Ordinal regression: A review and a taxonomy of models

Tutz, Gerhard

Veröffentlichungsversion / Published Version
Zeitschriftenartikel / journal article

Empfohlene Zitierung / Suggested Citation:
Tutz, G. (2022). Ordinal regression: A review and a taxonomy of models. *WIREs Computational Statistics*, *14*(2), 1-28. https://doi.org/10.1002/wics.1545

Received: 10 July 2020 | Revised: 2 November 2020 | Accepted: 28 November 2020

DOI: 10.1002/wics.1545

ADVANCED REVIEW

WIREs COMPUTATIONAL STATISTICS WILEY

Ordinal regression: A review and a taxonomy of models

Gerhard Tutz

Ludwig-Maximilians-Universität München Akademiestraße 1, Munchen, Germany

Correspondence
Gerhard Tutz, Ludwig-Maximilians-Universität München Akademiestraße 1, Munchen 80799, Germany.
Email: tutz@stat.uni-muenchen.de

Abstract

Ordinal models can be seen as being composed from simpler, in particular binary models. This view on ordinal models allows to derive a taxonomy of models that includes basic ordinal regression models, models with more complex parameterizations, the class of hierarchically structured models, and the more recently developed finite mixture models. The structured overview that is given covers existing models and shows how models can be extended to account for further effects of explanatory variables. Particular attention is given to the modeling of additional heterogeneity as, for example, dispersion effects. The modeling is embedded into the framework of response styles and the exact meaning of heterogeneity terms in ordinal models is investigated. It is shown that the meaning of terms is crucially determined by the type of model that is used. Moreover, it is demonstrated how models with a complex category-specific effect structure can be simplified to obtain simpler models that fit sufficiently well. The fitting of models is illustrated by use of a real data set, and a short overview of existing software is given.

This article is categorized under:
Statistical Models > Fitting Models
Data: Types and Structure > Categorical Data
Statistical Models > Generalized Linear Models

KEYWORDS

adjacent categories model, cumulative model, hierarchically structured models, ordinal regression, proportional odds model, sequential model

1 | INTRODUCTION

Ordered categorical regression, or simply ordinal regression, aims at exploiting the ordering in the responses to obtain simply structured models. In particular, McCullagh's seminal article (McCullagh, 1980) stimulated research in the area and made ordinal regression a widely used tool that avoids the pitfalls of using ANOVA-type models on ordered categorical data. The distinct limitations of the use of ordinary least squares approaches to modeling ordinal responses have been outlined, for example, by Agresti (2010). Nowadays a multitude of models and corresponding software are available, for an overview of classical models see, for example, Agresti (2010, 2013), Tutz (2012). Long and Freese (2006) and Williams and Quiroz (2020) gave overviews with a focus on social science applications.

Some basic ordinal models assume that an unobserved latent variable underlies the ordinal response variable. The observed variable is considered as a categorization of the underlying continuous response. The approach yields simple models but makes assumptions that are not really needed for the construction of ordinal models. Alternative models are derived from sequential decision processes, which make assumption on the process that yields the final outcome. Although we will consider such motivations for ordinal models we aim at characterizing ordinal models in a general way by investigating how models can be constructed from simpler, in particular binary models. It also makes explicit which binary models are contained in basic ordinal models.

Based on the construction principle a taxonomy of ordinal models is derived that covers the most prominent models in common use and also includes more recent developments. The derivation of a taxonomy includes a survey of the alternative ways how to model ordinal responses. Particular attention is given to the modeling of additional heterogeneity as, for example, dispersion effects, which has been somewhat neglected and only recently investigated more closely. Accounting for additional heterogeneity can avoid biased estimates of the effects of explanatory variable, but also provides additional information on the effects of explanatory variables. Although several models that account for heterogeneity have been proposed it is not always clear what exactly is modeled. One of the objectives is to clarify the nature of the heterogeneity that is captured by these models. We also consider hierarchical models, which seem not to be sufficiently developed, although they have several advantages and it is straightforward to include heterogeneity effects. Though hierarchical models have been considered in item response analysis their potential has not yet been exploited sufficiently in ordinal regression. The flexible class of hierarchical models is presented in a structured way and embedded into the taxonomy. The taxonomy is completed by including the more recently proposed class of mixture models that account for uncertainty in a different way.

In Section 2, basic ordinal models are reviewed and described as composed from binary models, which yields a preliminary taxonomy. In Section 2, it is demonstrated that unobserved heterogeneity may yield misleading results. It is shown how heterogeneity can be modeled and the meaning of heterogeneity in alternative models is clarified. In Section 4, category-specific parameter structures are investigated. Section 5 is devoted to the wide class of hierarchically structured models with a focus on symmetric models. The finite mixture approach to modeling heterogeneity is considered in Section 6. Finally some further areas of current research are briefly mentioned.

2 | BASIC MODELS

One can distinguish between three basic models, the cumulative model, which can be derived from a latent continuous variable, the sequential model, which is a process model, and the adjacent categories model, which is strongly related to nominal models. In the following these models are characterized by the binary models that are contained in these models. With $Y \in \{1, ..., k\}$ denoting an ordinal response the interesting binary responses that can be used to obtain a simple characterization of the models are the *split variables*

$$Y_r = 1 \text{ if } Y \geq r, \text{ and } Y_r = 0, \text{otherwise},$$

which split the response categories into the subsets $\{1, ..., r-1\}$ and $\{r, ..., k\}$. The variable Y_r simply distinguishes between low and high response categories with low and high referring to $Y < r$ and $Y \geq r$. The link between the response and the split variables is given by

$$Y = r \quad \Leftrightarrow \quad (Y_1 ..., Y_k) = (1, ..., 1, 0, ..., 0),$$

where Y_r is the last of the sequence of binary variables with a value 1. The vectors $(Y_1 ..., Y_k)$ form a so-called Guttman space, all members have the form (1, ..., 1, 0, ..., 0), in which a sequence of ones is followed by a sequence of zeros. They can be seen as a specific vector-valued representation of the response in k categories.

2.1 | Cumulative models

Cumulative models as considered by McCullagh (1980) are typically derived from the assumption of an underlying latent regression model with a continuous response. Let Y^* be an underlying latent variable which follows a regression

model $Y^* = \boldsymbol{x}^T\boldsymbol{\beta} - \varepsilon$, where ε is a noise variable with continuous distribution function $F(.)$. Instead of observing Y^* one observes a coarser categorical version determined by $Y = r \Leftrightarrow \theta_{r-1} \leq Y^* \leq \theta_r$, where $-\infty = \theta_0 < \theta_1 < ... < \theta_k = \infty$ are thresholds on the latent scale. Simple derivation yields the *cumulative model*

$$P(Y \geq r|\boldsymbol{x}) = F\left(\beta_{0r} + \boldsymbol{x}^T\boldsymbol{\beta}\right), \quad r = 1, ..., k, \tag{1}$$

where $\beta_{0r} = -\theta_{r-1}$. The class of cumulative models comprises many alternative models since $F(.)$ can be chosen as any strictly increasing distribution function.

Alternatively the cumulative model can be seen as a collection of binary response models. It is equivalent to postulating that for the split variables the models

$$P(Y_r = 1|\boldsymbol{x}) = F\left(\beta_{0r} + \boldsymbol{x}^T\boldsymbol{\beta}\right), \quad r = 1, ..., k \tag{2}$$

hold. Thus, the cumulative model can be constructed from models (2). Conversely, if the cumulative model holds one obtains the binary models for the split variables. In both cases one has to postulate that intercepts are ordered, that is, $\beta_{0k} < ... < \beta_{01} = \infty$. It is crucial that the binary response models (2) have to hold *simultaneously* for the binary responses Y_2, ..., Y_k. The main link between the binary models is that one assumes that the effect of explanatory variables captured in $\boldsymbol{x}^T\boldsymbol{\beta}$ is the same in all of the models.

The link to the split variables makes explicit that the cumulative model essentially *compares groups of categories*. It is a model that simultaneous compares categories that result from splitting the categories into {1, ..., $r-1$} and {r, ..., k}. In the construction no reference to latent variables is needed.

The most widely used model is the so-called *proportional odds model*, which uses the logistic distribution $F(\eta) = \exp(\eta)/(1 + \exp(\eta))$, yielding

$$\text{logit}P(Y \geq r|\boldsymbol{x}) = \theta_r + \boldsymbol{x}^T\boldsymbol{\beta}.$$

2.2 | Sequential models

Sequential models can be derived from the assumption that 1, ..., k are reached successively. They reflect the successive transition to higher categories in a stepwise fashion. The *sequential model* has the form

$$P(Y \geq r \mid Y \geq r-1, \boldsymbol{x}) = F\left(\beta_{0r} + \boldsymbol{x}^T\boldsymbol{\beta}\right), \quad r = 2, ..., k, \tag{3}$$

where $F(.)$ is again a distribution function.

The model can be seen as a step model with steps representing the transition to higher categories. Let, the process start in category 1. The decision between category {1} and categories {2, ..., k} is determined in the first step by a dichotomous response model $P(Y \geq 2|\boldsymbol{x}) = F(\beta_{01} + \boldsymbol{x}^T\boldsymbol{\beta})$. If $Y = 1$, the process stops. If $Y \geq 2$, the second step is a decision between category {2} and categories {3, ..., k} and is determined by $P(Y \geq 3 \mid Y \geq 2, \boldsymbol{x}) = F(\beta_{02} + \boldsymbol{x}^T\boldsymbol{\beta})$. In general in the rth step the decision between category {r} and categories {$r+1$, ..., k} is modeled by the binary model given in Equation (3). In addition to the stepwise modeling it is only assumed that the decision between the category reached and higher categories is determined by the same binary model that has response function $F(.)$. The collection of steps yields the sequential model.

The sequential character of the model is visualized in Figure 1, which shows the successive steps for four categories. The first step distinguishes between categories {1} and {2,3,4}. Given the choice was in favor of the categories {2,3,4}, in the next step it is distinguished between categories {2} and {3, 4}. The splitting continues until end nodes are reached. An essential characteristic of the model is that it is *conditional*, it uses binary models as building blocks to specify the occurrence of the event $Y \geq r$ *given* $Y \geq r-1$.

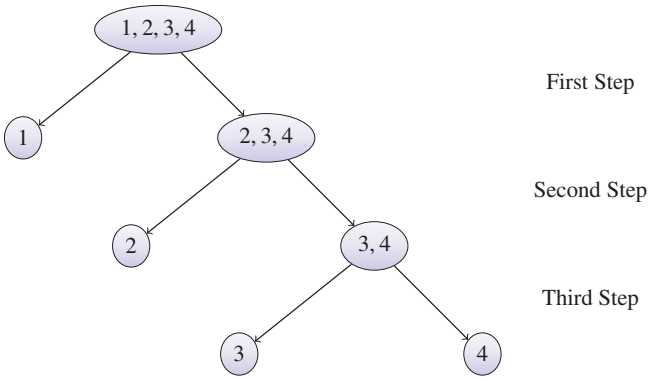


FIGURE 1 The sequential model as a hierarchically structured model

The model can also be represented by split variables. It is easily seen that the model is equivalent to assuming that for the split variables the binary models

$$P(Y_r = 1 \mid Y_{r-1} = 1, ..., Y_2 = 1, \boldsymbol{x}) = F(\beta_{0r} + \boldsymbol{x}^T\boldsymbol{\beta}), \quad r = 2, ..., k,$$

hold. Since the binary variables are from a Guttman space the models can also be given by

$$P(Y_r = 1 \mid Y_{r-1} = 1, \boldsymbol{x}) = F(\beta_{0r} + \boldsymbol{x}^T\boldsymbol{\beta}), \quad r = 2, ..., k, \tag{4}$$

which uses the simpler condition $Y_{r-1} = 1$. Both representations (3) and (4) show that the binary models contained in the sequential model are *conditional models*, in contrast to the binary models contained in cumulative type models.

The most prominent member of the family of models is the *continuation ratio model*, which results from using the logistic distribution function,

$$\log\left(\frac{P(Y \geq r-1 \mid \boldsymbol{x})}{P(Y = r \mid \boldsymbol{x})}\right) = \beta_{0r} + \boldsymbol{x}^T\boldsymbol{\beta}.$$

The model directly compares the categories {r , ..., k} to category r-1.

Sequential models were considered by McCullagh (1980), Läärä and Matthews (1985), Armstrong and Sloan (1989), Tutz (1991), and Ananth and Kleinbaum (1997). Their strong link to discrete survival modeling is investigated, for example, in Tutz and Schmid (2016).

2.3 | Adjacent categories models

The *adjacent categories model* has the basic form

$$P(Y \geq r \mid Y \in \{r-1, r\}, \boldsymbol{x}) = F(\beta_{0r} + \boldsymbol{x}^T\boldsymbol{\beta}), \quad r = 2, ..., k. \tag{5}$$

Since $P(Y \geq r \mid Y \in \{r-1, r\}, \boldsymbol{x}) = P(Y = r \mid Y \in \{r-1, r\}, \boldsymbol{x})$ it specifies the probability of observing category r given the response is in categories $\{r-1, r\}$. The split variables representation of the model is given by

$$P(Y_r = 1 \mid Y_{r-1} = 1, Y_{r+1} = 0, \boldsymbol{x}) = F(\beta_{0r} + \boldsymbol{x}^T\boldsymbol{\beta}), \quad r = 2, ..., k. \tag{6}$$

The binary response models, which are contained in the adjacent category model, are *conditional* models. The same holds for the sequential model, however, in contrast to the sequential model the adjacent categories model is not a

hierarchical model. That means that the choice between categories is not determined in a hierarchical way that can be represented as a tree as is possible for the sequential model (Figure 1). The reason is that the conditions in the binary models are overlapping, they contain, for example, $Y \in \{1, 2\}$ and $Y \in \{2, 3\}$, which share the response category 3.

