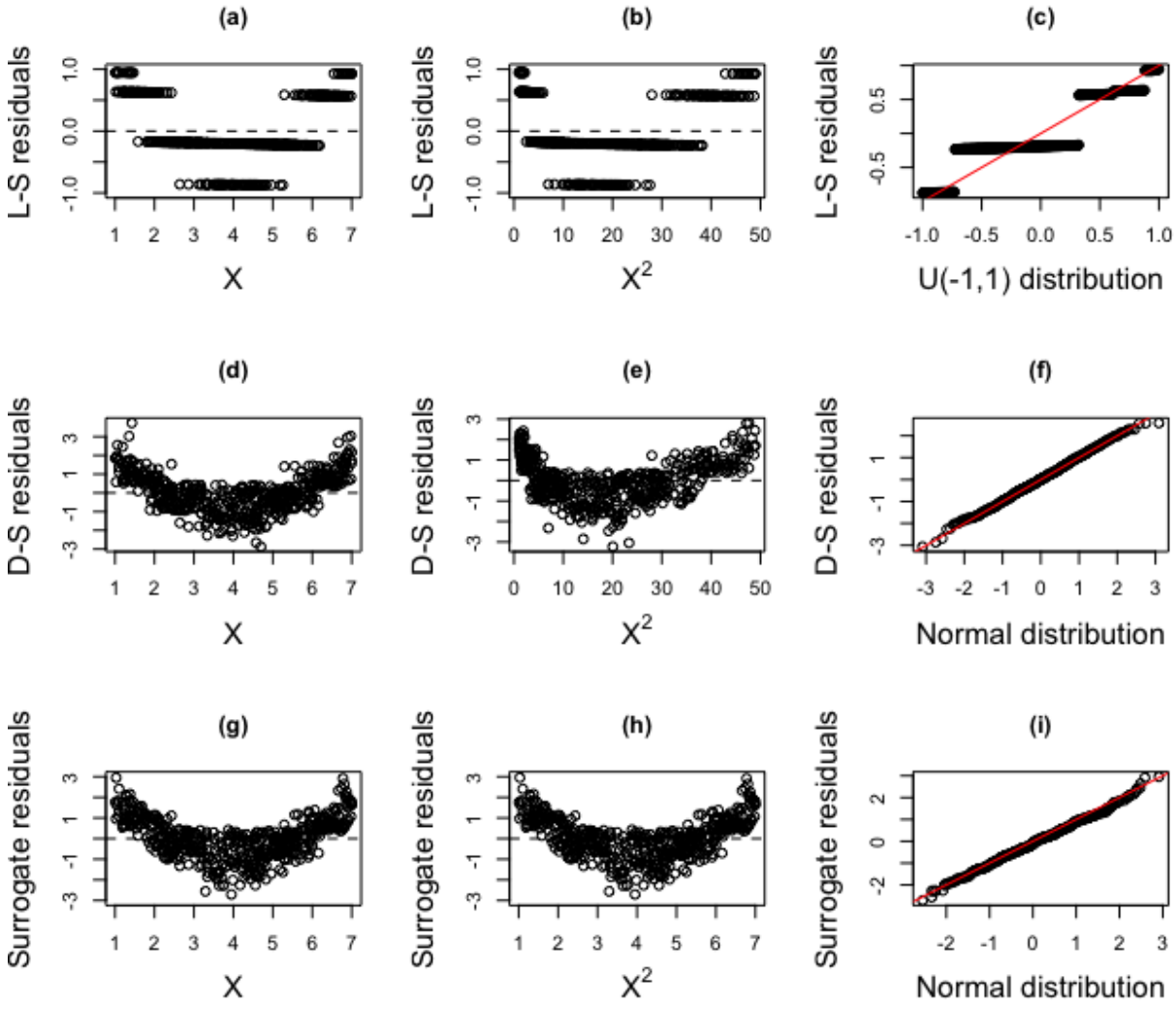

Figure 5.22: Comparison of L-S, D-S and surrogate residuals for the *Quadratic-Linear-Linear subscenario* and a cumulative probit fit.

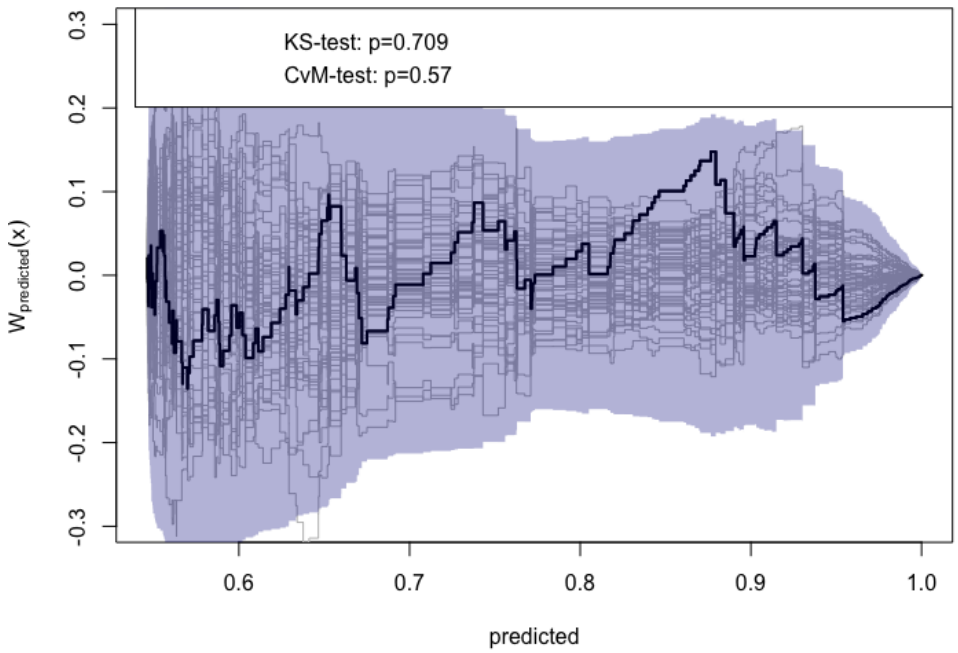

Figure 5.23: Cumulative residual process of the quadratic model in the *Quadratic-Quadratic-Quadratic* subscenario with equidistant thresholds $\boldsymbol{\alpha} = (0, 4, 8)$, and residuals ordered by the predicted response. The grey curves are 50 realisations from the null model. The transparent purple area defines a 95% prediction band for all the simulated processes.

We have used plots of residuals vs. covariate as a diagnostic tool to examine model misspecification in x (Liu et al., 2009) and they have proven to be quite informative *a priori*. However, as the true variance of individual residuals is unknown it can be difficult to decide whether the residual plot indicates a reasonable specification of the mean or not (Holst, 2015). Arbogast and Lin (2005), and Su and Wei (1991) recommend using cumulative sums of the residuals over predicted values or the covariate of interest, in order to check for functional misspecification in x. The function `cum.residuals` from the `R` package `timereg` computes these residuals for survival analysis models. We need to specify a grouping of the data that is used for cumulating residuals, which must have the same size as data and be ordered in the same way. The `R` package `gof` presents an implementation of model diagnostics for the GLM based on aggregates of the residuals where the asymptotic behaviour under the null is imitated by simulations. We present an example for the *Subscenario 3.1. Quadratic-Quadratic-Quadratic* in Figure 5.23 (see Appendix D for other scenarios' results) which shows a good functional specification of our covariate. While these plots using the function `plot.cumres` are outside the scope of our study, we argue they might be a more user-friendly

alternative.

Scenario 4. Missing covariate

Finally, as expected, density plots do not seem very sensitive to misspecifications either, as can be seen from the results from this scenario shown in Subsection 5.3.2.2, where particularly both plots for D-S and surrogate residuals for the wrong model with unexplained variance fail to show very clear differences from the right model plots.

5.4 Case study: Connectedness to nature

We revisit Case study 1 in Section 4.3 to assess the residuals for the models representing the effect of covariates *connectedness to nature* (*CNS*) and *experience* (shopping or nature) and their interaction on the ordinal response variable *pleasantness* with $C = 3$ category levels.

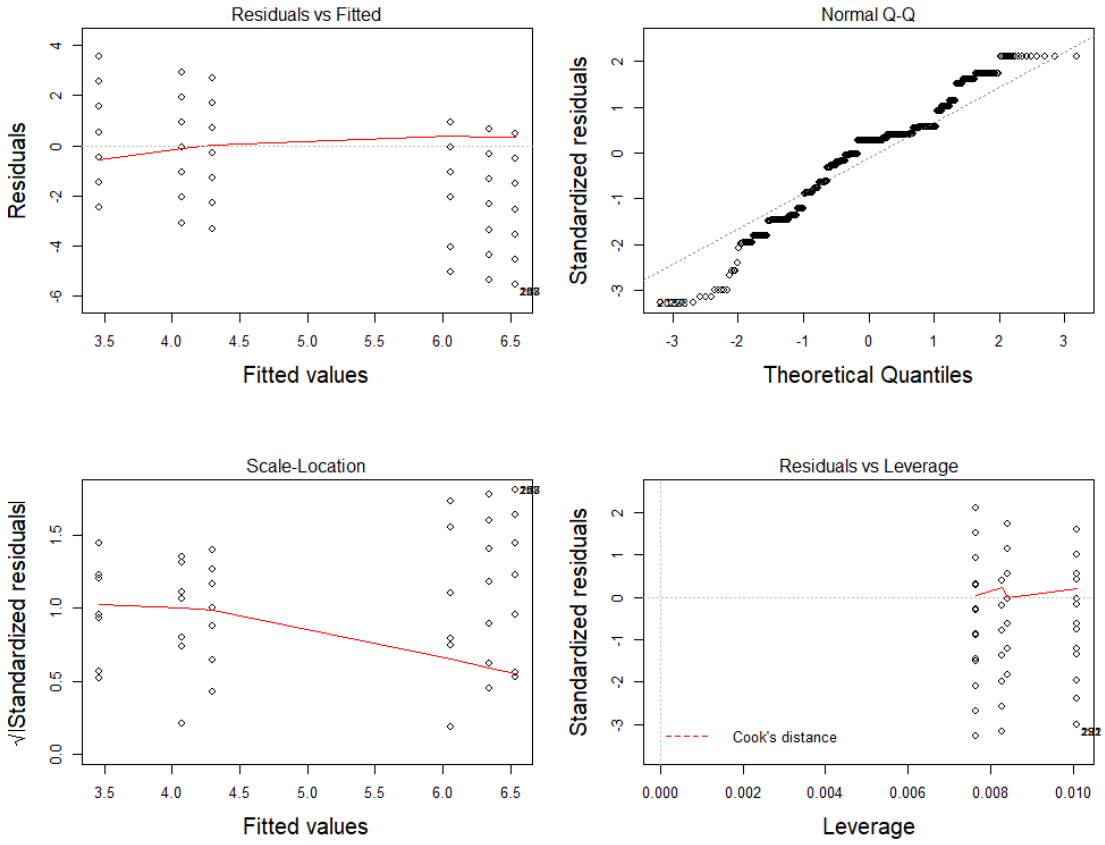

Figure 5.24: Connectedness to nature: traditional residual plots for the linear model. Dashed lines represent reference values. Red lines represent smoothers.

We fit a linear model

$$Y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \delta x_1 \times x_2, \tag{5.24}$$

and firstly calculate traditional residuals which we represent by means of residuals vs fitted values and Q-Q plots, which are shown in Figure 5.24. In addition to the Q-Q plot not reflecting normality of the residuals, none of the other three plots are easily interpretable given their discrete nature.

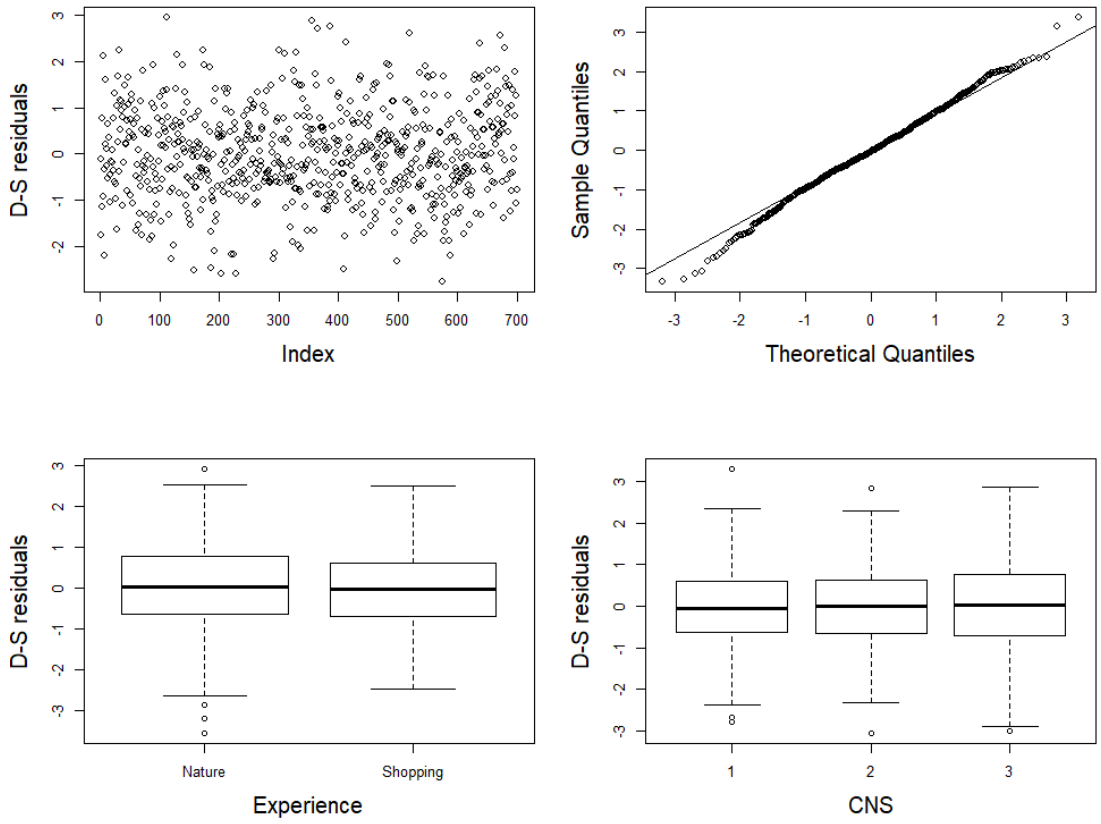

Figure 5.25: Connectedness to nature: D-S residual plots for PPOM with symmetric thresholds.

We then propose a more appropriate model for the data, a PPOM whose selection was described in Section 4.3. This model takes into account the ordered nature of the ordinal response an we can define it such that

$$\text{logit}(P(Y_i \leq j)) = \alpha_j + \sum_{k=1}^{p} \beta_k x_{ik} + \sum_{k=1}^{q} \gamma_{jk} z_{ik}, \quad i = 1, \ldots, n; j = 1, \ldots, C-1, \tag{5.25}$$

We denote the covariate *CNS* as x_1, our relaxed covariate *experience* as z_1, with $n = 357$, $C = 7$, and a symmetric threshold structure for α_j. We can then write (5.25) specifically for this model as

$$\text{logit}(P(Y \leq j)) = \alpha_j + \beta_1 x_1 + \gamma_{j1} z_1, \quad j = 1, \ldots, 6, \tag{5.26}$$

with resulting estimates $\hat{\boldsymbol{\alpha}} = (-2.6, -2.1, 0.4, 1.4)$ and $\hat{\boldsymbol{\beta}}_1 = (0.22, 0.04)$ - with reference category *CNS1*. Given the limited number of available functions, only D-S residuals are available for PPOMs with symmetric thresholds (see Figure 5.25).

These plots already show an improvement over the linear model (e.g., see Q-Q plot) and are all interpretable, therefore we gain more information on the appropriateness of the model.

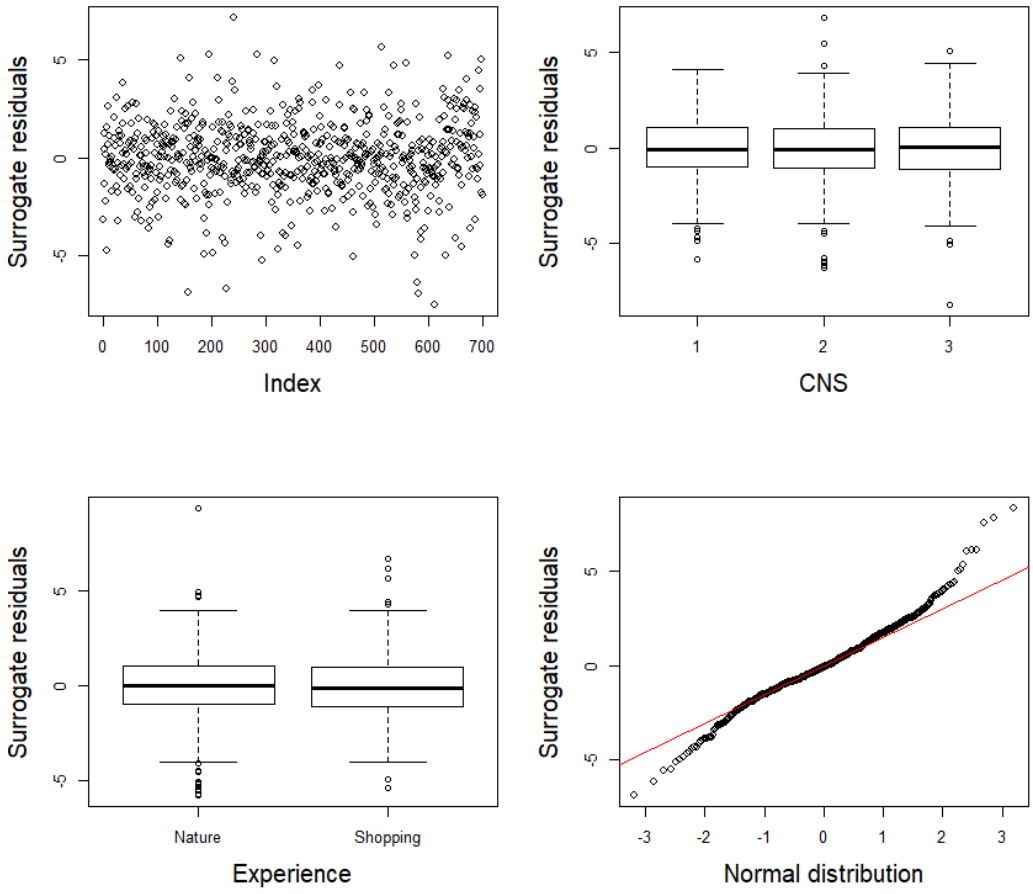

Figure 5.26: Connectedness to nature: surrogate residual plots for PPOM.

Alternatively, we can calculate the surrogate residuals for the PPOM without a specific threshold structure as shown below (see Figure 5.26).

We observe a heavy tailed pattern in the Q-Q plot but we can not disentangle whether it is caused by the threshold misspecification or the type of residual.

Another straightforward misspecification that we can check for is one where we ignore the interaction clearly present in the data between *CNS* and *experience* (as can be seen in Figure 2.1). When we model the data instead following a simpler model such that

$$\text{logit}(P(Y \leq j)) = \alpha_j + \beta_1 x_1 + \beta_2 x_2, \tag{5.27}$$

we find *a priori* that neither D-S nor surrogate residuals show clear deviations from the patterns found for the models including the interactions (see Figure 5.27 and Figure 5.28).

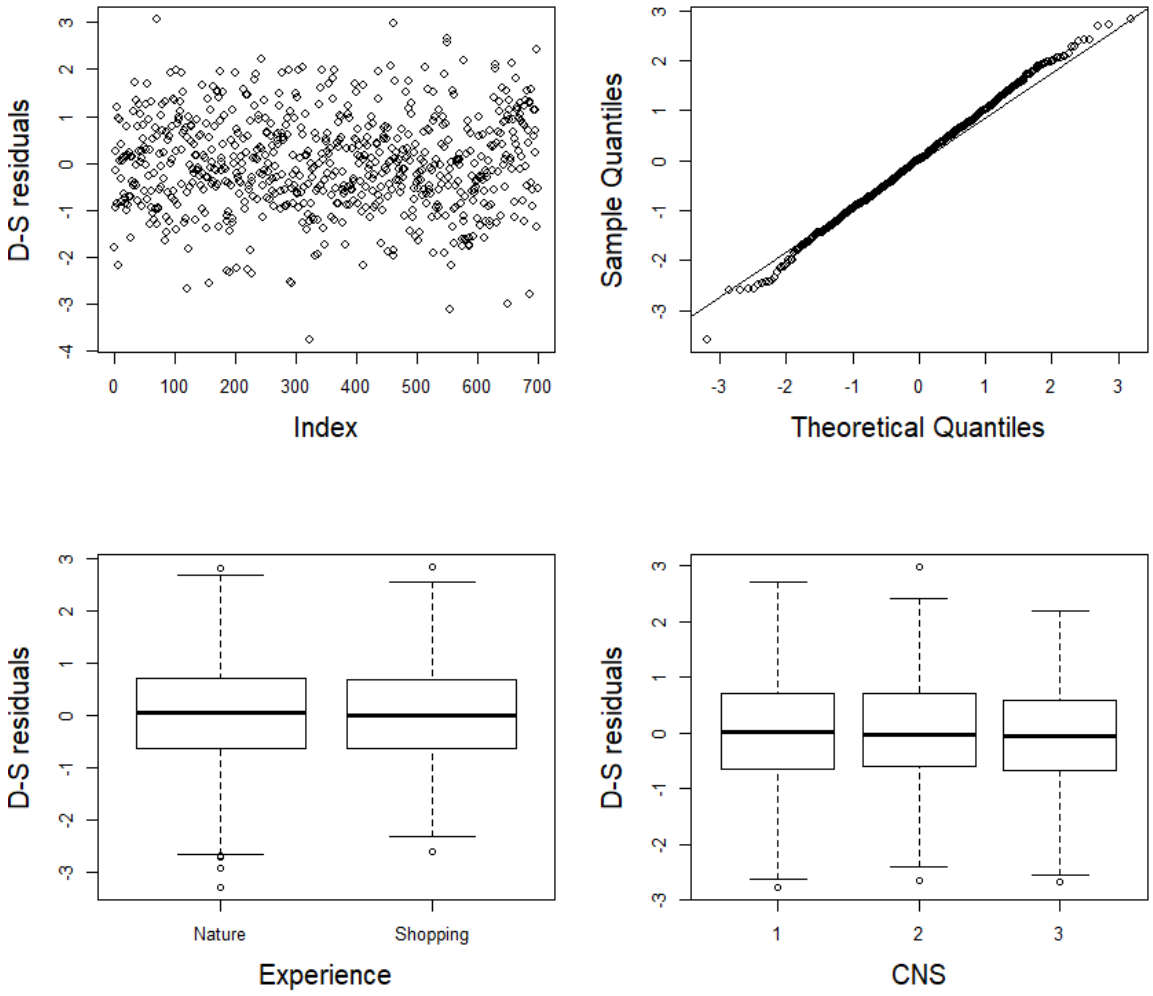

Figure 5.27: Connectedness to nature: D-S residual plots for PPOM with no interaction and symmetric thresholds.

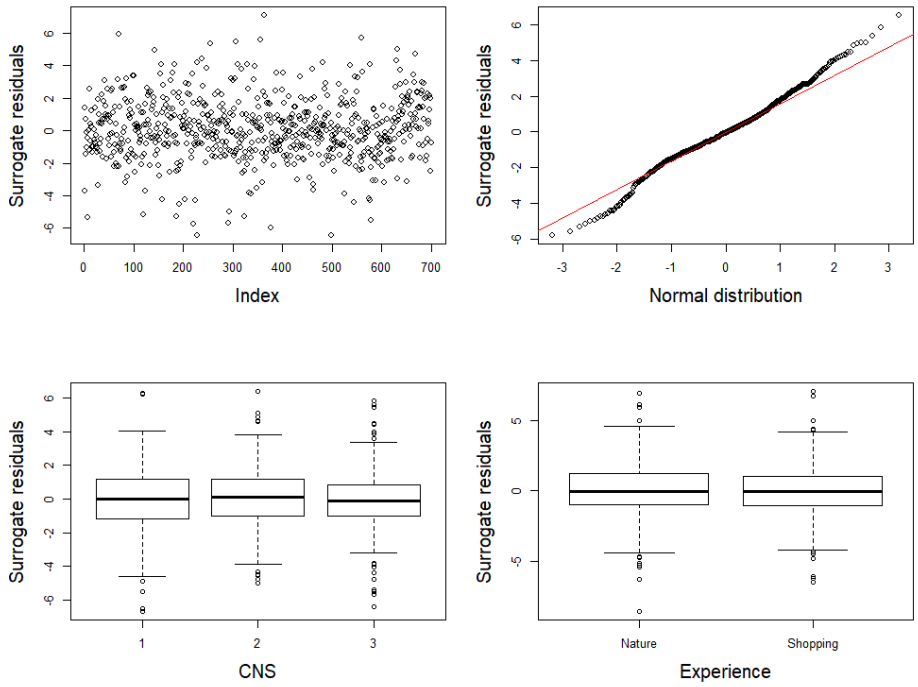

Figure 5.28: Connectedness to nature: surrogate residual plots for PPOM with no interaction and symmetric thresholds.

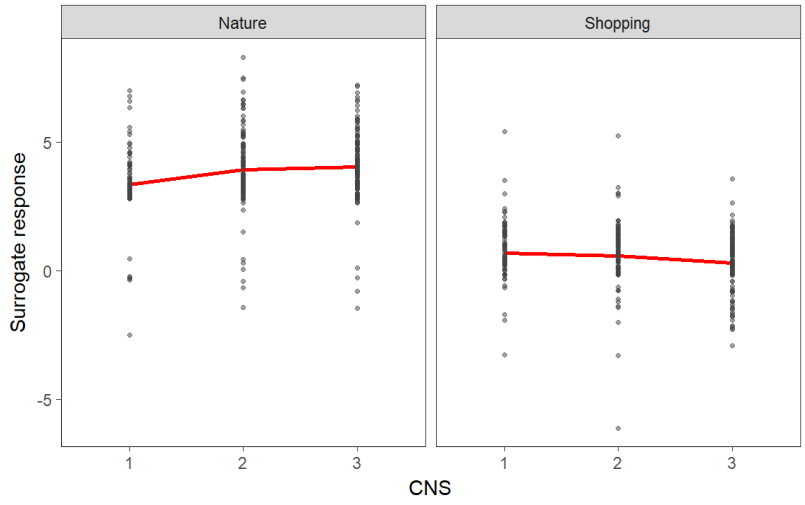

Figure 5.29: Connectedness to nature: D-S residual plots for PPOM with interaction $CNS \times$ *experience* and symmetric thresholds (left: nature, and right: shopping experience). Red lines connect the mean values for the residuals at each category of the factor.

However, if we look at the surrogate response at the *experience* category level, we find

more consistent patterns with the initially found interaction. Unfortunately, given the limitations of the `sure` package, we can only model a POM version with symmetric thresholds (see Figure 5.29).

In summary, due to the fact that software for residual diagnostics of ordinal models is still in development, we could not calculate the corresponding L-S and surrogate residuals for the symmetric threshold structure on the best fitting model found in Subsection 4.3.1. However, from the D-S residuals alone we can see that the ordinal fit is a better fit than the linear model. We can also conclude that, as well as the advantages highlighted in Section 4.3 regarding estimates and inference, by appropriately modelling the response variable as ordinal we can gain greater insights into the model fit (or lack of fit).

5.5 Conclusions

We have found that the nice properties proposed by Liu and Zhang (2018) for ordinal residuals (L-S, D-S, and surrogate residuals) seem to hold both for fitted POMs and PPOMs independently of whether the data was originally simulated from a POM or a PPOM, and that misspecification in terms of odds proportionality is not reflected (at least it is not patent visually) on any of these ordinal residuals behaviour.

With our simulation study we have assessed the effects on the different ordinal residuals patterns of certain model misspecifications. We have built on work by Liu and Zhang (2018) by focusing on CLMs, by including D-S residuals and by studying the effect of different threshold structures (when possible), and by reaching overall conclusions beyond specific solutions for unique runs. The case study has also allowed us to study the impact of not modelling a present interaction, something that is not reflected in any of the three types of residuals.

While none of the ordinal residuals are particularly sensitive to the PO misspecification, different threshold structure specifications at both the latent variable level and the modelling level do have a visible effect on these residuals. Q-Q plots have not proven to be very useful tools to identify misspecifications for any of the three types of ordinal residuals, with those corresponding to L-S residuals being the worst plots of all with the exception of the Scenario 3 Q-L-L where they consistently detect the ignored quadratic pattern for the 10,000 runs, which is not reflected in the corresponding Q-Q plots for either D-S or surrogate residuals. In general, L-S residuals appear to be not always interpretable given the presence of discrete patterns in the plots. While we have not found strong evidence to claim surrogate residuals are infallible at spotting ordinal model misspecifications, D-S and surrogate residuals are overall more informative and consistent than L-S residuals. They do however show a tendency

towards goodness of fit despite misspecifications in the models, having found little deviation from the accurate normal pattern for cases that should show lack of fit (e.g., when assuming Cauchy errors in *Scenario 2*). They have also shown slightly different graphical patterns for our scenarios, from which we can infer that they are somehow distinct. For instance, we have highlighted in the quadratic scenario (*Scenario 3*) that surrogate residuals appear to be slightly more sensitive to model misspecifications (i.e., more accurate) than D-S residuals.

This simulation study has not taken into account two potential identification problems arising from the latent variable generation process with an impact on estimable parameters of the cumulative model. The first identification problem arises from the fact that the threshold is absorbed in the model constant α, so one can alternatively let $\alpha = 0$ and estimate the thresholds, or simulate data with the constant α and hen estimate thresholds α_j (as explained in Scenario 2). The second identification problem is caused by the scale of the latent variable Y^* not being identifiable: if we rescale both the latent response and the thresholds (i.e. multiply by a positive constant k), we would get the same probabilities. As a consequence, the scale of the latent response is not estimated but it is fixed at a conventional value (for example, the scale is fixed at $\phi^2/3$ in logit models corresponding to the standard logistic distribution). Thus, the estimated parameters from the ordinal model are the 'true' parameters of the latent model rescaled by the (unknown) residual variance. This results in estimated parameters from logit models with different covariates not being comparable due to a different rescaling. In presence of heteroscedastic errors dependent on observed covariates (e.g., Scenario 5), the indeterminacy of the scale issue becomes very problematic (see Section 7.4 of Greene and Hensher (2010) for a detailed explanation).

Some authors (Agresti, 2007) claim that residuals in the context of Generalised Linear Models (GLM) and in particular of CLMs are neither easy to interpret nor particularly useful or informative (e.g., Cai and Tsai (1999) suggest that extensive assessment of various types of residuals would be needed) and that the models should be assessed in terms of predictive performance instead (e.g., by means of ROC curves plotting sensitivity vs (1-specificity) for all possible collapsing of the C categories; Toledano and Gatsonis, 1996), or one should use Bayesian approaches that could provide extra information complementing the graphical assessment of the residuals. We argue that provided an accurate definition for ordinal response model residuals is agreed, they would become a very useful tool to complement model selection which is particularly challenging in this area. While we have found many limitations on the diagnostic assessment of complex ordinal models, we aim to fix this limitation in the near future and would ideally aim to build open software functions in `R`. Ultimately, this systematic review of ordinal residuals has also contributed to the argument expressed in the previous chapter highlighting that by ignoring the ordered nature of the

response variable and fitting a linear model (with subsequent linear residuals), we might be missing out relevant information in model misspecifications.

Chapter 6

Solutions to negative fitted category probabilities in PPOMs

6.1 Introduction

As we introduced in Subsection 3.3.2 a PPOM for an ordinal response variable Y_i with C ordered categories, can be defined by the following expression

$$\begin{aligned} \text{logit}(P(Y_i \leq j)) &= \alpha_j + \sum_{k=1}^{p} \beta_k x_{ik} + \sum_{k=1}^{q} \gamma_{jk} z_{ik} = \\ &= \alpha_j + \sum_{k=1}^{p} \beta_k x_{ik} + \sum_{k=1}^{q} (\gamma_k + u_{jk}) z_{ik}, \\ &\quad i = 1, \ldots, n; j = 1, \ldots, C-1, \end{aligned} \tag{6.1}$$

where $-\infty < \alpha_1 < \alpha_2 < \ldots < \alpha_{C-1} < \infty$; $\gamma_{jk} = \gamma_k + u_{jk}$ with $\gamma_{1k} = 0$ for all $k = 1, \ldots, q$, β_k corresponds to the p covariates for which we assume the PO holds and are also referred to as 'global effects' (Poßnecker and Tutz, 2016), while we get parameters γ_{jk} for the q covariates for which we relax the PO assumption, and are also known as 'category-specific effects' (Poßnecker and Tutz, 2016). Each individual parameter γ_{jk} can be additively broken down into a fixed component γ_k and an individual component u_{jk}, i.e., $\gamma_{jk} = \gamma_k + u_{jk}$ where u_{jk} represents the deviation of γ_{jk} from the "typical" value γ_k in the population for response category j.

Despite clear arguments by some authors supporting the fact that PPOMs are "often a superior alternative [to PO models]" (Williams, 2016) and a particularly better and more accurate alternative when the PO assumption is violated by some or all of the covariates", POMs are still the most commonly used ordinal response models. This might be due to the

fact that, as Williams (2016) states, "the use of PPO models has itself been problematic or at least sub-optimal." Another potential reason is the fact that the latent variable interpretation of cumulative link models is no longer valid for PPOMs (Greene and Hensher, 2010, Walker, 2016). In this chapter we describe the most common drawbacks of PPOMs as reported in the literature, and propose solutions to these.

6.2 Issues with partial proportional odds models

6.2.1 Lack of parsimony

PPOMs lead to added complexity when the PO assumption does not hold for more than a few of the variables in the model. "If several variables violate the assumption, then the PPOM offers little in the way of parsimony" (Williams, 2016). The presence of a high number of parameters γ_{jk} also might lead to over-parameterisation and can result in a non-trivial interpretation, which in our opinion could be a major reason for limited uptake of these models. The latent variable rationale for interpretation of these models is problematic because in the case of a PPOM we get more than one estimate of the latent variable Y_i^*, each one associated with each of the different slopes (Williams, 2016). We explore different alternatives to overcome this issue of interpretation and highlight the model selection capabilities of our final proposed solution.

6.2.2 Non-convergence

According to Tutz and Scholz (2003), "the more flexible non-proportional odds model or PPOM have the disadvantage that common estimation procedures as Fisher scoring often fail to converge. Then neither estimates nor test statistics for the validity of PPOMs are available. With the existence of maximum likelihood estimates for the non-proportional model being questionable the further analysis of the data is questionable."

We explore whether this statement is accurate by running a simulation to determine the rate of lack of convergence. Some authors claim that for those settings for which it happens too often, then perhaps it is not worth fixing as it might be showing that the PPOM is really not a good fit for the specific data (Tutz and Scholz, 2003), but we hope that our proposed approaches suggested in Section 6.3 might also help solve this issue.