The most widespread model is again the model that uses the logistic distribution function. Then logits are built locally for adjacent categories of the form

$$\log\left(\frac{P(Y=r|\boldsymbol{x})}{P(Y=r-1|\boldsymbol{x})}\right) = \beta_{0r} + \boldsymbol{x}^T\boldsymbol{\beta}, \quad r = 2,...,k, \tag{7}$$

yielding the probabilities

$$P(Y=r|\boldsymbol{x}) = \frac{\exp\left(\sum_{l=2}^{r}\{\beta_{0l} + \boldsymbol{x}^T\boldsymbol{\beta}\}\right)}{\sum_{s=1}^{k}\exp\left(\sum_{l=2}^{s}\{\beta_{0l} + \boldsymbol{x}^T\boldsymbol{\beta}\}\right)}, \quad r = 1,...,k.$$

The representation (7) shows that the model directly compares adjacent categories. The model can also be seen as a submodel of the nominal multinomial logit model

$$\log\left(\frac{P(Y=r|\boldsymbol{x})}{P(Y=1|\boldsymbol{x})}\right) = \gamma_{0r} + \boldsymbol{x}^T\boldsymbol{\gamma}_r, \quad r = 1,...,k, \tag{8}$$

where $\gamma_{01} = 0$, $\boldsymbol{\gamma}_1^T = (0,...,0)$. If the general multinomial logit model (8) holds one can derive

$$\log\left(\frac{P(Y=r|\boldsymbol{x})}{P(Y=r-1|\boldsymbol{x})}\right) = \beta_{0r} + \boldsymbol{x}^T\boldsymbol{\beta}_r, \quad r = 2,...,k, \tag{9}$$

where $\boldsymbol{\beta}_r = \boldsymbol{\gamma}_r - \boldsymbol{\gamma}_{r-1}$, $\beta_{0r} = \gamma_{0r} - \gamma_{0,r-1}$. One obtains the adjacent categories model (7) by assuming $\boldsymbol{\beta}_2 = ... = \boldsymbol{\beta}_k = \boldsymbol{\beta}$. This assumption implies that the model uses the ordering of the categories. The more general model (9) does not use the order, it simply is an alternative representation of the nominal multinomial logit model. More recently, Dolgun and Saracbasi (2014) investigated under which conditions the dependence of the parameters on the categories can be simplified to $\boldsymbol{\beta}_2 = ... = \boldsymbol{\beta}_k$. A slightly weaker assumption is $\boldsymbol{\beta}_r = \alpha_r\boldsymbol{\beta}$, $r = 2, ..., k$, where $\alpha_r > 0$ are scaling constants. The model is equivalent to Anderson's stereotype model (Anderson, 1984), which was also considered by Greenland (1994), and, more recently, by Fernandez et al. (2019). The adjacent categories logit model may also be considered as the corresponding regression model that is obtained from the row-column (RC) association model considered by Goodman (1981a, 1981b), Kateri (2014).

2.4 | Common structure

The characterization of models by the binary models that are embedded and form the building blocks yields a simple (preliminary) taxonomy of models. While the binary models contained in the cumulative type model are unconditional models for groups of categories, the adjacent categories models compare categories conditionally and the sequential models compare categories and groups of categories conditionally. By focusing on the conditioning one obtains the structure given in Figure 2.

Table 1 shows the form the models have for the ordinal response Y in k categories as well as for the split variables. The last column shows the representation of the logistic models, which have a particularly simple form. It should be noted that all models are represented in a way that the increase in $\boldsymbol{x}^T\boldsymbol{\beta}$ means a tendency to higher categories. Alternative representations are in common use, in particular the cumulative model is often given in the form $P(Y \le r) = F(\beta_{0r} + \boldsymbol{x}^T\boldsymbol{\beta})$.

One should be aware that the structure shown in Figure 2 actually represents families of models since one can use quite different link functions in the binary models. Most common choices are the logit and probit link, but also minimum and maximum extreme value distributions have been used, see, for example, Agresti (2010), Tutz (2012), and Christensen (2015). Peyhardi et al. (2015) gave a careful investigation of the relationship among ordinal models with

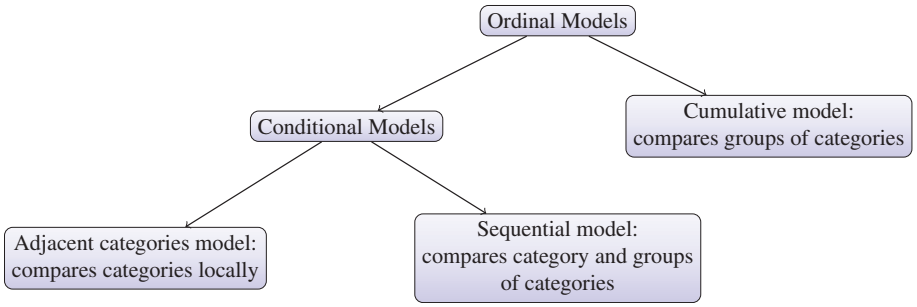


FIGURE 2 Structure of ordinal latent trait models

TABLE 1 Overview of ordinal models

	Conditional representation using response $\boldsymbol{P(.) = F(\beta_{0r} + x^T\beta)}$	**Conditional representation using split variables** $\boldsymbol{P(.) = F(\beta_{0r} + x^T\beta)}$	**Category representation logistic version** $\boldsymbol{\log(.) = \beta_{0r} + x^T\beta}$
Cumulative	$P(Y \geq r)$	$P(Y_r = 1)$	$\log\left(\frac{P(Y \geq r)}{P(Y < r)}\right)$
Adjacent categories	$P(Y = r \mid Y \in \{r-1, r\})$	$P(Y_r = 1 \mid Y_{r-1} = 1, Y_{r+1} = 0)$	$\log\left(\frac{P(Y = r)}{P(Y = r-1)}\right)$
Sequential	$P(Y \geq r \mid Y \geq r-1)$	$P(Y_r = 1 \mid Y_{r-1} = 0)$	$\log\left(\frac{P(Y \geq r)}{P(Y = r-1)}\right)$

different link functions and derived invariance properties for the models. For simplicity, in applications we use the logistic versions only.

2.4.1 | Interpretation of parameters

The interpretation of parameters depends on the model that is used. By using the representation as binary models one can use the interpretation from these (conditional) models. Interpretation becomes especially simple for the logistic version of the models. Then the parameter of the j-th variable, β_j, is directly linked to the change in specific odds if x_j increases by one unit. The corresponding odds are:

- the cumulative odds $\gamma_r(\boldsymbol{x}) = \frac{P(Y \geq r|\boldsymbol{x})}{P(Y < r|\boldsymbol{x})}$, which compare the categories $\{r, ..., k\}$ to $\{1, ..., r-1\}$, in the cumulative logistic model,
- the conditional sequential odds $\gamma_r(\boldsymbol{x}) = \frac{P(Y \geq r|Y \geq r-1, \boldsymbol{x})}{P(Y = r-1|Y \geq r-1, \boldsymbol{x})} = \frac{P(Y \geq r|\boldsymbol{x})}{P(Y = r-1|\boldsymbol{x})}$, which compare the categories $\{r, ..., k\}$ to the category $r-1$, in the sequential logistic model (continuation ratio model),
- the local odds $\gamma_r(\boldsymbol{x}) = \frac{P(Y = r|Y \in \{r-1, r\}, \boldsymbol{x})}{P(Y = r-1|Y \in \{r-1, r\}, \boldsymbol{x})} = \frac{P(Y = r|\boldsymbol{x})}{P(Y = r-1|\boldsymbol{x})}$, which compare the categories r and $r-1$, in the adjacent categories logistic model.

More concise, if variable x_j is increased by one unit the odds $\gamma_r(\boldsymbol{x})$ change by the factor e^{β_j} when all other variables are kept fixed. Thus, e^{β_j} can be directly interpreted as the odds ratio that compares the corresponding odds with value $x_j + 1$ in the j-th variable to the odds with value x_j in the j-th variable, when all other variables are kept fixed,

$$e^{\beta_j} = \frac{\gamma_r(x_1, ..., x_j + 1, ..., x_p)}{\gamma_r(x_1, ..., x_j, ..., x_p)}.$$

Interpretation is simplified by the fact that the change in odds ratio does not depend on r, variables change all odds in the same way. For example, in the adjacent categories model the change in the odds $P(Y = r|\boldsymbol{x})/P(Y = r-1|\boldsymbol{x})$ obtained by increasing the j-th variable by one unit is the same for all pairs of adjacent categories. For more on the interpretation of effects and odds see, for example, McCullagh (1980), Agresti (2010), and Tutz (2012).

It has to be emphasized that the odds that are used to interpret parameters are conditional odds for sequential and adjacent categories models. If, for example, the odds used in the sequential model are presented as $P(Y \geq r|\boldsymbol{x})/P(Y = r - 1|\boldsymbol{x})$ the conditioning is hidden, however, interpretation refers to the transition to higher categories *given* specific categories have been reached, the corresponding odds are the *local* odds, $P(Y \geq r \mid Y \geq r-1, \boldsymbol{x})/P(Y = r-1 \mid Y \geq r-1, \boldsymbol{x})$.

For models with other link functions than the logistic one can use alternative measures that are often simpler to interpret than the model parameters themselves, see Agresti and Kateri (2017) and Agresti and Tarantola (2018). One approach to investigate the effect of explanatory variables that has been propagated in particular in the social sciences uses so-called marginal effects. One can, for example, examine the effect of a quantitative variable x_j by considering the rate of change for specific categories at particular values of x_j, that is, one considers $\partial P(Y = r|\boldsymbol{x} = \boldsymbol{x}^*)/\partial x_j$, which is the rate of change of category r when other variables are fixed at certain values $\boldsymbol{x}^*$ (Agresti & Tarantola, 2018). Although usually referred to as marginal effect it is rather a conditional effect because of the conditioning on $\boldsymbol{x}^*$. Consequently, curves that show the rate of change depend on the chosen value of the other variables. One way to deal with this problem is to compute the marginal effect with every explanatory variable set at its mean to obtain the *marginal effect at the mean*, or to compute the marginal effect at each of the sample values and then averaging them to obtain the *average marginal effect*. More details on marginal effects measures are given in Long (1997), Long and Freese (2006), Williams (2012). Long and Mustillo (2018) showed how this approach can be used to compare groups in binary regression models.

2.5 | Illustrating application

For illustration, we use data from the German Longitudinal Election Study, which is a long-term study of the German electoral process (Rattinger et al., 2014). The data consist of 2036 observations and originate from the pre-election survey for the German federal election in 2017 and are concerned with political fears. In particular, the participants were asked: "How afraid are you due to the use of nuclear energy? The answers were measured on Likert scales from 1 (not afraid at all) to 7 (very afraid). The explanatory variables in the model are *Abitur* (high school leaving certificate, 1: Abitur/A levels; 0: else), *Age* (age of the participant), *EastWest* (1: East Germany/former GDR; 0: West Germany/former FRG), *Gender* (1: female; 0: male), *Unemployment* (1: currently unemployed; 0: else).

Table 2 shows the estimated parameters of the logit versions of the three models. It is seen that although parameter estimates are quite different in all three models the same variables are found to have an impact on the response. The log-likelihood values are −3,772.140 (cumulative model), −3,772.692 (adjacent categories model), and −3,776.712 (sequential model), that means, in terms of the goodness of fit the models are well comparable, in particular the cumulative and the adjacent categories model show almost the same fit.

It is not uncommon that all three models yield similar results in terms of goodness of fit and relevance of explanatory variables. The main difference is that they use differing odds but the simple linear structure in the explanatory term is the same. If the goodness of fit is comparable one can choose the odds that one prefers for the interpretation. In particular, the odds for the cumulative and the adjacent categories model, which tend to yield very similar results, are simple to explain to practitioners. The sequential model is somewhat different, since its odds refer to a sequence of transitions, which is appropriate in particular if the response might be seen as the result of a process, for example, in categorized time of unemployment where long-term unemployment can only be observed if the person was previously short-term unemployed. In cases like that one observes the end of a process.

3 | ACCOUNTING FOR ADDITIONAL HETEROGENEITY

The basic models considered in the previous section are simple to apply but often do not meet the requirements of the modeling task at hand. The main reason is that, the parameterization focuses on location but ignores potential heterogeneity in the population. One specific form of heterogeneity is unobserved dispersion, which, if ignored, may yield strongly misleading results. In the following approaches to model heterogeneity are considered. We will use the general framework of response styles to embed more traditional approaches as well as the approaches that have been proposed more recently. The models used are from the given taxonomy but the focus is on parameterization, which is modified to address potential heterogeneity. We first consider the simplest case, namely binary regression. The problems that may arise if heterogeneity is ignored are already seen from this simple case, and it is useful for generalizations.

TABLE 2 Estimates of logit versions of ordinal regression models for fears of the use of nuclear energy

	Cumulative model				Adjacent categories			
	Estimate	**Std. error**	***z* value**	**Pr (>|z|)**	**Estimate**	**Std. error**	***z* value**	**Pr (>|z|)**
Age	0.0162	0.0021	7.492	0.0000	0.0049	0.0007	6.946	0.0000
Gender	0.5819	0.0789	7.368	0.0000	0.1967	0.0260	7.559	0.0000
Unemployment	−0.0290	0.2483	−0.117	0.9077	−0.0139	0.0794	−0.176	0.8602
EastWest	−0.5144	0.0845	−6.087	0.0000	−0.1671	0.0274	−6.089	0.0000
Abitur	−0.0456	0.0813	−0.561	0.5755	−0.0190	0.0264	−0.720	0.4716

Sequential model				
	Estimate	**Std. error**	***z* value**	**Pr (>|z|)**
Age	0.0136	0.0016	8.460	0.0000
Gender	0.3620	0.0585	6.183	0.0000
Unemployment	0.1241	0.1844	0.673	0.5008
EastWest	−0.3180	0.0624	−5.090	0.0000
Abitur	−0.0457	0.0604	−0.757	0.4491

3.1 | Ignoring variance heterogeneity

Let a latent regression model be given by $Y_i^* = \beta_0 + \boldsymbol{x}_i^T\boldsymbol{\beta} - \sigma_i\varepsilon_i$, where σ_i now depends on the specific observation i. In the simplest case one has $\sigma_i = \exp(z_i\gamma)$, where z_i is an indicator variable, which takes the value one for group 1 (e.g., males) and the value zero for group 0 (e.g., females). By assuming that the observed response is determined by $Y_i = 1$ if $Y_i^* \geq 0$ one obtains the binary response model

$$\begin{aligned} P(Y_i = 1|\boldsymbol{x}_i) &= F\left(\beta_{0r}/\sigma + \boldsymbol{x}_i^T(\boldsymbol{\beta}/\sigma)\right) \quad \text{for observations from group 1, and} \\ P(Y_i = 1|\boldsymbol{x}_i) &= F\left(\beta_{0r} + \boldsymbol{x}_i^T\boldsymbol{\beta}\right) \qquad \text{for observations from group 0.} \end{aligned} \tag{10}$$

Although the effects are the same in the underlying regression model, effects of covariates in the binary models differ between the groups. One has $\boldsymbol{\beta}/\sigma$ in group 1 and $\boldsymbol{\beta}$ in group 0. If, for example, $\sigma = 0.5$ the effect strength in the binary model in group 1 is twice the effect strength in group 0. The dependence on the group is simply ignored if one sets $\sigma_i = 1$, which is typically assumed in categorical regression to obtain identifiability. It means that in both groups the same scaling is used, although different ones are needed.

Thus, ignoring variance heterogeneity may yield rather misleading effect strengths when comparing parameters. The effect has been demonstrated in particular by Allison (1999) who considered an example with the binary response being the promotion to an associate professor from the assistant professor level. It turned out that the number of published articles had a much stronger effect for male researchers than for female researchers, which seems rather unfair, but could be due to variance heterogeneity.

The distortion of effects is not restricted to binary models, it occurs also in ordinal models. In addition, the problem does not arise only when effects of separate fits are compared, it is also present if one fits a closed model in which group effects are included as covariates. Allison's article stimulated extensive research that aimed at avoiding such misleading results, see, for example, Williams (2009), Mood (2010), Karlson et al. (2012), Breen et al. (2014), Rohwer (2015), and Tutz (2019).

3.2 | Modeling variance heterogeneity: The location-scale model

One way to introduce variance heterogeneity is to model it explicitly as depending on covariates. In the cumulative type model this approach has been used by McCullagh (1980). If the underlying latent variable is given by $Y^* = \boldsymbol{x}^T\boldsymbol{\beta}$ - $\sigma\varepsilon$ with $\sigma = \exp(\boldsymbol{z}^T\boldsymbol{\gamma})$, where $\boldsymbol{z}$ is an additional vector of covariates, and one assumes the category boundaries approach, that is, $Y = r \Leftrightarrow \theta_{r-1} \leq Y^* \leq \theta_r$ one obtains the *location-scale model*

$$P(Y \geq r|\boldsymbol{x}, \boldsymbol{z}_i) = F\left(\frac{\beta_{0r} + \boldsymbol{x}^T\boldsymbol{\beta}}{\exp(\boldsymbol{z}^T\boldsymbol{\gamma})}\right). \tag{11}$$

The model contains two terms that specify the impact of covariates, the location term $\beta_{0r} + \boldsymbol{x}^T\boldsymbol{\beta}$ and the variance or scaling term $\exp(\boldsymbol{z}^T\boldsymbol{\gamma})$, which derives from the "variance equation" $\sigma = \exp(\boldsymbol{z}^T\boldsymbol{\gamma})$. If $\boldsymbol{x}$ and $\boldsymbol{z}$ are distinct the interpretation of the $\boldsymbol{x}$-variables is the same as in the proportional odds model.

The location-scale model was introduced by McCullagh (1980) and considered by Nair (1987) and Hamada and Wu (1990). In the social sciences, the model is also known as *heterogeneous choice model* or *heteroscedastic logit model* (Alvarez & Brehm, 1995; Williams, 2009). The logistic version is also related to the logistic response model with proportionality constraints proposed by Hauser and Andrew (2006) and extended by Fullerton and Xu (2012).

It is important to note that the scaling component does not necessarily represent variance heterogeneity. If one derives the model from an underlying continuous regression model one typically thinks of variance heterogeneity, however, the parameters of the proportional odds model can also be interpreted without reference to an underlying continuous response. For simplicity, let us consider the case where $z \in \{0, 1\}$ is binary denoting two groups (male/female). Then the effect strength of the j-th variable in the location scale model is given by β_j/e^{γ} for observations from group 1 and β_j for observations from group 0. That means gender is an *effect modifying variable*. It changes the effect of the covariates as a function of gender without any reference to an underlying continuous response. It can easily be embedded into the general framework of varying coefficient models proposed by Hastie and Tibshirani (1993).

3.3 | Modeling heterogeneity: The location-shift model

An alternative way to model heterogeneity, which has some advantages, specifies that covariates modify the thresholds. Instead of allowing the variance in the underlying continuous response to vary across groups of individuals one assumes that the intercepts, which refer to thresholds, vary across individuals.

Let the intercept (or thresholds) β_{0r} in the cumulative model (1) be replaced by

$$\beta_{0r} + (k/2 - r + 1)\boldsymbol{z}^T\boldsymbol{\alpha}, \tag{12}$$

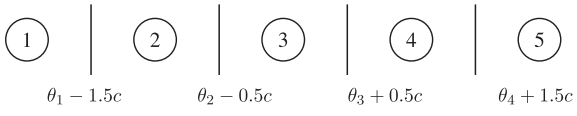
to obtain the *location-shift model*

$$P(Y \geq r|\boldsymbol{x}) = F\left(\beta_{0r} + (k/2 - r + 1)\boldsymbol{z}^T\boldsymbol{\alpha} + \boldsymbol{x}^T\boldsymbol{\beta}\right), \quad r = 2, \ldots, k, \tag{13}$$

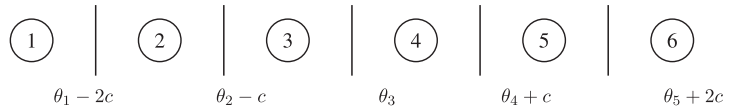
which (in a more elementary form) has been proposed by Tutz and Berger (2017). The scaling $(k/2 - r + 1)$ in the linear term $(k/2 - r + 1)\boldsymbol{z}^T\boldsymbol{\alpha}$ is chosen such that the difference between adjacent predictors $\eta_r = \beta_{0r} + (k/2 - r + 1)\boldsymbol{z}^T\boldsymbol{\alpha} + \boldsymbol{x}^T\boldsymbol{\beta}$ has the form

$$\eta_r - \eta_{r+1} = \beta_{0r} - \beta_{0,r+1} + \boldsymbol{z}^T\boldsymbol{\alpha}.$$

That means the difference between adjacent predictors is widened or shrunk by $\boldsymbol{z}^T\boldsymbol{\alpha}$. The effect of adding $\boldsymbol{z}^T\boldsymbol{\alpha}$ becomes obvious if one considers the thresholds on the latent variables. The thresholds are given by $\theta_r = -\beta_{0,r+1}$. Adding the term yields the new thresholds $\tilde{\theta}_r = \theta_r - (k/2 - r + 1)\boldsymbol{z}^T\boldsymbol{\alpha}$, and differences $\tilde{\theta}_r - \tilde{\theta}_{r-1} = \theta_r - \theta_{r-1} + \boldsymbol{z}^T\boldsymbol{\alpha}$. For illustration, let us consider concrete examples with odd and even numbers of categories. For $k = 5$ one obtains with $c = \boldsymbol{z}^T\boldsymbol{\alpha}$ the thresholds.



For $k = 6$ one obtains the thresholds.



If $z \in \{0, 1\}$ is a simple indicator variable, yielding $\boldsymbol{z}^T\boldsymbol{\alpha} = z\alpha$, one obtains for positive α that the difference between adjacent categories are widened by α in group 1 indicating more concentration in middle categories than in group 0. For negative α the difference is shrunk indicating more concentration in the extreme categories in group 1. Therefore, the parameter α indicates a tendency to middle categories. It explicitly models how covariates change the response behavior of respondents concerning a tendency to middle or extreme categories with regard to the tendency to extreme or middle categories.

Typically large α and therefore more concentration in middle categories means smaller variation of responses, and small α, with more concentration in extreme categories, means higher dispersion. Thus, Tutz and Berger (2017) described the heterogeneity effects in the location-shift model as dispersion effects, which was supported by applications in which the location-scale model and the location-shift model showed comparable goodness of fit. However, the modeled effects are not exactly the same, differences in interpretation will be investigated later.

Location-scale and location-shift models use different parameterizations but both model types aim at separating the location from response behavior that is not determined by location. While the location-scale model uses a multiplicative structure (motivated by variance heterogeneity in the underlying continuous response) yielding a dispersion effect, the location-shift model uses an additive structure (motivated by the shifting of thresholds) yielding a tendency to middle or extreme categories.

The location-shift model has several advantages over the location-scale model. It is a (multivariate) generalized linear model. Therefore, all the inference techniques, including diagnostic tools and asymptotic results that have been shown to hold for this class of models can be used. Also selection of variables can be done within that framework by using regularization methods as the lasso. In contrast, the location-scale model with its multiplicative structure is not a generalized model and such general inference tools seem not yet available. More important than the finer differences between the modeling approaches is that one should account for potential heterogeneity in some form since otherwise results may be untrustworthy. In particular,

ignoring variance heterogeneity might yield spurious effects of explanatory variables as described in Section 3.1,
estimates of the location parameters can be strongly biased if heterogeneity is present but not modeled (see simulation results in Tutz and Berger(2017)),
modeling of heterogeneity provides additional information about the effects of covariates, they show, for example, which groups of persons have strong variability, and which groups can be considered as homogeneous, see the following example.

3.4 | Nuclear fear data

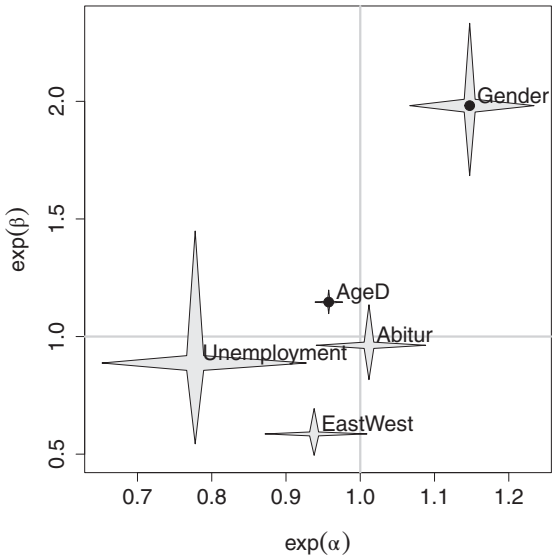
Table 3 shows the estimates of the location-scale and location-shift model for the nuclear fear data. The upper part shows the effects of variables in the location term, the lower part shows the parameters in the heterogeneity term. The names of the variables in the latter term are appended by"H." It is seen that heterogeneity effects should not be neglected. At least three of the explanatory variables show significant effects. For example, the estimates for the location-scale model show that females tend to have lesser dispersion than men, also younger people show lesser variability than older people in both models. In accordance, the location-shift model shows that females have a stronger tendency to middle categories than men, and the same holds when older people are compared to younger ones. The p-value of the heterogeneity variable EastWest is smaller in the location-scale model than in the location-shift model, but in both models it is not far from 0.05. It should be noted that the estimates of the location term have changed by including heterogeneity effects. Comparison with Table 2 shows that effects are stronger if one accounts for heterogeneity, the increase is more distinct in the location-scale model.

A visualization of the parameter estimates of the location-shift model can be obtained by using star plots (Tutz & Berger, 2016). Figure 3 shows the tuples $\left(e^{\hat{\alpha}}, e^{\hat{\beta}}\right)$ for the linear effects of the location-shift model. The first value, $e^{\hat{\alpha}}$, represents the heterogeneity effect on the odds, for values larger than one there is a tendency to middle categories, for values smaller than one there is a stronger tendency to extreme categories than in the simple proportional odds model.

TABLE 3 Estimates of location-scale model and location-shift model for nuclear energy fear data

	Location scale-model				Location shift-model			
	Estimate	**Std. error**	**z value**	**Pr (>|z|)**	**Estimate**	**Std. error**	**z value**	**Pr (>|z|)**
Age	0.0201	0.0030	6.670	0.0000	0.0136	0.0022	−6.115	0.0000
Gender	0.7467	0.1115	6.697	0.0000	0.6839	0.0830	−8.236	0.0000
Unemployment	−0.1159	0.4307	−0.269	−0.7881	−0.1193	0.2500	−0.477	0.6331
EastWest	−0.6152	0.1162	−5.294	0.0000	−0.5344	0.0862	6.196	0.0000
Abitur	−0.0656	0.0996	−0.659	0.5102	−0.0378	0.0840	0.451	0.6519
AgeH	0.0054	0.0011	4.573	0.0000	−0.0043	0.0009	−4.375	0.0000
GenderH	−0.1603	0.0438	−3.659	0.0002	0.1374	0.0371	3.697	0.0002
UnemploymentH	0.3477	0.1540	2.257	0.0239	−0.2514	0.0898	−2.800	0.0051
EastWestH	0.0982	0.0465	2.111	0.0347	−0.0641	0.0373	−1.717	0.0859
AbiturH	−0.0188	0.0443	−0.426	0.6702	0.0116	0.0371	0.314	0.7535

FIGURE 3 Effect stars for location-shift model



The second value, $e^{\hat{\beta}}$, represents the location effect on the odds. For values larger than one high response categories are favored, for values smaller than one low response categories are favored. Also pointwise 95% confidence intervals are included, represented by stars with the horizontal and vertical lengths corresponding to the confidence intervals of $e^{\hat{\alpha}}$ and $e^{\hat{\beta}}$, respectively. If stars cross the horizontal or/and vertical lines the corresponding effects cannot be considered as significant. For example, the star for Abitur crosses both lines since both effects are not significant, while the star for Age is very small and far away from both lines, indicating strong significance. It should be noted that the plotted effect of Age refers to age measured in decades (age/10), otherwise the star would be too close to the "zero effect" point (1,1).

3.5 | Generalized models and response styles

A general framework, in which the modeling of heterogeneity can be embedded, is the modeling of response styles. Response styles have been described as a consistent pattern of responses that is independent of the content of a response (Johnson, 2003). They are commonly used to describe an individual's tendency to choose a certain kind of response category, for example extreme or middle categories, irrespective of the content-related response. The heterogeneity considered here may be seen as response style in a wider sense. It does not describe specific response behavior on

the individual level but on the group level (or determined by a linear term). Moreover, it applies in regression settings, in which just one observation per person is available.

General treatments of response styles are found in particular in the social sciences and item response theory, where more than one measurements per person are available. An overview was given by Van Vaerenbergh and Thomas (2013), see also Messick (1991), Baumgartner and Steenkamp (2001), Bolt and Newton (2011), and Johnson (2003).

In the next sections we investigate which choice behavior is modeled by specific heterogeneity effect terms by considering extreme values of parameters. It will be seen that the modeled behavior depends on the model that is used. Different model types can yield different response styles for the same term.