6.2.3 Negative fitted category probabilities

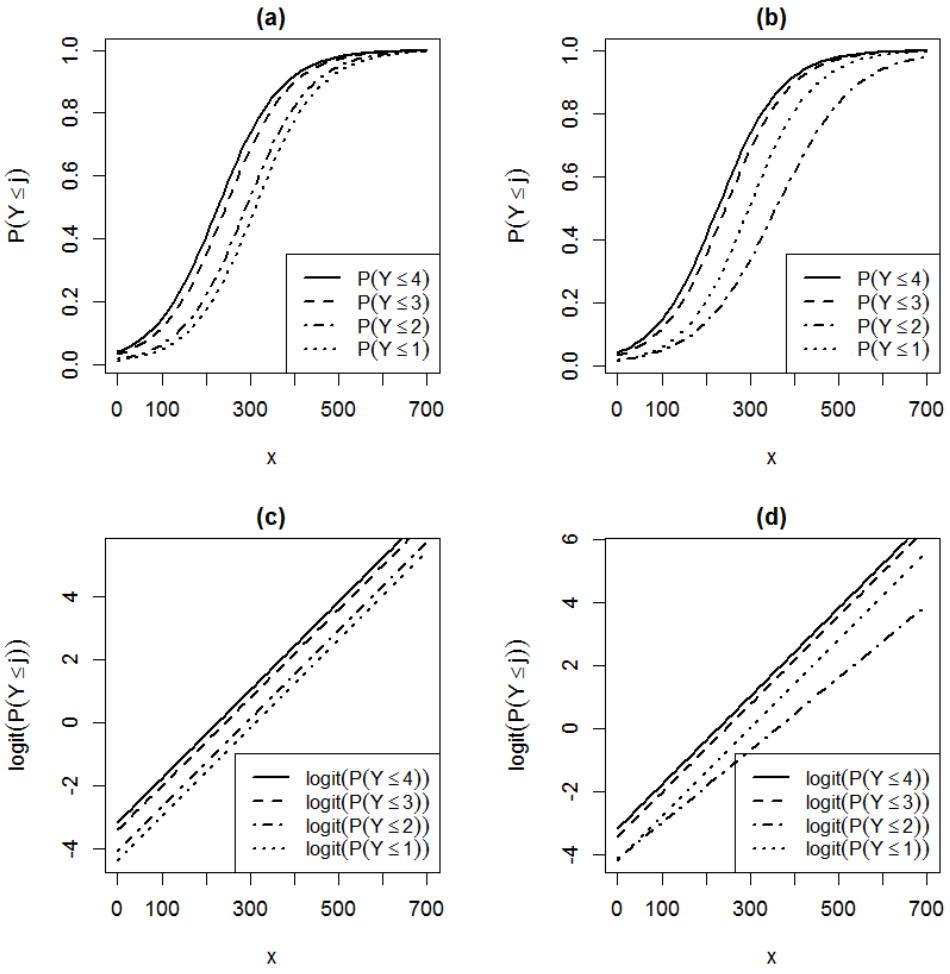

Figure 6.1: Cumulative probabilities corresponding to the categories of an ordinal response variable with 5 levels, for a POM on the (a) probability scale, and on the (c) logit scale, and a PPOM on the (b) probability scale, and on the (d) logit scale. Note that for (c), the corresponding lines on the logit scale are parallel (thus have equal slopes) but this is not the case for (d) where we get crossing.

PPOMs can produce negative fitted category probabilities (McCullagh and Nelder, 1989). As described by Hedeker et al. (2006), for an ordered response variable with C categories, "the effects on the cumulative log odds, namely $x_{ij}^T \alpha_j$ result in $C-1$ non-parallel regression lines". These regression lines must eventually intersect (McCullagh and Nelder, 1989) for some values of x_{ij}. "Negative fitted values are then unavoidable for some values of x_{ij}, though perhaps not in the observed range. If such intersections occur in a sufficiently remote region of the x-space, this flaw in the model need not be serious." For instance, "for x_{ij} variables contrasting two levels of a covariate (e.g., gender coded as 0 or 1), this crossing of regression

lines occurs outside the range of admissible values (i.e., < 0 or > 1). However, if the covariate is continuous, this crossing can occur within the range of the data" (Hedeker et al., 2006), and we would have negative fitted values for the response probabilities (see Figure 6.1 (b)). While one could argue that this might happen as a consequence of the model being misspecified, the PPOM is a solution to the PO misspecification, and hopefully our proposal in Section 6.3 would be an intermediate solution.

Proposed solutions to this issue suggested in the literature (e.g., Hedeker et al. (2006)) include the use of $C-1$ dummy-coded variables corresponding to the C levels of the ordinal response variable, to replace the problematic continuous variable. We argue that this solution might result in a dummy variable trap. The dummy variable trap is a scenario in which the independent variables are multicollinear (a scenario in which two or more variables are highly correlated; in simple terms one variable can be predicted from the others). Creating dummy variables from a continuous one will also cause information loss, so this is not an ideal solution. Williams (2016) suggests combining categories of the response variable, particularly for those categories with few observations which are likely to be causing the PO violation. This should not affect the results of the CLM as these models are invariant to collapsing of categories as stated in Chapter 3, but it is not recommended as mentioned in Chapter 1, and we argue that it is an *ad hoc* solution that will not necessarily work for all data sets.

Although Williams (2014) reports that such issues are rare, we have found that they can be in fact more common than expected, particularly in cases of separation and quasi-complete separation (Albert and Anderson, 1984), that is cases in which the response variable totally or partially splits the covariate (or a combination of covariates, also referred to as the Hauck-Donner effect). Both separation and quasi-complete separation are common problems in ordinal response models. Unlike in ordinary least-squares regression for modelling a normally distributed response, when a logistic model perfectly or nearly perfectly predicts the response (that is, separates or quasi-separates the response levels), unique MLEs do not exist. Some of the estimates of the model parameters are therefore non-unique and infinite. This is a common consequence of the data being sparse, meaning that not all response levels are observed in each of the covariate settings, which often happens with small data sets or when the event is rare or a response option is unlikely to be chosen (see Figure 6.2 where category 2 responses are very low, which can be due to social desirability bias for instance in the case of questionnaires).

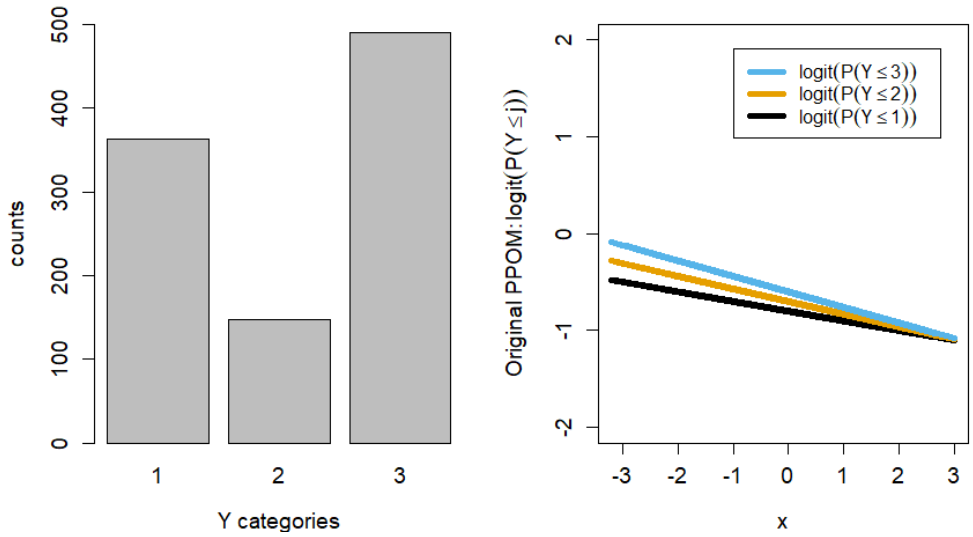

Figure 6.2: PPOM example with crossing of predicted probabilities (left) due to sparse data (right).

We propose two alternative solutions: firstly, by means of a Lasso penalisation; and secondly, through a constrained version of the model's log-likelihood. We compare the effectiveness of the proposed solutions via two case studies.

6.3 Proposed solutions

We propose the following alternatives to the previously mentioned solutions for a PPOM defined such that

$$\text{logit}(P(Y_i \leq j)) = \alpha_j + \sum_{k=1}^{p} \beta_k x_{ik} + \sum_{k=1}^{q} \gamma_{jk} z_{ik}, i = 1, \ldots, n; j = 1, \ldots, C-1, \qquad (6.2)$$

where z_{ik} are the covariates for which some or all coefficients vary by category (here γ_{ik} from $k = 1$ to $k = q$, where for $j = 1$, $\gamma_{1k} = 0$ for all k for identifiability) while the covariates x_{ik} remain constant for all response categories (here β_k from $k = 1$ to $k = p$). q represents the number of covariates for which the PO assumption does not hold, and p is the number of covariates for which the PO holds.

6.3.1 Lasso penalisation

Firstly, in order to address convergence problems caused by over-parameterisation, we assess the efficacy of the Lasso estimation method (Tibshirani, 1996) in fixing the crossing of

regression lines and ultimately determining which covariates satisfy the PO assumption.

Let L be the original log-likelihood as defined in Section 3.4 and let us define $\gamma_{jk} = \gamma_k + u_{jk}$ to avoid cases in which it might be that the Lasso penalises a variable for which the PO assumption does not hold either. Given the tuning or shrinkage parameters $\lambda_k \geq 0$, the corresponding log-likelihood for the CLM can be defined as

$$L^* = L + \sum_{k=1}^{q} \lambda_k \sum_{j=1}^{C} |u_{jk}|, \tag{6.3}$$

For those cases for which it is assumed that $u_{jk} = 0 \ \forall j$ for a given k, then the terms for covariate z_K are reduced to γ_k, i.e. the PO assumption holds for that variable.

Optimal tuning parameters can be determined by leave-one-out cross-validation (CV) which minimises the Kullback-Leibler discrepancy (Zahid et al., 2015) defined such that

$$KL = \sum_{j=1}^{C} \sum_{i=1}^{n} \pi_{ij} \log \left(\frac{\pi_{ij}}{\hat{\pi}_{ij}} \right), \tag{6.4}$$

where $\hat{\pi}_{ij}$ is the estimated probability of response j for observation i and $0 \times \log(0) = 0$ by convention. For a given λ, the model is fitted n times to get the parameter estimates for each data set formed from leaving out each row. The 'errors' between each of the n observations and the fit of each of the n models are calculated, and then by summing these errors up which constitute the CV error for that λ, and which we then minimise over the λs.

Common penalisation approaches use penalisation to obtain smooth effects of covariates (Eilers and Marx, 1996, Ruppert, 2002). The choice of this particular penalty is based upon work by Gertheiss and Tutz (2008) showing "the usefulness of simple penalisation techniques for ordered categorical covariates". The authors followed Land and Friedman (1997) and Tibshirani et al. (2005) fused Lasso, and introduced a Lasso-type penalty on differences of adjacent regression coefficients (also referred to by the authors as a form or 'vertical' smoothing in Tutz and Scholz (2003), in contrast to 'horizontal' smoothing across the values of covariates).

One of the advantages of this approach is that the Lasso is also an automatic model-variable selection strategy (Zhao and Yu, 2006), which will decide for which variables the PO assumption will hold and for which ones it will be violated. It does so by shrinking off coefficients to 0, i.e., dropping those variables from your model. However, it is well-known that the Lasso penalty provides a sparse but biased solution (Xue and Qu, 2017), that it pays a price in predictive discrimination for trying to do variable selection, and that if two covariates are highly correlated but only one is a true driver of the response variable, Lasso

can end up dropping the true driver at random (Zou and Hastie, 2005) which would not be appropriate for making predictions for a population where those two covariates are not highly correlated, or at the very least would require adjusting the tuning factor (Hebiri and Lederer, 2013). In addition to these limitations, unlike ML estimates, Lasso estimates are not generally invariant under general linear transformations of the parameters for CLMs, and are sensitive to the choice of optimality criterion and cross-validation to determine the tuning parameter. For these reasons, these estimates can only be an *ad hoc* solution to the problem.

Software availability

There are not many popular software packages for regularized regression for ordinal models (Wurm et al., 2017).We have found for instance that the popular `R penalized` package does not include options for ordinal models, whereas `ordinalNet` implements a combined version of ridge and lasso (called elastic net) with k-fold cross-validation. Both `glmnetcr` and `glmnet` libraries are only suitable for fitting CRMs instead of CLMs. We have used the function `cvglmnet` in `glmnet` to determine the optimal λ value for our optimisation of the log-likelihood. In `SAS`, regularization methods for generalized linear mixed models are implemented in `GLMSELECT`.

Variable selection for generalized linear mixed models via Lasso-penalized estimation is available in the `R` package `glmmLasso` (Groll, 2017) which includes the cumulative logit models (POM only).

6.3.2 Constrained log-likelihood

Secondly, in an attempt to achieve a more direct solution, we propose a geometric reformulation of the model which guarantees that class probabilities will be non-negative (i.e., regression lines will not overlap within the stated limits).

Consider the PPOM

$$\text{logit}(P(Y \leq j)) = \alpha_j + \mathbf{x}^T\boldsymbol{\beta} + \mathbf{z}^T\boldsymbol{\gamma_j}, \tag{6.5}$$

where $\boldsymbol{x}^T$ are the covariates for which the PO assumption holds, and $\boldsymbol{z}^T$ those for which we relax the assumption. Without loss of generality we can assume $0 \leq z_k \leq 1$ for $k = 1, \ldots, q$. To avoid negative category probabilities we require

$$P(Y \leq j) \leq P(Y \leq j + 1) \tag{6.6}$$

for $j = 1, \ldots, C-2$. Since logit is a monotonically increasing function (6.6) is equivalent to

$$\text{logit}(P(Y \leq j)) \leq \text{logit}(P(Y \leq j+1)). \tag{6.7}$$

In combination with (6.5), this gives us the constraints that

$$\alpha_j + \mathbf{z}^T \boldsymbol{\gamma_j} \leq \alpha_{j+1} + \mathbf{z}^T \boldsymbol{\gamma_{j+1}}, \tag{6.8}$$

for $j = 1, \ldots, C-2$. Let $r = 2^q$ and Z be the $r \times q$ matrix where each row gives each unique combination of lower (0) and upper (1) limits for the values of $\mathbf{z} = (z_1, \ldots, z_q)$. The justification for setting these limits is that non-parallel lines will necessarily overlap somewhere and so we have to define a range within which we guarantee no crossing. The constraints are then

$$\alpha_j \mathbf{1}_r + \mathbf{Z}\boldsymbol{\gamma_j} \leq \alpha_{j+1}\mathbf{1}_r + \mathbf{Z}\boldsymbol{\gamma_{j+1}}, \tag{6.9}$$

for $j = 1, \ldots, C-2$, where the inequalities are applied elementwise.

Suppose the complete vector of parameters is $\boldsymbol{\theta} = (\boldsymbol{\beta}, \alpha_1, \ldots, \alpha_{C-1}, \boldsymbol{\gamma_1}, \ldots, \boldsymbol{\gamma}_{C-1})$, where $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}_j$ are $p \times 1$ and $q \times 1$ vectors. Then the log-likelihood is maximised subject to the following constraint

$$\begin{pmatrix} 0_{r\times p} & -\mathbf{1}_r & \mathbf{1}_r & \mathbf{0}_r & \ldots & \mathbf{0}_r & \mathbf{0}_r & -Z & Z & 0_{r\times q} & \ldots & 0_{r\times q} & 0_{r\times q} \\ 0_{r\times p} & \mathbf{0}_r & -\mathbf{1}_r & \mathbf{1}_r & \ldots & \mathbf{0}_r & \mathbf{0}_r & 0_{r\times q} & -Z & Z & \ldots & 0_{r\times q} & 0_{r\times q} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0_{r\times p} & \mathbf{0}_r & \mathbf{0}_r & \mathbf{0}_r & \ldots & -\mathbf{1}_r & \mathbf{1}_r & 0_{r\times q} & 0_{r\times q} & 0_{r\times q} & \ldots & -Z & Z \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta} \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_{C-1} \\ \boldsymbol{\gamma_1} \\ \boldsymbol{\gamma_2} \\ \vdots \\ \boldsymbol{\gamma_{C-1}} \end{pmatrix} \geq \begin{pmatrix} \mathbf{0}_r \\ \mathbf{0}_r \\ \vdots \\ \mathbf{0}_r \end{pmatrix} \tag{6.10}$$

where the inequalities are applied elementwise.

We consider the cases for one and two Z variables. Firstly, suppose $q = 1$, then $r = 2$ and our limits $z_{min} = 0$ and $z_{max} = 1$ such that

$$Z = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \tag{6.11}$$

and the constraints are

$$\begin{aligned} \alpha_j &\leq \alpha_{j+1} \\ \alpha_j + \boldsymbol{\gamma_j} &\leq \alpha_{j+1} + \boldsymbol{\gamma_{j+1}}, \end{aligned} \tag{6.12}$$

for $j = 1, \ldots, C-2$, which can be seen graphically as imposing the constraints that the regression lines do not overlap (see Figure 6.3).

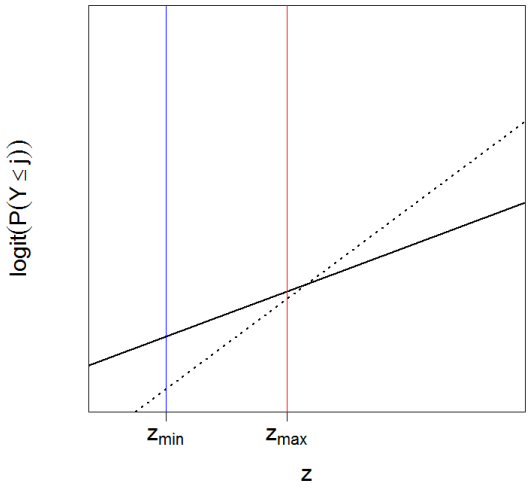

Figure 6.3: PPOM crossing regression lines (black) and constraints setting to prevent this crossing (blue and red lines).

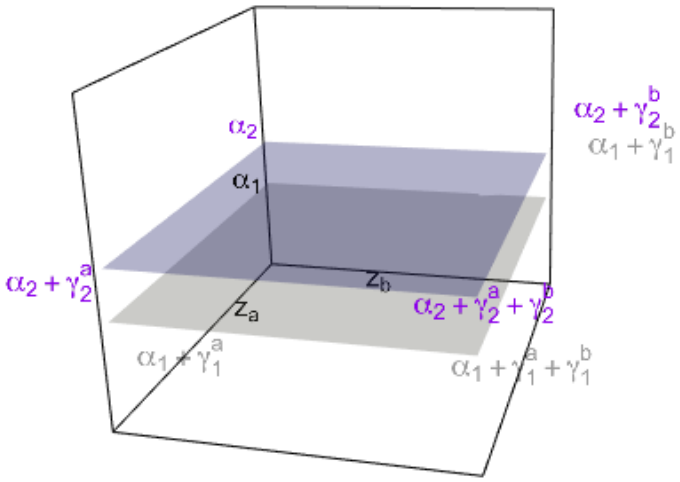

Figure 6.4: Constraints for a PPOM with two covariates z_a and z_b.

Secondly, suppose $q = 2$ (i.e., two covariates z_a. z_b), then $r = 4$, and the constraints are

$$\begin{aligned} \alpha_j &\leq \alpha_{j+1} \\ \alpha_j + \boldsymbol{\gamma_j^a} &\leq \alpha_{j+1} + \boldsymbol{\gamma_{j+1}^a} \\ \alpha_j + \boldsymbol{\gamma_j^b} &\leq \alpha_{j+1} + \boldsymbol{\gamma_{j+1}^b} \\ \alpha_j + \boldsymbol{\gamma_j^a} + \boldsymbol{\gamma_j^b} &\leq \alpha_{j+1} + \boldsymbol{\gamma_{j+1}^a} + \boldsymbol{\gamma_{j+1}^b}, \end{aligned} \tag{6.13}$$

which can be seen graphically as imposing the constraints that the planes' corners do not overlap (see Figure 6.4).

In all cases we maximise the likelihood subject to the constraints using constrained optimisation.

Software availability

In `Stata`, there exists some form of imposition of parallel lines constraints via the `autofit` option in `gologit2` which may also help because it reduces the likelihood of non-parallel lines intersecting. However, to the best of our knowledge, there are no `R` libraries or any other software available to run models with our proposed constrained solution. In order to apply our solution, we can use the constrained optimisation function `constrOptim`.

6.4 Simulation study

We simulate a simple ordinal model with an ordinal response variable Y with 4 categories, $x \sim N(0, 1)$, and $n = 1,000$. We fit a CLM with Y as ordinal response and x as covariate, and we consider 6 scenarios; 3 where we generate data from a POM and 3 from a PPOM, and we compare the results for the default log-likelihood, the constrained log-likelihood solution and the Lasso penalisation, for both POM and PPOM fits. POM forces the lines corresponding to the predictions in the linear predictor scale to be always parallel (either if the model is well or wrongly specified). PPOMs might present either crossing of lines or not for both well and wrongly specified models. We therefore expect the following patterns *a priori* and our ultimate aim is to fix the issues when we get crossing on correct PPOMs:

(i) For both cases in which we fit the default POM to data generated from a POM and data generated from a PPOM, we expect the regression lines to be always parallel.

(ii) For the cases in which we fit the newly parameterised and Lasso POMs to data generated from a POM and data generated from a PPOM, we expect the lines not to cross.

(iii) When we fit a default PPOM to data generated from a POM, we anticipate the lines are less likely to cross but it could still happen, potentially because the model is wrongly specified.

(iv) When we fit a default PPOM to data generated from a PPOM, we anticipate the lines can cross due to uncertainty in estimation.

(v) More specifically, for the newly constrained model the lines will not cross by definition, whereas the Lasso penalty should be reducing the risk of crossing occurring.

For data generated from a POM

$$\text{logit}(P(Y_i \leq j) = \alpha_j + \beta x_i, \tag{6.14}$$

with $x_i \sim N(0,1)$, $i = 1, \ldots, 1000$, $j = 1, 2, 3$, $\boldsymbol{\alpha} = (-0.8, -0.7, -0.6)$, and $\beta = -0.2$, with original logits and Y categories distribution as shown in Figure 6.5, we do not observe crossing when we fit a POM, as expected, although we note that for the POM data fitted as PPOM, the regression line $P(Y \leq 2)$ (shown in Figure 6.6 in yellow) crosses the corresponding $P(Y \leq 1)$ line (in black).

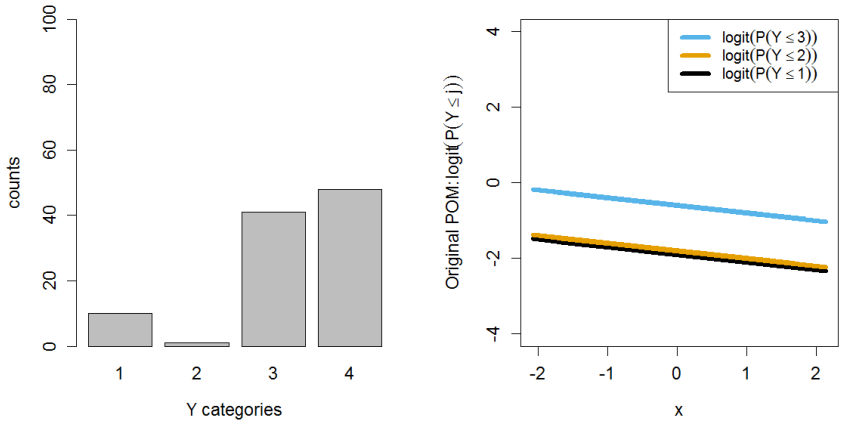

Figure 6.5: POM-generated data.

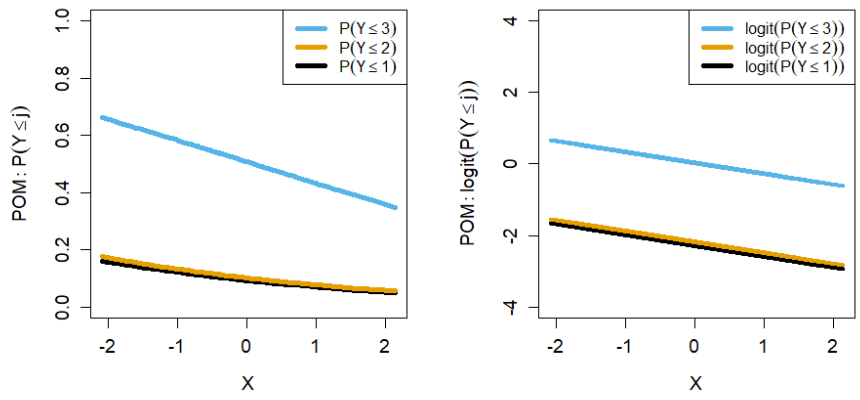

Figure 6.6: POM-generated data fitted as POM on the probability (left) and the logit (right) scales.

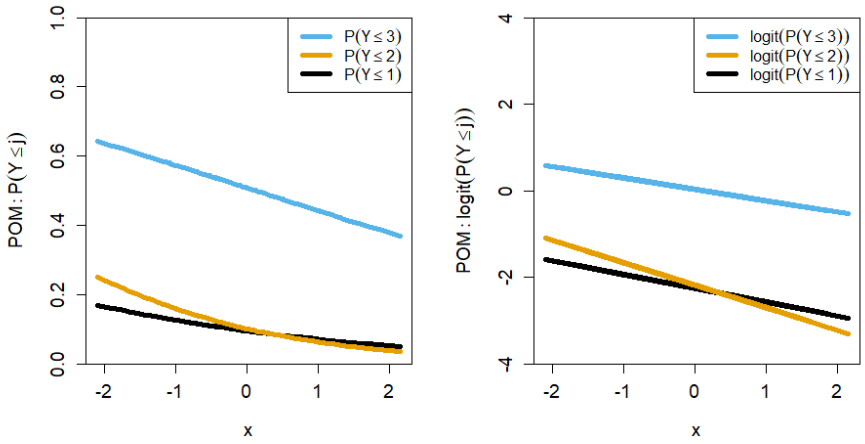

Figure 6.7: POM-generated data fitted as PPOM on the probability (left) and logit (right) scales.

For data generated from a PPOM,

$$\text{logit}(P(Y_i \leq j) = \alpha_j + \beta_j x_i, \tag{6.15}$$

with $x_i \sim N(0,1)$, $i = 1, \ldots, 1000$, $j = 1, 2, 3$, $\boldsymbol{\alpha} = (-0.8, -0.7, -0.6)$, and $\boldsymbol{\beta} = (0.1, 0.13, 0.16)$, with original logits and Y categories distribution shown in Figure 6.8.

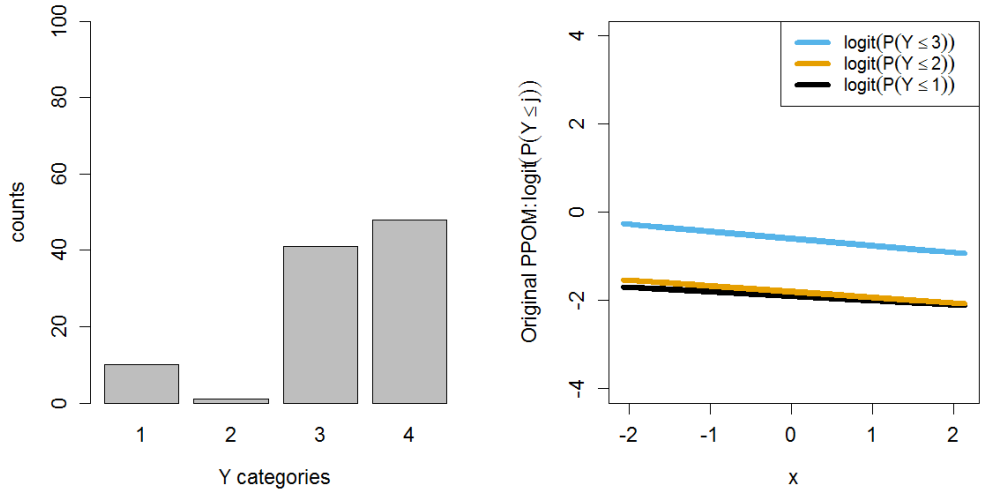

Figure 6.8: PPOM-generated data.

there is also evidence of crossing of the predicted probabilities $P(Y \leq 1)$ and $P(Y \leq 2)$ for the PPOM model (see Figure 6.9 left).

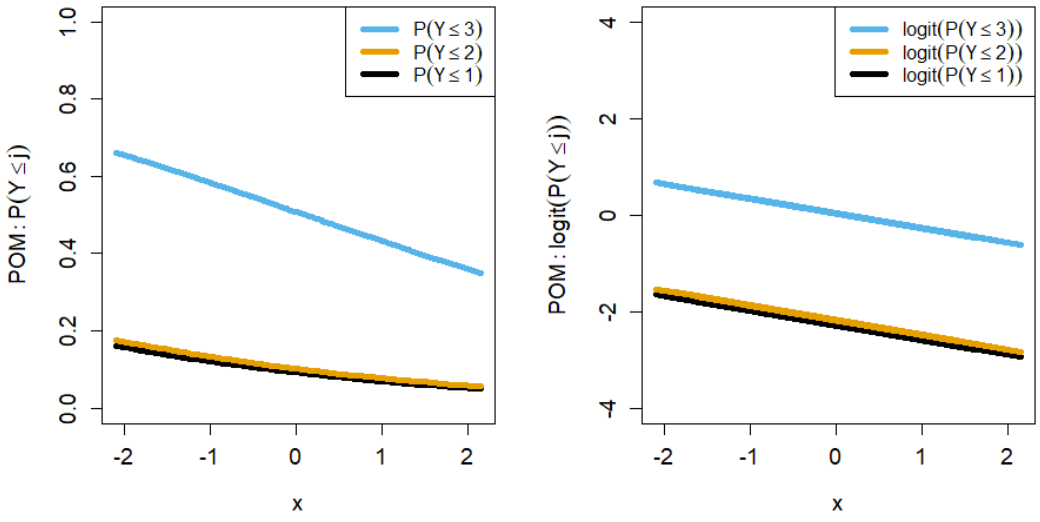

Figure 6.9: PPOM-generated data fitted as POM on the probability (left) and logit (right) scales.

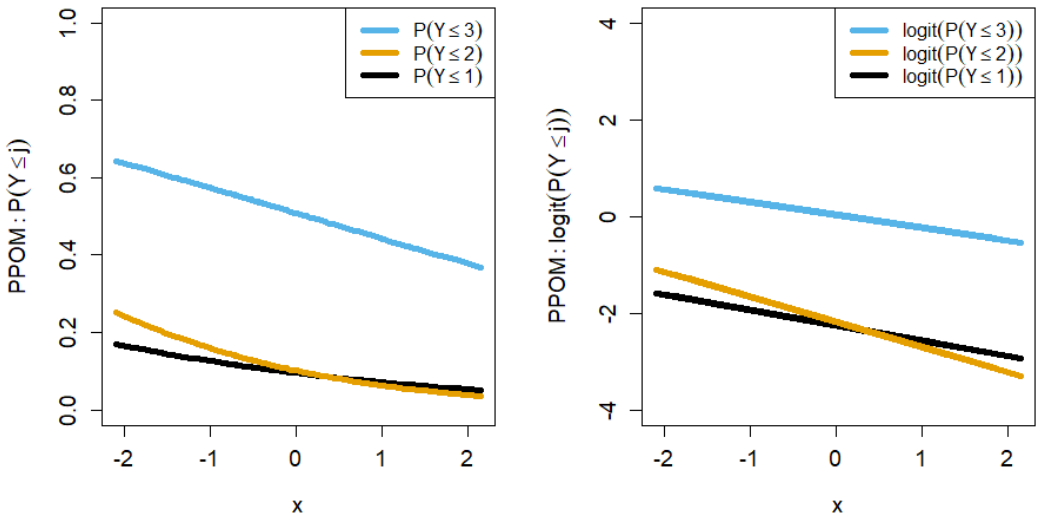

Figure 6.10: PPOM-generated data fitted as PPOM. Predictions on the probability scale (left) and the logit scale (right)

Both the Lasso approach ($\lambda = 0.04$) and our constrained approach fix the crossing lines issue as shown in Figures 6.11 and 6.12 (where the lines thickness has been reduced for ease of visualisation).

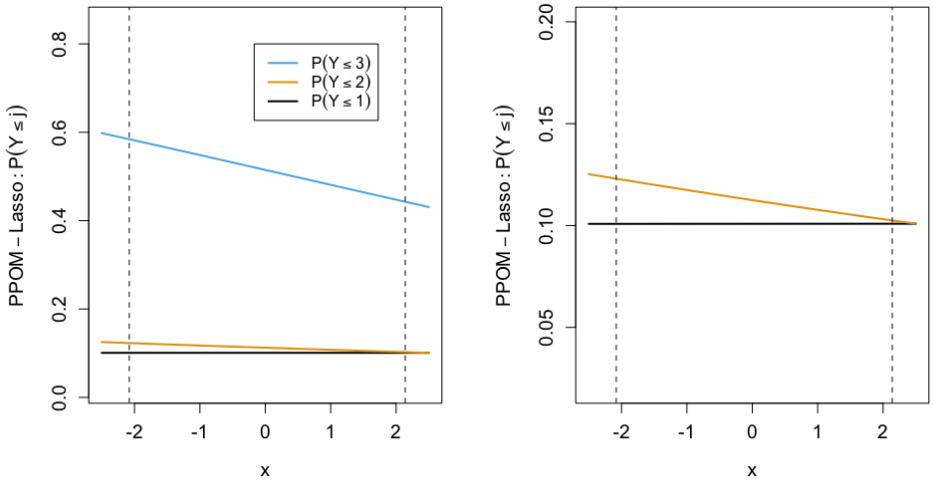

Figure 6.11: Lasso ($\lambda = 0.04$) - Predicted probabilities for PPOM-generated data fitted as PPOM plotted on the probability scale (left) and zoomed-in (right) to visualise the position of the crossing (outside of the range of our data).

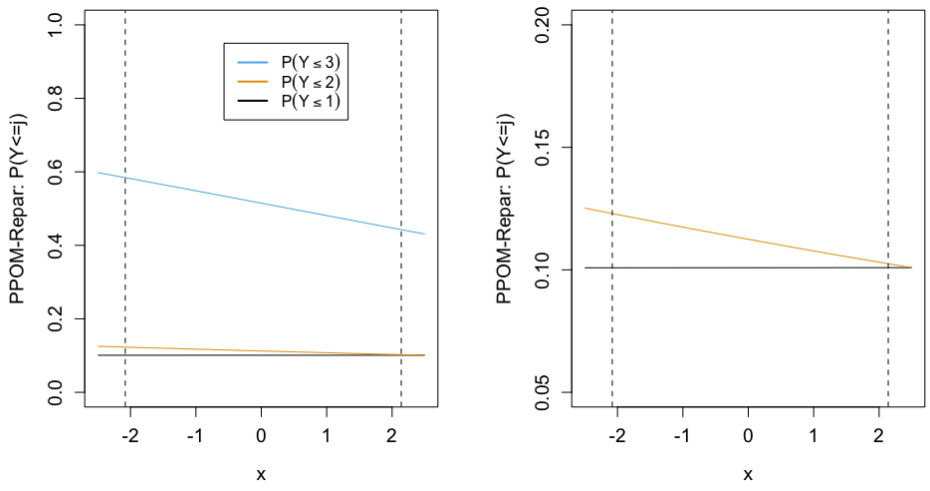

Figure 6.12: Constrained log-likelihood - Predicted probabilities for PPOM-generated data fitted as PPOM plotted on the probability scale (left) and zoomed-in (right) to visualise the position of the crossing (outside of the range of our data).

We replicate the case in which we fit a PPOM to data generated from a PPOM for two covariates (see Figure 6.13) such that

$$\text{logit}(P(Y_i \leq j)) = \alpha_j + \beta_j x_i + \gamma_j z_i, \tag{6.16}$$

with $x_i \sim N(0,1)$, $z_i \sim N(0.1, 1.1)$, $i = 1, \ldots, 1000$, $j = 1, 2, 3$, $\boldsymbol{\alpha} = (-1.9, -1.8, -0.6)$, $\boldsymbol{\beta} = (0.1, 0.13, 0.16)$, and $\boldsymbol{\gamma} = (0.1, 0.03, 0.26)$.

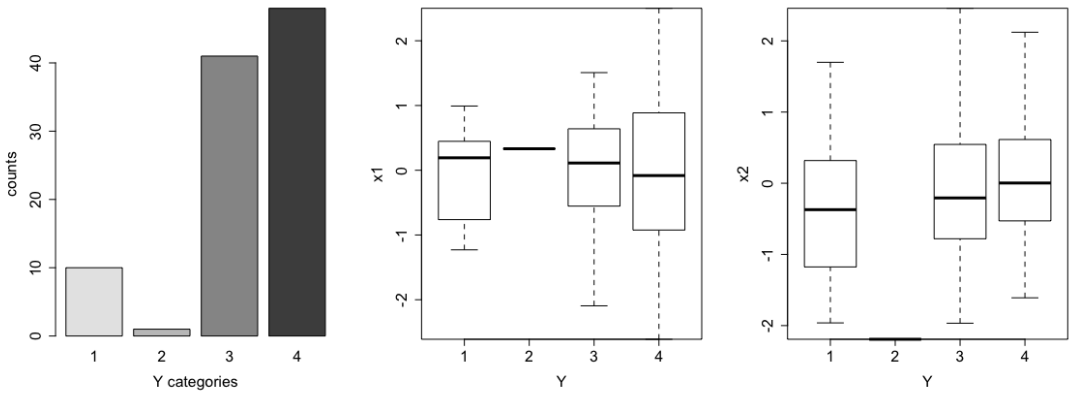

Figure 6.13: PPOM generated response variable Y (left) as a function of 2 covariates x_1 (middle) and x_2 (right).