3.5.1 | Extensions of cumulative models

Let us again consider the *location-scale model*, which contains the term $\exp(\boldsymbol{z}^T\boldsymbol{\gamma})$ in the denominator of the predictor. The essential modifying term is the "standard deviation" given by $c = \boldsymbol{z}^T\boldsymbol{\gamma} = z_1\gamma_1 + ... + z_m\gamma_m$. In particular, extreme values show what exactly is implied by the inclusion of this term. It can be shown that for $c \to -\infty$ the probability for one of the response categories becomes 1, all other probabilities become zero. If $c \to \infty$ one obtains full concentration in the extreme categories, more concrete, one gets for symmetric distribution functions $P(Y = 1) = P(Y = k) = 0.5$. That means, for positive z_j large parameters γ_j indicate a tendency to a distinct response while small parameters γ_j indicate a tendency to extreme categories. If, for example, z_j is an indicator variable with $z_j = 1$ representing female responders positive values γ_j indicate that females have more distinct preferences than males. Females tend to choose specific categories in rating scale responses, with the categories that are chosen being determined by the location term, while males tend to choose one of the extreme categories. Thus, a response style is modeled that covers the continuum between a distinct choice and a choice of extreme categories. The response style is in accordance with the dispersion concept, small dispersion means a deliberate choice while large dispersion means extreme categories, however, no reference to latent traits is needed if one sees it as modeling of response behavior that is characterized by a distinct choice or strong variability.

It should be noted that the concept of response styles used here is not restricted to questionnaires where people consciously choose one of the categories. If, for example, the categories refer to income brackets and $z_j = 1$ again represents females, large parameters γ_j indicate that females are more concentrated in specific response categories than males. Males show stronger heterogeneity, they have more probability mass in extreme categories. Thus, the response style captured in the multiplicative term refers more general to a tendency to stronger or weaker dispersion.

In the *location-shift model* the effect-modifying term is the *additive* term $c = \boldsymbol{z}^T\boldsymbol{\alpha}$. If $c \to \infty$ the probability is concentrated in the middle categories, if c is small the probability mass is in the extreme categories 1 and k. Thus, c determines the tendency to middle or extreme categories. This is a response style that is slightly different from the style obtained by the location-scale model. Both models are able to capture the tendency to extreme response categories. They differ in the other extreme, the tendency to middle categories or one specific category, which is not necessarily the middle category. Since both are able to model the tendency to extreme categories they often yield similar fits and can be used to identify which individuals show strong dispersion.

3.5.2 | Other models with response style

In the previous section, extensions of the cumulative model have been considered. But the link between ordinal models and binary models investigated in the previous section allows to construct more general ordinal models that account for heterogeneity or varying effects. In all models given in Table 1 one can replace the predictor $\eta_r = \beta_{0r} + \boldsymbol{x}^T\boldsymbol{\beta}$ by $(\beta_{0r} + \boldsymbol{x}^T\boldsymbol{\beta})/\exp(\boldsymbol{z}^T\boldsymbol{\gamma})$ to obtain a scaled version, or by $\beta_{0r} + \boldsymbol{x}^T\boldsymbol{\beta} - (k/2 - r)\boldsymbol{z}^T\boldsymbol{\alpha}$ to obtain the shift version. For example, the *scaled adjacent-categories model* has the form

$$P(Y \geq r \mid Y \in \{r-1, r\}, \boldsymbol{x}) = F\left(\frac{\beta_{0r} + \boldsymbol{x}^T\boldsymbol{\beta}}{\exp(\boldsymbol{z}^T\boldsymbol{\gamma})}\right), \quad r = 2, ..., k,$$

the *shifted adjacent-categories model* is given by

$$P(Y \geq r \mid Y \in \{r-1, r\}, \boldsymbol{x}) = F\left(\beta_{0r} + \boldsymbol{x}^T\boldsymbol{\beta} + (k/2 - r + 1)\boldsymbol{z}^T\boldsymbol{\alpha}\right), \quad r = 2, ..., k,$$

In this shifted version positive values of the term $c = \boldsymbol{z}^T\boldsymbol{\alpha}$ increase the probabilities of higher categories for $r = 1, ..., m$ but decrease them for $r = m+1, ..., k$ (k odd, m=(k+1)/2). Thus, c determines if middle categories or extreme categories are preferred. For $c = \boldsymbol{z}^T\boldsymbol{\gamma} \to \infty$ one obtains $\pi_m \to 1$ and therefore a tendency to the middle category while $c \to -\infty$ entails $\pi_2, ..., \pi_{k-1} \to 0$ and therefore a preference of the extreme categories.

In the scaled version, the modeled effect is different. One obtains for large $c = \boldsymbol{z}^T\boldsymbol{\gamma}$ that specific categories obtain very high probability whereas for c small ($c \to \infty$) one obtains the discrete uniform distribution $P(Y = 1) = ... = P(Y = k) = 1/k$. It is tedious to show in the general case which specific categories are chosen when $c \to \infty$. For illustration, we consider the simple case $k = 3$ and $\beta_{02} > \beta_{03}$. Then, for $c \to \infty$, one obtains $P(Y = 1) = 1$ if $\boldsymbol{x}^T\boldsymbol{\beta} < -\beta_{02}$, $P(Y = 2) = 1$ if $-\beta_{02} < \boldsymbol{x}^T\boldsymbol{\beta} < -\beta_{03}$, and $P(Y = 3) = 1$ if $-\beta_{03} < \boldsymbol{x}^T\boldsymbol{\beta}$. If $\beta_{02} < \beta_{03}$, one obtains $P(Y = 1) = 1$ if $\boldsymbol{x}^T\boldsymbol{\beta} < -(\beta_{03} + \beta_{02})/2$ and $P(Y = 3) = 1$ if $\boldsymbol{x}^T\boldsymbol{\beta} > -(\beta_{03} + \beta_{02})/2$. That means it depends on the value $\boldsymbol{x}^T\boldsymbol{\beta}$ and the thresholds which category is preferred. Overall, the response style in scaled versions of adjacent categories models represents a continuum between a distinct response (a person with large c prefers a specific category) and a random response, that is, each category gets the same probability. The latter has been described in the literature as *noncontingent response style*, which is found if persons have a tendency to respond carelessly, randomly, or nonpurposefully (Baumgartner & Steenkamp, 2001; Van Vaerenbergh & Thomas, 2013). It may also be seen as uncertainty in response behavior. Alternative approaches to modeling uncertainty will be considered in Section 6.

The response style contained in the shifted version of the adjacent categories model is the same as the response style in the shifted version of the cumulative model. Both are able to model the tendency to extreme categories and therefore strong dispersion. In contrast, the scaled versions are different, the response style in the cumulative model captures dispersion while the response style term in the adjacent categories model represents the degree of uncertainty. The response styles arising from the modifications of the linear predictor are summarized in Table 4. A simplified version of the shifted adjacent categories model was proposed by Tutz and Berger (2016), the scaled adjacent categories model seems to have not yet been investigated.

4 | MODELS WITH CATEGORY-SPECIFIC EFFECTS

The basic models from Section 2 can be made more flexible by using a more complex parameterization that yields models often referred to as nonproportional odds models. These models are widely used if the basic models show poor fit. Although the models seem to use quite different parameterizations than the heterogeneity modeling considered in the previous section, there are strong links between the modeling approaches, which are considered in the following.

Basic models can be extended to more flexible models by allowing for category-specific effects of explanatory variables. More concise, the linear predictor $\eta_r = \beta_{0r} + \boldsymbol{x}^T\boldsymbol{\beta}$ in the basic models is replaced by the predictor

$$\eta_r = \beta_{0r} + \boldsymbol{x}^T\boldsymbol{\beta}_r,$$

in which the effects of covariates, $\boldsymbol{\beta}_r^T = \left(\beta_{1r}, ..., \beta_{pr}\right)$, depend on r and therefore may vary across categories. Of course it is possible that only some of the variables have category-specific effects while the rest of the variables have so-called global effects, that is, effects that do not vary across categories.

In particular, extensions of the cumulative logistic model with category-specific effects have been considered in the literature. The resulting *nonproportional odds model* and *partial proportional odds model* have been investigated extensively, see, for example, by Brant (1990), Peterson and Harrell (1990), Bender and Grouven (1998), Cox (1995), Kim (2003), and Liu et al. (2009). Distinguishing between variables that are category-specific and variables that have

TABLE 4 Response styles in cumulative and adjacent categories models

	Cumulative	Adjacent categories
Location-shift ($(k/2 - r)\boldsymbol{z}^T\boldsymbol{\alpha}$)	Middle versus extreme categories	Middle versus extreme categories
Location-scale ($\boldsymbol{x}^T\boldsymbol{\beta}/e^{\boldsymbol{z}^T\gamma}$)	Distinct category versus extreme categories	Distinct category versus uniform

global effects can be obtained by tests that investigate if a specific variable is global, see, for example, Peterson and Harrell (1990). For the interpretation of effects in the generalized ordered logit models, as models with category-specific effects are often called in the social sciences, see also Williams (2016), Hedeker and Mermelstein (1998).

The nonproportional odds model typically shows a better fit to data but has some disadvantages. One of them is that the simple interpretation of parameters gets lost. Also, severe restrictions are postulated. While the simple proportional odds model only postulates the ordering of the intercepts $\beta_{02} \geq ... \geq \beta_{0k}$ the extended version postulates $\beta_{02} + \boldsymbol{x}^T\boldsymbol{\beta}_2 \geq ... \geq \beta_{0k} + \boldsymbol{x}^T\boldsymbol{\beta}_k$ for all values $\boldsymbol{x}$, which can severely restrict the possible values of explanatory variables. Even if estimates exist, in future observations with more extreme values in the explanatory variables the estimated probabilities can be negative. For problems with the model see also Walker (2016) who even argues that it is impossible to generalize the cumulative class of ordered regression models in ways consistent with the spirit of generalized cumulative regression models.

We do not think that extensions to category-specific effects are in principle useless. They can provide a better way to explore how explanatory effects determine responses and yield models that show better fit, however, we think that they are rarely useful in the general form considered above. The most relevant information can be obtained in a simpler way by using a much sparser parameterization. The basis for the simplification is the link between models with category-specific parameters and the modeling of heterogeneity.

Let us consider the shift versions of heterogeneity models, which, in the most general case $\boldsymbol{x} = \boldsymbol{z}$ uses the predictor $\beta_{0r} + \boldsymbol{x}^T\boldsymbol{\beta} + (k/2 - r + 1)\boldsymbol{z}^T\boldsymbol{\alpha}$, which can be rewritten as

$$\beta_{0r} + \boldsymbol{x}^T(\boldsymbol{\beta} + (k/2 - r + 1)\boldsymbol{\alpha}) = \beta_{0r} + \boldsymbol{x}^T\boldsymbol{\beta}_r,$$

where $\boldsymbol{\beta}_r = \boldsymbol{\beta} + (k/2 - r + 1)\boldsymbol{\alpha}$ is a category-specific effect. Thus, the location-shift model is a model with category-specific effects, but typically with a much sparser parameterization. For $k = 3$, the model with category-specific effects and the shift-version model are even equivalent since each model with category-specific parameters can be reparameterized by $\boldsymbol{\beta} = (\boldsymbol{\beta}_2 + \boldsymbol{\beta}_3)/2$, $\boldsymbol{\alpha} = \boldsymbol{\beta}_2 - \boldsymbol{\beta}_3$. While the parameter $\boldsymbol{\beta}$ contains the location effect, the parameter $\boldsymbol{\alpha}$ represents the response style effect. That means there is a reparameterization of the model with category-specific effects that has easy-to-interpret parameters.

In the general case, one obtains the simple hierarchy given in Figure 4 for the logistic cumulative model, however, the same hierarchy applies to all extended versions of the basic models. This allows to use likelihood ratio tests to investigate if the general model with category-specific effects can be reduced to the location-shift model, which typically contains much fewer parameters. With p explanatory variables the full model contains $(k - 1)(p + 1)$ parameter while the location-shift model contains $k - 1 + 2p$ parameters, which yields the difference $(k - 3)p$. If one has in a questionnaire, for example seven categories the difference is $4p$, which already for a moderate number of explanatory variables yields a much sparser model. The test itself it easily performed, one compares the likelihhood of the model with category-specific effects to the likelihood of a global effects model with two sets of predictors, $\boldsymbol{x}$ and $\boldsymbol{x}^T(k/2 - r + 1)$ (see Figure 4).

Of course, the sparser model can only be used if the tests are in favor of the restricted model. But this is the case in many applications. Typically, if the model with global effect does not fit well, it suffices to fit the shift version of the model to improve the fit distinctly (see also the application in the following). It is not necessary to use the general model with category-specific effects since tests that compare the full model and the location-shift model turn out to be nonsignificant, for various examples, see Tutz and Berger (2020). One might even conclude that the nonproportional

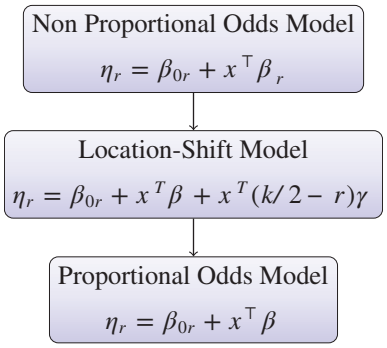


FIGURE 4 Model hierarchy

TABLE 5 Model hierarchy

	Log-likelihood	Number of parameters	Deviance	*df*
Nonproportional odds model	−3,738.21	36		
Location-shift model	−3,749.66	16	22.90	20
Proportional odds model	−3,772.69	11	46.06	5

model is widely dispensable, only in very extreme data settings it might still be useful to apply this parameter intensive model. The shift model has the additional advantage that parameters have a straightforward simple interpretation, which is not the case in the most general model with separate parameters for all categories.

4.1 | Nuclear fear data

Table 5 shows the fit of the models in the hierarchy given in Figure 4. The deviance is used to test if models may be simplified. More concrete, the given deviances are conditional deviances (or differences of deviances) that test if there is a significant difference between the sub model $\tilde{M}$ and the model M. It is seen that the difference between the location-shift model and the general nonproportional odds model is not significant (deviance 22.90 on 20 *df*). Thus, the model may be simplified to the location-shift model. Further simplification seems not warranted because the difference between the proportional odds model and the location-shift model is highly significant (deviance 46.06 on 5 *df*). Therefore, the location-shift model captures the essential structure in the data that has to be modeled (beyond the effects modeled by the proportional odds model), but the general nonproportional odds model is not needed. This is in particular fortunate since the nonproportional odds model always contains many parameter to be interpreted, in the present case 30 (without intercepts) in contrast to the location-scale model with only 10 (without intercepts), two for each variable.