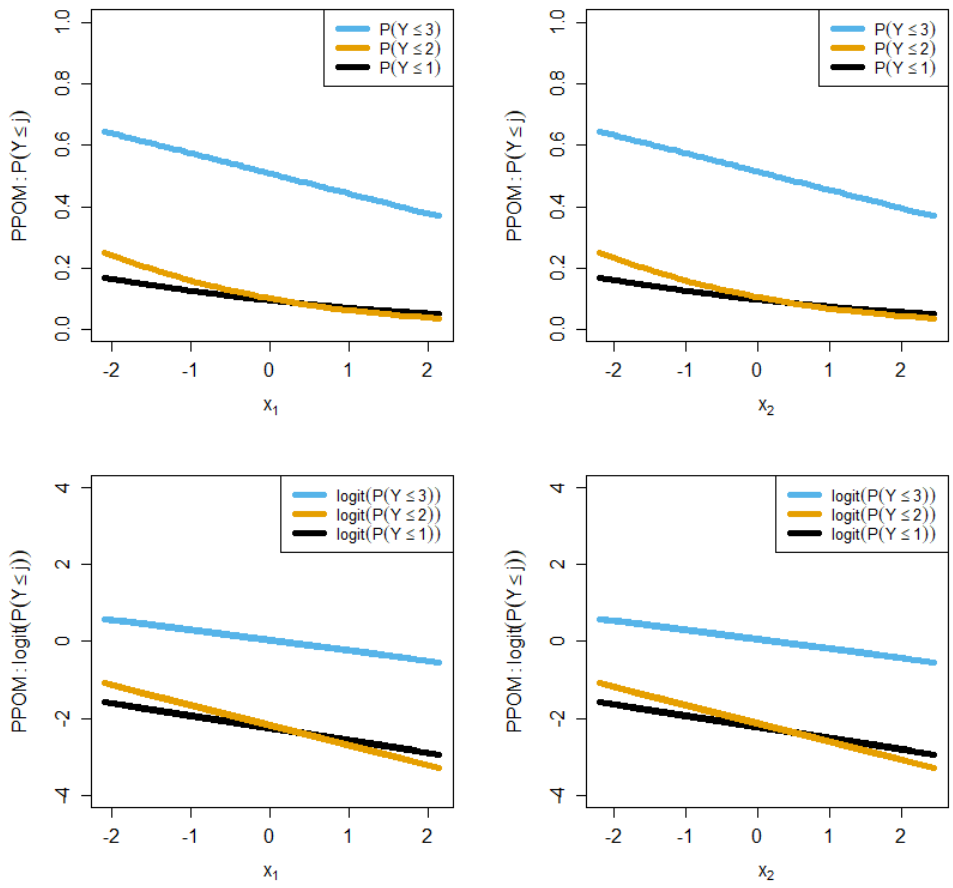

Figure 6.14: Centered-term predicted probabilities for PPOM-generated data as PPOM on the probability (top) and logit (bottom) scale, for covariates x_1 (left) and x_2 (right).

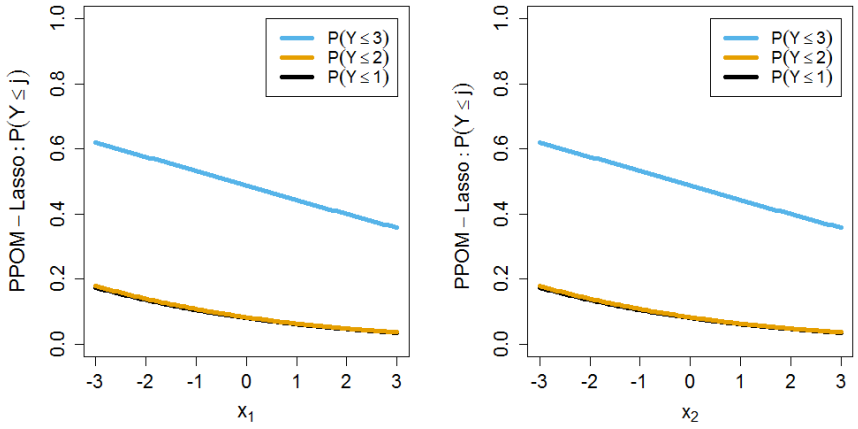

Figure 6.15: Lasso ($\lambda = 0.01$) - predicted probabilities for PPOM-generated data fitted as PPOM for covariates x_1 (left) and x_2 (right).

Similarly to the case with one covariate, we confirm that crossing is present for the two simulated covariates (see Figure 6.15).

Crossing for the PPOM predicted probabilities vs x_1 and x_2 does not get fixed with the Lasso approach ($\lambda = 0.01$) while it does with our constrained log-likelihood approach (see Figure 6.16; $\hat{\boldsymbol{\alpha}} = (-1.52, -1.05, -1.05)$ and $L = 105.26$).

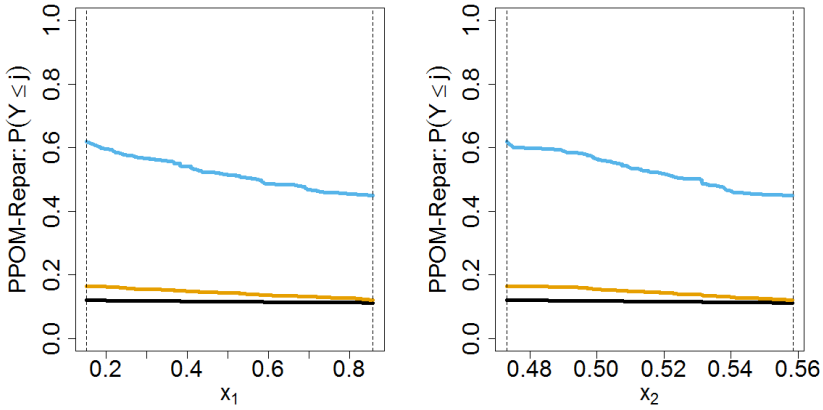

Figure 6.16: Constrained log-likelihood - predicted probabilities for PPOM-generated data fitted as PPOM for scaled covariates x_1 (left) and x_2 (right).

6.5 Case studies

6.5.1 Pro-environmental attitudes

In addition to the simulation study, we apply our constrained solution to data from the "Scottish Environmental Attitudes and Behaviours Survey" (Ipsos MORI Scotland and Scottish Government, 2009; described in Chapter 2). Within this data set we obtain crossing of regression lines when we model *educational attainment* versus *age* via a PPOM (see Figure 6.17).

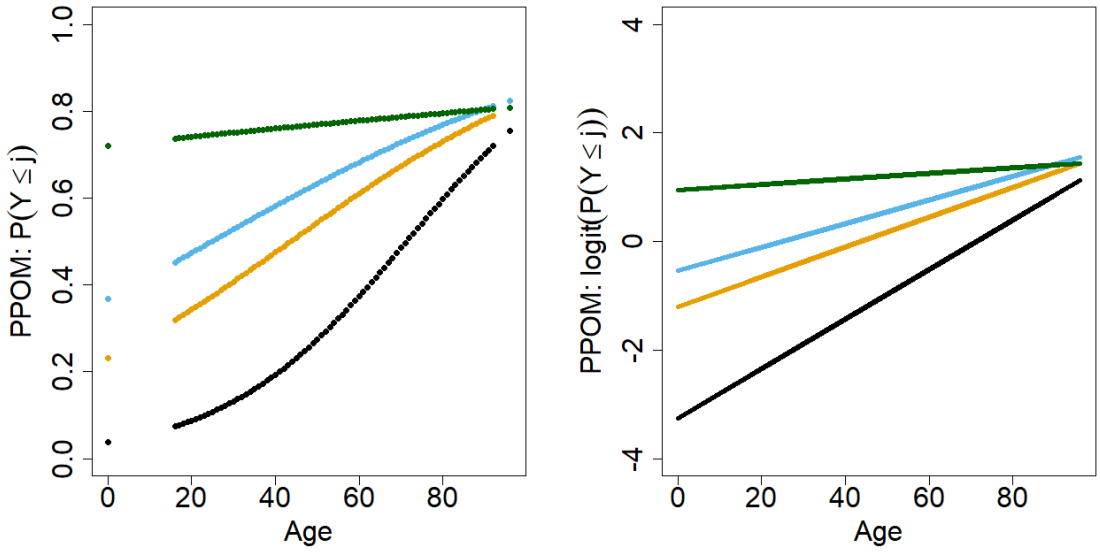

Figure 6.17: PPOM for Pro-environmental attitudes case study.

We can define the PPOM for this case study as follows

$$\text{logit}(P(Y_i \leq j)) = \alpha_j + \gamma_j x_i = \alpha_j + (\gamma + u_j)x_i, \tag{6.17}$$

where $i = 1, .., n$, $j = 1, \ldots, 4$, and we have relaxed the PO assumption for the covariate *age*.

Predicted cumulative probabilities from the model with the corresponding penalised log-likelihood are shown in Figure 6.18 (choice of λ and optimal value justified graphically in Appendix E).

We have considered as starting values for the threshold and regression parameters those from the corresponding POM. We have determined the values for u_{ij} as those parameters for the $C-1$ ordinary logistic regression models resulting from progressively dichotomising our ordinal response (e.g., for $j = 1$, the model is a logistic regression model where the binomial response is divided into those observations falling in category 1 or less, $Y_i \leq 1$, and those falling in a higher category than 1, $Y_i > 1$). For our mental impairment example, we would get the following results if we used a set of logistic regressions and a CLM (POM; see Table 6.1).

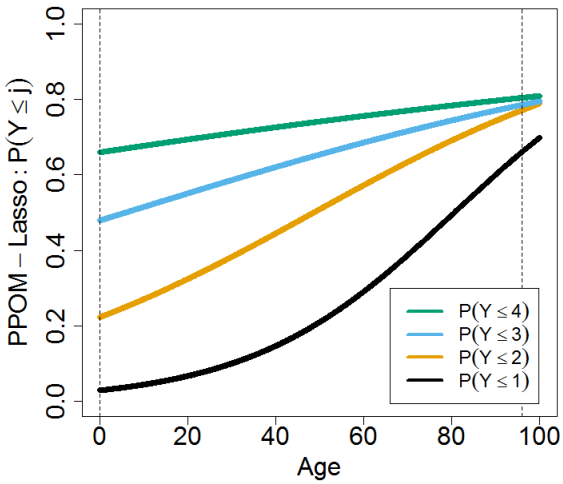

Figure 6.18: Lasso ($\lambda = 0.002$) - PPOM for Pro-environmental attitudes case study.

Table 6.1: Comparison of ordinary logistic regressions and POM.

	OLR		**POM**	
j	Intercept	Age	α_j	Age
1	-3.790	0.051	-2.502	0.028
2	-1.416	0.030	-1.234	
3	-0.028	0.014	-0.611	
4	0.674	0.008	-0.181	

We also consider the linear constraints approach such that we get the predicted cumulative probabilities shown in Figure 6.19 which do not cross.

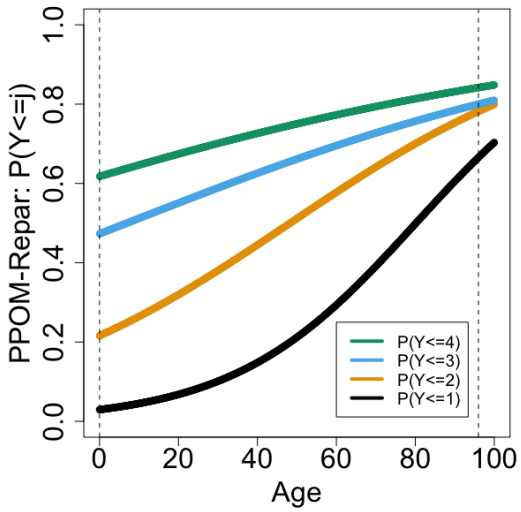

Figure 6.19: Constrained log-likelihood PPOM for pro-environmental attitudes case study.

We have found for this example that both solutions fix the crossing issue. However, as we have specified before, our constrained approach is a more systematic and efficient technique than the Lasso penalty.

6.5.2 Retinopathy

We have found in Chapter 3 that we get crossing of predicted regression lines (see Figure 3.2) for the PPOM of *left eye retinopathy severity* with covariates *systolic blood pressure* and *left eye refraction index*. We now assess the effectiveness of the proposed methods to fix this issue for this specific model.

6.5.2.1 Comparison of Lasso and constrained solution

We have run the Lasso penalised PPOM (find the assessment for the choice of optimal λ and log-likelihood in Appendix E), and have obtained the following results (see Figure 6.20) which still show signs of crossing.

Alternatively, our constrained approach is computationally more efficient and guarantees no crossing (see Figure 6.21 where the predictions do not cross).

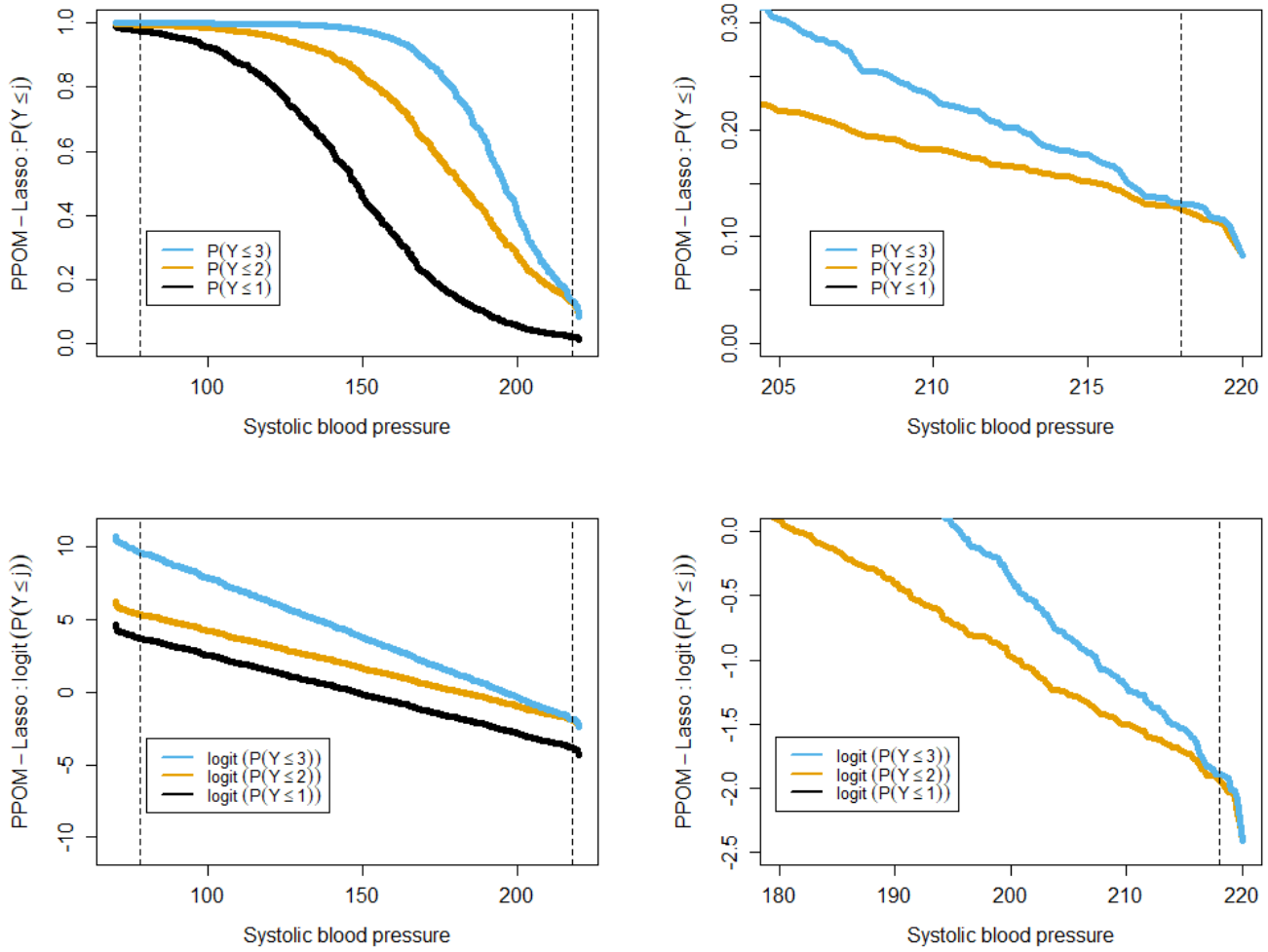

Figure 6.20: Centered-term predictions for PPOM of *left eye severity of retinopathy* as a function of *systolic blood pressure* and *left eye refraction index* with Lasso penalisation ($\lambda = 0.01$), on the probability scale (top) and logit scale (bottom). Zoomed-in versions of the plots have been included on the right for ease of visualisation.

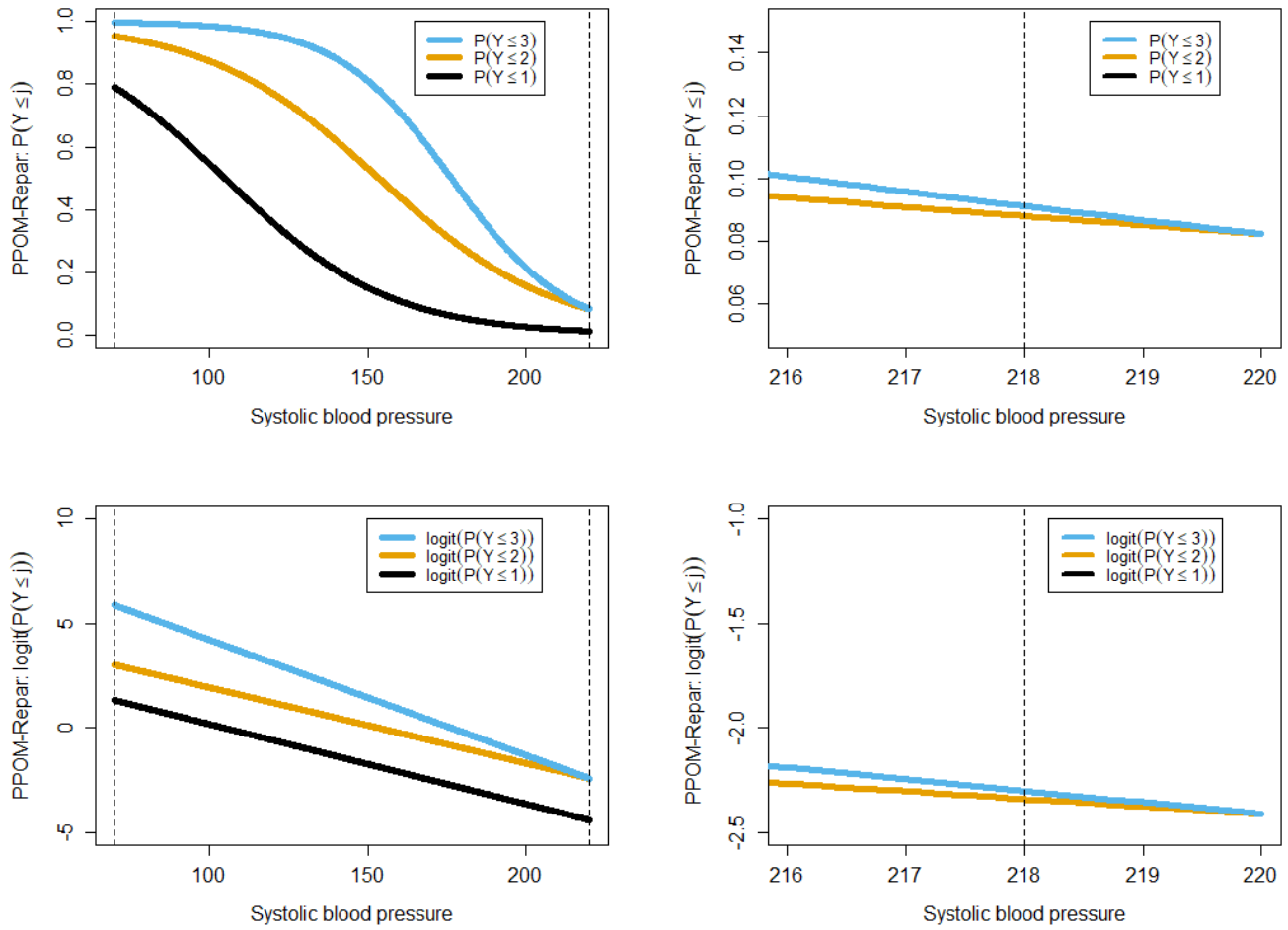

Figure 6.21: PPOM of *left eye severity of retinopathy* as a function of *systolic blood pressure* and *left eye refraction index* with constrained log-likelihood, on the probability scale (top) and logit scale (bottom). Zoomed-in versions of the plots have been included on the right for ease of visualisation.

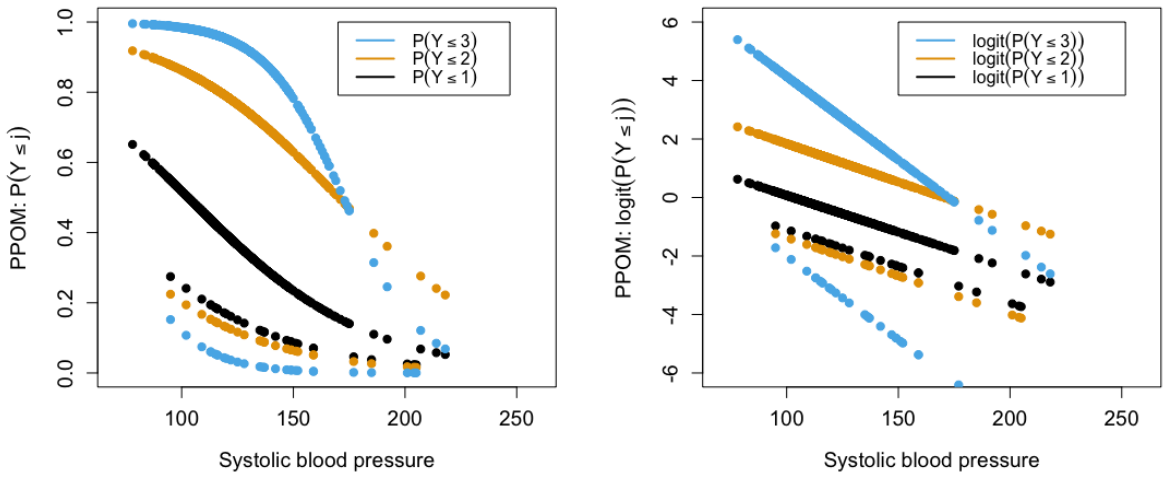

Figure 6.22: PPOM of *left eye severity of retinopathy* with the PO assumption relaxed for both *systolic blood pressure* and *left eye refraction index*.

We have found that a more flexible version of that model, for which as well as relaxing the PO assumption for the *systolic blood pressure*, we relax it for the *left eye refraction index*, issues still arise (see Figure 6.22 for the predicted probabilities with respect to the blood pressure; given that *left eye refraction index* is a dummy variable, representation with respect to that covariate is not straightforward).

6.5.2.2 Model selection

Procedure

We apply stepwise selection which is an iterative process whereby the AIC of the models is computed and the model that yields the lowest AIC is retained for the next iteration. The covariate that leads to a minimum AIC when dropped, is dropped for the next iteration, until there is no significant drop in AIC. In addition to this, we compare the AICs for all the PO/PPO combinations. We can visualise this procedure by means of 'best subsets' plots where the various shortlisted models are represented with the corresponding measures of fit.

Case studies

Our constrained log-likelihood approach allows us to determine the best fitting model, by indicating for which variables we should relax the PO assumption. We show the results of this model selection for two of our case studies.

Connectedness to nature

We start by looking at model selection for this case study with only 2 covariates, *CNS* and *Experience*.

Table 6.2: Connectedness to nature data set - constrained log-likelihood model selection summary. All models contain an intercept therefore first model is intercept-only.

PO variables	PPO variables	Log-likelihood	AIC
		-1192.76	2397.51
{*CNS*}		-1192.74	2401.48
{*Experience*}		-1047.04	2108.07
{*CNS,Experience*}		-1047.02	2122.04
	{*CNS*}	-1179.73	2395.47
	{*Experience*}	-1260.35	2544.70
{*CNS*}	{*Experience*}	-1036.30	2110.61
{*Experience*}	{*CNS*}	-1199.36	2426.73
	{*CNS,Experience*}	-1246.03	2540.06

We find the results in Table 6.2 when applying our constrained log-likelihood method, whereby the best fit is the POM with the PO assumption holding for covariate *Experience* (AIC=2108.07).

Retinopathy

Automated model selection procedures for ordinal response data models are not available in `R`. While stepwise or forward selection methods are unavailable in `Stata` (Williams, 2006), the `gologit2`'s `autofit` option provides a kind of backward selection technique for PPOMs. We apply this iterative method to the retinopathy data set considering *lerl* as ordinal response variable and the other 17 variables in the data set as covariates. We find that for a desired 5% significance level the model that best fits the data with *lerl* as response variable is such that

$$\text{logit}(P(Y_{i,lerl} \leq j)) = \alpha_j + \sum_{k=1}^{15} \beta_k x_{ik} + \sum_{k=1}^{2} \gamma_{jk} z_{ik}, \tag{6.18}$$

where $i = 1, \ldots, 720$ and $j = 1, 2, 3$. The PO assumption holds for variables

$$\boldsymbol{x} = \{rme, lme, rre, lre, riop, liop, age, diab, gh, dbp, bmi, pr, sex, prot, rerl\},$$

and is relaxed for $\boldsymbol{z} = \{sbp, dose\}$, with $\hat{\boldsymbol{\alpha}} = (-20.08, -11.07, -7.13)$ and $L = -442.36$. This selection method has the limitation that it only compares different PO/PPO specifications

for a model with all the covariates. Further to (6.18), we have found that we can reduce the number of PO covariates stepwise (see Figure F.1 in Appendix F) which leads us to the following model

$$\text{logit}(P(Y_{i,lerl} \leq j)) = \alpha_j + \sum_{k=1}^{7} \beta_k x_{ik} + \sum_{k=1}^{2} \gamma_{jk} z_{ik}, \tag{6.19}$$

where $i = 1, \ldots, 720$ and $j = 1, 2, 3$. The PO assumption holds for variables

$$\boldsymbol{x} = \{lme, lre, diab, gh, dbp, prot, rerl\},$$

and is relaxed for $\boldsymbol{z} = \{sbp, dose\}$, with $\hat{\boldsymbol{\alpha}} = (-19.57, -10.72, -6.85)$ and $L = -444.01$.

However, we argue that the validity of the `Stata` automatic model selection is limited as models with negative predictive probabilities give a warning but still run and provide particularly high values therefore would not be selected by the routine. We propose instead the use of our constrained log-likelihood to guide the selection as we show next. When we run our iterative model selection routine for the same data set we find that the PPOM suggested above presents a higher AIC than the corresponding POM (see Table 6.3 for a summary of related models) defined as

$$\text{logit}(P(Y_{i,lerl} \leq j)) = \alpha_j + \sum_{k=1}^{9} \beta_k x_{ik}, \tag{6.20}$$

where $i = 1, \ldots, 720$ and $j = 1, 2, 3$. The PO assumption holds for all the covariates

$$\boldsymbol{x} = \{lme, lre, diab, gh, dbp, prot, rerl, sbp, dose\},$$

with $\hat{\boldsymbol{\alpha}} = (-0.02, 4.11, 7.48)$.

Table 6.3: Retinopathy data set - constrained log-likelihood model selection summary. All models contain an intercept therefore first model is intercept-only.

PO variables	PPO variables	Log-likelihood	AIC
{*lee,lre,diab,gh,dbp,prot,rerl*}	{*sbp,dose*}	-459.61	955.21
{*lee,lre,diab,gh,dbp,prot,rerl,sbp,dose*}		-448.30	924.60
{*lee,lre,diab,gh,dbp,prot,rerl,dose*}	{*sbp*}	-466.26	964.53
{*lee,lre,diab,gh,dbp,prot,rerl,sbp*}	{*dose*}	-466.31	964.62

6.6 Conclusions

We have shown the limitations of PPOMs and have addressed issues associated with negative predictions. In particular, we have compared a common regularisation approach (Lasso penalisation) to our constrained log-likelihood solution.

Lasso penalisation does not guarantee the crossing not to be present, and it is not as systematic an approach as our proposed solution. In addition to this limitation, it requires a choice of shrinkage parameter, which limits its user-friendliness. Results can be highly sensitive to this choice too. At the same time, in theory the lasso can help guide us to whether the PO assumption for a variable should be used or not.

Our proposed constrained approach is a more systematic and computationally efficient approach, and has proven to fix the issue of getting negative fitted category probabilities consistently for the examples under study. This solution enables us to do model selection in a consistent, accurate manner as shown in the connectedness to nature and retinopathy case studies. Future work will be devoted to the exploration of its performances and statistical properties, by means of a larger and well-structured simulation study which will include visiting work by Espinosa and Hennig (2018) who build a constrained maximum likelihood to fit proportional odds with monotonicity constraints applied to the coefficients of ordinal predictors of an ordinal response and the associated `R` package `crov` fitting the proposed constrained model. Further simulations are also planned to ascertain advantages of the constrained log-likelihood over Lasso in more extreme cases and for ordinal mixed models, to ensure there are not convergence issues.

Chapter 7

Discussion

This thesis has researched past and current developments in ordinal response models, from both a methodological and an applied perspective. By exploring the specific limitations in the modelling and goodness of fit diagnostics, and examining the interpretation challenges of these models, we have arrived at a better understanding of the particular reasons for their relatively low uptake in socio-economic research and paved the way to overcoming this issue through the promotion of good practice and with the addition of a novel solution to one of the ordinal models' limitations which ultimately will serve as a user-friendly model selection tool. In this chapter we will discuss how this has been achieved in more detail and set out some thoughts on further work.

We have studied ordinal response models via the latent variable approach with an emphasis on CLMs. These appear to be the preferred 'generic' options in the literature and have been described in Chapters 1 and 3. The latent variable approach takes into account the ordered nature of the response variable and provides an intuitive interpretation in terms of an underlying continuous variable (Agresti, 2010). Despite some criticisms on the interpretation of the latent parameters (e.g., Boes and Winkelmann, 2006), we have argued in Chapter 1 that latent constructs are prominent in socio-economic analysis and therefore can be particularly appealing and easily understandable by researchers in the area. CLMs are based on cumulative response probabilities for each response category, which also reflects the ordered nature of the response. We consider by default the logit link function for its intuitive interpretation in terms of odds ratios, and its popular use. CLMs do not make any specific assumption on the data generation process, therefore they are more generic and flexible than continuation ratio, sequential and adjacent category models (further described in Chapter 3). These models are also more suitable than multinomial logit models because multinomial models do not reflect the ordered (and not necessarily equally spaced) nature of the categories of the response variable and can result in an excessive number of parameters

(Williams, 2016) as we have discussed in Chapters 3 and 6. Alternatives such as CUB models (Iannario and Piccolo, 2012) which consider a different approach from the latent, and Bayesian techniques (widely described in Gill, 1993) are out of the scope of this thesis.

An exhaustive framework for the application of these models within areas of research where they have been most prominent (e.g., throughout decades of medical research) does not exist and this is reflected in the associated applied literature which includes varied levels of interpretation of CLMs (e.g., see Sasidharan and Menendez, 2014 for what we believe is an incomplete application). This prevalence is however reflected in the abundance of openly available medical data sets with ordered responses. We have taken this into account by studying a case study from that field of research (in particular, from ophthalmology) as described in Chapter 2. In addition, we have concentrated upon examples from the socio-economic area in order to better understand some of the potential issues researchers in this field encounter. For instance, our three socio-economic examples (all of them associated to data from questionnaires on environmental topics and also described in Chapter 2) have shown the need for further research into multilevel or mixed effects structures that incorporate subject effects and nested levels of variability (see Chapters 4 and 5). We have introduced these models in Chapter 3 but further assessment and application have fallen outside the scope of this thesis.

Our literature review (Chapter 3), has detailed the current state of the art in terms of the advantages and limitations of CLMs. We have set out how they have been used in socio-economics and in other areas of research, establishing the context for a better understanding of what has failed in certain areas and succeeded in others. We have covered both POMs which incorporate the restrictive PO assumption that rarely holds for all the covariates in the model (Lall et al., 2002), and PPOMs which relax this assumption for some or all of the covariates. Through the course of this detailed analysis we have found that POMs are widespread in the socio-economic literature irrespectively of whether this PO assumption holds for all the covariates (in some cases, it is not even reported whether it has been tested; e.g., Brumback et al., 2012) which can lead to biased estimates. Other common practices in the area that this thesis has brought to light, are either not reporting the threshold parameters (e.g., Mayer and Foster, 2015 for a PPOM) or highlighting their limited importance (e.g, Kuesten et al., 2017 who claim that they "have less practical meaning than the factors"). Our research raises concerns that this is sometimes done to avoid interpretation challenges, and we have provided guidance on how to address this *a priori* daunting task (see Appendix A).

Our software review (included in Chapter 3) has shown the restricted capabilities of the multiple `R` libraries implementing ordinal response models. This is apparent in our two chosen libraries. While only `ordinal` allows different threshold specifications, it does not

have enough flexibility to model complex PPOMs in contradistinction to `VGAM`. None of the libraries, our review reveals, include ordinal-specific residual diagnostics. We have also mentioned libraries from commercial software (i.e., `SAS` and `Stata`) but our focus on open software puts these resources beyond the remit of this current study.

A clear theme emerging from both reviews is the lack of agreement on the use of ordinal models. This state of affairs constitutes one of the challenges where we have sought to make progress, specifically leading us to develop the initial piece of research both on the adequacy of these models and on the effects on inference from inadequate modelling as described in Chapter 4. We have worked on a simulation study and two of the case studies to show the advantages of an appropriate ordinal modelling of ordered response data, which has highlighted differences both at the intercept and the slope level in terms of significance and parameter estimates. This study has shown that by fitting 'simpler' linear models to ordinal response data, researchers might be reaching the wrong conclusions which might invalidate research evidence in specific cases, and it has highlighted the practical relevance of an appropriate ordinal modelling.

Following our work on the advantages and quality of ordinal models (Chapter 4), we have identified a gap in the assessment of these models. The most complete methodological compilation of ordered response models (Agresti, 2010) for instance only mentions goodness of fit measures and it disregards graphical diagnostics. Our literature review has been able to show how recent literature is starting to address this topic (Li and Shepherd, 2012, Shepherd et al., 2016, Liu and Zhang, 2018). We have assessed the available residual diagnostics for CLMs and have shifted the focus to three specific residuals (i.e., D-S, L-S, and surrogate residuals). We have completed a simulation study comparing these residuals and exploring other questions regarding graphical checks of different types of model misspecification and overall goodness of fit. We have reported appropriate and informative summary measures and results from the simulation in Chapter 5 and Appendix D, which have shown that in terms of graphical diagnostics, surrogate residuals are probably the most sensitive residuals and L-S residuals fail to provide an interpretable view of the patterns in the data, given the presence of discrete bands. Our simulation study has looked to extend work by Liu and Zhang (2018) by considering D-S residuals and equidistant, symmetric and unconstrained threshold structures. As a consequence and when possible, we have used the authors' simulated data and threshold values, which are not necessarily consistent across scenarios. Our study has revealed new avenues for future research whereby a homogeneous choice of thresholds and simulated variables across scenarios has the potential to allow further comparisons and lead to wider conclusions, and this will be covered in future research. We are also now in a position, as a result of the progress we have made, to identify the need for further work to

determine more specific benefits of surrogate over D-S residuals.

In Chapter 6 we have discussed limitations of PPOMs and proposed solutions to the issue of negative fitted category probabilities. We argue that solutions to this issue in the literature (e.g., combining response categories as proposed by Williams, 2016) are not necessarily effective for all data, and could cause information loss. We have initially imposed a Lasso penalisation and more systematically, defined a framework for a constrained log-likelihood that avoids the crossing of the corresponding regression lines. Our investigations have shown that the former method does not always solve the problem while the latter method guarantees the issue will not appear. We have also presented by means of two small case studies, its capabilities as an efficient model selection tool. Having made this methodological advance, automation of this novel approach in `R` will be a relatively straightforward next step.

We have inferred from the literature in the area that there is a distinct lack of consistency across research areas and between individual researchers both in model specification (e.g., PO or PPO, threshold structures) and diagnostics. This is reflected in the varied impact of these models in the applied literature. We further contend that this might itself be causing avoidance of these models if researchers think it is not strictly necessary or the trade-off complexity-benefits seems unbalanced towards the former. In those cases, researchers take a linear model approach, where by means of a standard regression, they model the response as a continuous variable with categories' scores as real values (Kramer et al., 2001). We show that ignoring the ordered nature of the data can have an impact on inference and might be hiding other forms of model misspecification. In this way we add to the literature, where substantial arguments have been made for improvement of statistical power (Capuano et al., 2007, Kosmidis, 2014), but no specific mention to inference and model misspecification diagnostics has been made.

We argue that despite often being ignored, PPOMs are a more flexible, attractive starting point for ordinal modelling than POMs and are the best fit for our data when the PO does not hold. Any apparent increasing complexity over POMs is compensated by the fact that PPOMs allow the data to drive the analysis in contrast to POMs which force a strong assumption. They also allow a more flexible interpretation because the effect of a covariate can be assumed to be different according to the different levels of the response. For instance, if the ordered categories represent the stages of an illness, certain external factors (measured by the covariates) might be more intrusive at onset or later stages. Overall, this is a more open representation of reality, which gives the researcher a richer insight on the relationships between the variables. We recognise however that with many covariates, the task might become insurmountable, and we identify this as an issue for future research. While there are some other limitations to these models (e.g., non-convergence as described in Chapter

6), we have successfully dealt with one of the most restrictive issues. By imposing linear constraints on the model, we have managed to avoid the presence of negative fitted category probabilities, which were initially causing lack of convergence for specific cases (e.g., quasi-separation) and limited model selection strategies.

In terms of diagnostics, we have confirmed that L-S residuals are not a convenient method to detect lack of fit and deviations from distributional assumptions for ordinal models as it is also reported in Greenwell et al. (2017). Given the discrete nature of these residuals, their graphical representations are not informative and it becomes difficult to reach conclusions both on distributional assumptions and goodness of fit. While we have only detected small differences between D-S and surrogate residuals, our simulation study has led us to view surrogate residuals as slightly more sensitive to misspecifications. This insight provides the basis for future comparison and closer scrutiny of this sensitivity as it has not been addressed by the authors defining surrogate residuals (Liu and Zhang, 2018) and requires further attention. We recommend the use of more than one method of graphical diagnostics to be able to account for the complex nature of ordinal response models and the different types of potential misspecifications. This includes a new suite of graphical tools known as cumulative residual processes (included in the `R` library `timereg`) that we would like to study further. This simulation study covering scenarios where different misspecifications have been considered, has also confirmed our argument that by ignoring the ordinal nature, we might be also covering further misspecifications which would not be necessarily apparent in the standard graphical diagnostics for the corresponding 'linear' residuals.

In summary, we have attempted to overcome three significant hurdles related to the lack of uptake of ordinal methods, as summarised in Chapters 4 to 6. This work builds firm foundations for a comprehensive, improved framework for ordinal response data modelling. In addition to those topics already highlighted in previous paragraphs, future research will look at those methods that have not been studied yet (e.g., CUB models and cumulative process diagnostics plots). It will also tackle implementation of PPOMMs with different threshold structures to assess whether these models can provide better estimates and more noteworthy differences. We also aim to arrange the work described in this thesis in the form of an `R` package that will allow for a straightforward application of these methods covering both modelling and diagnostics, which therefore could increase uptake of these models.

Appendices

Appendix A

Ordinal response models implementation in R

Although there are several libraries in `R` (R Core Team, 2018) which can implement some of the models reported in this paper (e.g., `vcrpart`, `polr`, `MCMCglmm` -only with probit link function-, `rms`, `mixcat`, `DPpackage`, `lcmm`, and `GMMBoost`), we specify below the functionalities of packages `ordinal` (Christensen, 2015) and `VGAM` (Yee, 2010), which include a substantially more user-friendly and complete implementation of ordinal response models via ML than some of the above mentioned packages.

A.1 Cumulative logit models

These CLMs can be easily run using the `ordinal` `clm` function (or `clm2`). Analogously, for a mixed model approach, we can use either `clmm` or `clmm2`. Both sets of functions consider the logit link function by default.

In `VGAM`, the function `vglm` with the command `family` includes the cumulative family with both `probit` and `logit` options for link function.

A.2 Proportional odds models

`nominal_test` and `scale_test` allow us to test for nominal and scale effects in `ordinal`, thus they are used in practice as a way to check the PO assumption or equal slopes across the response categories thresholds. However, these functions are not available for the cases of cumulative link mixed models (i.e., `clmm`, `clmm2` function) so in those cases a basic likelihood ratio test approach would serve the purpose.

Within `vglm` and the `cumulative` family in `VGAM`, we can state that the `parallel` assumption is `TRUE` for POMs.

A.3 Partial proportional odds models

`nominal` allows us to relax one or more of the independent variables to have different regression parameters according to the response category in order to define a more general version of the proportional odds model.

Within `vglm` and the `cumulative` family in `VGAM`, we can state that the `parallel` assumption is `FALSE` for PPOMs, and specify the variable for which we are relaxing the PO assumption.

A.4 Types of thresholds

As stated in 3.3.4, there are different threshold structure specifications that can be defined by design. In `clm` (and `clmm`) models these are included by using the command threshold with options `flexible` (unconstrained), `symmetric`, `symmetric2` (symmetric around 0; not allowed in older versions `clm2` and `clmm2`) and `equidistant`.

We illustrate the case of a response variable with 7 categories, for which, if we set the thresholds to be unconstrained, we would need 6 parameters ($C-1$, where C is the number of ordered categories). For symmetric thresholds, 4 parameters would be needed to accurately report the results from the ordinal model (*central1*, *central2*, *spacing1*, *spacing2*; a visual interpretation of these parameters is shown in Figure A.1).

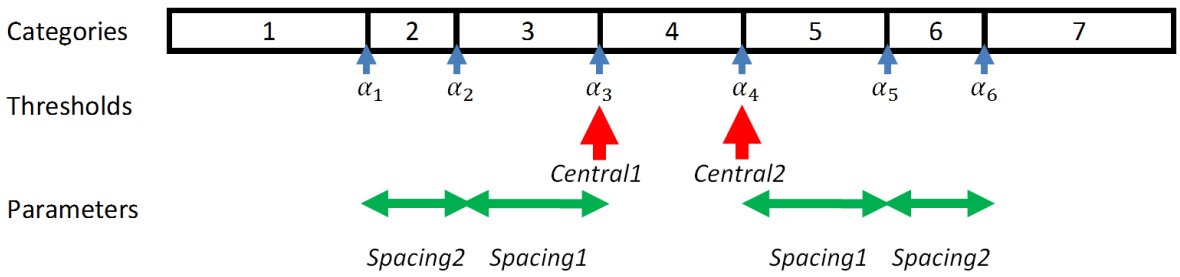

Figure A.1: Symmetric threshold coefficients interpretation.

`clm.model$Theta` contains the actual positions of the thresholds on the latent scale as derived from Figure 35 and as shown in Table 3.9 (note that similar calculations can be

Table A.1: Case study 1: Example of determination of threshold values in the latent scale for *CNS1* and *nature*.

-Spacing2+ Central1	**-Spacing1+ Central1**	**Central1**	**Central2**	**Spacing1+ Central2**	**Spacing2+ Central2**
-1.419+ -2.080= -3.499	-0.383+ -2.080= -2.463	-2.080	-1.591	0.383+ -1.591= -1.209	1.419+ -1.591= -0.173

derived for the slopes). Given a specific model, these values will be necessarily similar for both constrained and unconstrained thresholds structure.

Finally, for equidistant thresholds, only 2 parameters are required (threshold1, spacing; a visual interpretation of these parameters is shown in Figure A.2). However, it could be argued that this equidistant structure would in reality be analogous to the continuous approach.

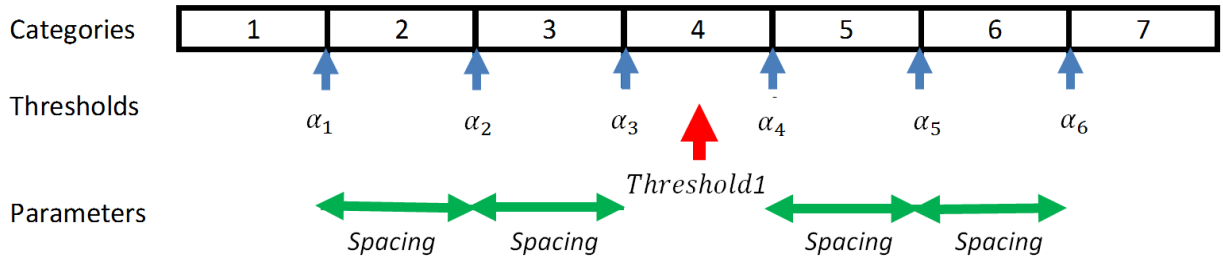

Figure A.2: Equidistant threshold coefficients interpretation.

A.5 Category probabilities

Given one of the theta or cumulative logit values as calculated in the previous subsection, we can proceed to translate this cumulative logit into the corresponding response category probability (the procedure to calculate these probabilities and the results for Case study 1 are shown in Table A.1). If we exponentiate the cumulative logit, we obtain what is known as cumulative odds ratio (the odds of at least jth level according to definitions in Materials and Methods). If we divide 1 by the sum of the cumulative odds and 1, we obtain the cumulative probability ($P(Y_i) \leq j$, as defined in the above mentioned section). Finally, when we subtract this probability from the one corresponding to a higher category, we obtain that categorys probability.

Alternatively, the function `predict` will provide the category probabilities directly in the case of CLMs. For cumulative logit mixed models, further calculations are required.

Table A.2: Case study 1: Example of determination of *pleasantness* response category predicted probabilities calculation for *CNS1* and *nature* (above) and *shopping* (below) experiences.

Nature	***Categories***						
	1	**2**	**3**	**4**	**5**	**6**	**7**
Cumulative logits	-	-3.499	-2.463	-2.080	-1.591	-1.209	-0.173
Cumulative odds ratios	-	0.030	0.085	0.125	0.204	0.299	0.842
Cumulative probabilities	1.000	0.971	0.922	0.889	0.831	0.770	0.543
Category probabilities	0.029	0.049	0.033	0.058	0.061	0.227	0.543
Shopping	***Categories***						
	1	**2**	**3**	**4**	**5**	**6**	**7**
Cumulative logits	-	-2.005	-1.215	-0.608	0.094	0.701	1.491
Cumulative odds ratios	-	0.135	0.297	0.544	1.099	2.016	4.442
Cumulative probabilities	1.000	0.881	0.771	0.647	0.477	0.332	0.184
Category probabilities	0.119	0.110	0.124	0.171	0.145	0.148	0.184

The function `summary` provides z-values and `drop1` with the option `test=Chi` provides Chi-squared tests and F-values. Additionally, the `anova` command can only be used within the ordinal package to compare different models (via likelihood ratio tests).

A.6 Residual diagnostics

Several `R` packages (e.g., `boral`, `mvabund`, `GAMLSS`, and `statmod`) include functions that calculate randomized quantile residuals for GLMs but a gap seems to exist in the specific case of packages fitting CLMs. Package `sure` (Greenwell et al., 2017) produces surrogate residuals. The package currently supports CLMs built using packages `MASS`, `ordinal`, `rms` and `VGAM`.

A.7 Regularisation and constrained log-likelihood

As mentioned in Chapter 6, there are not specific `R` packages for regularisation of CLMs. `cvglmnet` has been used for tuning parameters determination. To the best of our knowledge, there are no `R` libraries available to run models with our proposed constrained log-likelihood.

Appendix B

Predicted probabilities

B.1 Case study 1

Table B.1: Case study 1: *pleasantness* response category predicted probabilities by *CNS* and *Experience*.

CNS=1	***Categories***						
Experience	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Nature	0.029	0.049	0.033	0.058	0.061	0.227	0.543
Shopping	0.119	0.110	0.124	0.171	0.145	0.148	0.184
CNS=2	***Categories***						
Experience	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Nature	0.015	0.025	0.018	0.033	0.037	0.164	0.708
Shopping	0.140	0.124	0.133	0.173	0.138	0.134	0.157
CNS=3	***Categories***						
Experience	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Nature	0.009	0.017	0.012	0.022	0.026	0.123	0.791
Shopping	0.222	0.164	0.150	0.164	0.111	0.093	0.096

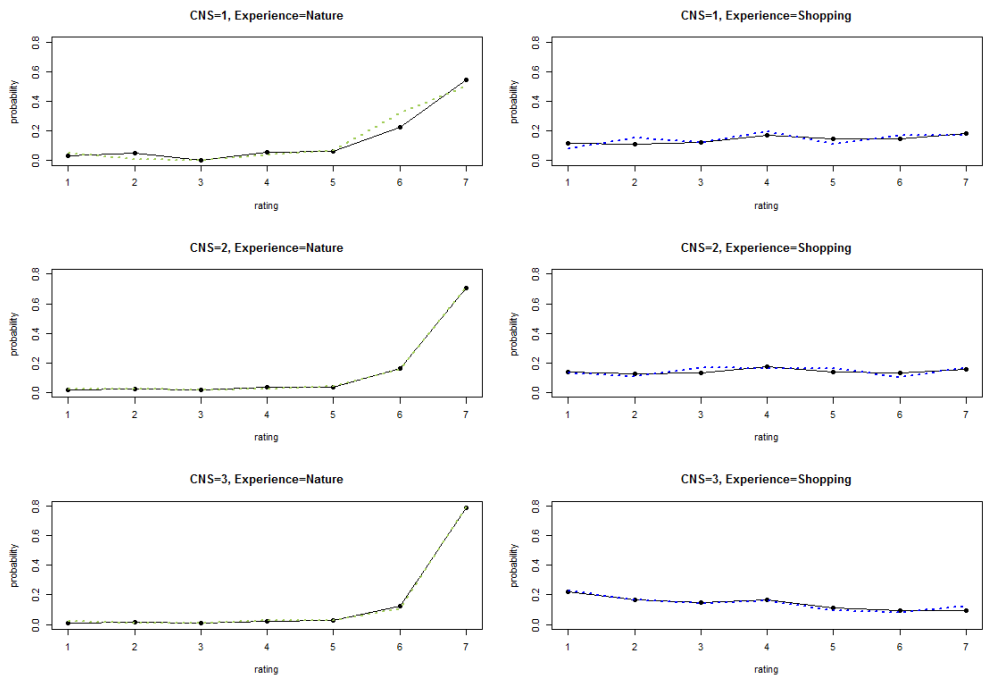

Figure B.1: Case study 1: prediction (continuous line) and actual distribution relative frequencies of response categories from the data (dotted line) for the POM.

B.2 Case study 2

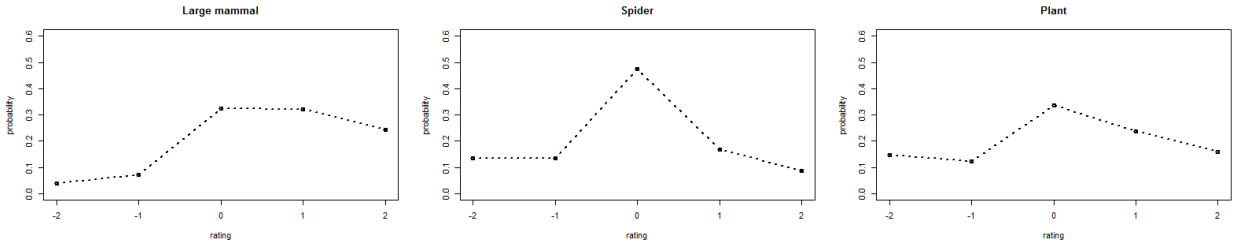

Figure B.2: Case study 2: overall relative frequencies of the three species' *desirability* response categories.

Table B.2: Case study 2: *desirability* response category predicted probabilities by *country* and *species*.

Austria	**Categories**					***Netherlands***	**Categories**				
Species	**-2**	**-1**	**0**	**1**	**2**	**Species**	**-2**	**-1**	**0**	**1**	**2**
Large mammal	0.056	0.000	0.774	0.141	0.030	***Large mammal***	0.004	0.000	0.693	0.284	0.018
Spider	0.108	0.000	0.747	0.140	0.005	***Spider***	0.045	0.000	0.804	0.111	0.040
Plant	0.109	0.000	0.600	0.223	0.069	***Plant***	0.316	0.000	0.618	0.057	0.009
Flanders	**Categories**					***Romania***	**Categories**				
Species	**-2**	**-1**	**0**	**1**	**2**	**Species**	**-2**	**-1**	**0**	**1**	**2**
Large mammal	0.000	0.000	0.257	0.396	0.347	***Large mammal***	0.000	0.000	0.143	0.174	0.683
Spider	0.031	0.000	0.830	0.121	0.018	***Spider***	0.387	0.000	0.588	0.025	0.000
Plant	0.006	0.000	0.358	0.408	0.229	***Plant***	0.381	0.000	0.571	0.049	0.000
France	**Categories**					***Scotland***	**Categories**				
Species	**-2**	**-1**	**0**	**1**	**2**	**Species**	**-2**	**-1**	**0**	**1**	**2**
Large mammal	0.010	0.000	0.361	0.298	0.332	**Large mammal**	0.030	0.000	0.717	0.181	0.072
Spider	0.063	0.000	0.664	0.219	0.055	***Spider***	0.045	0.000	0.754	0.151	0.050
Plant	0.043	0.000	0.578	0.329	0.050	***Plant***	0.112	0.000	0.617	0.234	0.037
Hungary	**Categories**					***Slovakia***	**Categories**				
Species	**-2**	**-1**	**0**	**1**	**2**	**Species**	**-2**	**-1**	**0**	**1**	**2**
Large mammal	0.000	0.000	0.352	0.403	0.245	***Large mammal***	0.004	0.000	0.332	0.332	0.332
Spider	0.115	0.000	0.783	0.080	0.022	***Spider***	0.109	0.000	0.781	0.085	0.025
Plant	0.006	0.000	0.303	0.382	0.309	***Plant***	0.091	0.000	0.591	0.242	0.076

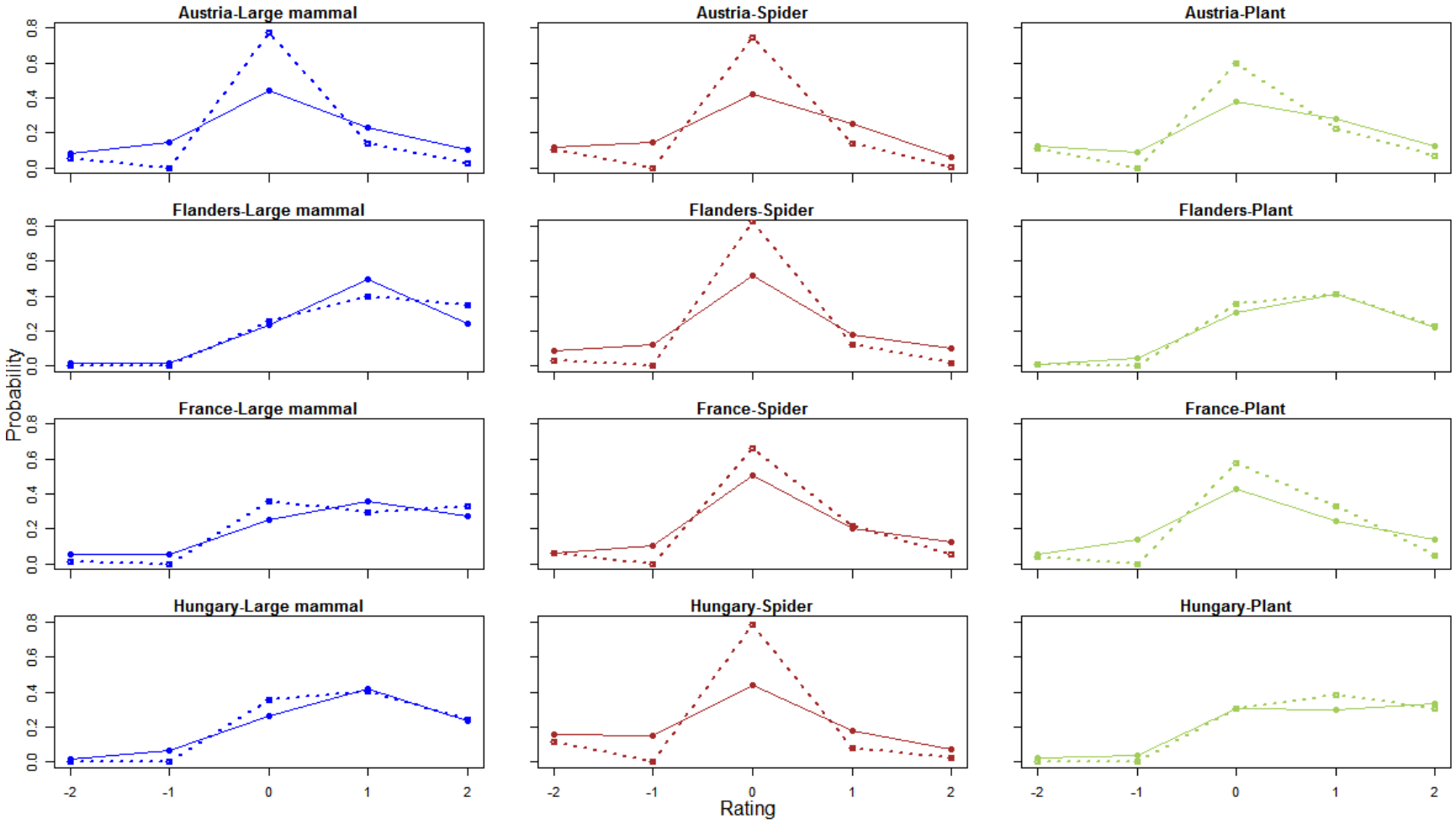

Figure B.3: Case study 2: prediction and actual distribution for the PPOMM.

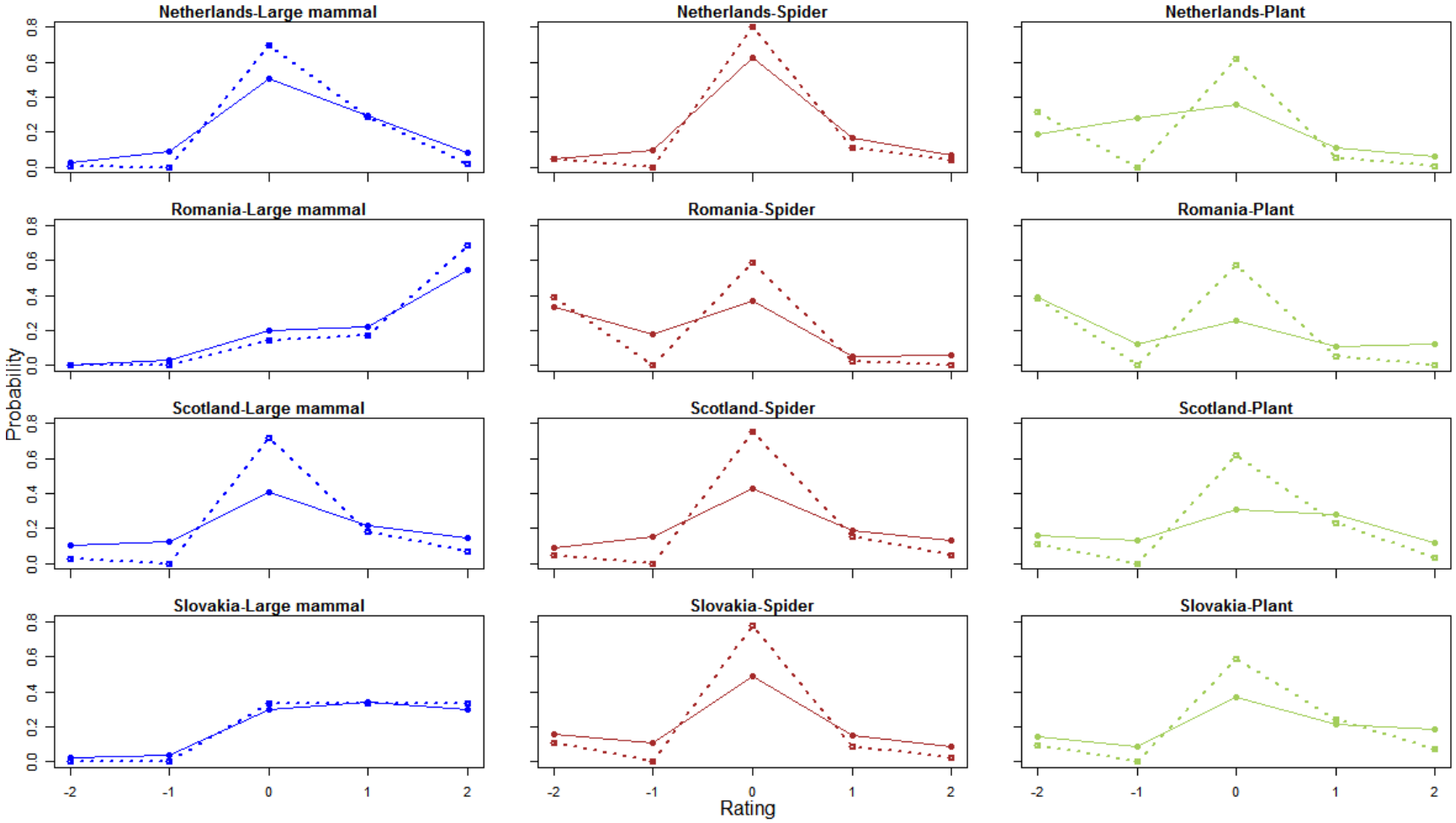

Figure B.4: Case study 2 (continued): prediction (dotted line) and actual distribution from the data (continuous line) for the PPOMM.

Appendix C

Traditional residuals

C.1 'Crude', standardised and studentised residuals

Crude or raw residuals were first defined by Cox and Snell (1968) for binomial data as

$$r_{ij} = Y_{ij} - \hat{Y}_{ij} \tag{C.1}$$

where $i = 1, \ldots, n$, $j = 1, \ldots, C-1$, and $\hat{Y}_{ij}$ is the fitted value for the ith observation, which is basically the difference between observed and fitted values. The function `resid` for `vglm` objects provides these 'crude' residuals when we specify `type=response`.

Standardised residuals are defined as the ratio of the 'crude' residuals divided by their standard error Agresti (2013) as shown in

$$sr_{ij} = r_{ij}/\sigma\big(r_{ij}\big), \tag{C.2}$$

where σ is the estimated standard error of the residuals under the assumption that the model holds. When the model holds, standardised residuals have approximate standard normal distributions.

These standardised and studentised residuals are generally preferred to 'crude' residuals (Andrews and Pregibon, 1978, Cook and Weisberg, 1982, Law and Jackson, 2017). However, when the response is not continuous and in particular ordered, the definition of residuals is not so evident, i.e., the simple decomposition $observation = fitted + residual$ does not hold (Di Iorio and Piccolo, 2009).

C.2 Pearson residuals

Together with Anscombe and deviance residuals, Pearson residuals were introduced to mimic the main assumptions of linear regression models in GLMs (McCullagh and Nelder, 1989). Based on the idea of subtracting off the mean and dividing by the standard deviation, these residuals are defined as

$$r_{ij} = \frac{Y_{ij} - \hat{\pi}_{ij}}{\sqrt{\hat{\pi}_{ij}(1 - \hat{\pi}_{ij})}} \tag{C.3}$$

They refer to the cumulative logits because it is their values that are being modelled (Upton, 2017). These residuals do not have the standard normal as their reference distribution.

The function `resid` for `vlgm` objects provides these 'crude' residuals when we specify `type=pearson`.

C.3 Cumulative residuals

Graphical diagnostics based on cumulative sums of residuals proposed initially by Toledano and Gatsonis (1996) for binary variables and extended to ordinal response variables by Liu et al. (2009).

C.4 Deviance residuals

Deviance residuals are defined such that

$$d_{ij} = s_{ij}\sqrt{-2Y_{ij}log\pi_{ij} + (1 - Y_{ij})\log(1 - \hat{\pi}_{ij})} \tag{C.4}$$

where $s_{ij} = \text{sign}\big(Y_{ij} - \hat{Y}_{ij}\big)$.

We can check deviance residuals for approximate normality and are easy to compute. Although they have many nice properties and are quite popular across a wide variety of models Pierce and Schafer (1986), they involve disjoint components (the deviance and the sign) and they are not naturally constructed for ordinal models Shepherd et al. (2016).

C.5 Adjusted deviance residuals

Adjusted deviance residuals are first defined by Pierce and Schafer (1986) and are obtained by making a first-order correction for the mean bias on the deviance residuals such that

$$ad_{ij} = d_{ij} + \frac{1 - 2\hat{\pi}_{ij}}{6\sqrt{n_i \hat{\pi}_{ij}(1 - \hat{\pi}_{ij})}} \tag{C.5}$$

As stated in (Upton, 2017), out of the previously described residuals, the distribution of the adjusted deviance residuals is often most nearly normal and yields surprisingly accurate approximations for even small n_i, provided that the fitted probabilities are not too close to 0 or 1.

C.6 Generalised residuals

Generalised residuals were originally designed for econometric models by Gourieroux et al. (1987), they assume a latent variable, and are shown to have properties similar to those of the traditional residuals for linear models. They assume that the observations are not serially correlated, which might be a strong assumption given that serial correlation is difficult to tackle in this type of models. They are highlighted by Iannario et al. (2017) as a particularly robust choice in the case of ordinal response models with the logit link function.

C.7 Score residuals

Score residuals are defined by Therneau et al. (1990) as

$$u_{ij} = x_i(Y_{ij} - \hat{\pi}_{ij}), \tag{C.6}$$

where $i = 1, \ldots, n$ and $j = 1, \ldots, C - 1$.

C.8 Partial residuals

Partial residuals are defined by Pruscha (1994) as

$$\hat{r}_i^{par} = \hat{\beta}_k x_{ik} + \frac{\big(Y_i - \hat{Y}_i\big)}{\hat{\pi}_{ij}\big(1 - \hat{\pi}_{ij}\big)}, \tag{C.7}$$

where$i = 1, \ldots, n$, $j = 1, \ldots, C - 1$, and $k = 1, \ldots, p$.

Bender and Benner (2000) recommend as graphical diagnostics, smoothed scatterplots with lowess -locally weighted scatterplot smoothing- as default smoothing technique), which are available in the `R` package `rms`.

C.9 Latent residuals

The first latent residual under consideration are defined by Albert and Chib (1995) in the binary response framework for outlying detection in symmetrical models such that

$$\varepsilon_i\big(z_i, \beta\big) = z_i - x_i\beta. \tag{C.8}$$

Appendix D

Additional results from residual simulations

We show further results to those reported in Chapter 3.5.

D.1 Scenario 1: PO misspecification

When the data is generated from **equidistant thresholds** $\boldsymbol{\alpha} = (-3.0, -1.5, 0, 1.5, 3.0)$, L-S residuals show discrete patterns which we cannot interpret in terms of fit, while D-S and surrogate residuals show the random scatter we would expect from a good model fit (see Figure D.1) and therefore fail to reflect the fact that we have forced the PO assumption on a PPOM.

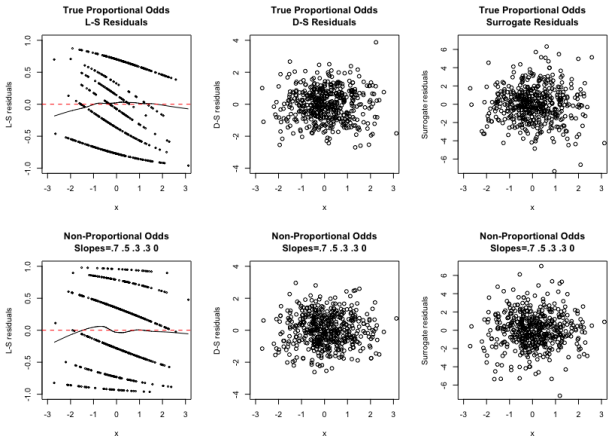

Figure D.1: Residuals for data generated from a POM with equidistant thresholds $\boldsymbol{\alpha} = (3.0, 1.5, 0, -1.5, -3.0)$ (top) and data generated from a PPOM (bottom): (a) L-S, (b) D-S, and (c) surrogate residuals.

The Q-Q plots in Figure D.2 show that the $U(-1, 1)$ distributional assumption for the L-S residuals do not hold as we get heavy tails. D-S residuals show a particularly good fit to the ideal normal distribution, and the normal distribution is less clear in the case of surrogate residuals than for the D-S residuals.

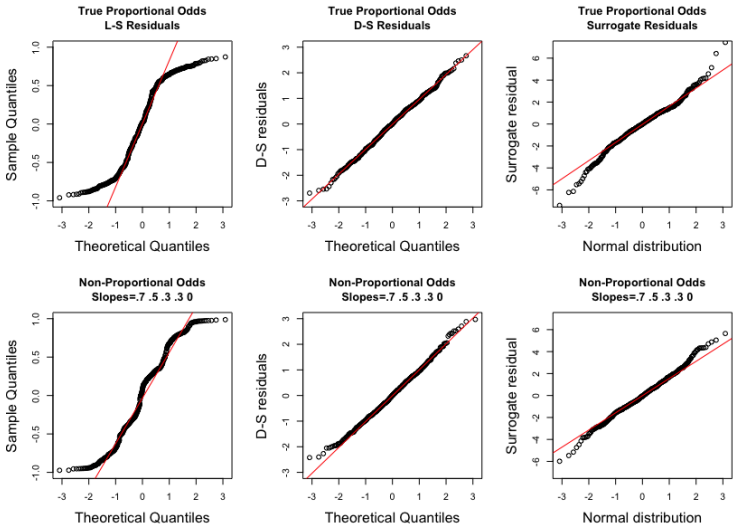

Figure D.2: Residuals for data generated from a POM with equidistant thresholds $\boldsymbol{\alpha} = (3.0, 1.5, 0, -1.5, -3.0)$ (top) and data generated from a PPOM (bottom): (a) L-S, (b) D-S, and (c) surrogate residuals.

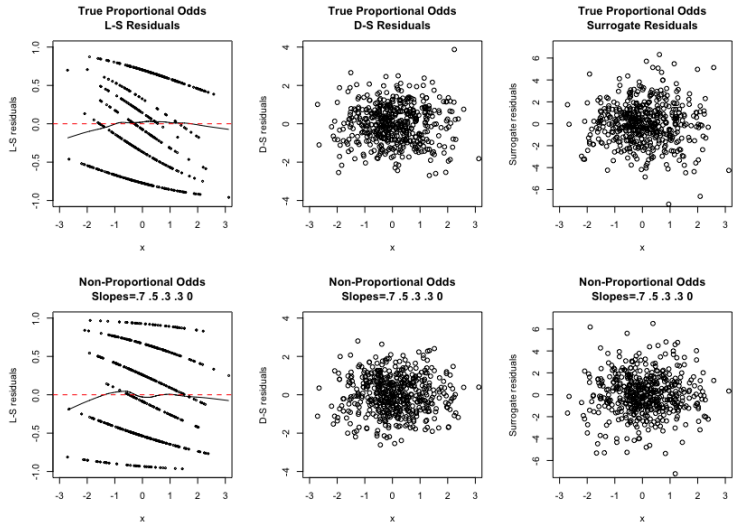

Figure D.3: Residuals vs covariate for POM for data generated from a POM (top) and data generated from a PPOM (bottom) with unconstrained thresholds $\boldsymbol{\alpha} = (-2.5, -1.0, 0, 0.5, 3.0)$: (a) L-S, (b) D-S, and (c) surrogate residuals.