5 | HIERARCHICALLY STRUCTURED MODELS

So far, the only conditional hierarchical model that has been considered is the sequential model. However, in particular for Likert-type items hierarchical models are very flexible and useful tools to obtain parsimonious parameterizations. In Likert-type items or Likert scales the categories 1, ..., k are assumed to be ordered and reflect agreement/disagreement or approval/disapproval of the respondent with respect to the value statement. In five-grade Likert scales, the grades are typically interpreted by strongly disagree, disagree, neutral (undecided), agree, and strongly agree. A specific trait of responses of that type is that response categories are naturally (by design) partitioned into groups of homogeneous categories, a partitioning that can be exploited in regression modeling.

In general, a hierarchical model is obtained by successively modeling the response in groups of response categories, with groups that typically are formed by collecting homogeneous categories. Let, the response categories $K = \{1, ..., k\}$ be subdivided into basic *ordered* sets S_1, ..., S_m, where $K = S_1 \cup ... \cup S_m$, and $r < s$ for $r \in S_j$, $s \in S_{j+1}$. In the first step one models the grouping variable defined by $Y_1 = g$ if $Y \in S_g$, for example, by using a cumulative model,

$$P(Y_1 \geq g|\boldsymbol{x}) = F\left(\boldsymbol{x}^T\boldsymbol{\beta}^{(1)}\right).$$

In the next step, the conditional responses given S_g are modeled by partitioning S_g into $S_{g1}, ..., S_{gm_g}$, where $S_g = S_{g1} \cup ... \cup S_{gm_g}$. If one uses again a cumulative model one has for the conditional response

$$P\left(Y \geq r|Y \in S_g, \boldsymbol{x}\right) = F\left(\boldsymbol{x}^T\boldsymbol{\beta}^{(S_g)}\right).$$

All subsets that contain more than one category can be subdivided further, and can be described by adding further indices to S_{gl} for each new partition.

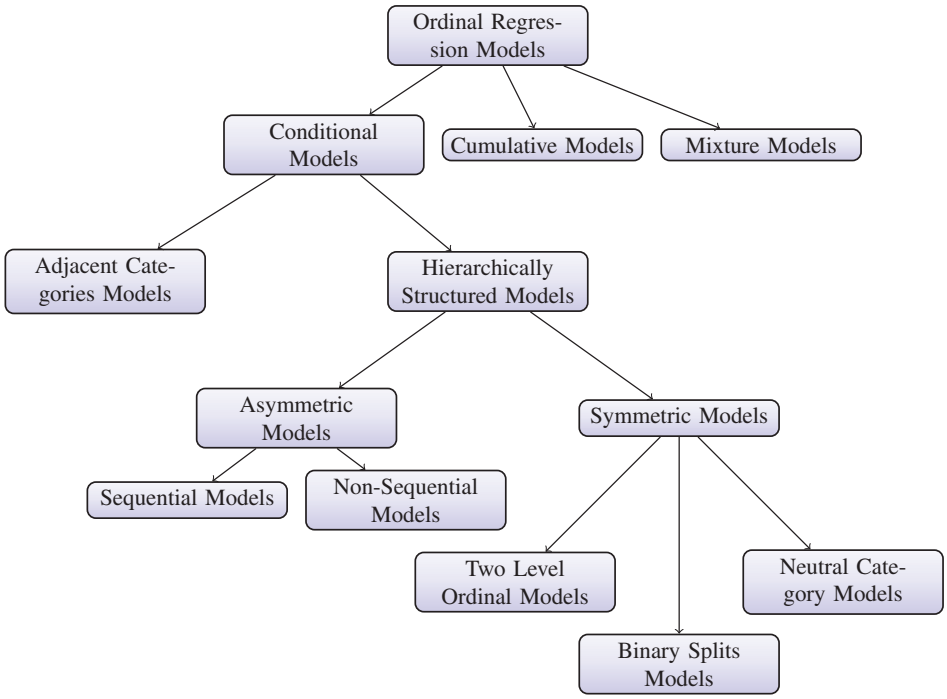


FIGURE 5 Taxonomy of ordinal models

The sequential model considered previously uses the partitioning $S_1 = \{1\}$, $S_2 = \{2, ..., k\}$, $S_{21} = \{2\}$, $S_{22} = \{3, ..., k\}$, and so on. Each set of categories is split into two subsets and a binary model is used to model the splits.

In Figure 5, we show the taxonomy of models that is obtained by including hierarchically structured models, which are specific conditional models. We distinguish between symmetric and asymmetric models. Symmetric models are defined by invariance, if a model for $Y \in \{1, ..., k\}$ is equivalent to the corresponding model for the inverse ordering, that is, for $\tilde{Y} = k + 1 - Y$, it is called a symmetric model. In the terminology of Peyhardi et al. (2015) they are invariant under permutation. The sequential model is a nonsymmetric model even if $F(.)$ is a symmetric distribution function. Thus it is found in Figure 5 as a sub model of nonsymmetric models. We do not consider alternative asymmetric models since the structuring is strongly determined by the specific application, see Peyhardi et al. (2016) for examples. Instead we consider modeling strategies that can be used for symmetrically structured data as Likert scales.

5.1 | Two level ordinal models

Let the number of categories in a Likert scale be even, and subdivided into disagreement categories $S_1 = \{1, ..., k/2\}$ and agreement categories $S_2 = \{k/2 + 1, ..., k\}$. When using these grouping, in the first step one has a binary response, and can specify

$$P(Y_1 \geq 2|\boldsymbol{x}) = P(Y \geq k/2 + 1|\boldsymbol{x}) = F\left(\beta_0^{(1)} + \boldsymbol{x}^T\boldsymbol{\beta}^{(1)}\right).$$

In the second step one can use, for example, cumulative models,

$$P\left(Y \geq r|Y \in S_g, \boldsymbol{x}\right) = F\left(\beta_{0r}^{(S_g)} + \boldsymbol{x}^T\boldsymbol{\beta}^{(S_g)}\right), g = 1, 2.$$

The hierarchical model assumes that the effect of explanatory variables on the initial decision between agreement and disagreement categories is determined by the parameter $\boldsymbol{\beta}^{(1)}$. If a respondent's answer is positive the effect of

explanatory variables on the degree of agreement is determined by $\boldsymbol{\beta}^{(S_2)}$, if the respondent's answer is negative it is $\boldsymbol{\beta}^{(S_1)}$. As Böckenholt and Meiser (2017) noted with reference to latent trait modeling, there is empirical evidence "that respondents arrive at an initial response based on retrieval processes and, subsequently, decide whether to edit this response and report a more positive or less revealing answer instead." This stage-wise process is explicitly captured by the model. It allows to investigate if the effects of explanatory variables are the same in the initial response and in the selection of a particular category. Since the models are generalized linear models one can use all the test procedures that are available for generalized linear models.

A particular strength of hierarchical models is that they are able to disentangle attitudinal measurements, that is, the location effects, from effects like dispersion or a tendency to extreme categories. It is straightforward to include response style effects by using the second level models

$$P(Y \geq r \mid Y \in S_1, \boldsymbol{x}, \boldsymbol{z}) = F\left(\beta_{0r}^{(S_1)} + \boldsymbol{x}^T\boldsymbol{\beta} + \boldsymbol{z}^T\boldsymbol{\alpha}\right),$$
$$P(Y \geq r \mid Y \in S_2, \boldsymbol{x}, \boldsymbol{z}) = F\left(\beta_{0r}^{(S_2)} + \boldsymbol{x}^T\boldsymbol{\beta} - \boldsymbol{z}^T\boldsymbol{\alpha}\right),$$

where, for simplicity, it is assumed that process parameters are the same. The crucial elements are the linear terms $\boldsymbol{z}^T\boldsymbol{\alpha}$, which are included with a positive sign for disagreement categories and a negative sign for agreement categories. That means $\boldsymbol{z}^T\boldsymbol{\alpha}$ represents the tendency to middle or extreme categories. For an uneven number of categories the basic structure remains the same, but at the first level one uses a model for three categories by grouping the categories into agreement categories, disagreement categories and the neutral category.

The model allows for a finely tuned investigation of covariate effects but often simplifies when effect strengths are the same across different stages of the process and some effects are nonsignificant. Examples of simple models with no location effect but non-neglectable dispersion were given by Tutz (1989), although the considered compound models were not embedded into the general framework of hierarchical models. Latent trait models of this form were used by Thissen-Roe and Thissen (2013).

5.2 | Symmetric binary splits model

Rather simply structured hierarchical models are obtained by using only binary splits that are symmetrically structured. A model of this type is visualized in Figure 6. In the first step the model distinguishes between agreement categories, in the second step it is modeled if the response is moderate or not (extremity I), and in the third step it is modeled if the response is extreme or not (extremity II). The models at each level are binary and have the form

$$P(Y \in S_1 \mid Y \in S, \boldsymbol{x}) = F\left(\beta_{0r}^{(S_1)} + \boldsymbol{x}^T\boldsymbol{\beta}^{(S_1)}\right), \tag{14}$$

where S is partitioned into S_1 and S_2, $S = S_1 \cup S_1$. Thus at each stage of the process one has a binary decision which separates the response in S_1 and S_2 given $Y \in S$. In latent trait modeling the binary decisions are considered to refer to mental queries and have been called pseudo items (Böckenholt and Meiser, 2017). The corresponding regression model contains many parameters if one assumes that each binary model has its own parameter. More parsimonious models are obtained by assuming that some of the parameters are identical. For example, one might assume that the tendency to more extreme categories is the same, that is, $\boldsymbol{\beta}^{(\{1,2\})} = \boldsymbol{\beta}^{(\{5,6\})}$ given {1,2,3} for $Y \in \{1, 2\}$ and {4,5,6} for $Y \in \{5, 6\}$. The most parsimonious model is obtained if one assumes $\boldsymbol{\beta}^{(S)} = \boldsymbol{\beta}$ for all S and specifies the binary models such that S_1 contains the higher categories.

Symmetric binary split models have been considered extensively in item response theory under the name *item response trees* (Böckenholt, 2017; Böckenholt & Meiser, 2017; De Boeck & Partchev, 2012; Khorramdel & von Davier, 2014; Meiser et al., 2019; Plieninger & Meiser, 2014). However, in item response theory the focus is on measuring latent traits, not the impact of explanatory variables. In contrast to the treatment here no covariates are included. Moreover, in item response trees typically in each split new person and trait parameters are specified. The consequence is that in most steps response styles are modeled but the tendency to higher categories gets lost. An exception is the parameterization given by Meiser et al. (2019), they include content related parameters in all steps of item response models.

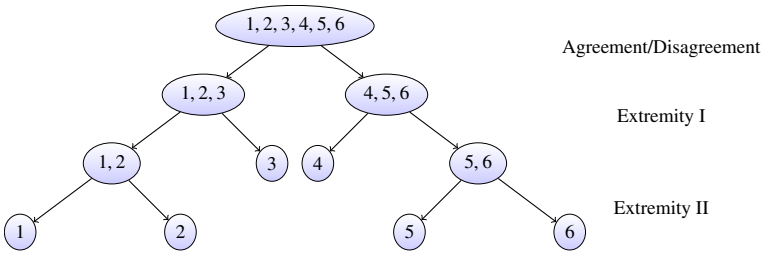


FIGURE 6 A tree for six ordered categories, categories 1,2,3 represent levels of disagreement, categories 4,5,6 represent levels of agreement

5.3 | The case of the neutral category

Likert scales with an odd number of categories include a neutral middle category. It is often unclear which role this neutral category plays. It can be part of the ordered scale or it can be viewed by the respondent as a "dumping ground" for unsure or nonapplicable response (Kulas et al., 2008). If it is used as a dumping ground and one fits an ordinal model that includes the middle category estimates will be distorted.

Hierarchically structured model offer a possibility to separate the neutral category and investigate if the impact of explanatory variables changes if the neutral category is separated. Let the set of response categories be partitioned into $S_1 = \{1, ..., m-1\}$, $S_2 = \{m\}$, $S_3 = \{m+1, ..., m\}$, where $m = (k+1)/2$ denotes the middle category. A model that separates the middle category within a hierarchical model is

$$P(Y = m|\boldsymbol{x}) = F\left(\beta_{0r}^{(m)} + \boldsymbol{x}^T\boldsymbol{\beta}^{(m)}\right),$$
$$P(Y \geq r|Y \in S_1 \cup S_3, \boldsymbol{x}) = F\left(\beta_{0r} + \boldsymbol{x}^T\boldsymbol{\beta}\right).$$

The first part of the model investigates the effect of covariates on the choice of the middle category by using a binary model, the second part investigates the effect of covariates if the respondent shows some degree of agreement or disagreement by using a cumulative model. It is straightforward to include dispersion in the second part of the model.

In the second part one can also use a model that lets effects be different in agreement or disagreement categories by using(comment! in the following formula replace Y \in S_1 by Y \in S_g)

$$P(Y \geq r|Y \in S_1, \boldsymbol{x}) = F\left(\beta_{0r}^{(S_g)} + \boldsymbol{x}^T\boldsymbol{\beta}^{(S_g)}\right), g = 1, 3.$$

Alternatively, one can use a tree-model for the categories $S_1 \cup S_3$. In both cases it is easy to include dispersion effects. For illustrations, see Tutz (2020).

5.4 | Fitting of hierarchical models

Binary splits models have the advantage that one can use software for binary response models. Since all responses are binary, and the structure is hierarchical, one only has to compute the design variables that correspond to specific binary responses. When constructing the design one has to account for the fact that one has various binary decisions. Let us consider the pseudo item $Y \in S_{j1}$ from the binary model $P\left(Y \in S_{j1}|Y \in S_j, \boldsymbol{x}\right) = F\left(\beta_{0r}^{(S_{j1})} + \boldsymbol{x}^T\boldsymbol{\beta}^{(Sj1)}\right)$, $S_{j1} \subset S_j$, which is the model (14), but with more indices. A pseudo item can be characterized by the set of categories S_j that is split. Let the sets be given by S_1..., S_m. Then the binary response when splitting S_j is defined by $Y^{S_j} = 1$ if $Y \in S_{j1}$ Thus, the explanatory variables linked to the binary response Y^{S_j} are $\left(\boldsymbol{x}_1^T, ..., \boldsymbol{x}_m^T\right)$, with $\boldsymbol{x}_j = \boldsymbol{x}$ and $\boldsymbol{x}_s = \boldsymbol{0}$ if $s \neq j$, where $\boldsymbol{0}$ is a vector of zeros.