For data generated from **unconstrained thresholds** $\boldsymbol{\alpha} = (-2.5, -1.0, 0, 0.5, 3.0)$, we

cannot interpret the scatterplots for L-S residuals either given the discrete bands, while both D-S and surrogate residuals show random scatters as we would have expected for a good fit (which is not the case; Figure D.3).

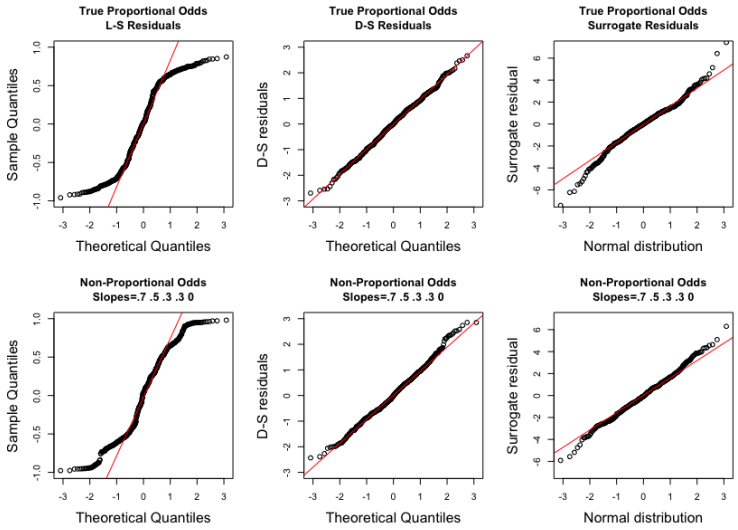

Figure D.4: Q-Q plots for POM residuals for data generated from a POM (top) and data generated from a PPOM (bottom) with unconstrained thresholds $\boldsymbol{\alpha} = (-2.5, -1.0, 0, 0.5, 3.0)$: (a) L-S, (b) D-S, and (c) surrogate residuals.

D.2 Scenario 2: Link misspecification

D.2.1 Equidistant thresholds example 1

D.2.1.1 D-S residuals

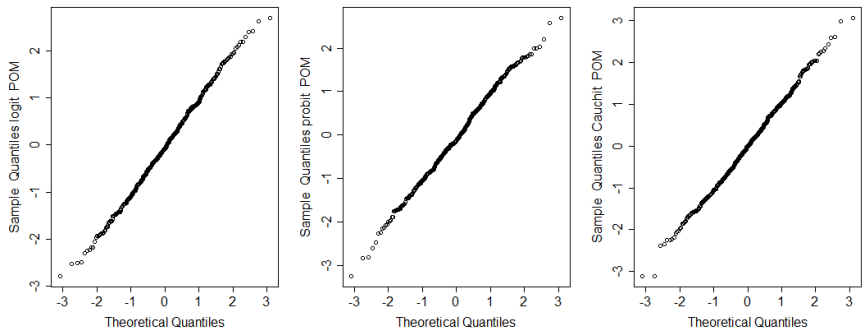

Figure D.5: Q-Q plots of the D-S residuals of POMs with: (a) logit, (b) probit, and (c) Cauchit link functions, with equidistant thresholds $\boldsymbol{\alpha} = (-16, -12, -8, -4)$ for data generated from a Cauchy random distribution.

For data generated with equidistant thresholds $\boldsymbol{\alpha} = (-16, -12, -8, -4)$, we find the following Q-Q plots for the D-S residuals of the logit, probit and Cauchit POMs, that do not show lack of normality for any of the link functions (see Figure D.5).

We compare the *ecdf* of the p-values under the null of Shapiro-Wilk tests for D-S residuals of cumulative link models with logit, probit and Cauchit link functions for the data generated from a **Cauchy** distribution (see Figure 5.11) and they all show normality. The test is in fact statistically significant at the 5% significance level for 97.4% of the 10,000 logit models and the probit models, while the percentage is only slightly higher (97.5%) for the Cauchit models.

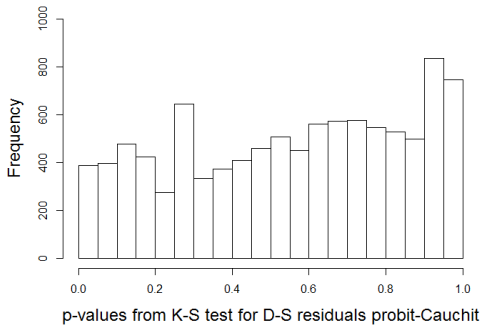

Figure D.6: p-values from Kolmogorov-Smirnov test of the two samples of D-S residuals for cumulative probit and Cauchit models of data derived from Cauchy distribution with equidistant thresholds $\boldsymbol{\alpha} = (-16, -12, -8, -4)$.

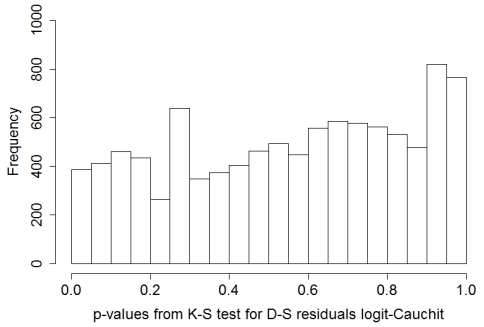

Figure D.7: p-values from Kolmogorov-Smirnov test of the two samples of D-S residuals for cumulative logit and Cauchit models of data derived from Cauchy distribution with equidistant thresholds $\boldsymbol{\alpha} = (-16, -12, -8, -4)$.

We find that the p-values for the K-S test of the two samples of D-S residuals (**probit and**

Cauchit) for data generated from a Cauchy distribution follow an approximately uniform distribution as shown in Figure D.6, therefore we cannot state that the D-S residuals of the probit models are reflecting the misspecification. Numerically, we find that 96.1% of the p-values are statistically significant at 5% significance level, which means that we cannot reject that there is a difference in distribution of the residuals for most of these probit and Cauchit models.

Similarly for the link function alternatives **logit and Cauchit**, we also obtain an approximately uniform distribution of the p-values as expected (see Figure D.7), where a 96.1% of the p-values are statistically significant at 5% significance level.

D.2.1.2 L-S residuals

The three *ecdf* plots corresponding to the three link specifications (see Figure D.8) show misspecification which reflects the fact that L-S residuals even fail to recognise a good model fit, which in this case would be Cauchit. The Shapiro-Wilk test is not statistically significant at the 5% significance level for any of the 10,000 logit, probit and Cauchit models.

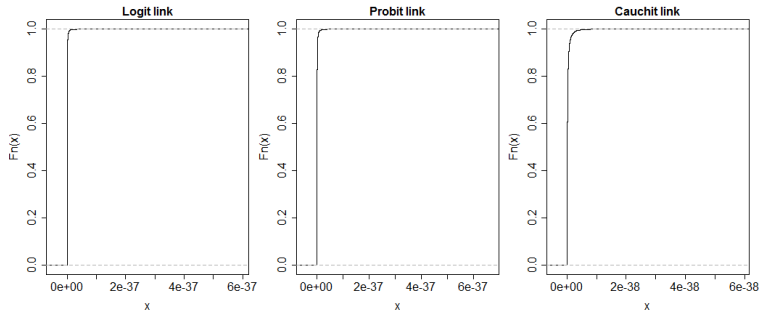

Figure D.8: *ecdf* of p-values from Shapiro-Wilk tests for L-S residuals of cumulative link models with: (a) logit, (b) probit, and (c) Cauchit link functions, with equidistant thresholds $\boldsymbol{\alpha} = (-16, -12, -8, -4)$ for data generated from a Cauchy random distribution.

The p-values from the K-S test for the comparison **probit and Cauchit** display a very skewed non-uniform pattern (see Figure D.9) reflecting issues from using these residuals for ordinal models. 8.1% of the p-values are statistically significant at 5% significance level.

Similarly, the p-values from the comparison **logit and Cauchit** can be seen in Figure D.10 showing a very high frequency for values close to 0. 9.3% of the p-values are statistically significant at 5% significance level.

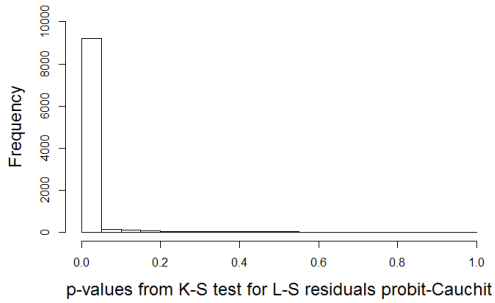

Figure D.9: p-values from Kolmogorov-Smirnov test of the two samples of L-S residuals for cumulative probit and Cauchit models of data derived from Cauchy distribution equidistant thresholds $\boldsymbol{\alpha} = (-16, -12, -8, -4)$.

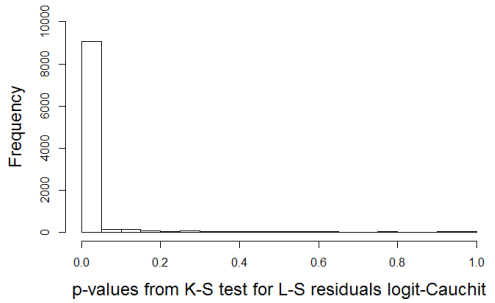

Figure D.10: p-values from Kolmogorov-Smirnov test of the two samples of L-S residuals for cumulative logit and Cauchit models of data derived from Cauchy distribution equidistant thresholds $\boldsymbol{\alpha} = (-16, -12, -8, -4)$.

D.2.1.3 Surrogate residuals

For surrogate residuals and equidistant thresholds example 1 with $\boldsymbol{\alpha} = (-16, -12, -8, -4)$, we find that, by comparing the *ecdf* of the p-values resulting from Shapiro-Wilk tests for surrogate residuals of cumulative link models with logit, probit and Cauchit link functions for the data generated from a Cauchy distribution (see Figure D.11), the *ecdf* of the p-values for the Shapiro-Wilk test of the surrogate residuals for the logit and Cauchit models are not accurate whereas those for the probit model seem to be the only ones for which the distributional assumptions hold, which is not consistent with the fact that the Cauchit model is the accurate specification, although it could be due to the fact that the generated model incorporates a quadratic term which might be confounding the issue. The test is statistically significant at the 5% significance level for 16.3% of the 10,000 logit models, 97.6% of the probit models, and 0%)for the Cauchit models.

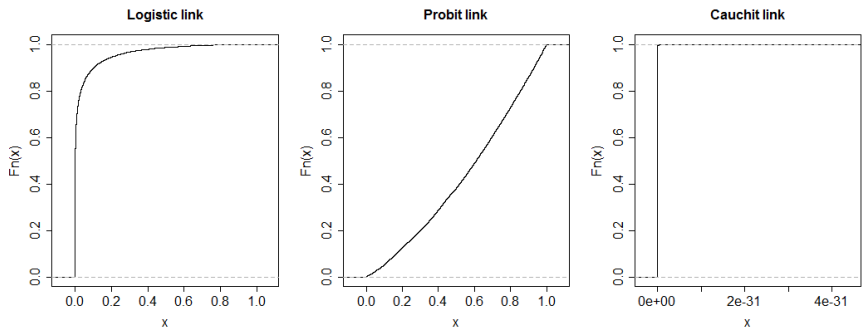

Figure D.11: *ecdf* of p-values from Shapiro-Wilk tests for surrogate residuals of cumulative link models with: (a) Logit, (b) Probit, and (c) Cauchit link functions, with equidistant thresholds $\boldsymbol{\alpha} = (-16, -12, -8, -4)$ for data generated from a quadratic model with Cauchy random distribution.

For a set of equidistant thresholds example 1 with $\boldsymbol{\alpha} = (-16, -12, -8, -4)$, the comparison of **probit-Cauchit** models results in convergence issues. The K-S p-values distribution from the comparison **logit-Cauchit** models is not uniform which also reflects issues (see Figure D.12). 77.3% of the p-values are statistically significant at the 5% significance level.

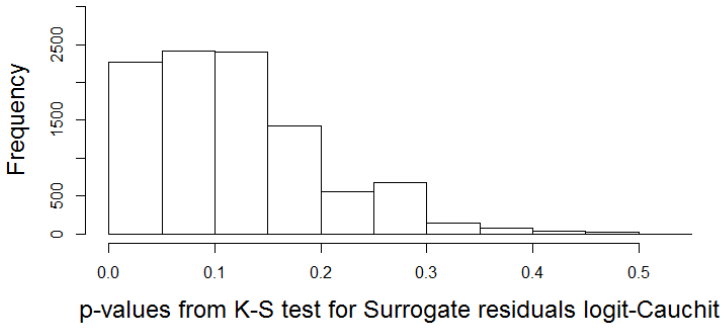

Figure D.12: p-values from Kolmogorov-Smirnov test of the two samples of surrogate residuals for cumulative logit and Cauchit models of data derived from Cauchy distribution with equidistant thresholds $\boldsymbol{\alpha} = (-16, -12, -8, -4)$.

D.2.2 Equidistant thresholds example 2

D.2.2.1 D-S residuals

For equidistant thresholds with $\boldsymbol{\alpha} = (0, 4, 8, 12)$, we find that, by comparing the *ecdf* of the p-values resulting from Shapiro-Wilk tests for D-S residuals of cumulative link models with

logit, probit and Cauchit link functions for the data generated from a Cauchy distribution (see Figure D.13), the D-S residuals for the probit and logit models are not accurate whereas the *ecdf* of the p-values for the Shapiro-Wilk test of the residuals for the Cauchit model follow a uniform distribution as we expected, and therefore are correct. The test is in fact statistically significant at the 5% significance level for 0% of the 10,000 logit models and the probit models, while the percentage is 97.1% for the Cauchit models, which reflects well the link misspecification.

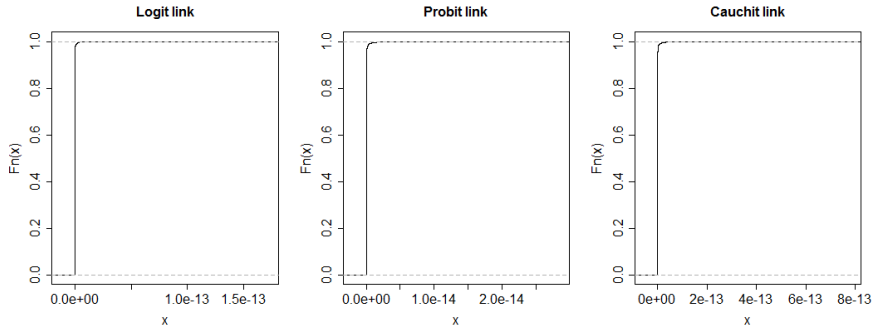

Figure D.13: *ecdf* of p-values from Shapiro-Wilk tests for D-S residuals of cumulative link models with: (a) logit, (b) probit, and (c) Cauchit link functions, with equidistant thresholds $\boldsymbol{\alpha} = (0, 4, 8, 12)$ for data generated from a Cauchy random distribution.

We find as shown in Figure D.14 a distribution which is not uniform for the p-values for the K-S test of the two samples of D-S residuals (**probit and Cauchit**) for data generated from a Cauchy distribution and we now find that 80% of the p-values are statistically significant at 5% significance level.

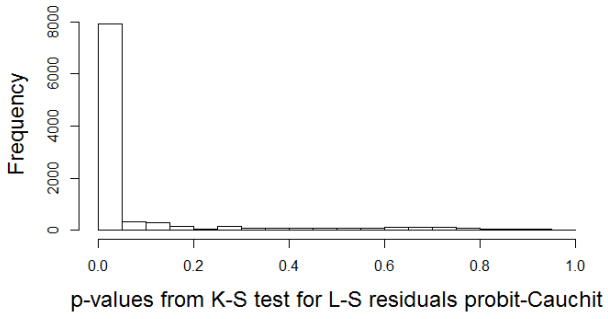

Figure D.14: p-values from Kolmogorov-Smirnov test of the two samples of D-S residuals for cumulative probit and Cauchit models of data derived from Cauchy distribution with equidistant thresholds $\boldsymbol{\alpha} = (0, 4, 8, 12)$.

For the link function alternative **logit and Cauchit**, the distribution of the p-values is closer to uniform (see Figure D.15) and this is reflected in a higher percentage of statistically significant results of the test (97%) at 5% significance level.

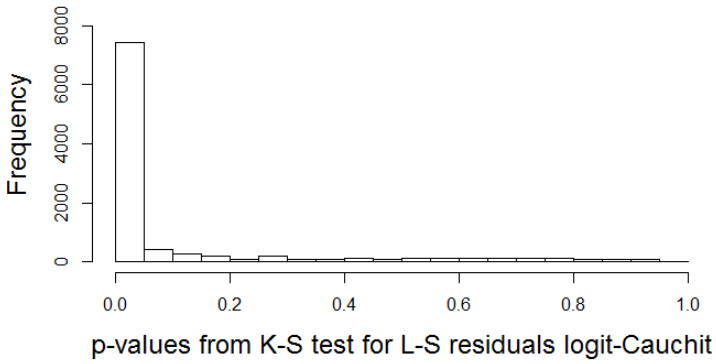

Figure D.15: p-values from Kolmogorov-Smirnov test of the two samples of D-S residuals for cumulative logit and Cauchit models of data derived from Cauchy distribution with equidistant thresholds $\boldsymbol{\alpha} = (0, 4, 8, 12)$.

D.2.2.2 L-S residuals

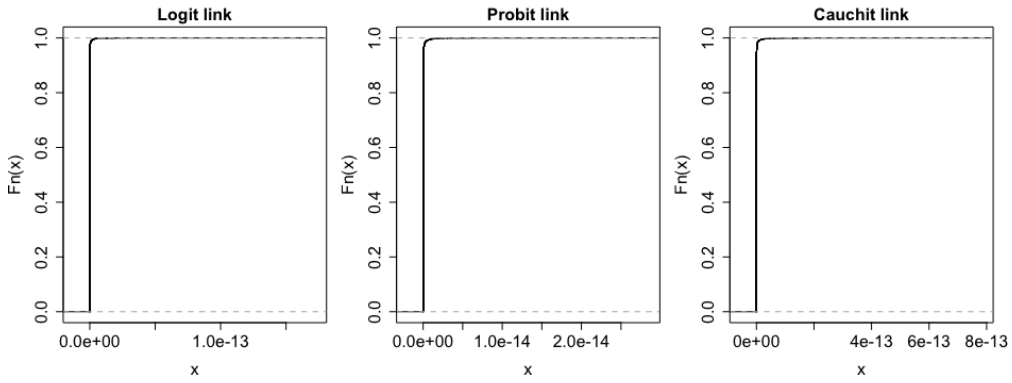

Figure D.16: *ecdf* of p-values from Shapiro-Wilk tests for L-S residuals of cumulative link models with: (a) logit, (b) probit, and (c) Cauchit link functions, with equidistant thresholds $\boldsymbol{\alpha} = (0, 4, 8, 12)$ for data generated from a Cauchy random distribution.

For the implementations of L-S residuals available for `rms` and `VGAM`, convergence issues arise when modelling the quadratic nature of the original data. Instead, we have modelled the dependence of the direct variable linearly with the independent variable.

The distributional assumptions for the three model specifications do not hold for either of the link specifications (see Figure D.16) and the Shapiro-Wilk test is not statistically significant at the 5% significance level for any of the 10,000 logit, probit and Cauchit models, thus not showing any sign of the link misspecification either. We argue this might be a consequence of L-S residuals bad performance with ordinal models.

The corresponding K-S test results for the comparison **probit-Cauchit** and **logit-Cauchit** are plotted in Figures D.17 and D.18, showing no uniform distribution consistently with the results above. 20.7% and 25.7% of the p-values are statistically significant at 5% significance level, respectively.

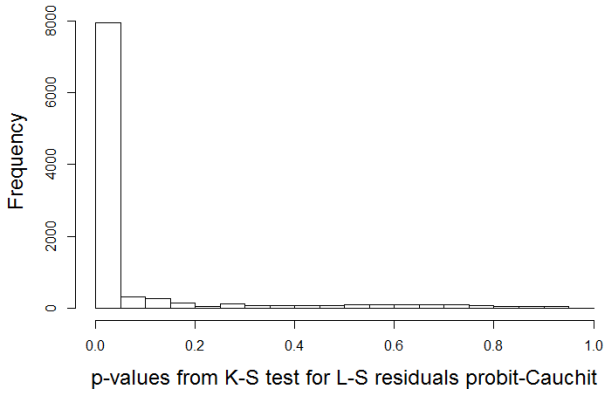

Figure D.17: p-values from Kolmogorov-Smirnov test of the two samples of L-S residuals for cumulative probit and Cauchit models of data derived from Cauchy distribution equidistant thresholds $\boldsymbol{\alpha} = (0, 4, 8, 12)$.

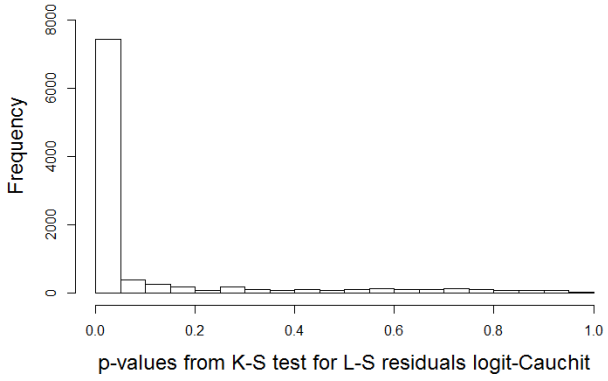

Figure D.18: p-values from Kolmogorov-Smirnov test of the two samples of L-S residuals for cumulative logit and Cauchit models of data derived from Cauchy distribution equidistant thresholds $\boldsymbol{\alpha} = (0, 4, 8, 12)$.

D.2.2.3 Surrogate residuals

For surrogate residuals and equidistant thresholds example 2 with $\boldsymbol{\alpha} = (0, 4, 8, 12)$, we find that, by comparing the *ecdf* of the p-values resulting from Shapiro-Wilk tests for surrogate residuals of cumulative link models with logit, probit and Cauchit link functions for the data generated from a Cauchy distribution (see Figure D.19), we find that the distributional assumptions do not hold for either the logit or the Cauchit model while they do for the cumulative probit model, which is not consistent with our expectations.

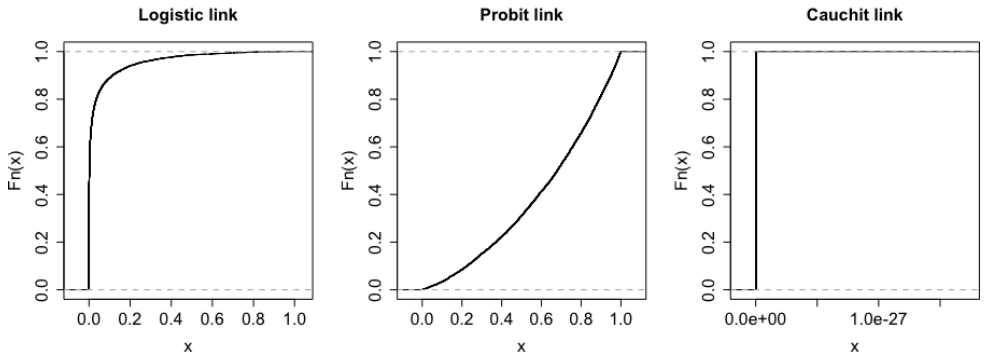

Figure D.19: *ecdf* of p-values from Shapiro-Wilk tests for surrogate residuals of cumulative link models with: (a) logit, (b) probit, and (c) Cauchit link functions, with equidistant thresholds $\boldsymbol{\alpha} = (0, 4, 8, 12)$ for data generated from a quadratic model with Cauchy random distribution.

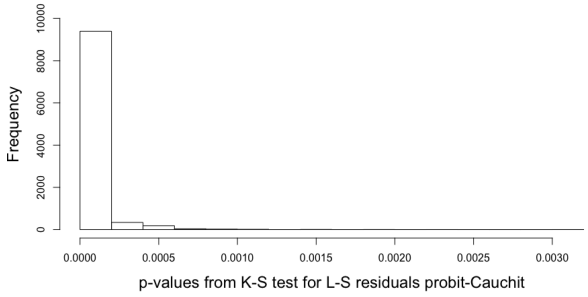

Figure D.20: p-values from Kolmogorov-Smirnov test of the two samples of surrogate residuals for cumulative probit and Cauchit models of data derived from Cauchy distribution with equidistant thresholds $\boldsymbol{\alpha} = (0, 4, 8, 12)$.

For a set of equidistant thresholds example 2 with $\boldsymbol{\alpha} = (0, 4, 8, 12)$, the K-S p-values from the comparison **probit-Cauchit** models are clearly not uniform as we expected (see Figure D.20).

For a set of equidistant thresholds example 2 with $\boldsymbol{\alpha} = (0, 4, 8, 12)$, the K-S p-values distribution from the comparison **logit-Cauchit** models is not uniform (see Figure D.21).

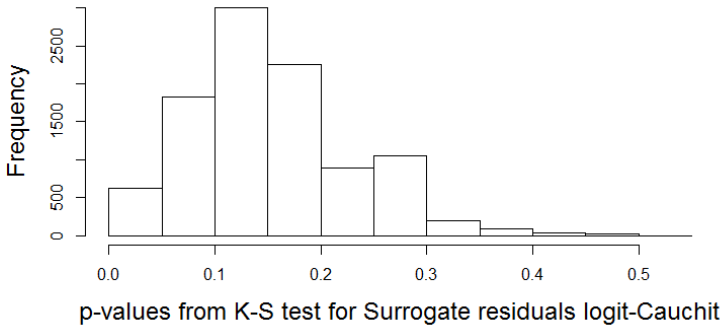

Figure D.21: *ecdf* of p-values from Shapiro-Wilk tests for D-S residuals of cumulative link models with: (a) Logit, (b) Probit, and (c) Cauchit link functions, with unconstrained thresholds $\boldsymbol{\alpha} = (-1.5, 0, 1, 3)$ for data generated from a quadratic model with Cauchy random distribution.

D.2.3 Unconstrained thresholds

D.2.3.1 D-S residuals

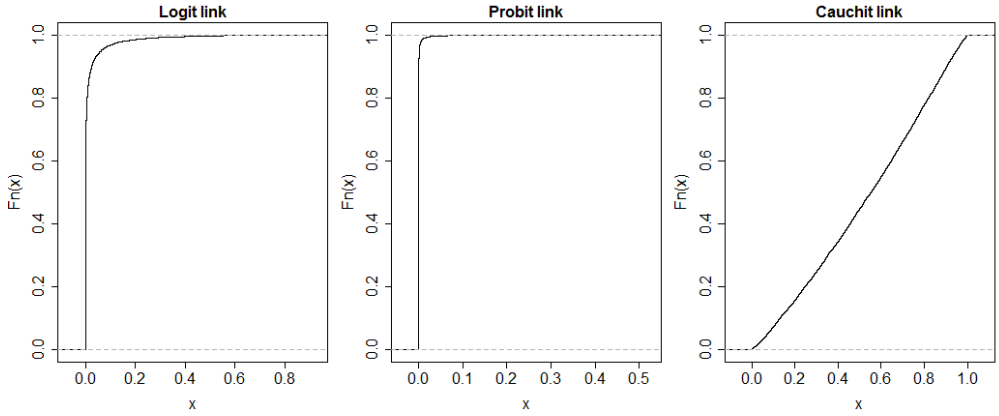

Figure D.22: *ecdf* of p-values from Shapiro-Wilk tests for D-S residuals of cumulative link models with: (a) logit, (b) probit, and (c) Cauchit link functions, with unconstrained thresholds $\boldsymbol{\alpha} = (-1.5, 0, 1, 3)$ for data generated from a quadratic model with Cauchy random distribution.

For unconstrained thresholds $\boldsymbol{\alpha} = (-1.5, 0, 1, 3)$, we find that, by comparing the *ecdf* of the p-values resulting from Shapiro-Wilk (S-W) tests for D-S residuals of cumulative link

models with logit, probit and Cauchit link functions for the data generated from a Cauchy distribution (see Figure D.22), the D-S residuals for the probit and logit models are not accurate whereas the *ecdf* of the p-values for the S-W test of the residuals for the Cauchit model follow a uniform distribution as we expected, and therefore are correct.

The S-W test is in fact statistically significant at the 5% significance level for 0% of the 10,000 logit models and the probit models, while the percentage is 96.8% for the Cauchit models which means the link misspecification is accurately reflected in the D-S residuals for this threshold structure.

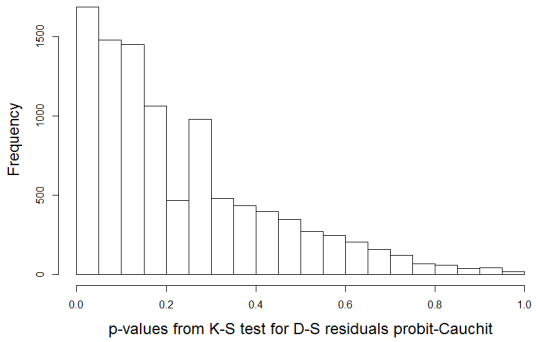

Figure D.23: p-values from Kolmogorov-Smirnov test of the two samples of D-S residuals for cumulative probit and Cauchit models of data derived from Cauchy distribution with unconstrained thresholds $\boldsymbol{\alpha} = (-1.5, 0, 1, 3)$.

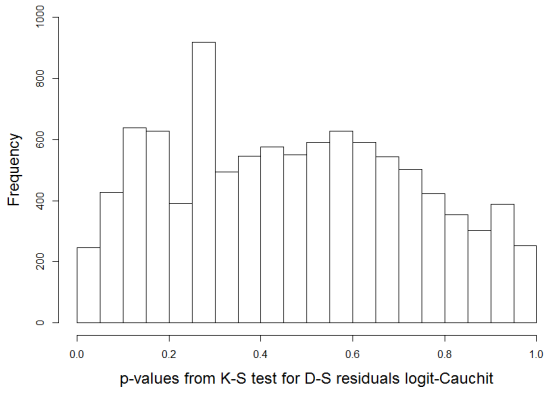

Figure D.24: p-values from Kolmogorov-Smirnov test of the two samples of D-S residuals for cumulative logit and Cauchit models of data derived from Cauchy distribution with unconstrained thresholds $\boldsymbol{\alpha} = (-1.5, 0, 1, 3)$.

For a set of unconstrained thresholds $\boldsymbol{\alpha} = (-1.5, 0, 1, 3)$, the K-S p-values from the

comparison **probit-Cauchit** models are clearly not uniform (see Figure D.23) and 83.2% of the p-values are statistically significant at 5% significance level.

For a set of unconstrained thresholds $\boldsymbol{\alpha} = (-1.5, 0, 1, 3)$, the K-S p-values' distribution from the comparison of **logit-Cauchit** models is approximately uniform and 97.5% of the p-values are statistically significant at the 5% significance level, which is consistent with our expectations (see Figure D.24).

D.2.3.2 L-S residuals

For L-S residuals and unconstrained thresholds $\boldsymbol{\alpha} = (-1.5, 0, 1, 3)$, we find that, by comparing the *ecdf* of the p-values resulting from Shapiro-Wilk tests for L-S residuals of cumulative link models with logit, probit and Cauchit link functions for the data generated from a Cauchy distribution (see Figure D.25), the D-S residuals for the probit and logit models are not accurate whereas the Cauchit model's *ecdf* do not follow a uniform distribution either and therefore still reflect a misspecification that does not exist. The test is not statistically significant at the 5% significance level for any of the 10,000 logit, probit and Cauchit models, thus not showing any sign of the link misspecification either. We argue this is a consequence of the L-S residuals not working very well for ordinal models, independently of the threshold structure and link function.

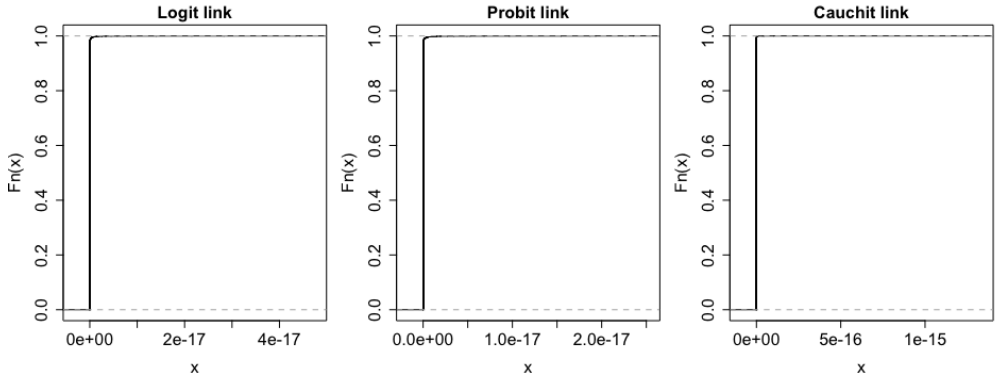

Figure D.25: *ecdf* of p-values from Shapiro-Wilk tests for L-S residuals of cumulative link models with: (a) logit, (b) probit, and (c) Cauchit link functions, with unconstrained thresholds $\boldsymbol{\alpha} = (-1.5, 0, 1, 3)$ for data generated from a quadratic model with Cauchy random distribution.

For a set of unconstrained thresholds $\boldsymbol{\alpha} = (-1.5, 0, 1, 3)$, the K-S p-values from the comparison **probit-Cauchit** models are clearly not uniform (see Figure D.26). 22.7% of the p-values are statistically significant at 5% significance level.

For a set of unconstrained thresholds $\boldsymbol{\alpha} = (-1.5, 0, 1, 3)$, the K-S p-values distribution from the comparison **logit-Cauchit** models is not uniform (see Figure D.27). 28.2% of the p-values are statistically significant at 5% significance level.

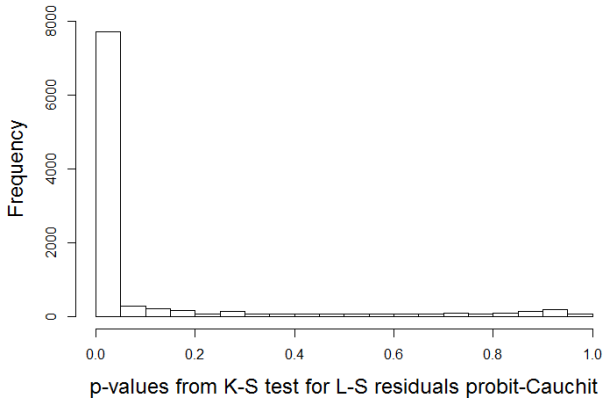

Figure D.26: p-values from Kolmogorov-Smirnov test of the two samples of L-S residuals for cumulative probit and Cauchit models of data derived from Cauchy distribution with unconstrained thresholds $\boldsymbol{\alpha} = (-1.5, 0, 1, 3)$.

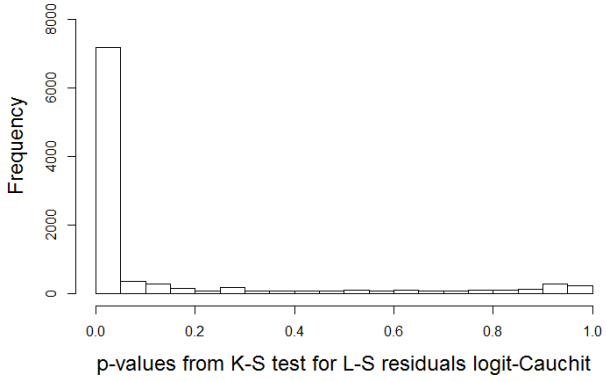

Figure D.27: p-values from Kolmogorov-Smirnov test of the two samples of L-S residuals for cumulative logit and Cauchit models of data derived from Cauchy distribution with unconstrained thresholds $\boldsymbol{\alpha} = (-1.5, 0, 1, 3)$.

D.2.3.3 Surrogate residuals

When we check distributional assumptions for surrogate residuals and unconstrained thresholds $\boldsymbol{\alpha} = (-1.5, 0, 1, 3)$, we find by comparing the *ecdf* of the p-values resulting from Shapiro-Wilk tests for surrogate residuals of cumulative link models with logit, probit and Cauchit

link functions for the data generated from a Cauchy distribution (see Figure D.28), that the D-S residuals for the probit and logit models reflect the misspecification whereas the *ecdf* of the p-values for the Shapiro-Wilk test of the residuals for the Cauchit model do not, as we expected. The test is statistically significant at the 5% significance level for 18.2% of the CLMs, 0% of the Cauchit models and 97.7% of the probit models.

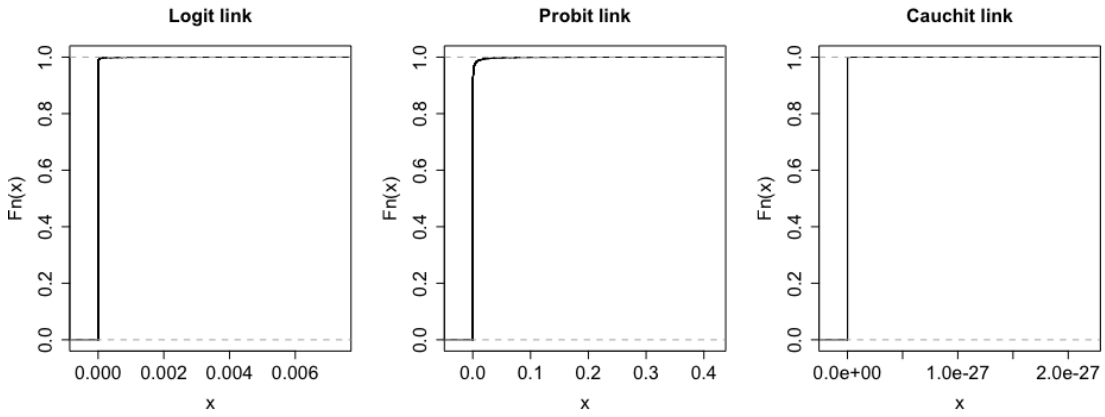

Figure D.28: *ecdf* of p-values from Shapiro-Wilk tests for surrogate residuals of cumulative link models with: (a) logit, (b) probit, and (c) Cauchit link functions, with unconstrained thresholds $\boldsymbol{\alpha} = (-1.5, 0, 1, 3)$ for data generated from a quadratic model with Cauchy random distribution.

For a set of unconstrained thresholds $\boldsymbol{\alpha} = (-1.5, 0, 1, 3)$, the K-S p-values from the comparison **probit-Cauchit** models are clearly not uniform (see Figure D.29). 0% of the p-values are statistically significant at 5% significance level.

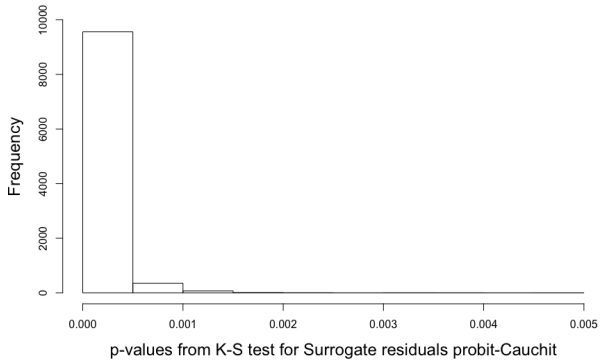

Figure D.29: p-values from Kolmogorov-Smirnov test of the two samples of surrogate residuals for cumulative probit and Cauchit models of data derived from Cauchy distribution with unconstrained thresholds $\boldsymbol{\alpha} = (-1.5, 0, 1, 3)$.

For a set of unconstrained thresholds $\boldsymbol{\alpha} = (-1.5, 0, 1, 3)$, the K-S p-values distribution from the comparison **logit-Cauchit** models is closer to uniform which is consistent with our expectations (see Figure D.30). This is reflected in a 90.3% of the p-values being statistically significant at 5% significance level.

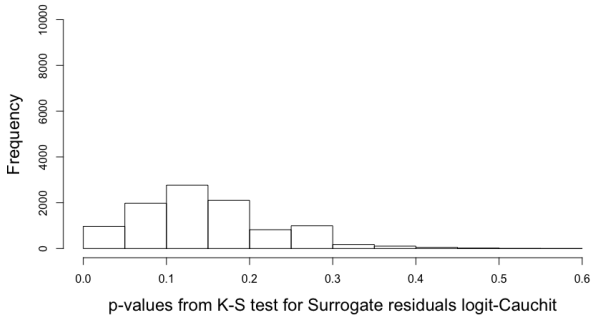

Figure D.30: p-values from Kolmogorov-Smirnov test of the two samples of surrogate residuals for cumulative logit and Cauchit models of data derived from Cauchy distribution with unconstrained thresholds $\boldsymbol{\alpha} = (-1.5, 0, 1, 3)$.

D.2.4 Symmetric thresholds

D.2.4.1 D-S residuals

For symmetric thresholds $\boldsymbol{\alpha} = (-36, -6, 34, 64)$, we find that, by comparing the *ecdf* of the p-values resulting from Shapiro-Wilk tests for D-S residuals of cumulative link models with logit, probit and Cauchit link functions for the data generated from a Cauchy distribution (see Figure D.31), the D-S residuals for the probit and logit models do not show the misspecification and show the same distribution as the *ecdf* of the p-values for the Shapiro-Wilk test of the residuals for the Cauchit model. The test is statistically significant at the 5% significance level for 99% of the 10,000 logit, probit and Cauchit models, thus not showing any sign of the link misspecification either.

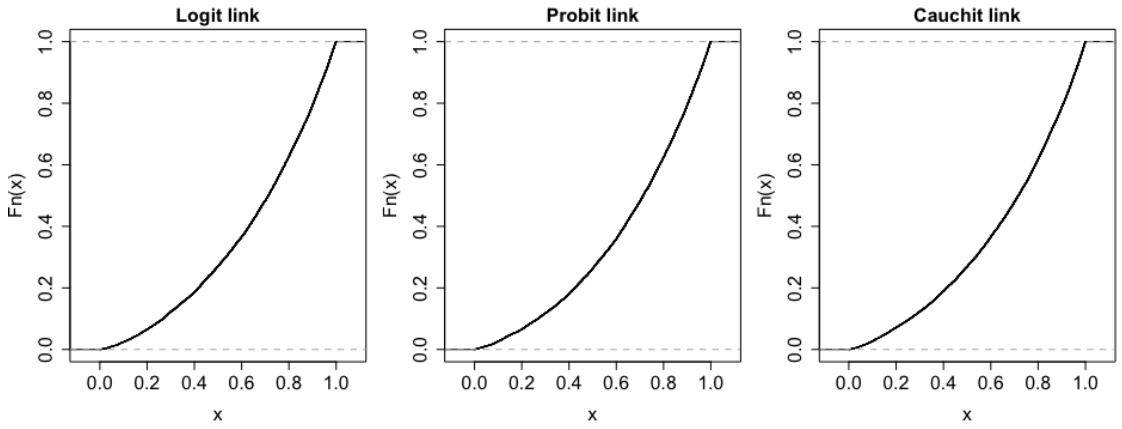

Figure D.31: *ecdf* of p-values from Shapiro-Wilk tests for D-S residuals of cumulative link models with: (a) logit, (b) probit, and (c) Cauchit link functions, with symmetric thresholds $\boldsymbol{\alpha} = (-36, -6, 34, 64)$ for data generated from a Cauchy random distribution.

For a set of symmetric thresholds $\boldsymbol{\alpha} = (-36, -6, 34, 64)$ the distribution of the K-S p-values from the comparison **probit-Cauchit** models are close to uniform and 96% of the p-values are statistically significant at 5% significance level.

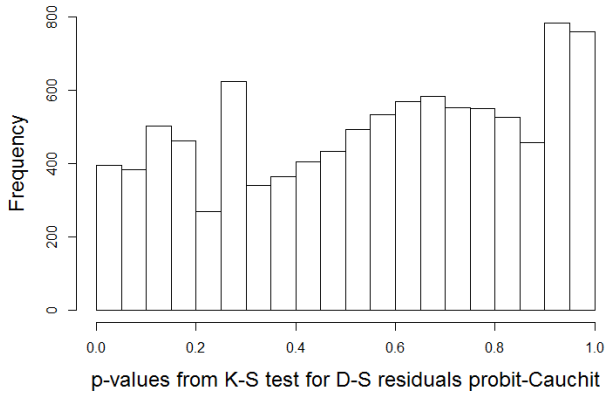

Figure D.32: p-values from Kolmogorov-Smirnov test of the two samples of D-S residuals for cumulative probit and Cauchit models of data derived from Cauchy distribution with symmetric thresholds $\boldsymbol{\alpha} = (-36, -6, 34, 64)$.

For a set of symmetric thresholds $\boldsymbol{\alpha} = (-36, -6, 34, 64)$, the K-S p-values distribution from the comparison **logit-Cauchit** models is approximately uniform (see Figure D.33) and 96% of the p-values are statistically significant at 5% significance level. Therefore, we cannot state that either of the comparisons show a clear difference between the link function misspecification and correct specification for this subscenario.

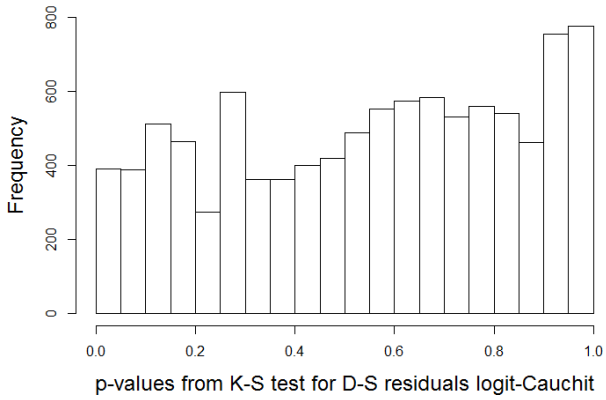

Figure D.33: p-values from Kolmogorov-Smirnov test of the two samples of D-S residuals for cumulative logit and Cauchit models of data derived from Cauchy distribution with symmetric thresholds $\boldsymbol{\alpha} = (-36, -6, 34, 64)$.

D.2.4.2 L-S residuals

For L-S residuals and symmetric thresholds $\boldsymbol{\alpha} = (-36, -6, 34, 64)$, we find that, by comparing the *ecdf* of the p-values resulting from Shapiro-Wilk tests for L-S residuals of cumulative link models with logit, probit and Cauchit link functions for the data generated from a Cauchy distribution (see Figure D.34), the L-S residuals for the logit and probit models are not accurate and the Cauchit model's *ecdf* does not follow the uniform distribution that we would expect, and therefore are not correct either. The test is not statistically significant at the 5% significance level for any of the 10,000 logit, probit and Cauchit models, thus not showing any sign of the link misspecification either.

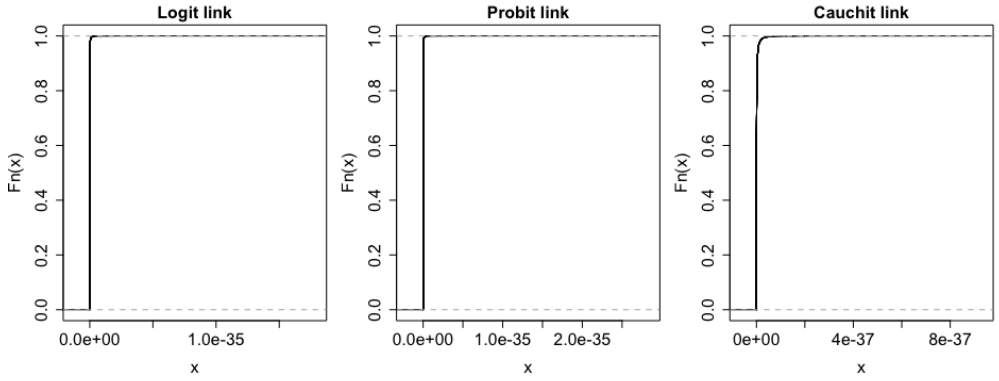

Figure D.34: *ecdf* of p-values from Shapiro-Wilk tests for L-S residuals of cumulative link models with: (a) logit, (b) probit, and (c) Cauchit link functions, with symmetric thresholds $\boldsymbol{\alpha} = (-36, -6, 34, 64)$ for data generated from a Cauchy random distribution.

For a set of symmetric thresholds $\boldsymbol{\alpha} = (-36, -6, 34, 64)$ and the K-S p-values from the comparison **probit-Cauchit** models are clearly not uniform (see Figure D.35). Only 4.3% of the p-values are statistically significant at 5% significance level.

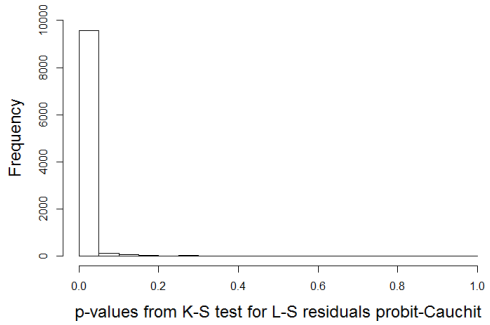

Figure D.35: p-values from Kolmogorov-Smirnov test of the two samples of L-S residuals for cumulative probit and Cauchit models of data derived from Cauchy distribution with symmetric thresholds $\boldsymbol{\alpha} = (-36, -6, 34, 64)$.

For a set of symmetric thresholds $\boldsymbol{\alpha} = (-36, -6, 34, 64)$, the K-S p-values distribution from the comparison **logit-Cauchit** models is again not uniform (see Figure D.36). Only 5.7% of the p-values are statistically significant at 5% significance level.

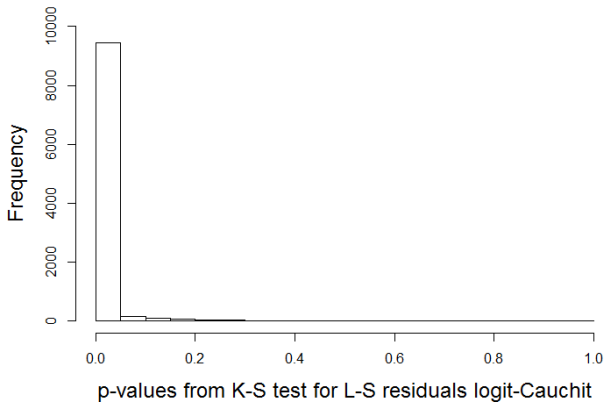

Figure D.36: p-values from Kolmogorov-Smirnov test of the two samples of L-S residuals for cumulative logit and Cauchit models of data derived from Cauchy distribution with symmetric thresholds $\boldsymbol{\alpha} = (-36, -6, 34, 64)$.

D.2.4.3 Surrogate residuals

For surrogate residuals we get very different patterns for the three link functions.

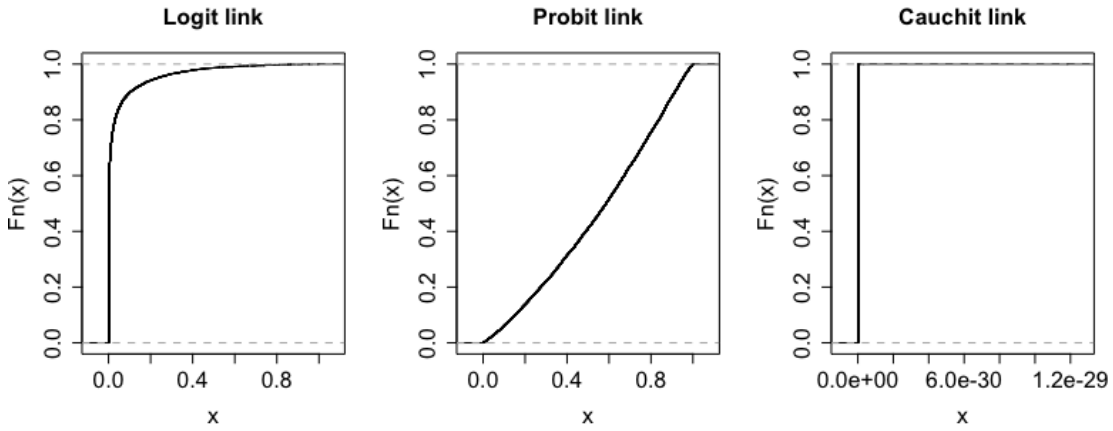

Figure D.37: *ecdf* of p-values from Shapiro-Wilk tests for surrogate residuals of cumulative link models with: (a) logit, (b) probit, and (c) Cauchit link functions, with symmetric thresholds $\boldsymbol{\alpha} = (-36, -6, 34, 64)$ for data generated from a Cauchy random distribution.

For a set of symmetric thresholds $\boldsymbol{\alpha} = (-36, -6, 34, 64)$ and the K-S p-values from the comparison **probit-Cauchit** models are clearly not uniform.

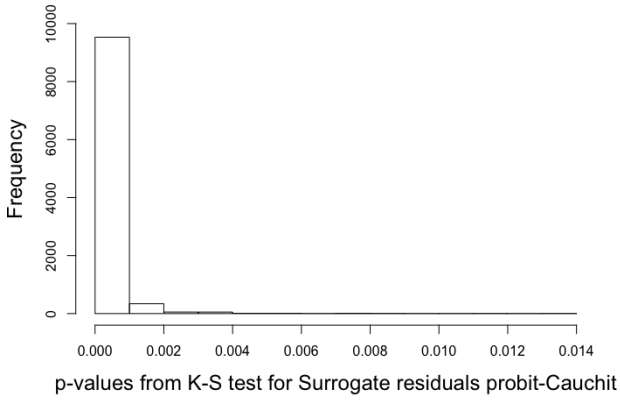

Figure D.38: p-values from Kolmogorov-Smirnov test of the two samples of surrogate residuals for cumulative probit and Cauchit models of data derived from Cauchy distribution with symmetric thresholds $\boldsymbol{\alpha} = (-36, -6, 34, 64)$.

For a set of symmetric thresholds $\boldsymbol{\alpha} = (-36, -6, 34, 64)$, the K-S p-values distribution from the comparison **logit-Cauchit** models is not uniform which means there are issues in the subscenario (see Figure D.39).

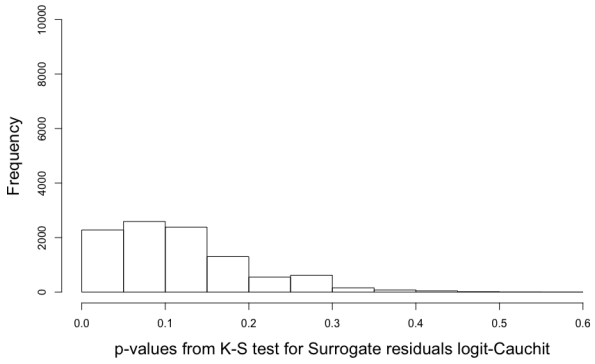

Figure D.39: p-values from Kolmogorov-Smirnov test of the two samples of surrogate residuals for cumulative logit and Cauchit models of data derived from Cauchy distribution with symmetric thresholds $\boldsymbol{\alpha} = (-36, -6, 34, 64)$.

D.3 Scenario 3: Quadratic model

We force another misspecification of the mean structure, by simulating data derived from a model with a quadratic term x^2 with normal errors and assessing the differences between the correct model and a misspecified model with only a linear term on x, where $x \sim U(1, 7)$ and $n = 500$. To this end, we generate our ordered response variable Y from a latent continuous variable $Y^* = 16 - 8x + x^2$ of the form shown in Figure D.40 (a). We assess the differences between the correct model (quadratic and probit link function) and the misspecified models (non-quadratic with probit and logit link functions alternatively). In order to summarise the different comparisons, we consider three levels; generating model, fitted model, and a third fitted model for the residuals, and consider only 5 combinations of special interest, where 'quadratic' refers to an ordinal model including both the linear (x) and the quadratic term (x^2), and 'linear' refers to a model with the linear term only (x):

Table D.1: Subscenarios for the quadratic model scenario.

	Generating model	Fitted model	Residuals fitted model
1	Quadratic	Quadratic	Quadratic
2	Quadratic	Quadratic	Linear
3	Linear	Linear	Linear
4	Quadratic	Linear	Quadratic
5	Quadratic	Linear	Linear

For the implementations of **L-S residuals** available for `rms` and `VGAM`, convergence issues arise when modelling the quadratic nature of the original data. Instead, we have modelled the dependence of the direct variable linearly with the independent variable. We therefore acknowledge that there will be some confounding of the link and quadratic misspecifications.

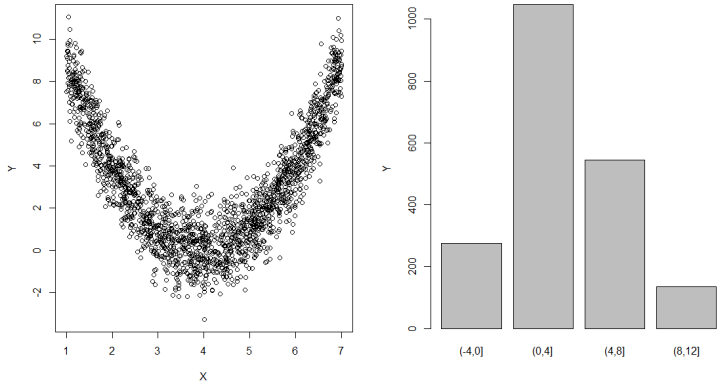

Figure D.40: Quadratic model (a) and ordinal response variable (b).

We show here the complete simulation results for **equidistant thresholds** (also **symmetric**). We consider our ordinal variable with categories defined by the cut-points $\boldsymbol{\alpha} = (0, 4, 8)$ as shown in Figure D.40 (b). This variable is generated from a cumulative probit model. We fit the relationship between this ordinal response variable and a covariate x via two cumulative models with logit and probit link functions respectively. The probit function is the inverse of the *cdf* of the standard normal distribution, so we would expect that the errors in the corresponding model would be down to missing the quadratic pattern in the data; while the model with the logit link function should present lack of fit issues both due to not describing the quadratic relationship and using the wrong link function.

As an initial step, we assess graphically whether the residuals are accurately reflecting the correct model specification. We do so by plotting the residuals against the covariate x for the correct quadratic cumulative probit model (see Figure D.41). For both D-S (d-f) and surrogate residuals (g-i) the random scatter in the plots versus the covariate and the Q-Q plot perfect diagonal fit show accurately the correct model specification and distributional assumption. However, L-S residuals (a-c) fail to reflect these. The first two plots show discrete patterns in the form of 4 quadratic curves, and the Q-Q plot shows departures from the expected uniform distribution $U(-1, 1)$ with heavy tails.

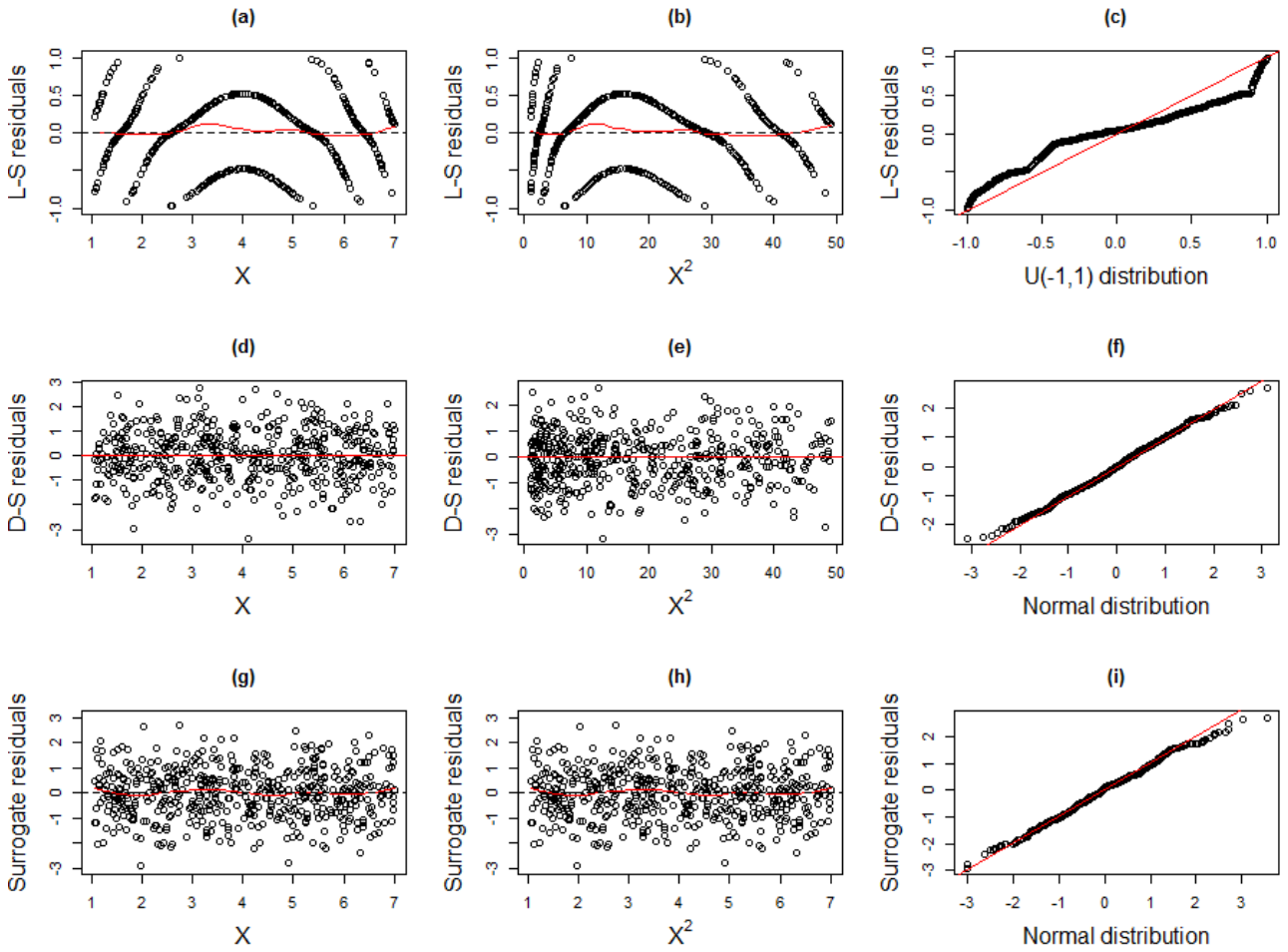

Figure D.41: L-S (a-c), D-S (d-f) and surrogate (g-i) residual plots for a correct quadratic cumulative probit model.

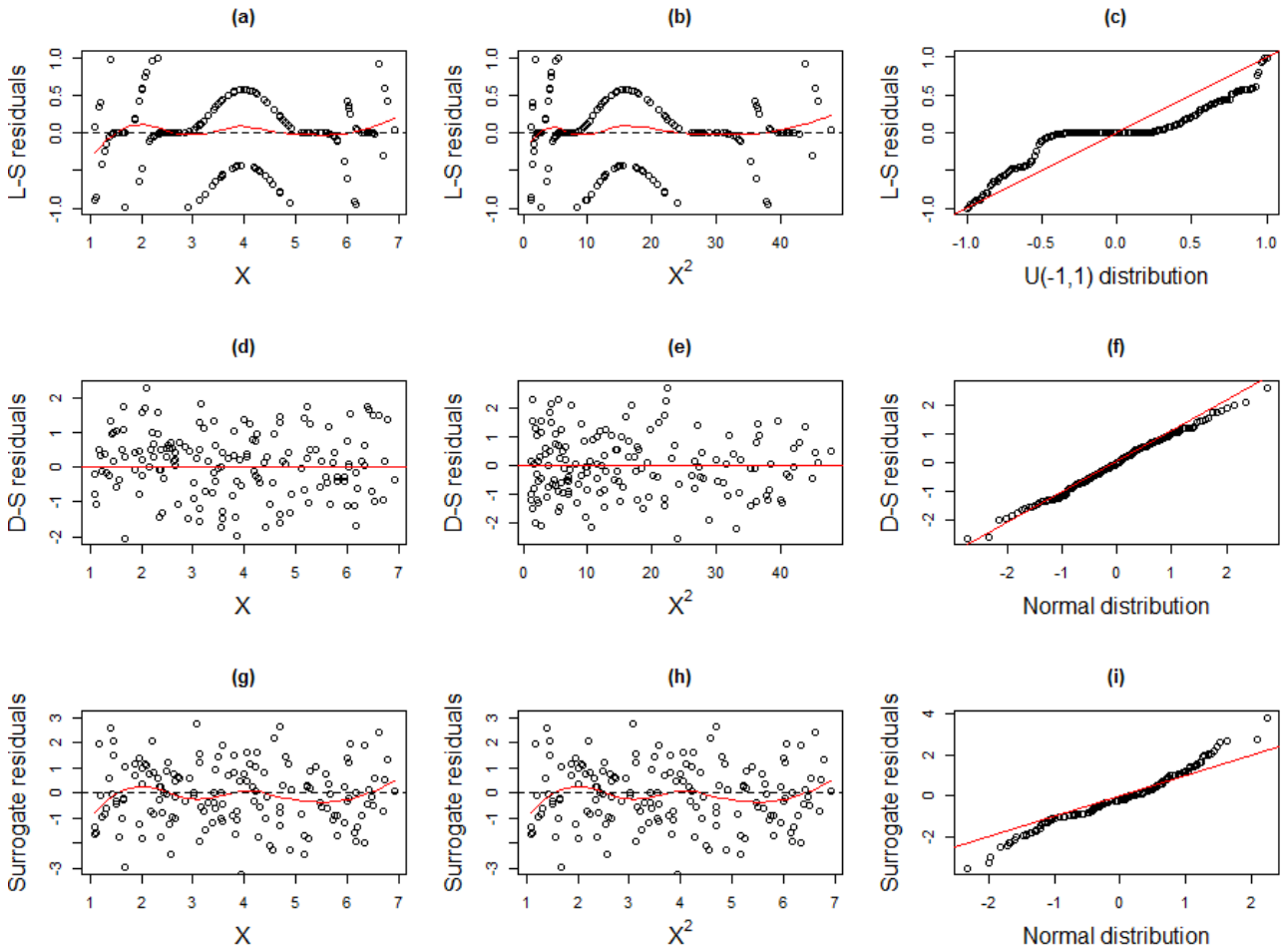

Figure D.42: L-S (a-c), D-S (d-f) and surrogate (g-i) residual plots for 'quadratic' CLM.

For the wrong quadratic CLM, we get signs of heteroscedasticity for D-S residuals as shown in Figure D.42 (e), non-normality for surrogate residuals as shown in (i), and abnormal discrete patterns for the L-S residuals as shown in (a).

1. Quadratic - Quadratic - Quadratic

The previous results only present the performance of the residuals based on one simulated data set but we have in fact simulated 10,000 data sets from the true model to ensure the reliability of the summary results. When we study the L-S residuals via a quadratic model $r = \alpha + \beta_1 x + \beta_2 x^2$, a summary of the p-values for the 10,000 runs for the coefficients β_1 and β_2 can be found in the following histograms for the L-S, D-S and surrogate residuals models (Figure D.43 for probit model and Figure D.44 for logit model). As expected, for the correct probit model smaller values (significant at the 5% significance level) are more frequent.

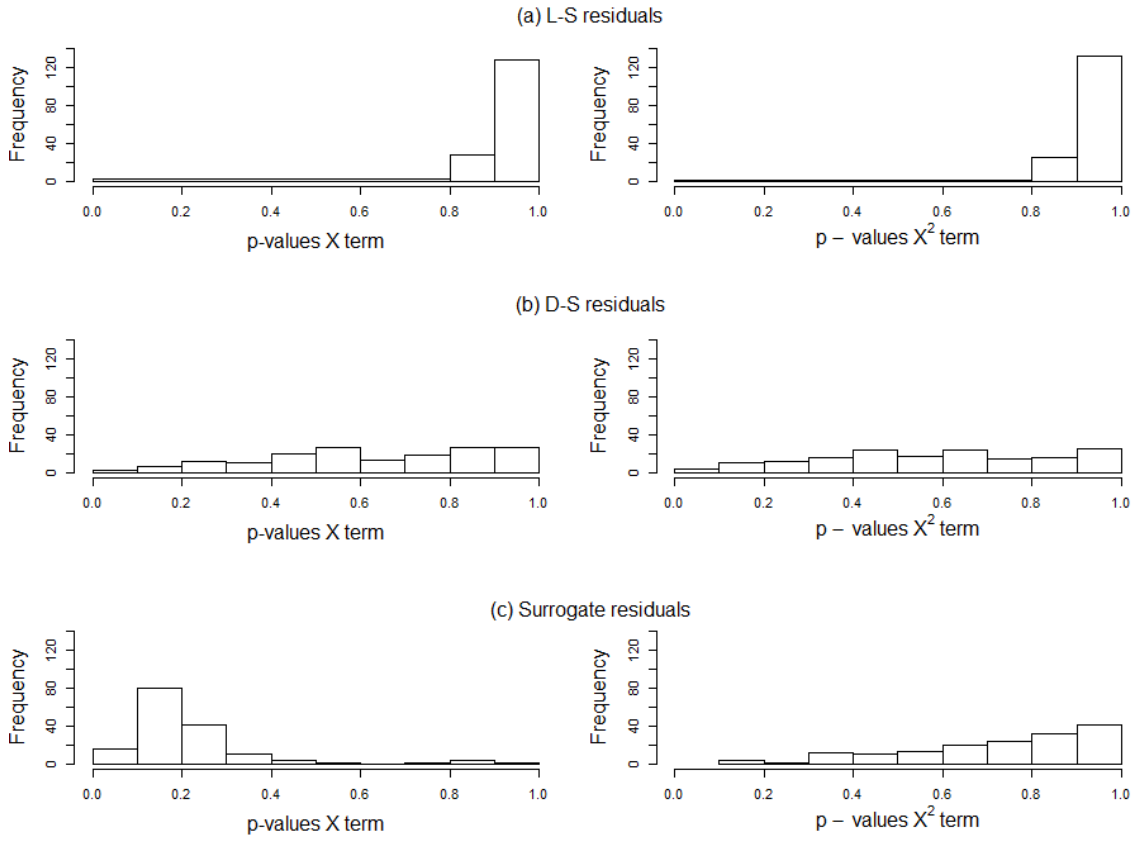

Figure D.43: p-values for the quadratic fit of the L-S, D-S, and surrogate residuals corresponding to the cumulative probit model.

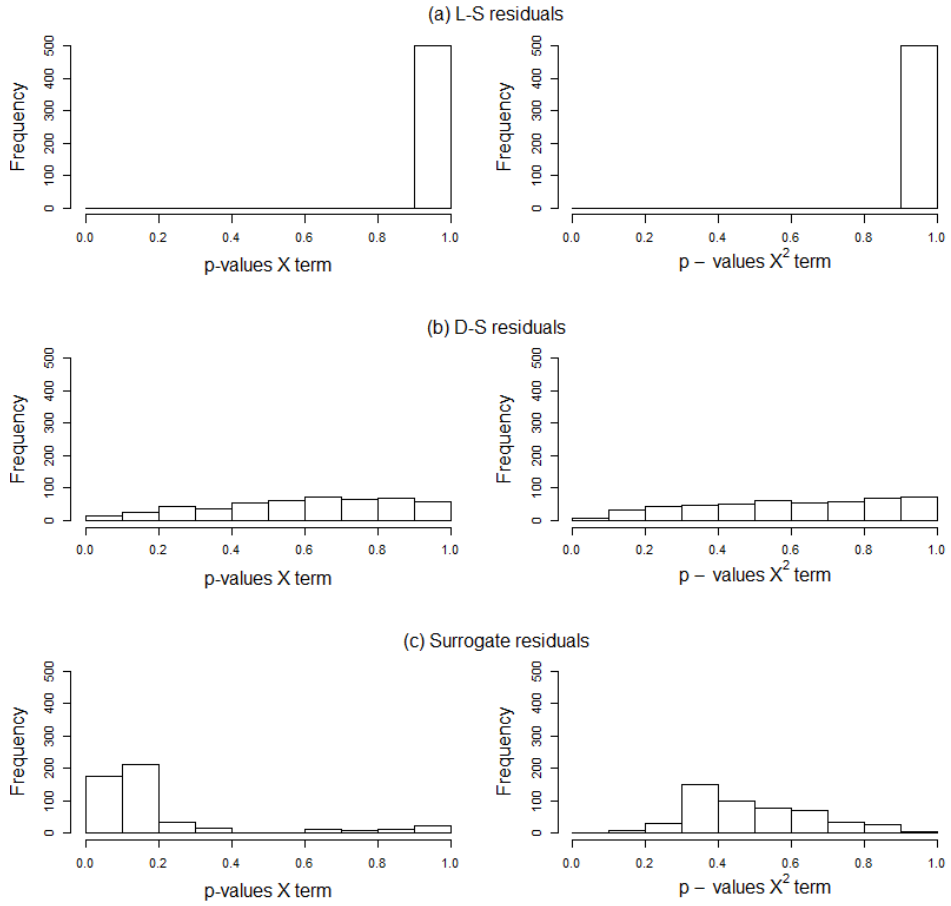

Figure D.44: p-values for the quadratic fit of the L-S, D-S, and surrogate residuals for the CLM.