In the general case with a mixture of binary and multi-categorical pseudo items specific software is needed. There are some exceptions. For example, in a 7-point Likert item the initial response in categories {1,2,3}, {4}, {5,6,7} has three

TABLE 6 Estimates of hierarchical model for fears of the use of nuclear energy

	Hierarchical model				Hierarchical model with response style			
	Estimate	**Std. error**	***z* value**	**Pr (>|z|)**	**Estimate**	**Std. error**	***z* value**	**Pr (>|z|)**
Age	0.0125	0.0017	7.251	0.0000	0.0104	0.0017	5.860	0.0000
Gender	0.4375	0.0633	6.907	0.0000	0.5056	0.0659	7.667	0.0000
Unemployment	0.0448	0.2033	0.221	0.8254	−0.0244	0.2034	−0.120	0.9044
EastWest	−0.3528	0.0677	−5.210	0.0000	−0.3556	0.0690	−5.150	0.0000
Abitur	−0.0282	0.0650	−0.434	0.6653	−0.0029	0.0670	−0.044	0.9648
AgeR					−0.0116	0.0026	−4.422	0.0000
GenderR					0.3077	0.0969	3.175	0.0015
UnemploymentR					−0.7592	0.3086	−2.460	0.0139
EastWestR					−0.1502	0.1015	−1.479	0.1392
AbiturR					0.1612	0.0973	1.657	0.0976

categories, and the response within {1,2,3}, {5,6,7} also has three categories. Then one can proceed in the same way as for binary splits by constructing the design variables that correspond to these pseudo items.

5.5 | Nuclear fear data

We restrict consideration to a simple two level model with cumulative components. Table 6 shows the estimates of the model with the logistic link function in the components. The left columns show the estimates of the basic hierarchical model, the right columns show the estimates if response style parameters that represent a tendency to middle or extreme categories are included (response style parameters have the ending R). It is seen from the left columns that the same variables show significant effects as in the simple cumulative model given in Table 2. Thus, the hierarchical model may be used as an alternative to the cumulative model. However, for this data set the fit of the cumulative model (log-likelihood −2772.14) is better than for the hierarchical model (log-likelihood −3782.21).

The inclusion of response style parameters does not strongly change the parameters of significant variables in the location term, but provides additional information. For example, women show a stronger tendency to middle categories than men. The response style effects are similar to the effects seen in the location-shift model, with which they can be directly compared (Figure 3). Goodness of fit is better for the location-shift model (log-likelihood −3749.66) than for the hierarchical model with response style (log-likelihood −3758.99).

6 | MIXTURE MODELS

A quite different class of ordinal models that has been proposed in the last decades are finite mixture models. Although originally the focus was on modeling uncertainty more recently extensions to alternative response styles have been considered. Mixture models are different from other ordinal models and therefore are included in the taxonomy in Figure 5 as a separate class of models.

The basic mixture model for ordinal responses including uncertainty has the form

$$P(Y_i = r|\boldsymbol{x}_i) = \pi_i P_M(C_i = r|\boldsymbol{x}_i) + (1-\pi_i)P_U(U_i = r), \tag{15}$$

where Y_i represents the observed response and C_i, U_i are unobserved random variables taking values from {1, ..., k}. The variable C_i represents the deliberate choice, that is, the content related response determined by the preferences of a person while U_i represents the uncertainty of the respondent arising from factors like amount of time devoted to the response, fatigue, partial understanding, and so on. What is observed is not the content related response but a response

that results from a mixture of content related response and uncertainty. Iannario and Piccolo (2010a) discuss extensively the logical foundations and psychological motivations of the mixture.

Essential components of the modeling strategy are:

- the model for the content related response $P_M(C_i = r| \boldsymbol{x}_i)$, which can be any ordinal model M,
- the model for uncertainty, which is typically chosen as a uniform discrete distribution, $P_U(U_i = r) = 1/k$,
- the mixture weights, which are determined by explanatory variables in the form $\text{logit}(\pi_i) = \boldsymbol{x}_i^T\boldsymbol{\gamma}$.

The model contains two sets of parameters, one for the content related response in the linear predictor $\boldsymbol{x}^T\boldsymbol{\beta}$ of the ordinal model for Y, and one in the specification of the mixture components,

$$\text{logit}(\pi_i) = \boldsymbol{x}_i^T\boldsymbol{\gamma}.$$

The latter is used to investigate which group of individuals tends to uncertainty.

The uncertainty mixture model has been propagated in a series of articles, including Piccolo (2003), D'Elia and Piccolo (2005), Iannario and Piccolo (2010b), Iannario (2012a), Iannario (2012b), Manisera and Zuccolotto (2014), Piccolo (2015). An extensive overview has been given more recently by Piccolo and Simone (2019). In most of the approaches the model for the content-related response was specified as a shifted *binomial model* as proposed in one of the early articles on uncertainty mixture models (Piccolo, 2003). This classical version was named CUB model for Combination of a discrete Uniform and a shifted Binomial random. A more general model, which allows for any ordinal model in the content related component, was considered by Tutz et al. (2017). It links the more conventional models as the cumulative and the adjacent categories model to uncertainty. In the CUB model literature the mixture has also been interpreted in a different way. Instead of assuming that a person comes from one of the two groups, content driven responders or responders affected by uncertainty, it is assumed that the response of each subject is a mixture between feeling and uncertainty (Piccolo & Simone, 2019).

Whatever the interpretation, the way uncertainty is modeled in mixture models of the form (15) differs crucially from the way uncertainty is modeled, for example, by the shifted adjacent categories model. In mixture models it is assumed that a person responds driven by content *or* randomly. In contrast, in the shifted adjacent categories model it is assumed that persons show uncertainty to a specific degree. The parameter $\boldsymbol{\alpha}$ compares (groups of) persons, which show differing degrees of uncertainty.

An advantage of the uncertainty mixture model over classical mixture models is that it clearly specifies the meaning of the mixture components. In classical mixture models for generalized linear models the components are typically left unspecified, see Greene and Hensher (2003); Grün and Leisch (2008); Breen and Luijkx (2010). It is assumed that the mixture components are from the same class of models. After fitting various models one selects a number of mixture components and tries to interpret the estimated parameters in the final model. However, one gets quite different parameter estimates if one fits two, three or four components, and choice of the right number of mixtures is difficult. It seems much better to use a mixture model that clearly specifies the meaning of the components. Then one uses a tool that is explicitly tailored to investigate a specific structure like uncertainty that may or may not be present in the data.

More recently alternative mixture models have been considered that allow to model response styles beyond uncertainty. Instead of using for U the discrete uniform distribution, which represents a rather extreme way of responding, one can include a preference for specific categories or a tendency to extreme or middle categories by using alternative distributions for U, see Gottard et al. (2016), Simone and Tutz (2018), and Tutz and Schneider (2019).

In general, mixture models are an interesting tool if one suspects heterogeneity in a population. However, estimation can be difficult since log-likelihoods are not concave and one might end up in local maxima. Typically large data sets are needed to obtain stable estimates. If one suspects uncertainty it might be preferable to include uncertainty explicitly in the predictor as is done in the adjacent categories model. Moreover, then one works within the generalized linear model framework. In mixture models it is assumed that a person is a content-related responder or the person chooses a category at random, showing maximal uncertainty. What one can estimate is the posterior probability of being a content-driven responder. In models that include uncertainty in the linear predictor

TABLE 7 Estimates of the adjacent categories model with an uncertainty term and the classical CUB model

	Adjacent categories location shift-model				**CUB model**		
	Estimate	Std. error	z value	Pr (>|z|)	Estimate	Std. error	z value
Age	0.0041	0.0007	5.629	0.0000	−0.0206	0.0026	−7.7484
Gender	0.2162	0.0274	7.871	0.0000	−0.4586	0.1083	−4.2329
Unemployment	−0.0264	0.0706	−0.375	0.7077	−10.4195	66.7125	−0.1561
EastWest	−0.1671	0.0275	−6.072	0.0000	0.3871	0.1620	2.3895
Abitur	−0.0149	0.0272	−0.548	0.5840	−0.0337	0.0998	−0.3378
AgeU	−0.0040	0.0007	−5.236	0.0000	−0.00418	0.0061	−0.6765
GenderU	0.0758	0.0285	2.657	0.0078	1.19185	0.2347	5.0773
UnemploymentU	−0.1983	0.0800	−2.478	0.0132	−1.63231	0.647	−2.5206
EastWestU	−0.0194	0.0292	−0.666	0.5055	−1.29106	0.2857	−4.5174
AbiturU	0.0111	0.0285	0.390	0.6963	−0.12217	0.2407	−0.5075

uncertainty of persons or groups of persons is determined by parameters in the predictor. A person's response is simultaneously determined by its location and uncertainty term. In the following we briefly compare these alternative strategies.

Nuclear fear data

Table 7 shows the estimates of the adjacent categories model with an uncertainty term and the classical CUB model. The parameters that model uncertainty have the ending"U." They refer to the mixture component in the CUB model and the uncertainty term in the adjacent categories model. Of course parameter values cannot be compared since the distributional assumptions are quite different. Concerning relevance of explanatory variables, the location term parameters are comparable, although for unemployment one obtains a rather extreme parameter value in the CUB model. For the effect of the explanatory variables on uncertainty one obtains differing effects, in particular for age (needed in the adjacent categories model, not significant in CUB model) and EastWest (needed in CUB, not significant in the adjacent categories model) one gets quite different results. Thus, the choice of the model determines which variables are found influential. We also fitted a CUB model that contains significant effects only, since the estimates of the other variables are almost unchanged the results are not given.

7 | FURTHER DEVELOPMENTS

In the following we briefly consider some further topics in ordinal modeling. Additive models allow to relax the assumptions on the predictor, regularization, and variable selection are important in high dimensional settings, in which one has to select the most important components of models. The last section is devoted to tree-based models, which are especially useful in prediction.

7.1 | Additive models

Most of the models considered so far are members of the generalized linear models family. The models are nonlinear because of the link function, but nonetheless they are parametric, because the effect of covariates is contained in the linear term $\boldsymbol{x}^T\boldsymbol{\beta}$. Often parametric models are too restrictive and nonparametric models are warranted.

A general class of models that are well developed for univariate responses are generalized additive models. The model class may be extended to ordinal models, which are multivariate. It is straightforward to replace in basic models the predictor $\eta_r = \beta_{0r} + \boldsymbol{x}^T\boldsymbol{\beta}$ by the additive predictor

$$\eta_r = \beta_{0r} + f_{(1)}(x_1) + \ldots + f_{(p)}(x_p),$$

where the $f_{(j)}(.)$ are unspecified functions. As in univariate models the unknown functions may be expanded in basis functions (Eilers & Marx, 1996), smoothing splines (Gu, 2002) or thin-plate splines (Wood, 2004). An overview of available methods for classical metric regression is found in Wood (2004). Ordinal models with additive predictors were considered by Yee (2010) within the framework of vector generalized additive models, for which also software is available.

Vector generalized additive models are useful tools that allow to use an additive predictor in basic models like the cumulative, sequential and adjacent categories model. However, additive versions of models are not yet available for models that contain more than a location effect. For example, one can obtain an additive modeling of response styles by using the shifting framework considered here. In the predictor

$$\eta_r = \beta_{0r} + f_{(1)}(x_1) + \ldots + f_{(p)}(x_p) + \text{sign}(k/2 - r)\left(g_{(1)}(x_1) + \ldots + g_{(q)}\right),$$

the smooth, unspecified functions $g_{(j)}(.)$ represent response styles that are potentially nonlinear. Models of this type allow for nonlinear location effects but also account for nonlinear dispersion effects. They aim at modeling the relevant features in a flexible way but forgo the problems that arise if one wants to model category-specific smooth effects as an extension of linear predictors with category-specific effects. Extensions of this sort were considered by Tutz (2003), but call for rather complex regularization methods to obtain estimates.

More flexible models can be obtained for all of the models given in Figure 5. Although the basic concept is simple, namely replacing the linear predictors by additive predictors, several problems have to be addressed. In mixture models identifiability issues might be relevant, in all models one has to carefully select tuning parameters that determine the degree of smoothness, and one has to decide which basis functions perform best. While software is available for additive versions of basic models, for most additive hierarchically structured models software has yet to be provided. An exception are symmetric models that use binary splits only, as, for example, the model specified by the tree in Figure 6. For these binary splits models software for additive binary models can be used after construction of the corresponding design matrix.

7.1.1 | Nuclear fear data

We briefly consider a model in which age is included as a smooth predictor, all other variables have linear effects, which is sensible because they are categorical variables. The first plot in Figure 7 shows the smooth effect of age if the cumulative logit model with additive effects is fitted. The second plot shows the effect if one fits a hierarchical two step model with cumulative components. The third plot shows the fit if one adds a tendency to middle or extreme categories. It is seen that the effect of age changes slightly across models but remains rather stable. We just focus on the effect of the metric variable age and do not give all the parameter estimates.

7.2 | Regularization and variable selection

Since the introduction of the lasso (Tibshirani, 1996) various regularization methods that are able to improve regression model coefficient estimation and prediction accuracy have been considered. In particular, variable selection tools are important if more predictors are available than can reasonably be included in a model. Selection methods that are derived from maximum likelihood estimates use the penalized log-likelihood $l_p(\boldsymbol{\beta}) = l(\boldsymbol{\beta}) - J_\lambda(\boldsymbol{\beta})$, where $l(\boldsymbol{\beta})$ is the usual log-likelihood, λ is a tuning parameter, and $J(\boldsymbol{\beta})$ is a penalty term. For basic ordinal models with predictor $\beta_{0r} + \boldsymbol{x}^T\boldsymbol{\beta}$ variable selection is obtained by using the lasso type penalty

$$J(\boldsymbol{\beta}) = \lambda \sum_{j=1}^{p} |\beta_j| .$$

The number of parameters that are selected depends on the tuning parameter λ, which typically is chosen in a data driven way. In the so-called elastic net penalty one uses $J(\boldsymbol{\beta}) = \lambda_1 \sum_j |\beta_j| + \lambda_2 \sum_j \beta_j^2$, which uses two tuning parameters.