2. Quadratic-Quadratic-Linear

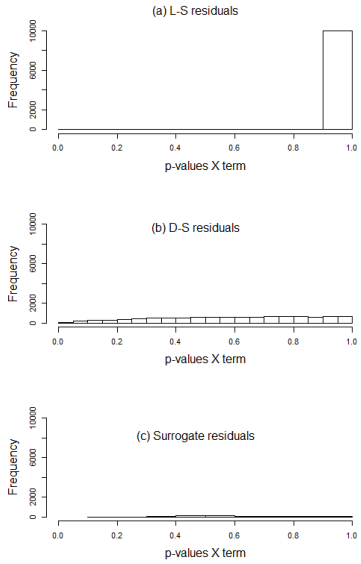

Figure D.45: p-values for the linear fit of the L-S, D-S, and surrogate residuals for the CLM.

In the *Quadratic-Quadratic-Linear subscenario* and CLM setting, we find that 0% of the p-values were lower than 0.05 for the L-S and surrogate residuals' models, and only 0.78% of the corresponding p-values for the D-S residuals were significant at that significance level.

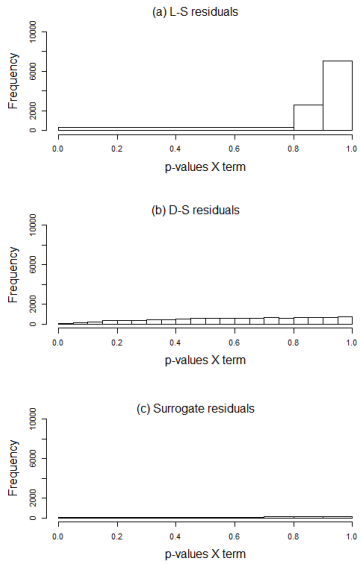

Figure D.46: p-values for the linear fit of the L-S, D-S, and surrogate residuals for the cumulative probit model.

3. Linear-Linear-Linear

On the one hand, we obtain the following residual patterns for the 'linear' ordinal model fitted to the data generated from a 'linear' CLM where only the D-S residuals reflect accurately the right modelling.

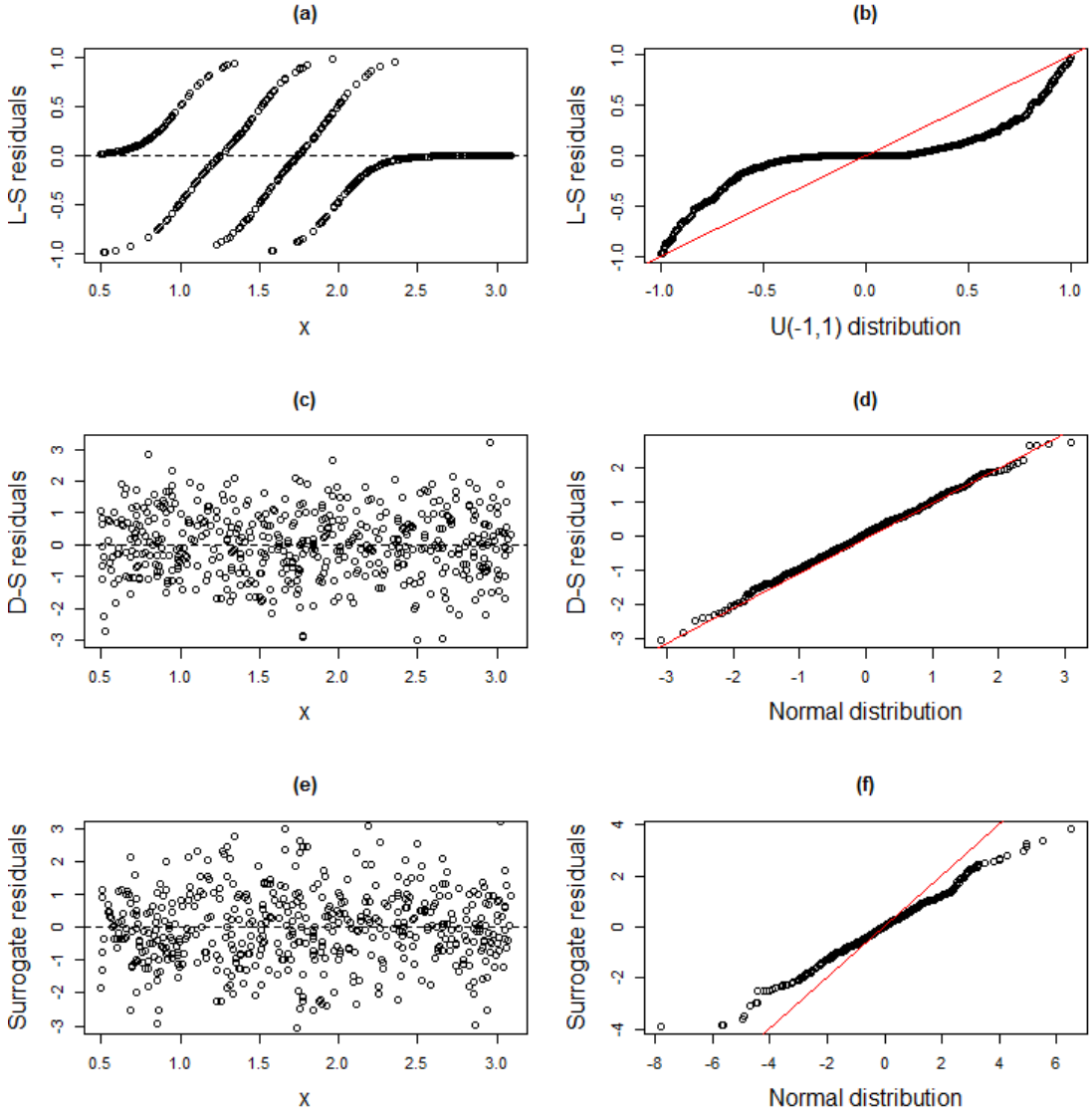

Figure D.47: L-S (a-b), D-S (c-d) and surrogate (e-f) residual plots for the 'linear' cumulative logit model.

On the other hand, for the wrong 'linear' model fitted to the data generated from a 'quadratic' model, we can notice different patterns for the cumulative probit model fit (see D.48) and the CLM (see D.49). In particular, the residuals vs covariate plots for both D-S

and surrogate residuals reflect the quadratic pattern that has not been specified in the model. L-S residuals show again discrete patterns that avoid a proper interpretation, most noticeable in the Q-Q plot. The link misspecification of the CLM is only particularly noticeable in the Q-Q plot for the surrogate residuals in D.49 (i).

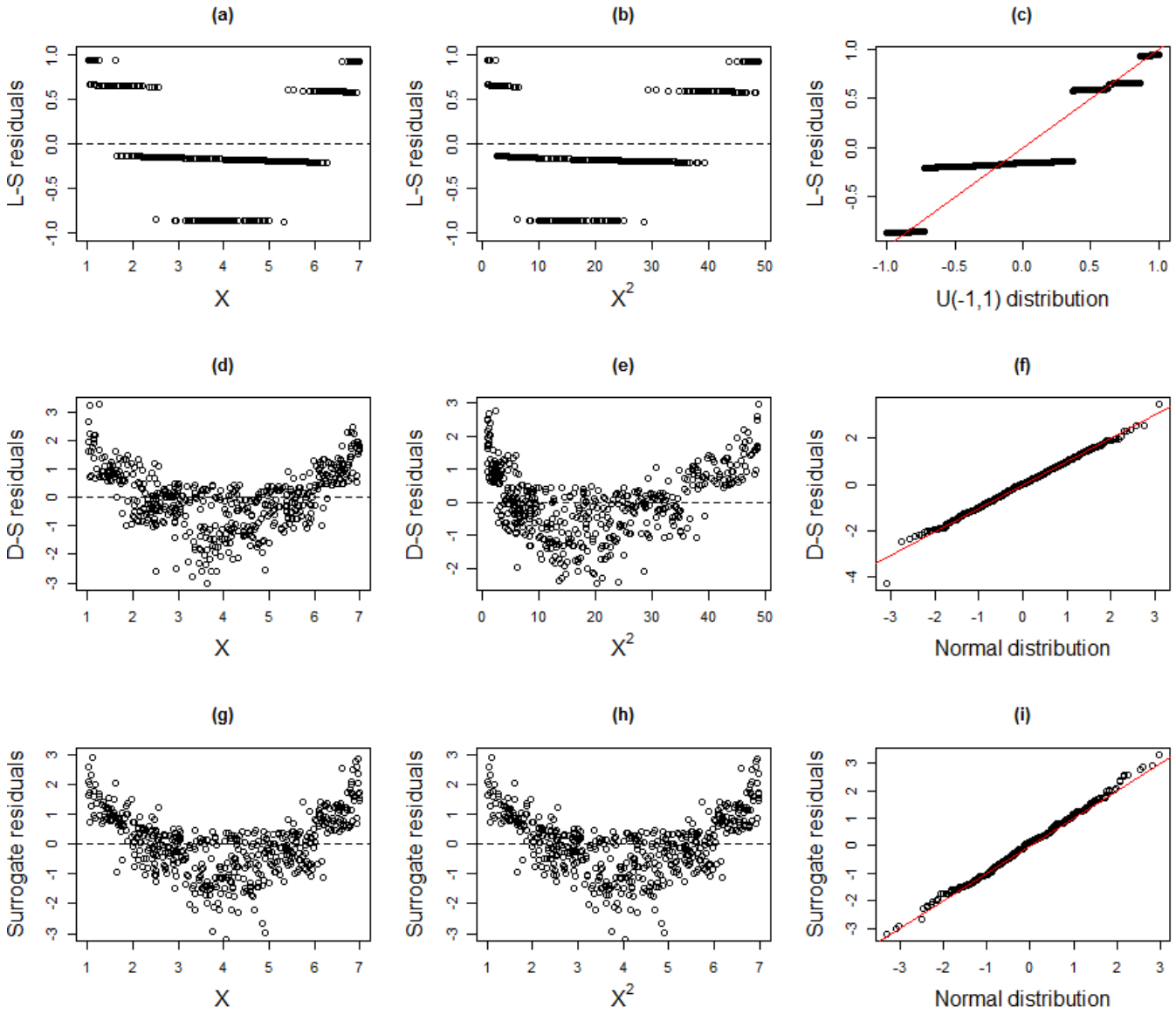

Figure D.48: L-S (a-c), D-S (d-f) and surrogate (g-i) residual plots for the 'linear' cumulative probit model.

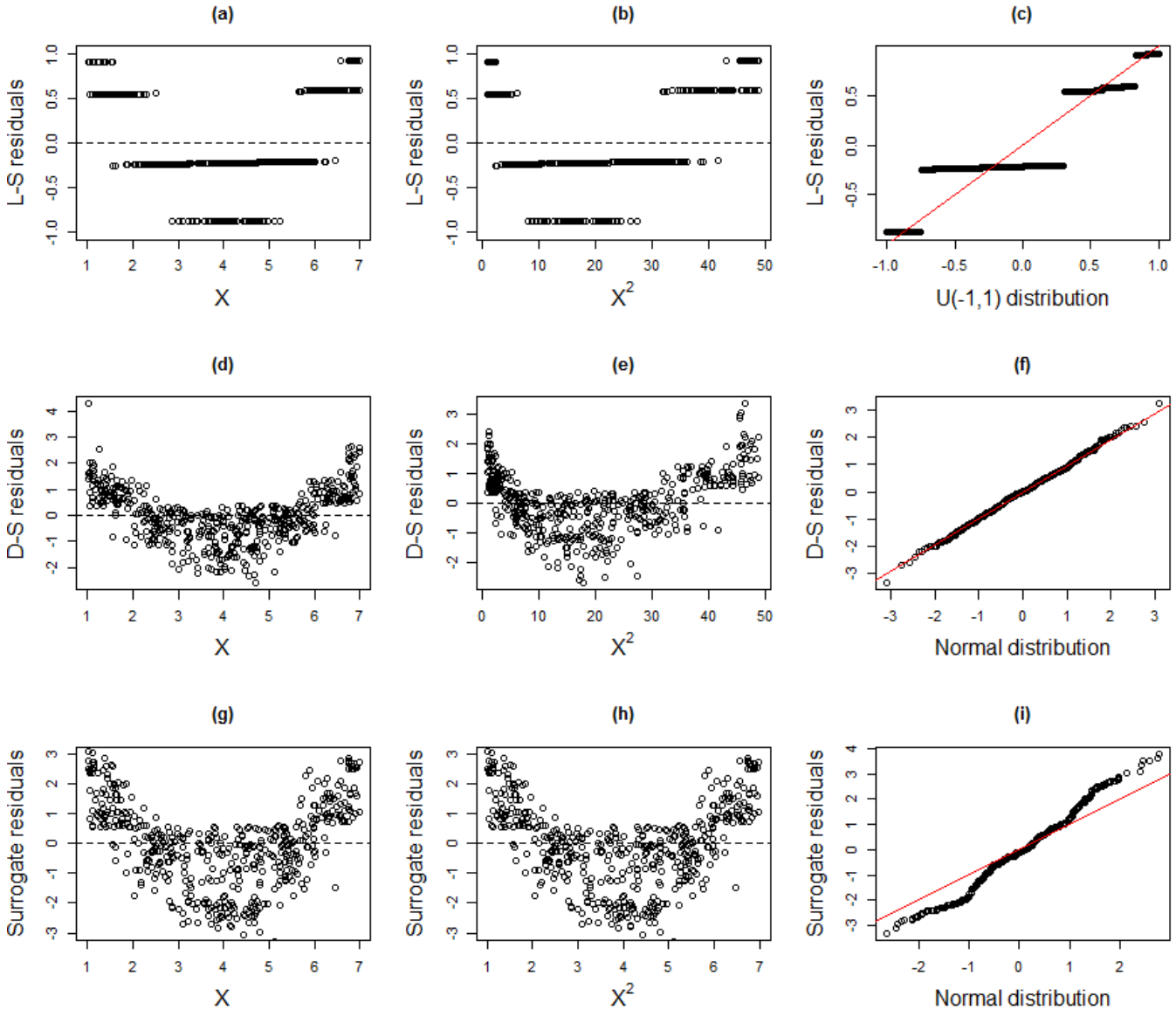

Figure D.49: L-S (a-c), D-S (d-f) and surrogate (g-i) residual plots for the 'linear' CLM.

4. Quadratic-Linear-Quadratic

When we fit a quadratic model to the residuals from the linear model such that $r = \alpha + \beta_1 x + \beta_2 x^2$, we obtain the following results.

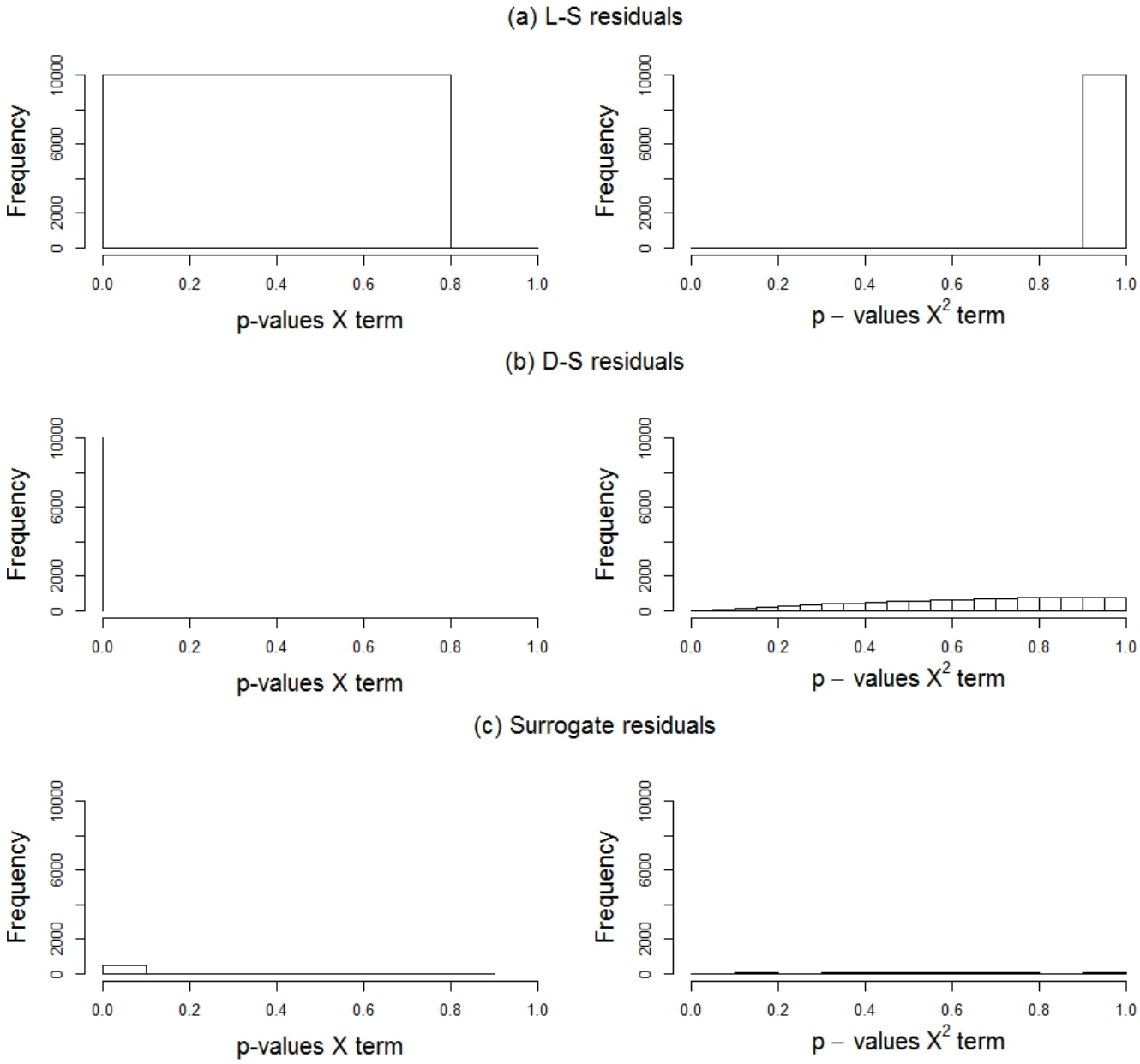

Figure D.50: p-values for the 'quadratic' fit of the L-S, D-S, and surrogate residuals for the 'linear' CLM.

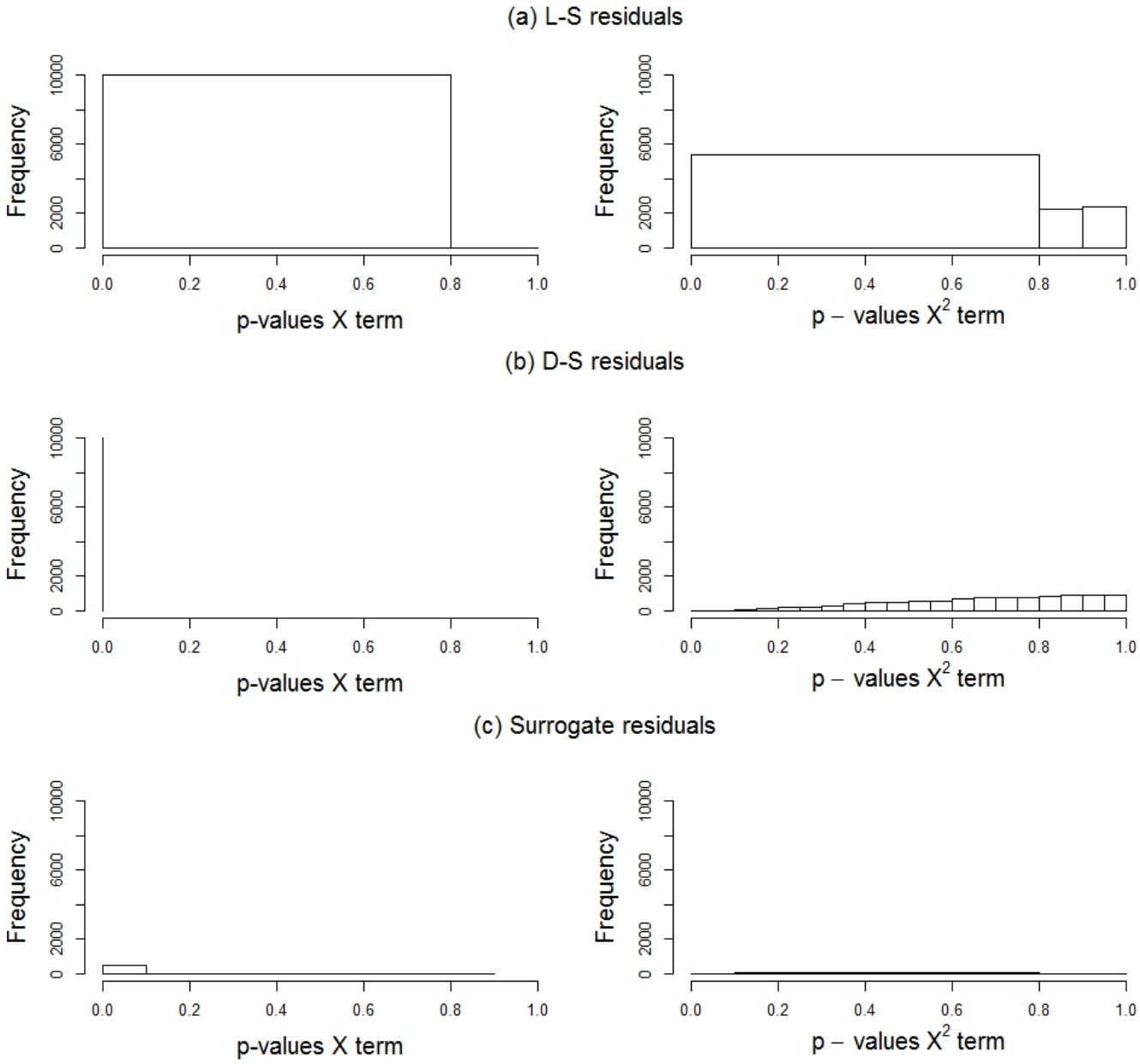

Figure D.51: p-values for the 'quadratic' fit of the L-S, D-S, and surrogate residuals for the 'linear' cumulative probit model.

5. Quadratic-Linear-Linear

When we fit a linear model to the residuals from the linear model such that $r = \alpha + \beta_1 x$, we obtain the following results.

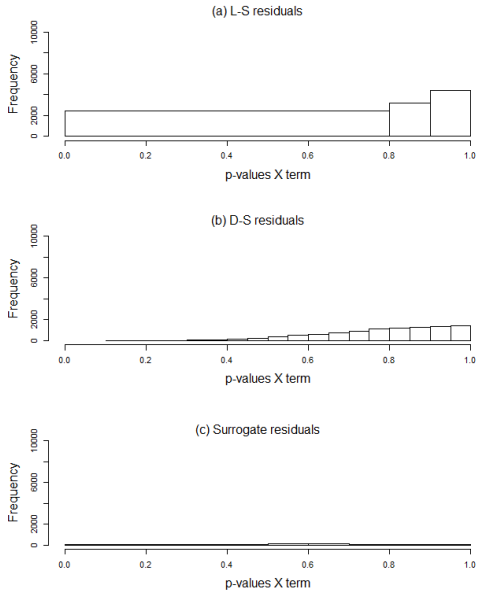

Figure D.52: p-values for the linear fit of the L-S, D-S, and surrogate residuals for the 'linear' cumulative probit model.

D.4 Scenario 4: Missing covariate

As well as testing this scenario for CLMs, we have checked the effects of this misspecification for cumulative probit models too, with a comparison of the results for different threshold structures shown in the following subsections.

D.4.1 Symmetric thresholds

For a symmetric structure with $\boldsymbol{\alpha} = (-2, 0, 2)$, we find that L-S residuals are not appropriate for graphical diagnostics of ordinal models, while D-S and surrogate residuals do not reflect the misspecification.

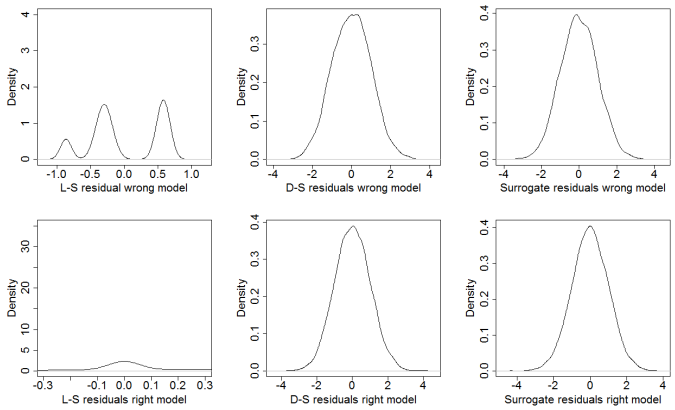

Figure D.53: Comparison of density plots for ordinal residuals with symmetric thresholds for wrong model missing one of the covariates (top) versus right model with both covariates (bottom).

D.4.2 Equidistant thresholds

For equidistant thresholds with $\boldsymbol{\alpha} = (0, 2, 4)$, we find that:

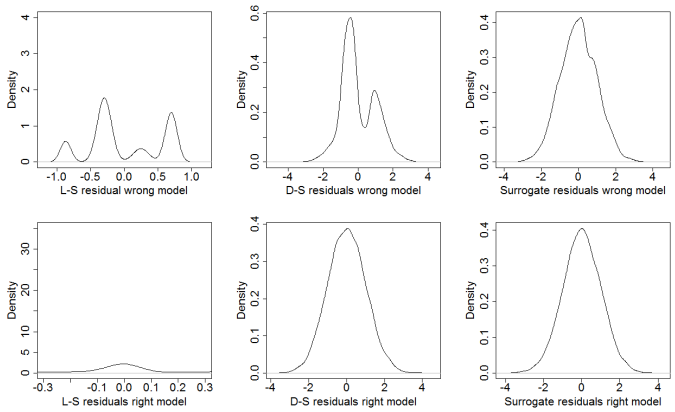

Figure D.54: Comparison of density plots for ordinal residuals with equidistant thresholds for wrong model missing one of the covariates (top) versus right model with both covariates (bottom).

D.4.3 Unconstrained thresholds

Finally, for an unconstrained structure with $\boldsymbol{\alpha} = (-1, 3, 4)$, we find no signs of the misspecification. Unusual patterns in the left plots are due to noise in L-S residuals (equivalent to discrete patterns in scatterplots).

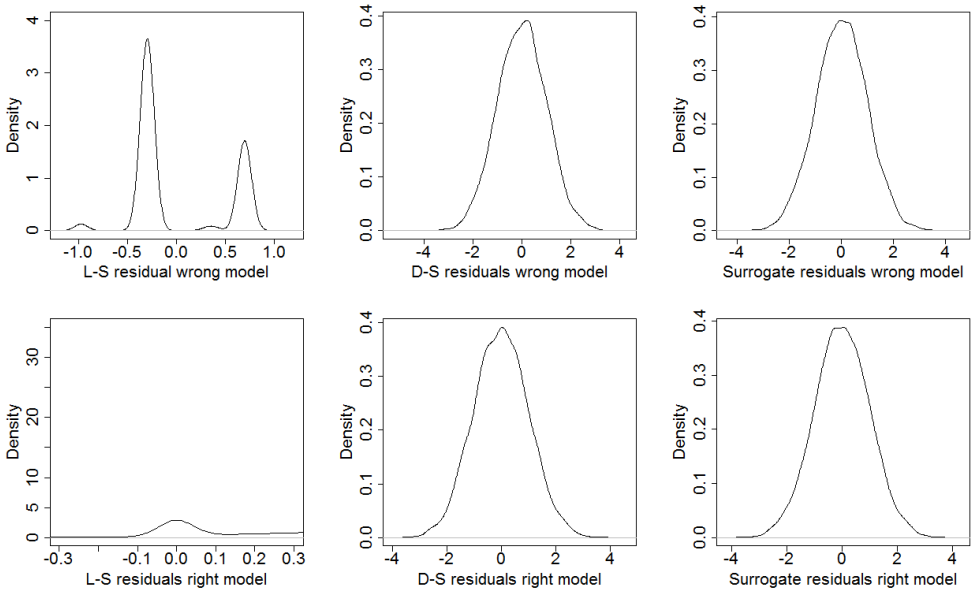

Figure D.55: Comparison of density plots for ordinal residuals with unconstrained thresholds for wrong model missing one of the covariates (top) versus right model with both covariates (bottom).

Appendix E

Regularisation methods

Pavlou et al. (2014) provide a nice summary of regularisation methods, specifically for low-dimensional data with few events, but also applicable to our models.

E.1 Linear Shrinkage Factor (LS)

According to Pavlou et al. (2014), we can reduce overfitting when fitting the model by using MLE and then shrinking the estimated regression coefficients by a common factor, which is called the LSF and can be obtained using bootstrapping. However, Harrell (2001) has claimed that using an LSF is not as good as building shrinkage into the estimation process by using penalised regression.

E.2 Ridge penalisation

Ridge (Cessie and Houwelingen, 1992) or l_2, uses a penalty proportional to the sum of squares of regression coefficients:

$$\hat{\beta} = argmax(l(\beta) - \lambda_2 \sum_{j=1}^{p} \|\beta_j\|_2) = argmax(l(\beta) - \lambda_2 \sum_{j=1}^{p} \beta_j^2) \tag{E.1}$$

E.3 Choice of lambda and optimal log-likelihood for the environmental attitudes data set

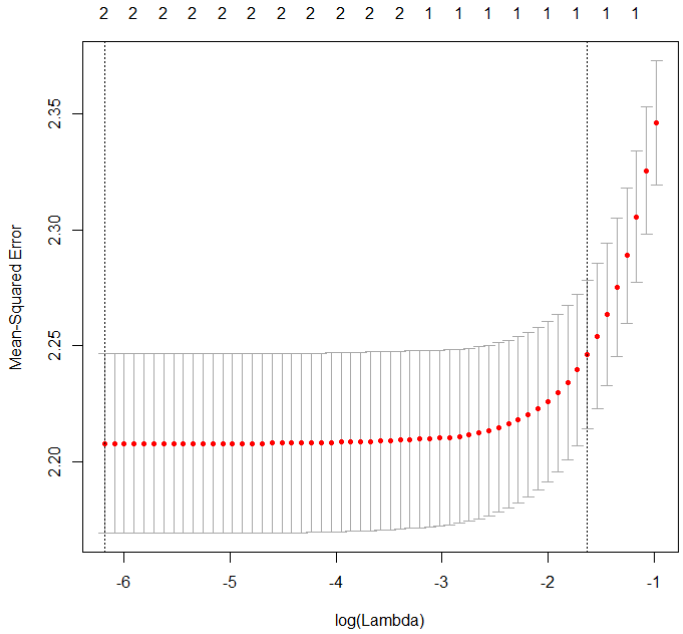

Figure E.1: Choice of optimal λ value for the optimisation.

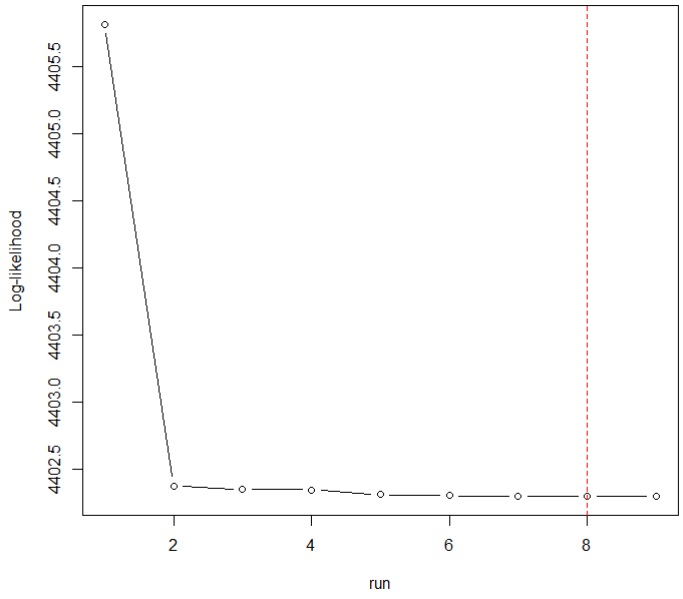

Figure E.2: Choice of optimal log-likelihood value.

E.4 Choice of lambda and optimal log-likelihood for the retinopathy data set

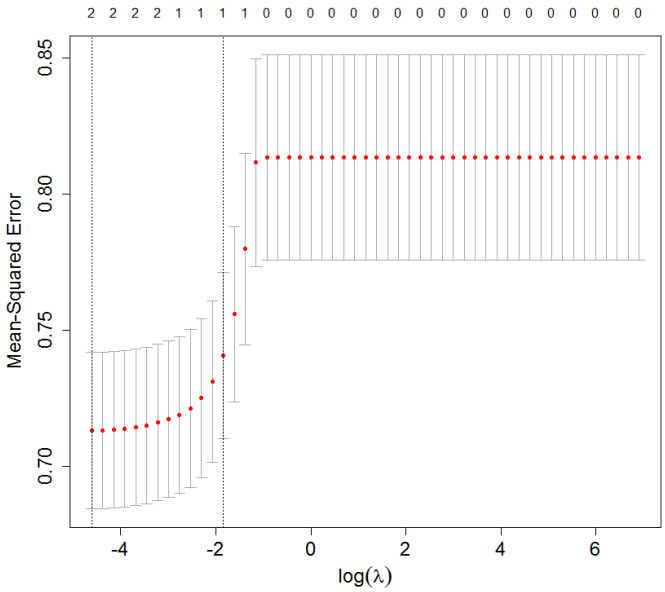

Figure E.3: Choice of optimal λ value for the optimisation.

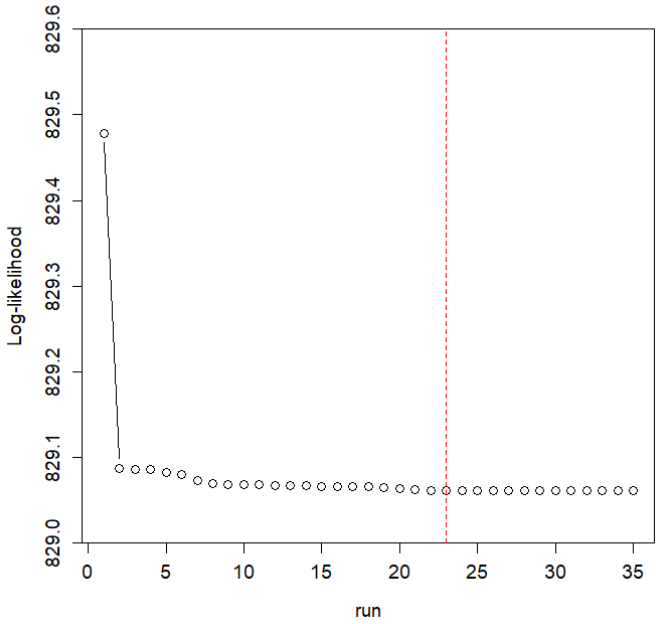

Figure E.4: Choice of optimal log-likelihood value.

Appendix F

Model selection

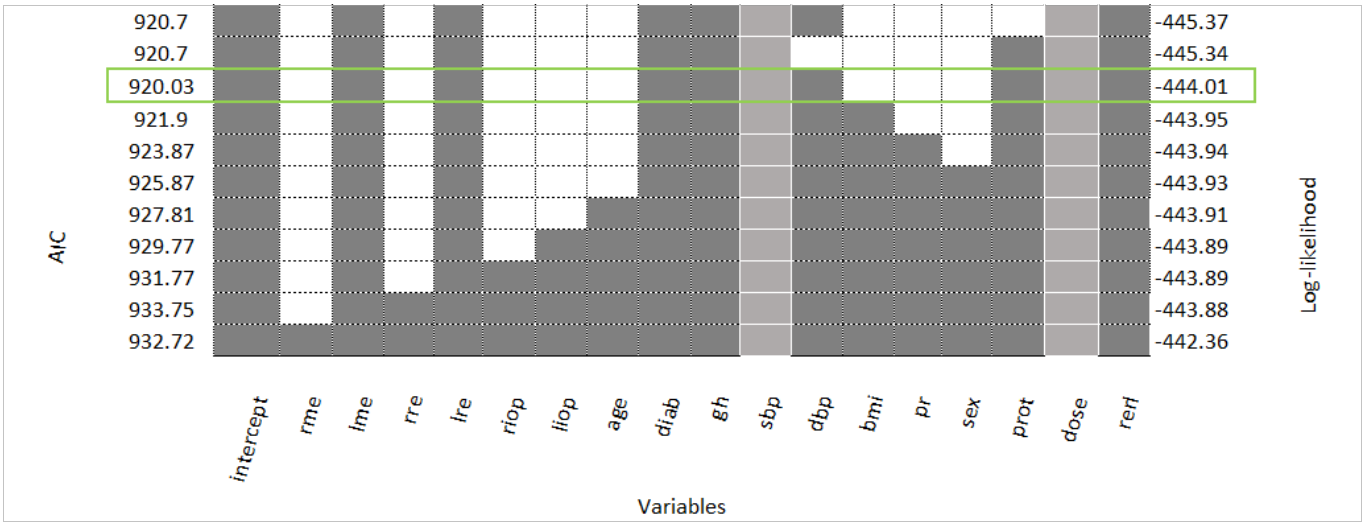

Figure F.1: Variable selection following PO/PPO model selection in Stata with final choice highlighted in green. Light grey boxes correspond to PPO variables.

Bibliography

M.N. Abreu, A.L. Siqueira, C.S. Cardoso, and W.T.. Caiaffa. Ordinal logistic regression models: application in quality of life studies. *Cadernos de Saude Publica, Rio de Janeiro*, 24(4):S581–S591, 2008.

M.N. Abreu, A.L. Siqueira, and W.T. Caiaffa. Ordinal logistic regression in epidemiological studies. *Revista de Saude Publica*, 43(1), 2009.

A. Agresti. Considerations in measuring partial association for ordinal categorical data. *Journal of the American Statistical Association*, 72:37–45, 1977.

A. Agresti. Measures of nominal-ordinal association. *Journal of the American Statistical Association*, 76:524–529, 1981.

A. Agresti. *An introduction to categorical data analysis, 2nd edition.* John Wiley & Sons Inc., New Jersey, 2007.

A. Agresti. *Analysis of ordinal categorical data.* John Wiley & Sons Inc., New Jersey, 2010. doi: http://dx.doi.org/10.1002/9780470594001.scard.

A. Agresti. *Categorical data analysis (3rd ed.).* John Wiley & Sons Inc., New Jersey, 2013.

A. Agresti. *Ordinal data.* Wiley StatsRef: Statistics Reference Online, 2015.

J. Aitchison and S.D. Silvey. The generalization of probit analysis to the case of multiple responses. *Biometrika*, 44:131–140, 1957.

A. Albert and J.A. Anderson. On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, 71(1):1–10, 1984.

J.H. Albert and S. Chib. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422):669–679, 1993.

J.H. Albert and S. Chib. Bayesian residual analysis for binary response regression models. *Biometrika*, 82:747–956, 1995.

P.D. Allison. Comparing logit and probit coefficients across groups. *Sociological Methods & Research*, 28:186–208, 1999.

T. Ameniya. Qualitative response models: a survey. *Journal of Economic Literature*, 19: 481–536, 1981.

C. Ananth and D. Kleinbaum. Regression models for ordinal responses: a review of methods and applications. *International Journal of Epidemiology*, 26:1323–1333, 1997.

J. A. Anderson and P.R. Philips. Regression, discrimination, and measurement models for ordered categorical variables. *Applied Statististics*, 30:22–31, 1981a.

J.A. Anderson. Regression and ordered categorical variables. *Journal of the Royal Statistical Society, Series B.*, 46:1–30, 1984.

J.A. Anderson and P.R. Philips. Regression, discrimination and measurement models for ordered categorical variables. *Journal of the Royal Statistical Society, Series C.*, 30(1): 22–31, 1981b.

D.F. Andrews and D. Pregibon. Finding the outliers that matter. *Journal of the Royal Statistical Society, Series B*, 40:85–93, 1978.

P.G. Arbogast and D.Y. Lin. Model-checking techniques for stratified case-control studies. *Statistics in Medicine*, 24:229–247, 2005.

K.J. Archer, J. Hou, Q. Zhou, K. Ferber, J.G. Layne, and A.E. Gentry. ordinalgmifs: An r package for ordinal regression in high-dimensional data settings. *Cancer Informatics*, 13: 187–195, 2014.

B. Armstrong and M. Sloan. Ordinal regression models for epidemiologic data. *American Journal of Epidemiology*, 129:191–204, 1989.

D.J. Bauer. A note on comparing the estimates of models for cluster-correlated or longitudinal data with binary or ordinal outcomes. *Psychometrika*, 74:97–105, 2009.

D.J. Bauer and S.K. Sterba. Fitting multilevel models with ordinal outcomes: performance of alternative specifications and methods of estimation. *Psychological Methods*, 16:373–390, 2011.

R. Bender and A. Benner. Calculating ordinal regression models in SAS and S-Plus. *Biometrical Journal*, 42(6):677–699, 2000.

R. Bender and U. Grouven. Ordinal logistic regression in medical research. *Journal of the Royal College of Physicians of London*, 31(5):546–551, 1997.

R. Bender and U. Grouven. Using binary logistic regression models for ordinal data with non-proportional odds.. *Journal of Clinical Epidemiology*, 51(10):809–816, 1998.

S. Boes and R. Winkelmann. Ordered response models. *Allgemeines Statistisches Archiv*, 90:165–180, 2006.

K.A. Bollen. Latent variables in psychology and the social sciences. *Annual Review of Psychology*, 53:605–634, 2002.

R. Brant. Assessing proportionality in the proportional odds model for ordinal logistic regression. *Biometrics*, 46(4):1171–1178, 1990.

R. Breen and R. Luijkx. Mixture models for ordinal data. *Sociological Methods & Research*, 39(1), 2010.

D.R. Brillinger and H.K. Preisler. *Maximum likelihood estimation in a latent variable problem in S. Karlin, T. Amemiya and L.A. Goodman, eds., Studies in Econometrics, Time Series and Multivariate Statistics*, volume 115. Academic, New York, 1983.

B.A. Brumback, A.B. Dailey, and H.W. Zheng. Adjusting for confounding by neighbourhood using a proportional odds model and complex survey data. *Practice of Epidemiology*, 175 (11):1133–1141, 2012.

R. Bürgin and G. Ritschard. Coefficient-wise tree-based varying coefficient regression with vcrpart. *Journal of Statistical Software*, 80(6), 2017.

Z. Cai and C.L. Tsai. Diagnostics for nonlinearity in generalized linear models. *Computational statistics & Data Analysis*, 29:445–469, 1999.

M.K. Campbel and A. Donner. Classification efficiency of multinomial logistic regression relative to ordinal logistic regression. *Journal of the American Statistical Association*, 84 (406):587–591, 1989.

M.K. Campbel, A. Donner, and K.M. Webster. Are ordinal models useful for classification? *Statistics in Medicine*, 10(3), 1991.

A.W. Capuano. *Constrained ordinal models with application in occupational and environmental health.* University of Iowa thesis dissertation., 2012.

A.W. Capuano and J.D. Dawson. The trend odds model for ordinal data. *Statistics in Medicine*, 32(13):2250–61, 2013.

A.W. Capuano, J.D. Dawson, and G.C. Gray. Maximizing power in seroepidemiological studies through use of the proportional odds model. *Influenza and Other Respiratory Viruses*, 1(3):87–93, 2007.

S.L. Cessie and J.C.V. Houwelingen. Ridge estimators in logistic regression. *Journal of the Royal Statistical Society, Series C*, 41(1):191–201, 1992.

A. Chesher and M. Irish. Residual analysis in the grouped and censored normal linear model. *Journal of Econometrics*, 34(1-2):33–61, 1987.

R.H. Christensen. *Sensometrics: Thurstonian and Statistical Models.* PhD thesis, Technical University of Denmark., 2012.

R.H. Christensen. Analysis of ordinal data with cumulative link models - estimation with the r-package ordinal, 2015. `https://cran.r-project.org/web/packages/ordinal/vignettes/clm_intro.pdf`.

S.R. Cole and C.V. Ananth. Regression models for unconstrained, partially or fully constrained continuation odds ratios. *International journal of epidemiology*, 30:1379–1382, 2001.

R. Colombi, S. Giordano, A. Gottard, and M. Iannario. Ordinal logistic regression in epidemiological studies. *Scandinavian Journal of Statistics*, 2018.

R. Cook and S. Weisberg. *Location-scale cumulative odds models for ordinal d ata: A generalized non-linear model approach.* Chapman & Hall/CRC., New York, NY, 1982.

C. Cox. Location-scale cumulative odds models for ordinal data: A generalized non-linear model approach. *Statistics in Medicine*, 14(11):1191–1203, 1995.

D.R. Cox and E.J. Snell. A general definition of residuals. *Journal of the Royal Statistical Society*, 30(2):248–275, 1968.

T. Craig, A. Fischer, and A. Lorenzo-Arribas. Shopping versus nature? an exploratory study of everyday experiences. *Frontiers in Psychology*, 9(9), 2018.

F. Di Iorio and M. Iannario. Residual diagnostics for interpreting CUB models. *Statistica*, 72:163–172, 2012.

F. Di Iorio and D. Piccolo. Generalized residuals in CUB models. *Quaderni di Statistica*, 11, 2009.

P.K. Dunn and G.K. Smyth. Randomized quantile residuals. *Journal of Computational and Graphical Statistics*, 5:236–244, 1996.

C. Dupont, J. Horner, C. Li, Q. Liu, and B. Shepherd. Presiduals: Probability-scale residuals and residual correlations, 2017. `https://CRAN.R-project.org/package=PResiduals`.

P.H.C. Eilers and B.D. Marx. Flexible smoothing with b-splines and penalties. *Statistical Science*, 11:89–121, 1996.

J. Engel. Polytomous logistic regression. *Stat Neerlandica*, 42:233–252, 1988.

J. Espinosa and C. Hennig. A constrained regression model for an ordinal response with ordinal predictors. *Statistics and Computing*, pages 1–22, 2018.

M.W. Fagerland and D.W. Hosmer. A generalized Hosmer-Lemeshow goodness-of-fit test for multinomial logistic regression models. *Stata Journal*, 12(3):447–453, 2012.

M.W. Fagerland and D.W. Hosmer. Tests for goodness of fit in ordinal logistic regression models. *Journal of Statistical Computation and Simulation*, 86(17):3398–3418, 2016.

M.W. Fagerland, D.W. Hosmer, and A. Bofin. Multinomial goodness of fit tests for logistic regression models. *Statistics in Medicine*, pages 4238–4253, 2008.

C. Feng, A. Sadeghpour, and L. Li. Randomized quantile residuals: an omnibus model diagnostic tool with unified reference distribution. 2017. `https://arxiv.org/pdf/1708.08527.pdf`.

A. Fielding. Why use arbitrary points scores?: ordered categories in models of educational progress. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 162(3): 303, 1999.

A. Fielding. Scaling for residual variance components of ordered category responses in generalised linear mixed multilevel models. *Quality and Quantity*, 38:425–433, 2004.

S.E. Fienberg. *The analysis of cross-classified categorical data*. The MIT Press., New Jersey, 1980.

D. Firth. Bias reduction of maximum likelihood estimates. *Biometrika*, 80(1):27–38, 1993.

A. Fischer, B. Bednar-Friedl, F. Langers, N. Gemana, K. Skogen, and M. Dumortier. Universal criteria for species convervation priorities? findings from a survey of public views across europe. *Biological Conservation*, 144:998–1007, 2011.

A.S. Fullerton. A conceptual framework for ordered logistic regression models.. *Sociological Methods & Research*, 38(2):306–347, 2009.

J. Gaito. Measurement scales and statistics: Resurgence of an old misconception. *Psychological Bulletin*, 87:564567, 1980.

A. Gelman, Y. Goegebur, F. Tuerlinckx, and I. van Mechelen. Diagnostic checks for discrete data regression models using posterior predictive simulations. *Applied Statistics*, 49(2): 247–268, 2000.

F. C Genter and V.T. Farewell. Goodness-of-link testing in ordinal regression models. *Canadian Journal of Statistics*, 13(1):37–44, 1985.

J. Gertheiss and G. Tutz. Penalized regression with ordinal predictors. *LMU Institute fur Statistik technical report*, 015, 2008.

J. Gill. *Bayesian Methods: A Social and Behavioral Sciences Approach, Third Edition.* Chapman & Hall/CRC Statistics in the Social and Behavioral Sciences, Boca Raton, 1993.

J. Gill and G. Casella. Nonparametric priors for ordinal bayesian social science models: specification and estimation. *Journal of the American Statistical Association*, 104(486), 2009.

L. A. Goodman. Simple models for the analysis of association in cross-classifications having ordered categories. *Journal of the American Statistical Association*, 74(367):537–552, 1979.

C. Gourieroux, A. Monfort, E. Renault, and A. Trognon. Generalised residuals. *Journal of Econometrics*, 34:5–32, 1987.

W.H. Greene. *Econometric Analysis 2nd edition.* Macmillan, New York, 1993.

W.H. Greene and D.A. Hensher. *Modelling ordered ordered choices: a primer.* Cambridge University Press, Cambridge, 2010.

S. Greenland. Alternative models for ordinal logistic regression. *Statistics in Medicine*, 13 (16):1665–1677, 1994.

B.M. Greenwell, A.J. McCarthy, B. Boehmke, and D. Liu. R package sure' version 0.2.0. *R journal*, 2017.

B.M. Greenwell, A.J. McCarthy, B.C. Boehmke, and D. Liu. Residuals and Diagnostics for Binary and Ordinal Regression Models: An Introduction to the sure Package. *The R Journal*, 10(1):381–394, 2018. `https://journal.r-project.org/archive/2018/RJ-2018-004/index.html`.

L. Grilli and C. Rampichini. *Multilevel models for ordinal data in Kenett, R.S. and Salini, S. Modern Analysis of Customer Surveys: with applications using R.* John Wiley & Sons Inc., West Sussex, 2012.

A. Groll. Package glmmlasso, 2017. `https://cran.r-project.org/web/packages/glmmLasso/glmmLasso.pdf`.

A. Guisan and F.E. Harrell. Ordinal response regression models in ecology. *Journal of Vegetation Science*, 11:617–626, 2000.

M. Guzman-Castillo, S. Brailsford, M. Luke, and H. Smith. A tutorial on selecting and interpreting predictive models for ordinal health-related outcomes. *Health Services and Outcomes Research Methodology*, 15:223–240, 2015.

F. Harrell. A comparison between the proportional odds and continuation ratio models for analyzing ordinal outcomes. Retrieved from http://biostat.mc.vanderbilt.edu/wiki/pub/Main/FHHandouts/asa96.pdf, 1996.

F. Harrell. rms regression modeling strategies rdocumentation. Retrieved from https://www.rdocumentation.org/packages/rms, 2018.

F.E. Harrell. *Regression Modelling Strategies: With applications to linear models, logistic regression, and survival analysis.* Springer, New York, 2001.

F.E. Harrell, P.A. Margolis, S. Gove, K.E. Mason, Mulholland E.K., Lehmann D., L. Muhe, Gatchalian S., and Eichenwald H.F. Development of a clinical prediction model for an ordinal outcome: the world health organization multicentre study of clinical signs and etiological agents of pneumonia, sepsis and meningitis in young infants. *Statistics in Medicine*, 17(8):909–944, 1998.

R.K. Hawkes. The multivariate analysis of ordinal measures. *American Journal of Sociology*, 76:908–926, 1971.

M. Hebiri and J. Lederer. How correlations influence lasso prediction. *IEEE Transactions on Information Theory*, 59(3), 2013.

D. Hedeker. Multilevel models for ordinal and nominal variables. In J. Deleeuw and E. Mejier, editors, *Handbook of Multilevel Analysis*, pages 237–274. Springer, 2007.

D. Hedeker, M. Berbaum, and R. Mermelstein. Location-scale models for multilevel ordinal data: between- and within-subjects variance modeling. *Journal of Probability and Statistical Science*, 4(1):1–20, 2006.

T. Heeren and R. D'Agostino. Robustness of the two independent samples t-test when applied to ordinal scaled data. *Statistics in Medicine*, 1:79–90, 1987.

G.P. Hoetker. Confounded coefficients: extending recent advances in the accurate comparison of logit and probit coefficients across groups. *SSRN Electronic Journal.*, 2004. `http://dx.doi.org/10.2139/ssrn.609104`.

K.K. Holst. Model diagnostics based on cumulative residuals: The r-package gof. *arXiv*, 2015. `https://arxiv.org/pdf/1507.01173.pdf`.

H.G. Hong and X. He. Prediction of functional status for the elderly based on a new ordinal regression model. *Journal of the American Statistical Association.*, 105(491):930–941, 2010.

D.W. Hosmer and S. Lemeshow. A goodness-of-fit test for the multiple logistic regression model. *Communications in Statistics.*, A10:1043–1069, 1980.

D.W. Hosmer and S. Lemeshow. *Applied logistic regression, 2nd edition.* John Wiley & Sons Inc., Hoboken, 2000.

M. Iannario and D. Piccolo. CUB models: statistical methods and empirical evidence. In Jan Fagerberg, David C. Mowery, and Richard R. Nelson, editors, *Modern analysis of customer surveys: with applications using R*, chapter 13, pages 231–258. John Wiley & Sons, Chichester, UK, 2012.

M. Iannario, A.C. Monti, D. Piccolo, and E. Ronchetti. Robust inference for ordinal response models. *Electronic Journal of Statistics*, 11(2):3407–3445, 2017.

D. Ilodigwe, G.D. Murray, N.F. Kassell, J. Torner, R.S.C. Kerr, A.J. Molyneux, and R.L. Macdonald. Sliding dichotomy compared with fixed dichotomization of ordinal outcome scales in subarachnoid hemorrhage trials clinical article. *Journal of Neurosurgery*, 118(1): 3–12, 2013.

Ipsos MORI Scotland and Scottish Government. Scottish environmental attitudes and behaviours survey, 2008. 2009. http://doi.org/10.5255/UKDA-SN-6249-1.

H. Ishwaran and C.A. Gatsonis. A general class of hierarchical ordinal regression models with applications to correlated roc analysis. *The Canadian journal of statistics*, 28(4): 731–750, 2000.

I. Jeliazkov and M.A. Rahman. *Binary and ordinal data analysis in economics: modelling and estimation, in Yang, Z.S. Mathematical Modeling with Multidisciplinary Applications.* John Wiley & Sons, Inc., New Jersey, 2012.

B. Jones and M. Sobel. Modeling direction and intensity in semantically balanced ordinal scales: an assessment of congressional incumbent approval. *American Journal of Political Science*, 44(1):174–185, 2000.

B. Jones and C. Westerland. Order matters (?): alternatives to conventional practices for ordinal categorical response variables. 2006. URL http://psfaculty.ucdavis.edu/bsjjones/ordermatters.pdf.

J.H. Kim. Assessing practical significance of the proportional odds assumption. *Statistics & Probability Letters*, 65:233–239, 2003.

T.R. Knapp. Treating ordinal scales as interval scales: an attempt to resolve the controversy. *Nursing research*, 39(2):121–123, 1990.

I. Kosmidis. Improved estimation in cumulative link models. *Journal of the Royal Statistical Society: Series B*, 76(1):169–196, 2014.

I. Kosmidis and D. Firth. Bias reduction in exponential family nonlinear models. *Biometrika*, 96(4):793–804, 2009.

S. Kramer, G. Widmer, B. Pfahringer, and M. De Groeve. Prediction of ordinal classes using regression trees. *Fundamenta Informaticae*, 47(1):1–13, 2001.

C. Kuesten, J. Bi, and Meiselman H.L. Analyzing consumers' profile of mood states (poms) data using the proportional odds model (pom) for clustered or repeated observations and r package 'repolr'. *Food Quality and Preference*, pages 38–49, 2017.

J. Kuha and C. Milss. On group comparisons with logistic regression models. *Sociological Methods & Research*, pages 1–28, 2018.

W.M. Jr Kuzon, M.G. Urbanchek, and S. McCabe. The seven deadly sins of statistical analysis. *Annals of Plastic Surgery*, 37:265–272, 1996.

R. Lall, M.J. Campbell, SJ. Walters, and K. Morgan. A review of ordinal regression models applied on health-related quality of life assessments. *Statistical Methods in Medical Research*, 162(3):49–67, 2002.

S.R. Land and J.H. Friedman. Variable fusion: a new adaptive signal regression method.. *Technical report 656. Department of statistics. Carnegie Mellon University Pittsburg.*, 1997.

J.M. Landwehr, D. Pregibon, and A.C. Shoemaker. Graphical methods for assessing logistic regression models. *Journal of the American Statistical Association*, 79(385):61–71, 1984.

M. Law and D. Jackson. Residual plots for linear regression models with censored outcome data: A refined method for visualizing residual uncertainty. *Communications in Statistics - Simulation and Computation*, 46(4):3159–3171, 2017.

C. Li and B.E. Shepherd. A new residual for ordinal outcomes. *Biometrika*, 99(2):473–480, 2012.

R. Likert. A technique for the measurement of attitudes. *Archives of Psychology*, 140:1–55, 1932.

K.C. Lin and Y.J. Chen. Assessing ordinal logistic regression models via nonparametric smoothing. *Communications in Statistics - Theory and Methods*, 37(6):917–930, 2008.

S.R. Lipsitz, G.M. Fitzmaurice, and G. Molenberghs. Goodness-of-fit tests for ordinal response regression models. *Journal of the Royal Statistical Society. Series C*, 45(2):175–190, 1996.

S.R. Lipsitz, G.M. Fitzmaurice, S.E. Regenbogen, D. Sinha, J.G. Ibrahim, and A.A. Gawande. Bias correction for the proportional odds logistic regression model with application to a study of surgical complications. *Applied Statistics*, 62(2):233–250, 2012.

D. Liu and H. Zhang. Residuals and diagnostics for ordinal regression models. *Journal of the American Statistical Association*, 113(522):845–854, 2018.

I. Liu, B. Mukherjee, T. Suesse, D. Sparrow, and S.K. Park. Graphical diagnostics to check model misspecification for the proportional odds regression model. *Statistics in Medicine*, 28:412–429, 2009.

J.S. Long. *Regression models for nominal and ordinal outcomes in Best, H. and Wolf, C. The SAGE Handbook of Regression Analysis and Causal Inference.* Sage, Thousand Oaks, CA, 2014.

J.S. Long and J. Freese. *Regression models for categorical dependent variables using stata (2nd ed.)* . Stata Press, Station, TX, 2006.

T. Lumley. Generalized estimating equations for ordinal data: A note on working correlation structures. *Biometrics*, 52(1):354–361, 1996.

M. Lunt. Stereotype ordinal regression. *Stata Technical Bulletin*, 10(61), 2001.

R.C. MacCallum, S. Zhang, K.J. Preacher, and D.D. Rucker. On the practice of dichotomization of quantitative variables. *Psychological Methods*, 7(1):19–40, 2002.

J.A.F. Machado and J.M.C. Santos-Silva. Quantiles for counts. *Journal of the American Statistical Association*, 100(472):1226–1237, 2005.

S.J. Machin and M.B. Stewart. Unions and the financial performance of british private sector establishments. *Journal of Applied Econometrics*, 5:327–350, 1990.

A. Mayer and M. Foster. Understanding recession and self-related health with the partial proportional odds model: an analysis of 26 countries. *PloS one*, 2015. doi: https://doi.org/10.1371/journal.pone.0140724.

P. McCullagh. Regression models for ordinal data (with discussion). *Journal of the Royal Statistical Society. Series B (Methodological)*, 42(2):109–142, 1980.

P. McCullagh and J.A. Nelder. *Generalized Linear Models, 2nd ed.* Chapman & Hall, London, 1989.

R.D. McKelvey and W. Zavoina. A statistical model for the analysis of ordinal level dependent variables. *Journal of Mathematical Sociology*, 4:103–120, 1975.

D. McParland and I.C. Gormley. *Clustering ordinal data via latent variable models, in Lausen, B., van den Poel, D. and Ultschs (eds.). Algorithms from and for Nature and Life: classification and data analysis.* Springer, New York, 2011.

C. Mood. Logistic regression: Why we cannot do what we think we can do, and what we can do about it. *European Sociological Review*, 26:67–82, 2010.

U. Munzel and F. Langer. A global view on parametric and nonparametric approaches to the analysis of ordered categorical data. *Biometrical Journal*, 46(1):7–18, 2004.

S.A. Murphy, A. J. Rossini, and A.W. van der Vaart. Maximum likelihood estimation in the proportional odds model. *Theory and Method*, 92(439):968–976, 1997.

G.D. Murray, D. Barer, S. Choi, H. Fernandes, B. Gregson, Kennedy R.L., A.I.R. Maas, A. Marmarou, A.D. Mendelow, E.W. Steyerberg, G.S. Taylor, G.M. Teasdale, and C. J. Weir. Design and analysis of phase iii trials with ordered outcome scales: The concept of the sliding dichotomy. *Journal of neurotrauma*, 22(5):511–517, 2005.

J. Nagler. Scobit: an alternative estimator to logit and probit. *American Journal of Political Science*, 38(1):230–255, 1994.

J.M. Neuhaus. Bias and efficiency loss due to misclassified responses in binary regression. *Biometrika*, 86(4):843–855, 1999.

N. Nooraee, G. Molenberghs, and E.R. van den Heuvel. Gee for longitudinal ordinal data: Comparing r-geepack, r-multgee, r-repolr, sas-genmod, spss-genlin. *Computational Statistics & Data Analysis*, 77:70–83, 2014.

G. Norman. Likert scales, levels of measurement and the laws of statistics. *Advances in Health Sciences Education*, 15:625–632, 2010.

R.M. O'Brien. Using rank-order measures to represent continuous variables. *Social Forces*, 61:144–155, 1982.

R.M. O'Brien. The relationship between ordinal measures and their underlying values: Why all the disagreement? *Quality and Quantity*, 19(3):265–277, 1985.

A.A. O'Connell and X. Liu. Model diagnostics for proportional and partial proportional odds models. *Journal of Modern Applied Statistical Methods*, 10(1):139–175, 2011.

J.G. Orme and T. Combs-Orme. *Multiple Regression With Discrete Dependent Variables.* Oxford University Press, New York, 2009.

N.R. Parsons, R.N. Edmonson, and S.G. Gilmour. A generalized estimating equation method for fitting autocorrelated ordinal score data with an application in horticultural research. *Journal of the Royal Statistical Society: Series C*, 55(4), 2006.

M. Pavlou, G. Ambler, S. Seaman, M. De Iorio, and R.Z. Omar. Review and evaluation of penalised regression methods for risk prediction in low-dimensional data with few events. *Statistics in Medicine*, 35:1159–1177, 2014.

B. Peterson and F.E. Harrell. Partial proportional odds models for ordinal response variables. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, 39(2):205–217, 1990.

J. Peyhardi, C. Trottier, and Y. Guedon. A new specification of generalized linear models for categorical responses. *Biometrika*, 102(4):889–906, 2015.

J. Peyhardi, C. Trottier, and Y. Guedon. Partitioned conditional generalized linear models for categorical responses. *Statistical Modelling*, 16(4):297–321, 2016.

D.A. Pierce and D.W. Schafer. Residuals in generalized linear models. *Journal of the American Statistical Association*, 81(396):977–986, 1986.

W. Poßnecker and G. Tutz. A general framework for the selection of effect type in ordinal regression. *Ludwig-Maximilians-Universitat Munchen Technical report*, (186), 2016.

D. Pregibon. *Data analytic methods for generalized linear models.* University of Toronto thesis dissertation., 1979.

D. Pregibon. Logistic regression diagnostics. *The Annals of Statistics*, 9(4):705–724, 1981.

H. Pruscha. Partial residuals in cumulative regression models for ordinal data. *Statistical Papers*, 35:273–284, 1994.

E. Pulkstenis and T.J. Robinson. Goodness-of-fit tests for ordinal response regression models. *Statistics in Medicine*, 23(6):999–1014, 2004.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2018. `http://www.R-project.org/`.

K.B. Rajan and X.H. Zhou. Semi-parametric area under the curve regression method for diagnostic studies with ordinal data.. *Biometrical Journal*, 54(1):143–156, 2012.

G. Ramaswami and R. Sukumar. Long-term environmental correlates of invasion by lantana camara (verbenaceae) in a seasonally dry tropical forest.. *PLOS one*, 8(10):e76995, 2013.

C.R. Rao. large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation. *Proceedings of the Cambridge Philosophical Society*, 44(3):50–57, 1948.

P. Resnick and H.R. Varian. Recommender systems. *Communications of the ACM*, 40(3): 56–58, 1997.

M. Rhemtulla, P.E. Brosseau-Liard, and V. Savalei. When can categorical variables be treated as continuous? a comparison of robust continuous and categorical sem estimation methods under suboptimal conditions. *Psychological Methods*, 17(3):354–373, 2012.

J.M. Robins, A. Vaart, and V. Venturai. Asymptotic distribution of p values in composite null models. *Journal of the American Statistical Association*, 95:1143–1156, 2000.

S.M. Rudolfer, P.C. Watson, and E. Lesaffre. Are ordinal models useful for classification? a revised analysis. *Journal of Statistical Computing and Simulation*, 52:105–132, 1995.

D. Ruppert. Selecting the number of knots for penalized splines. *Journal of Computational and Graphical Statistics*, 11(4):735–757, 2002.

S.E. Saffari, A. Love, M. Fredrikson, and O. Smedby. Regression models for analyzing radiological visual grading studies–an empirical comparison. *BMC Medical Imaging*, pages 15–49, 2015.

F. Samejnima. Estimation of latent ability using a pattern of graded scores. *Psychometrika*, Monograph Supplement 17, 1969.

SAS Institute. *Stata Statistical Software: Release 14.* SAS Institute, College Station, TX, 2008.

L. Sasidharan and M. Menendez. Partial proportional odds model - an alternate choice for analyzing pedestrian crash injury severities. *Accident Analysis and Prevention*, 72: 330–340, 2014.

B.E. Shepherd, C. Li, and Q Liu. Probability-scale residuals for continuous, discrete, and censored data. *The Canadian Journal of Statistics*, 44(4):463–479, 2016.

J.S. Simonoff. *Analysing categorical data.* Springer-Verlag, New York, 2003.

J.M. Singer, F.Z. Poleto, and P. Rosa. Parametric and nonparametric analyses of repeated ordinal categorical data. *Biometrical Journal*, 46(4):460–473, 2004.

A. Skrondal and S. Rabe-Hesketh. *Generalized latent variable modeling: Multilevel, longitudinal and structural equation models.* Chapman & Hall/CRC, Boca Raton, FL, 2004.

G. Smyth, Y. Hu, P. Dunn, B. Phipson, and Y. Chen. R package statmod version 1.4.29. *Statistical Modeling. CRAN*, 2017.

E.J. Snell. A scaling procedure for ordered categorical data. *Biometrics*, 20(3):592–607, 1964.

Stata Corp. *Stata Statistical Software: Release 14.* StataCorp LP, College Station, TX, 2015.

S.S. Stevens. On the theory of scales of measurement. *Science*, 103(2684):677–680, 1946.

W.L. Stevens. Fiducial limits of the parameter of a discontinuous distribution. *Biometrika*, 37:117–129, 1950.

J.Q. Su and L.J. Wei. A lack-of-fit test for the mean function in a generalized linear model. *Journal of American Statistical Association*, 86(414):420–426, 1991.

T.M. Therneau, P.M. Grambsch, and T.R. Fleming. Martingale-based residuals for survival models. *Biometrika*, 77:147–160, 1990.

D. Thissen and L. Steinberg. A taxonomy of item response models. *Psychometrika*, 51(4): 567–577, 1986.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society B*, 58(1):267–288, 1996.

R. Tibshirani, Rosset S. Zhu J. Sanders, M., and Kneight K. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society B*, 67:91–108, 2005.

A.Y. Toledano and C. Gatsonis. Ordinal regression methodology for roc curves derived from correlated data. *Statistics in Medicine*, 15(16):1807–1826, 1996.

V. Torra, J. Domingo-Ferrer, J.M. Mateo-Sanz, and M. Ng. Regression for ordinal variables without underlying continuous variables. *Information sciences*, 176:465–474, 2006.

G. Tutz. Sequential item response models with an ordered response. *British Journal of Mathematical and Statistical Psychology*, 43:39–55, 1990.

G. Tutz. Sequential models in ordinal regression. *Computational statistics & data analysis*, 11:275–295, 1991.

G. Tutz. *Regression for Categorical Data.* Cambridge University Press, Cambridge, 2012.

G. Tutz and T. Hechenbichler. Aggregating classifiers with ordinal response structure. *Journal of Statistical Computation and Simulation*, 75(5):391–408, 2005.

G. Tutz and T. Scholz. Ordinal regression modelling between proportional odds and non-proportional odds. *Sonderforschungsbereich*, 386(304), 2003.

G.J.G Upton. *Categorical data analysis by example.* John Wiley & Sons, Hoboken, N.J., 2017.

A.P.N. Van der Jagt, T. Craig, J. Anable, M.J. Brewer, and D.G. Pearson. Unearthing the picturesque: The validity of the preference matrix as a measure of landscape aesthetics. *Landscape and Urban Planning.*, 124:1–3, 2014.

R.W. Walker. On generalizing cumulative ordered regression models. *Journal of Modern Applied Statistical Methods*, 15(2):455–474, 2016.

R. Williams. Oglm: Stata module to estimate ordinal generalized linear models. 2006. `http://econpapers.repec.org/software/bocbocode/s453402.htm`.

R. Williams. Using heterogeneous choice models to compare logit and probit coefficients across groups. *Sociological Methods & Research*, 37(4):531–559, 2009.

R. Williams. Fitting heterogeneous choice models with oglm. *The Stata Journal*, 10(4): 540–567, 2010.

R. Williams. Gologit2/oglm troubleshooting. Retrieved from http://www3.nd.edu/ rwilliam/gologit2/tsfaq.html, 2014.

R. Williams. Understanding and interpreting generalized ordered logit models. *The Journal of Mathematical Sociology*, 40(1):7–20, 2016.

C. Winship and R.D. Mare. Regression models with ordinal variables. *American Sociological Review*, 49:512–525, 1984.

M.J. Wurm, P.J. Rathouz, and B.M. Hanlon. Regularized ordinal regression and the ordinalnet r package. 2017. `https://arxiv.org/abs/1706.05003`.

F. Xue and A. Qu. Variable selection for highly correlated predictors. *Arxiv*, 2017. `https://arxiv.org/pdf/1709.04840.pdf`.

T.W. Yee. The vgam package for categorical data analysis. *Journal of Statistical Software*, 32(10), 2010.

G. U. Yule and M.G. Kendall. *An introduction to the theory of statistics.* Charles Griffin and Co, Ltd., London, 1950.

F.M. Zahid, S. Ramzan, and C. Heumann. Regularized proportional odds models. *Journal of Statistical Computation and Simulation*, 85(2):251–268, 2015.

Q. Zhang and E. Haksing Ip. Generalized linear model for partially ordered data. *Statistics in Medicine*, 31:56–68, 2010.

P. Zhao and B. Yu. On model selection consistency of lasso. *Journal of Machine Learning Research*, 7:2541–2563, 2006.

H. Zou and T. Hastie. Regularisation and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B.*, 67:301–320, 2005.