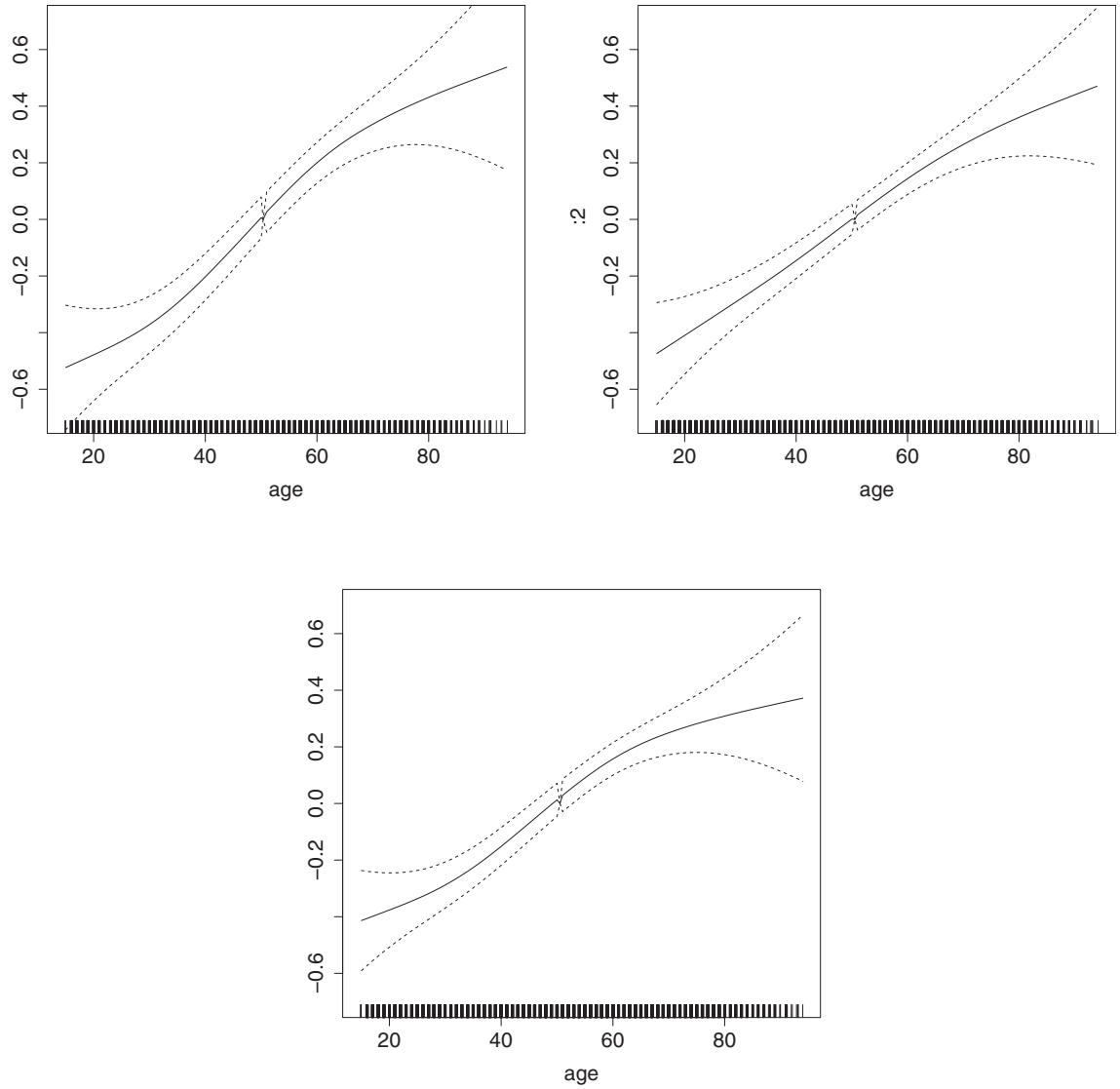


FIGURE 7 Smooth effect of age for the cumulative model, the hierarchical two step model, the hierarchical two step model with response style

For basic ordinal models Archer and Williams (2012) used the L_1 penalty and Archer and Williams (2012) used elastic net versions. Both methods are made available as program packages (Archer et al., 2014; Wurm et al., 2017). Also the M-estimators, which yield robust inference for classical ordinal response models, proposed by Iannario et al. (2017), can be seen as regularization based methods.

Selection and regularization is harder in more complex parameterizations. If the predictor contains category-specific effects with predictor $\beta_{0r} + \boldsymbol{x}^T\boldsymbol{\beta}_r$, one has to distinguish several levels, effects can be category-specific, global or negligible. Regularization methods that are able to distinguish between these levels will include fusion terms that distinguish between global and category-specific effects. First steps toward addressing the complexity of the selection problem were taken by Pössnecker and Tutz (2016). In the light of the discussion on the need for category-specific effects (Section 4) it seems more appropriate to restrict consideration to the selection of the most important effects, that is, location and response style effects like dispersion. If the predictor has the form $(\beta_{0\mathrm{r}} + x^{\mathrm{T}}\,\beta)/\exp\,(x^{\mathrm{T}}\,\gamma)$ penalties have to account for both effect types by using, for example,

$$J(\boldsymbol{\beta}) = \lambda_1 \sum_{j=1}^{p} |\,\beta_j\,| \; + \lambda_2 \sum_{j=1}^{p} |\,\gamma_j\,| \,.$$

Some care is needed in the selection of the tuning parameters λ_1, λ_2 since they determine which parts of the predictor are considered relevant. Similar problems occur when selecting variables in mixture models, which also contain different components that contribute to the complexity, namely the effects on the mixture probabilities and effects on the location. In the case of mixture model an additional difficulty is that one has to rely on the EM algorithm when estimating (penalized) parameters. For both models regularization methods seem not too have been developed sufficiently, and are not available in program packages.

7.3 | Tree-based models and random forests

All the models considered here can be used to investigate the impact of explanatory variables on an ordered response. Consequently, they can be used to predict ordinal outcomes for new observations. Applications of parametric models to predict have some tradition, they were considered, for example, by Rudolfer et al. (1995), Campbell and Donner (1989), Campbell et al. (1991), and Anderson and Phillips (1981). However, if prediction is the objective, parametric models are typically not be the best choice. It has been shown that in classification nonparametric alternatives like random forests are hard to beat. The same is to be expected for classification of ordinal data.

Recursive partitioning or simply trees, may be seen as a nonparametric way of investigating the effect of explanatory variables with a focus on interactions. The basic concept is very simple: by binary recursive partitioning the predictor space is partitioned into a set of rectangles and on each rectangle a simple model (e.g., a constant) is fitted. The most popular versions are CART (Breiman et al., 1984) and conditional inference trees, abbreviated by CTREE (Hothorn et al., 2006). Several approaches to handle ordinal responses have been proposed and some software packages are available. The packages rpartOrdinal (Archer, 2010) as well as the improved version rpartScore (Galimberti et al., 2012) are based on the Gini impurity function, Janitza and Boulesteix (2016) focus on variable selection and variable importance measures.

A problem with existing software is that they assume that scores are assigned to the ordered categories of the response. The assignment of scores can be warranted in some cases, in particular if ordinal responses are built from continuous variables by grouping. It is rather artificial and arbitrary in genuine ordinal response data, for example, if the response represents ordered levels of severeness of a disease. When using scores implicitly a metrically scaled response is assumed, which is not what a tree for ordinal responses should do. Trees that take the ordinal scale level seriously seem not yet available. A strategy that can be used is to use an ordinal model but replace the linear predictor by a tree. This modeling strategy yields a hybrid tree, which brings together parametric models and trees. Approaches of this type have been used to investigate varying coefficients, see Berger et al. (2019). They differ from recursive partitioning methods that fit models in subsets of the predictor space as proposed by Zeileis et al. (2008), and used in ordinal mixture models by Cappelli et al. (2019). The latter are able to identify subsets in which models differ most strongly but do not focus on separating populations which show responses in high or low categories.

If several trees are combined one obtains a random forest, see Breiman (2001) for the general concept of random trees. In random forests the main objective is prediction, not the detection of interactions. Then, the use of assigned scores can be considered justified, if it serves the purpose of obtaining the best prediction. For available software, see Section 8.

8 | CONCLUDING REMARKS AND SOFTWARE

Let us make some comments on the modeling approaches that were brought together in the previous sections.

- The given taxonomy yields a structured overview on ordinal models that includes the wide class of hierarchically structured models.
- Basic models as the cumulative, adjacent categories, and sequential model often yield similar results in terms of the impact of explanatory variables. In particular, for practitioners who are mainly interested in detecting, which variables have an effect on the response they are often well comparable.
- It is worthwhile to include additional heterogeneity effects, which are treated here as response styles in a wider sense. They typically provide better fit to the data and additional information on the effects of explanatory variables. If they are ignored estimates may be biased.

- Category-specific effects, which were treated extensively in the literature, can often be replaced by much simpler models that contain an heterogeneity term yielding much simpler and easy to interpret models.
- Hierarchical models offer a simple alternative modeling strategy that is able to incorporate response style effects in a straightforward way. They are very flexible tools that may also be used to investigate the use of the neutral category in Likert type responses. Their great potential to model ordinal response data has not yet been exploited and is only sketched in the present article.
- Mixture models are an alternative to model heterogeneity, in particular concerning uncertainty that varies in the population. In contrast to modeling heterogeneity by linear or multiplicative effects they assume that persons are from specific classes of responders.

Available software includes the following:

- *Basic models*. Basic models as the proportional odds model, the adjacent categories model and the sequential models can be fitted by using *vglm* from the package *VGAM* (Yee, 2010, 2015). The very flexible program also allows the fitting of models with category-specific effects and to reverse the order of categories. The proportional odds model can also be fitted with function *lrm* from the package *Design* and the function *polr* from the *MASS* library. Attention has to be paid to the algebraic signs of the coefficients. In Stata the program gologit2 can be used to fit ordered categories logit models with global and category-specific effects (Williams, 2006). In SAS one can use PROC GENMOD.
- *Models with additional heterogeneity*. The location-scale model can be estimated by using the function *clm* from the R package *ordinal* (Christensen, 2015), the Stata function oglm (Williams, 2010) or PROC NLIN when working with SAS. Location-shift versions of the cumulative and adjacent categories model can be fitted by using the R package *ordDisp*, which has been used in Section 2. Bayesian location-scale models can be fitted by using the function *clm* from the package *brms* (Bürkner, 2017).
- *Mixture models*. The package CUB (Iannario et al., 2020) fits finite mixture models with a binomial response distribution but also allows for other distribution models. It has been used to obtain estimates in Section 7. The package FastCUB (Simone, 2020) performs best-subset selection.
- *Variable selection*. The package *ordinalgmifs* (Archer et al., 2014) selects variables by using the lasso penalty, the package *ordinalNet* (Wurm et al., 2017) uses elastic net penalties.
- *Additive models*. The *VGAM* package (Yee, 2010) allows to fit multinomial and ordinal additive models. Cumulative models with identity link may also be fitted with *mgcv* (Wood, 2015).

ACKNOWLEDGMENT

Open access funding enabled and organized by Projekt DEAL.

CONFLICT OF INTEREST

The author has declared no conflicts of interest for this article.

ORCID

Gerhard Tutz https://orcid.org/0000-0002-6628-3539

RELATED WIREs ARTICLES

Learning ordinal data
Generalized linear models
Log-linear modelling

REFERENCES

Agresti, A. (2010). *Analysis of ordinal categorical data* (2nd ed.). New York: Wiley.

Agresti, A. (2013). *Categorical data analysis* (3rd ed.). New York: Wiley.

Agresti, A., & Kateri, M. (2017). Ordinal probability effect measures for group comparisons in multinomial cumulative link models. *Biometrics*, *73*(1), 214–219.

Agresti, A., & Tarantola, C. (2018). Simple ways to interpret effects in modeling ordinal categorical data. *Statistica Neerlandica*, *72*(3), 210–223.

Allison, P. D. (1999). Comparing logit and probit coefficients across groups. *Sociological Methods & Research*, *28*(2), 186–208.

Alvarez, R. M., & Brehm, J. (1995). American ambivalence towards abortion policy: Development of a heteroskedastic probit model of competing values. *American Journal of Political Science*, *39*, 1055–1082.

Ananth, C. V., & Kleinbaum, D. G. (1997). Regression models for ordinal responses: A review of methods and applications. *International Journal of Epidemiology*, *26*, 1323–1333.

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

Anderson, J. A., & Phillips, R. R. (1981). Regression, discrimination and measurement models for ordered categorical variables. *Applied Statistics*, *30*, 22–31.

Archer, K., & Williams, A. (2012). L 1 penalized continuation ratio models for ordinal response prediction using high-dimensional datasets. *Statistics in Medicine*, *31*(14), 1464–1474.

Archer, K. J. (2010). Rpartordinal: An R package for deriving a classification tree for predicting an ordinal response. *Journal of Statistical Software*, *34*(7).

Archer, K. J., Hou, J., Zhou, Q., Ferber, K., Layne, J. G., & Gentry, A. E. (2014). ordinalgmifs: An r package for ordinal regression in high-dimensional data settings. *Cancer Informatics*, *13*, CIN–S20806.

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

Baumgartner, H., & Steenkamp, J.-B. E. (2001). Response styles in marketing research: A cross-national investigation. *Journal of Marketing Research*, *38*(2), 143–156.

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

Berger, M., Tutz, G., & Schmid, M. (2019). Tree-structured modelling of varying coefficients. *Statistics and Computing*, *29*(2), 217–229.

Böckenholt, U. (2017). Measuring response styles in likert items. *Psychological Methods*, *22*, 69–83.

Böckenholt, U., & Meiser, T. (2017). Response style analysis with threshold and multi-process IRT models: A review and tutorial. *British Journal of Mathematical and Statistical Psychology*, *70*(1), 159–181.

Bolt, D. M., & Newton, J. R. (2011). Multiscale measurement of extreme response style. *Educational and Psychological Measurement*, *71*(5), 814–833.

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

Breen, R., Holm, A., & Karlson, K. B. (2014). Correlations and nonlinear probability models. *Sociological Methods & Research*, *43*(4), 571–605.

Breen, R., & Luijkx, R. (2010). Mixture models for ordinal data. *Sociological Methods and Research*, *39*, 3–24.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, J. C. (1984). *Classification and regression trees*. Monterey, CA: Wadsworth.

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28.

Campbell, M. K., & Donner, A. P. (1989). Classification efficiency of multinomial logistic-regression relative to ordinal logistic-regression. *Journal of the American Statistical Association*, *84*(406), 587–591.

Campbell, M. K., Donner, A. P., & Webster, K. M. (1991). Are ordinal models useful for classification? *Statistics in Medicine*, *10*, 383–394.

Cappelli, C., Simone, R., & Di Iorio, F. (2019). cubremot: A tool for building model-based trees for ordinal responses. *Expert Systems with Applications*, *124*, 39–49.

Christensen, R. H. (2015). Analysis of ordinal data with cumulative link models—Estimation with the R-package ordinal. R-Package version, 1–31.

Cox, C. (1995). Location-scale cumulative odds models for ordinal data: A generalized non-linear model approach. *Statistics in Medicine*, *14*, 1191–1203.

D'Elia, A., & Piccolo, D. (2005). A mixture model for preference data analysis. *Computational Statistics & Data Analysis*, *49*, 917–934.

De Boeck, P., & Partchev, I. (2012). Irtrees: Tree-based item response models of the glmm family. *Journal of Statistical Software*, *48*(1), 1–28.

Dolgun, A., & Saracbasi, O. (2014). Assessing proportionality assumption in the adjacent category logistic regression model. *Statistics and its Interface*, *7*(2), 275–295.

Eilers, P. H. C., & Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, *11*, 89–121.

Fernandez, D., Liu, I., & Costilla, R. (2019). A method for ordinal outcomes: The ordered stereotype model. *International Journal of Methods in Psychiatric Research*, *28*, e1801.

Fullerton, A. S., & Xu, J. (2012). The proportional odds with partial proportionality constraints model for ordinal response variables. *Social Science Research*, *41*(1), 182–198.

Galimberti, G., Soffritti, G., Maso, M. D.(2012). Classification trees for ordinal responses in R: The rpartscore package. *Journal of Statistical Software*, *47*(i10).

Goodman, L. A. (1981a). Association models and canonical correlation in the analysis of cross-classification having ordered categories. *Journal of the American Statistical Association*, *76*, 320–334.

Goodman, L. A. (1981b). Association models and the bivariate normal for contingency tables with ordered categories. *Biometrika*, *68*, 347–355.

Gottard, A., Iannario, M., & Piccolo, D. (2016). Varying uncertainty in CUB. *Advances in Data Analysis and Classification*, *10*(2), 225–244.

Greene, W., & Hensher, D. (2003). A latent class model for discrete choice analysis: Contrasts with mixed logit. *Transportation Resesearch, Part B*, *39*, 681–689.

Greenland, S. (1994). Alternative models for ordinal logistic regression. *Statistics in Medicine*, *13*, 1665–1677.

Grün, B., & Leisch, F. (2008). Identifiability of finite mixtures of multinomial logit models with varying and fixed effects. *Journal of Classification*, *25*, 225–247.

Gu, C. (2002). *Smoothing splines ANOVA models*. New York: Springer-Verlag.

Hamada, M., & Wu, C. F. J. (1990). A critical look at accumulation analysis and related methods. *Technometrics*, *32*, 119–130.

Hastie, T., & Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society B*, *55*, 757–796.

Hauser, R. M., & Andrew, M. (2006). 1. Another look at the stratification of educational transitions: The logistic response model with partial proportionality constraints. *Sociological Methodology*, *36*(1), 1–26.

Hedeker, D., & Mermelstein, R. J. (1998). A multilevel thresholds of change model for analysis of stages of change data. *Multivariate Behavioral Research*, *33*(4), 427–455.

Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics*, *15*, 651–674.

Iannario, M. (2012a). Hierarchical CUB models for ordinal variables. *Communications in Statistics-Theory and Methods*, *41*(16–17), 3110–3125.

Iannario, M. (2012b). Modelling *shelter* choices in a class of mixture models for ordinal responses. *Statistical Methods and Applications*, *21*, 1–22.

Iannario, M., Monti, A. C., Piccolo, D., Ronchetti, E. (2017). Robust inference for ordinal response models. *Electronic Journal of Statistics*, *11* (2), 3407–3445.

Iannario, M., & Piccolo, D. (2010a). A new statistical model for the analysis of customer satisfaction. *Quality Technology & Quantitative Management*, *7*(2), 149–168.

Iannario, M., & Piccolo, D. (2010b). Statistical modelling of subjective survival probabilities. *Genus*, *66*, 17–42.

Iannario, M., Piccolo, D., & Simone, R. (2020). CUB: A class of mixture models for ordinal data. R Package Version 1.1.4, Available from http://cran.r-project.org/package=cub.

Janitza, S. T. G., & Boulesteix, A.-L. (2016). Random forests for ordinal responses: Prediction and variable selection. *Computational Statistics and Data Analysis*, *96*, 57–73.

Johnson, T. R. (2003). On the use of heterogeneous thresholds ordinal regression models to account for individual differences in response style. *Psychometrika*, *68*(4), 563–583.

Karlson, K. B., Holm, A., & Breen, R. (2012). Comparing regression coefficients between same-sample nested models using logit and probit: A new method. *Sociological Methodology*, *42*(1), 286–313.

Kateri, M. (2014). *Contingency table analysis. Methods and implementation using R*, Aachen: . Springer.

Khorramdel, L., & von Davier, M. (2014). Measuring response styles across the big five: A multiscale extension of an approach using multinomial processing trees. *Multivariate Behavioral Research*, *49*(2), 161–177.

Kim, J.-H. (2003). Assessing practical significance of the proportional odds assumption. *Statistics & Probability Letters*, *65*(3), 233–239.

Kulas, J. T., Stachowski, A. A., & Haynes, B. A. (2008). Middle response functioning in likert-responses to personality items. *Journal of Business and Psychology*, *22*(3), 251–259.

Läärä, E., & Matthews, J. N. (1985). The equivalence of two models for ordinal data. *Biometrika*, *72*, 206–207.

Liu, I., Mukherjee, B., Suesse, T., Sparrow, D., & Park, S. K. (2009). Graphical diagnostics to check model misspecification for the proportional odds regression model. *Statistics in Medicine*, *28*(3), 412–429.

Long, J. S. (1997). Regression models for categorical and limited dependent variables. *Advanced quantitative techniques in the social sciences* (vol. 7, p. 219).

Long, J. S., & Freese, J. (2006). *Regression models for categorical dependent variables using Stata*, College Station, TX: . Stata Press.

Long, J. S., & Mustillo, S. A. (2018). Using predictions and marginal effects to compare groups in regression models for binary outcomes. *Sociological Methods & Research*. https://doi.org/10.1177/0049124118799374

Manisera, M., & Zuccolotto, P. (2014). Modeling rating data with nonlinear cub models. *Computational Statistics & Data Analysis*, *78*, 100–118.

McCullagh, P. (1980). Regression model for ordinal data (with discussion). *Journal of the Royal Statistical Society B*, *42*, 109–127.

Meiser, T., Plieninger, H., & Henninger, M. (2019). IRTree models with ordinal and multidimensional decision nodes for response styles and trait-based rating responses. *British Journal of Mathematical and Statistical Psychology*, *72*, 501–516.

Messick, S. (1991). Psychology and methodology of response styles. In R. E. Snow & D. E. Wiley (Eds.), *Improving inquiry in social science: A volume in honor of lee J. Cronbach* (pp. 161–200). Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Mood, C. (2010). Logistic regression: Why we cannot do what we think we can do, and what we can do about it. *European Sociological Review*, *26*(1), 67–82.

Nair, V. N. (1987). Chi-squared-type tests for ordered alternatives in contingency tables. *Journal of the American Statistical Association*, *82*, 283–291.

Peterson, B., & Harrell, F. E. (1990). Partial proportional odds models for ordinal response variables. *Applied Statistics*, *39*, 205–217.

Peyhardi, J., Trottier, C., & Guédon, Y. (2015). A new specification of generalized linear models for categorical data. *Biometrika*, *102*, 889–906.

Peyhardi, J., Trottier, C., & Guédon, Y. (2016). Partitioned conditional generalized linear models for categorical responses. *Statistical Modelling*, *16*(4), 297–321.

Piccolo, D. (2003). On the moments of a mixture of uniform and shifted binomial random variables. *Quaderni di Statistica*, *5*, 85–104.

Piccolo, D. (2015). Inferential issues on CUBE models with covariates. *Communications in Statistics-Theory and Methods*, *44*(23), 5023–5036.

Piccolo, D., & Simone, R. (2019). The class of CUB models: Statistical foundations, inferential issues and empirical evidence (with discussions and a rejoinder). *Statistical Methods and Applications*, *28*, 389–493.

Plieninger, H., & Meiser, T. (2014). Validity of multiprocess IRT models for separating content and response styles. *Educational and Psychological Measurement*, *74*(5), 875–899.

Pössnecker, W. & Tutz, G. (2016). *A general framework for the selection of effect type in ordinal regression*. Technical report, technical report 186. Department of Statistics LMU.

Rattinger, H., Roßteutscher, S., Schmitt-Beck, R., Weßels, B., & Wolf, C. (2014). Pre-election cross section (GLES 2013). *GESIS Data Archive, Cologne ZA5700 Data file Version 2.0.0*.

Rohwer, G. (2015). A note on the heterogeneous choice model. *Sociological Methods & Research*, *44*(1), 145–148.

Rudolfer, S. M., Watson, P. C., & Lesaffre, E. (1995). Are ordinal models useful for classification? A revised analysis. *Journal of Statistical Computation Simulation*, *52*(2), 105–132.

Simone, R. (2020). FastCUB: Fast EM and best-subset selection for CUB models for rating data, R package version 0.0.2, Available from https://cran.r-project.org/package=fastcub.

Simone, R., & Tutz, G. (2018). Modelling uncertainty and response styles in ordinal data. *Statistica Neerlandica*, *72*(3), 224–245.

Thissen-Roe, A., & Thissen, D. (2013). A two-decision model for responses to likert-type items. *Journal of Educational and Behavioral Statistics*, *38*(5), 522–547.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society B*, *58*, 267–288.

Tutz, G. (1989). Compound regression models for categorical ordinal data. *Biometrical Journal*, *31*, 259–272.

Tutz, G. (1991). Sequential models in ordinal regression. *Computational Statistics & Data Analysis*, *11*, 275–295.

Tutz, G. (2003). Generalized semiparametrically structured ordinal models. *Biometrics*, *59*, 263–273.

Tutz, G. (2012). *Regression for categorical data*, Cambridge: . Cambridge University Press.

Tutz, G. (2019). Modelling heterogeneity: On the problem of group comparisons with logistic regression and the potential of the heterogeneous choice model. *Advances in Data Analysis and Classification*, *14*, 517–542. https://doi.org/10.1007/s11634-019-00381-8

Tutz, G. (2020). Hierarchical models for the analysis of Likert scales in regression and item response analysis. *International Statistical Review*. https://doi.org/10.1111/insr.12396

Tutz, G., & Berger, M. (2016). Response styles in rating scales - simultaneous modelling of content-related effects and the tendency to middle or extreme categories. *Journal of Educational and Behavioral Statistics*, *41*, 239–268.

Tutz, G., & Berger, M. (2017). Separating location and dispersion in ordinal regression models. *Econometrics and Statistics*, *2*, 131–148.

Tutz, G. & Berger, M. (2020). Non proportional odds models are widely dispensable - sparser modeling based on parametric and additive location-shift approaches. Technical report, Availablr ftom http://arxiv.org/abs/2006.03914.

Tutz, G., & Schmid, M. (2016). *Modeling discrete time-to-event data*, Switzerland: . Springer-Verlag.

Tutz, G., & Schneider, M. (2019). Flexible uncertainty in mixture models for ordinal responses. *Journal of Applied Statistics*, *46*, 1582–1601.

Tutz, G., Schneider, M., Iannario, M., & Piccolo, D. (2017). Mixture models for ordinal responses to account for uncertainty of choice. *Advances in Data Analysis and Classification*, *11*(2), 281–305.

Van Vaerenbergh, Y., & Thomas, T. D. (2013). Response styles in survey research: A literature review of antecedents, consequences, and remedies. *International Journal of Public Opinion Research*, *25*(2), 195–217.

Walker, R. W. (2016). On generalizing cumulative ordered regression models. *Journal of Modern Applied Statistical Methods*, *15*(2), 28.

Williams, R. (2006). Generalized ordered logit/partial proportional odds models for ordinal dependent variables. *Stata Journal*, *6*(1), 58–82.

Williams, R. (2009). Using heterogeneous choice models to compare logit and probit coefficients across groups. *Sociological Methods & Research*, *37*(4), 531–559.

Williams, R. (2010). Fitting heterogeneous choice models with oglm. *Stata Journal*, *10*(4), 540–567.

Williams, R. (2012). Using the margins command to estimate and interpret adjusted predictions and marginal effects. *The Stata Journal*, *12*(2), 308–331.

Williams, R. (2016). Understanding and interpreting generalized ordered logit models. *The Journal of Mathematical Sociology*, *40*(1), 7–20.

Williams, R. A., & Quiroz, C. (2020). Ordinal regression models. In P. Atkinson, S. Delamont, A. Cernat, J. Sakshaug, & R. Williams (Eds.), *SAGE research methods foundations* (pp. 51–73).

Wood, S. (2015). Package mgcv. R Package Version 1, 29.

Wood, S. N. (2004). Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association*, *99*, 673–686.

Wurm, M. J., Rathouz, P. J., & Hanlon, B. M. (2017). Regularized ordinal regression and the ordinalnet R package. *arXiv Preprint arXiv*, *1706*, 05003.

Yee, T. (2010). The VGAM package for categorical data analysis. *Journal of Statistical Software*, *32*(10), 1–34.

Yee, T. W. (2015). *Vector generalized linear and additive models: With an implementation in R*, New York, NY: . Springer.

Zeileis, A., Hothorn, T., & Hornik, K. (2008). Model-based recursive partitioning. *Journal of Computational and Graphical Statistics*, *17*(2), 492–514.

How to cite this article: Tutz G. Ordinal regression: A review and a taxonomy of models. *WIREs Comput Stat*. 2022;14:e1545. https://doi.org/10.1002/wics.1545